Maria Novella Campagnoli è Professoressa Associata presso il Dipartimento di Giurisprudenza dell'Università degli Studi di Roma Tor Vergata, dove insegna: *Fondamenti della sostenibilità e strategie digitali*; *Introduzione alle tecniche e alla metodologia di ricerca informatica*; *Teoria e pratica dei diritti umani*. È membro della Commissione Presidio di Qualità di Ateneo; del Collegio dei docenti del Dottorato di Ricerca in Diritto Pubblico; del Gruppo di Gestione Assicurazione Qualità (AQ) del Corso di Laurea (DITES); del Comitato Unico di Garanzia di Ateneo; del Consiglio Scientifico del Centro Studi Giuridici "Di-con-per Donne". Dirige (con Massimo Farina) le collane "Orizzonte JUS" e "DirICTo". È autrice di svariate monografie, contributi e articoli scientifici pubblicati (in italiano, inglese, francese e portoghese) in collane e riviste di settore.

Piero Marra è Ricercatore di Filosofia del diritto presso il Dipartimento di Giurisprudenza dell'Università degli Studi di Bari Aldo Moro. Collabora alle attività didattiche e di ricerca nei Dipartimenti giuridici della medesima Università. È membro del progetto di ricerca FAIR – Future AI Research, per il quale ha tenuto diversi corsi di formazione, anche nell'ambito del Dottorato in "Computer Science and Mathematics". È titolare, inoltre, del corso in *New Technology and Philosophy of Law* presso l'Ateneo barese. Ha dedicato diversi studi ai temi della metodologia procedurale e alle relative implicazioni teoretiche e pratiche dinanzi ai mutamenti contemporanei del diritto, occupandosi di argomentazione giuridica e di logica del discorso giudiziale anche con riguardo all'uso dell'intelligenza artificiale nel processo.

Nein, gerade Tatsachen gibt es nicht, nur Interpretationen
No, non esistono fatti, ma solo interpretazioni

Friedrich Nietzsche, *Nachgelassene Fragmente* (1879-1881)

ISBN 979-12-5965-605-6

copia omaggio

23

INTELLIGENZA ARTIFICIALE E METAMORFOSI DEL DIRITTO a cura di M.N. Campagnoli e P. Marra

Unità del sapere giuridico

Quaderni di scienze penalistiche e filosofico-giuridiche

Collana diretta da
Antonio Incampo, Vito Mormando, Adolfo Scalfati

Intelligenza artificiale e metamorfosi del diritto

a cura di Maria Novella Campagnoli e Piero Marra

CACUCCI EDITORE
BARI

La rivoluzione digitale sta ampliando gli scenari applicativi e plasmando la stessa realtà sociale, grazie al crescente potere informazionale derivante dalla correlazione tra grandi masse di dati e i relativi metadati. Con il cambiamento radicale della realtà, muta anche l'esperienza giuridica. D'altronde, se non sono pochi i vantaggi legati all'intelligenza artificiale e alle tecnologie connesse, molti sono anche i rischi, soprattutto nella prospettiva dei diritti. Per questa ragione, l'opera riflette sulle molteplici implicazioni metamorfiche connesse a un'interazione, sempre più stretta e simbiotica, tra l'intelligenza umana e quella artificiale nell'esperienza giuridica. Il risultato è un'opera corale, che spazia dall'approccio teorico a quello pratico, attenzionando problematiche gius-filosofiche, logiche, sociologiche e psicologiche.

The digital revolution is expanding application scenarios and shaping social reality itself, thanks to the growing informational power deriving from the correlation between large masses of data and related metadata. With the radical change in reality, the legal experience is also changing. On the other hand, while there are many advantages associated with artificial intelligence and related technologies, there are also significant risks, especially from a rights perspective. For this reason, the work reflects on the multiple metamorphic implications of an increasingly close, symbiotic interaction between humans and artificial intelligence in the legal experience. The result is a choral work that ranges from theoretical to practical, focusing on legal-philosophical, logical, sociological, and psychological issues

Antonio Bruscella, *Deflagrazione di antico e moderno* (2020)
acrilico su tavola e collage, 100cm×70cm
collezione privata

ISBN 979-12-5965-605-6

23

a cura di M.N. Campagnoli e P. Marra

INTELLIGENZA ARTIFICIALE E METAMORFOSI DEL DIRITTO

copia omaggio

Unità del sapere giuridico

Quaderni di scienze penalistiche e filosofico-giuridiche

Collana diretta da
Antonio Incampo, Vito Mormando, Adolfo Scalfati

Intelligenza artificiale e metamorfosi del diritto

a cura di Maria Novella Campagnoli e Piero Marra

CACUCCI EDITORE
BARI

Unità del sapere giuridico
Quaderni di scienze penalistiche e filosofico-giuridiche

Collana diretta da
Antonio Incampo, Vito Mormando, Adolfo Scalfati

Comitato scientifico

Paolo Ferrua, Angiola Filipponio, Cristina Hermida del Llano, Heribert Franz Köck, Oliviero Mazza, Mariano Menna, Francesco Palazzo, Carlo Enrico Paliero, Luigi Pannarale, Otto Pfersmann, Aldo Regina, Herbert Schambeck, Giorgio Spangher, Gunther Teubner, Nicola Triggiani, Daniela Vigoni, Wojciech Żełaniec.

Comitato di redazione

Filippo Bottalico, Marilena Colamussi, Enzo Dell'Andro, Lucia Iandolo Pisanelli, Piero Marra, Maria Antonella Pasculli, Porzia Teresa Persio, Guglielmo Siniscalchi, Maurizio Sozio.

I volumi inseriti nella collana, ritenuti preliminarmente ammissibili dai Direttori con l'ausilio del Comitato Scientifico, sono sottoposti alla revisione tra pari [*peer review*] secondo i regolamenti attualmente vigenti per le pubblicazioni scientifiche.

I saggi contenuti in questo volume, rivisti ed aggiornati, sono apparsi per la prima volta sulla rivista "L'Ircocervo. Prima rivista elettronica italiana di Metodologia giuridica, Teoria generale del diritto e Dottrina dello Stato", nel fascicolo 1/2024 intitolato *Artificial Intelligence and Neuro-cognitive Sciences in Law: From Symbiosis to Substitution*. Si ringraziano il Comitato di Direzione della rivista per averne concesso la riedizione e i dott.ri Loredana Alò, Leonardo Zingaro e Giacomo D'Elia per l'opera di editing e redazione dell'indice dei nomi.

Il volume, in accesso aperto, è pubblicato con i fondi dell'Università degli Studi di Bari Aldo Moro afferenti allo Spoke 6 (Symbiotic AI) del progetto FAIR – Future AI Research (PE00000013), nell'ambito del programma MUR del PNRR finanziato dal NextGenerationEU.

Intelligenza artificiale e metamorfosi del diritto

Artificial Intelligence and Metamorphosis of Law

a cura di
Maria Novella Campagnoli e Piero Marra

CACUCCI EDITORE
BARI

© 2026 Cacucci Editore – Bari
Via Nicolai, 39 – 70122 Bari – Tel. 080/5214220
http://www.cacucciditore.it e-mail: info@cacucci.it

Ai sensi della legge sui diritti d'Autore e del codice civile è vietata la riproduzione di questo libro o di parte di esso con qualsiasi mezzo, elettronico, meccanico, per mezzo di fotocopie, microfilms, registrazioni o altro, senza il consenso dell'autore e dell'editore.

Indice

Introduzione 9

I.
IA e pensiero giuridico: scenari e rischi etico-giuridici

Agata C. Amato Mangiameli, *Intelligenza artificiale generativa: brevi note* 19

Massimo Farina, *L'arte di non degenerare: ecosistemi tracciabili dell'IA tra degrado informativo e rischio creativo* 33

Lucio Franzese, *Sulla insufficienza dell'Intelligenza Artificiale e della legge* 59

Piero Marra, *Un'idea simbiotica della giustizia digitale* 71

Francesco Romeo, *La produzione culturale dell'artificiale: prodromi di uno studio metaepistemologico su diritto ed informazione* 97

Francisco José Santamaría Ramos, *Implicaciones éticas y legales relacionadas con el uso creciente de la IA* 123

II.
Nuovi diritti e ricadute processuali

Wanda D'Avanzo, *Intelligenza artificiale e sanità digitale. Il diritto* by design *e il ruolo del* legal design *per la protezione dei dati personali* 171

Vincenzo Bruno Muscatiello, Ex Machina. *Il diritto penale, all'improvviso* 197

Lorenzo Pulito, *Previsioni algoritmiche e valutazioni prognostiche per la sostituzione della pena nella prospettiva della* symbiotic artificial intelligence 229

Paolo Sommaggio, Shan Ali, *Cognitive Liberty, Neurolaw, and Digital Governance: Integrating Neuroscience and Blockchain in the Era of Neuro-Rights* 257

Radosław Strugała, *Liability for Damage Caused to the Data Subject in the Event of GDPR Violation "by Artificial Intelligence": Between Legal Discrepancies and Policy Questions* 285

III.

Neuroscienze, s/oggettività e modelli relazionali

Salvatore Amato, *Macchinazioni. Soggettività, agentività ed entità esotiche* 307

Sveva D'Agosto, *La costruzione della figura femminile fra neuroni specchio e intelligenza artificiale simbiotica* 333

Teresa María García-Berrio, *Ética de la virtud para una IA compasiva: Ontología, subjetividad y condición humana en la era digital* 347

Olimpia G. Loddo, *L'agire sociale ai tempi dell'intelligenza artificiale. Il concetto di "istituzione socio-digitale"* 395

Paolo Moro, *Persona elettronica. Una finzione giuridica per l'intelligenza artificiale* 411

Guido Saraceni, *Intelligenza Artificiale e benessere psicologico* 435

Maurizio Sozio, *Il deficit fronetico dello "*Jus ex machina*"* 457

Autori 475

Indice dei nomi 477

Introduzione

Da quando, nel 1956, al Dartmouth College venne coniato il termine "intelligenza artificiale" (*artificial intelligence* o *AI*, nell'acronimo inglese), i presupposti teorici della disciplina che se ne occupa non sono mutati. Non a caso, sviluppi tecnici dell'*AI*, come il "*machine learning*" e il "*deep learning*", traggono origine proprio dal dibattito sorto negli anni '60 e '80. I risultati di quegli studi sono ancor oggi alla base della ricerca e delle recenti applicazioni in *AI*, sino agli attuali sistemi *general purpose* che, combinati con differenti tecniche, evocano il pericolo di un'imminente sostituzione dell'umano da parte dell'artificiale. La rivoluzione digitale, infatti, ha operato in maniera silente, ampliando gli scenari applicativi e plasmando la stessa realtà sociale, grazie al crescente potere informazionale derivante dalla correlazione di grandi masse di dati, e dei corrispondenti metadati.

Con il cambiamento radicale della realtà e della sua percezione da parte del soggetto, muta anche l'esperienza giuridica, mentre la scienza e la filosofia del diritto non possono fare altro che ripensare nuove categorie dogmatiche o riadattare le vecchie per comprenderne le conseguenze. D'altronde, se non sono pochi i vantaggi legati all'intelligenza artificiale e alle tecnologie connesse, molti sono anche i rischi, soprattutto nella prospettiva dei diritti. Per questa ragione, il volume intende riflettere sulle tante implicazioni metamorfiche connesse a un'interazione, sempre più stretta e simbiotica, fra l'intelligenza umana e artificiale nell'esperienza giuridica. Il risultato è un'opera corale, che spazia dall'approccio teorico a quello pratico, attenzionando problematiche gius-filosofiche, logiche, sociologiche e psicologiche.

Il volume si articola così in tre sezioni tra loro complementari e sinergiche, a dimostrarne la multidisciplinarietà. Alcune sezioni ospitano contributi riconducibili al progetto *FAIR – Future AI Research* nell'ambito del programma MUR del PNRR finanziato dal NextGenerationEU e al relativo *transversal project* dedicato a "Legal and Ethical Design of Trustworthy AI Systems" a cui afferiscono diversi Spokes. Ma c'è spazio anche per altri progetti, a riprova della trasversalità del tema: un capitolo è il risultato delle attività di ricerca del progetto *Start Up* denominato "Holmes" (*Harmonizing Ownership and Legal Measures for Ethical AI Systems*) sul tema del rapporto tra intelligenza artificiale generativa e diritto d'autore, l'altro del progetto *Artefatti normativi e disegni normativi: indagine sulla regolamentazione non linguistica (NAND)*, anche questo quale programma PNRR finanziato dal NextGenerationEU. La riferibilità ai diversi progetti è stata di volta in volta specificata in ogni saggio. Il risultato è un composito itinerario di ricerca, nel quale, le linee di indagine, di volta in volta proposte, coniugano la pluralità argomentativa con il rigore metodologico. Dalla ripresa delle questioni teoriche di ampio respiro si passa alla trattazione e alla problematizzazione di questioni più settoriali concernenti l'amministrazione della giustizia, il processo e la pena; si riserva, poi, ampio spazio alla disamina di temi specifici ed emergenti, come il rapporto uomo/macchina, la necessità di ripensare la soggettività, il mutamento della socialità e il benessere psicologico, per terminare con una riflessione sulla tutela dei (nuovi) diritti.

i) La prima sezione – *IA e pensiero giuridico: scenari e rischi etico-giuridici* – si apre con il contributo di Agata C. Amato Mangiameli, *Intelligenza artificiale generativa: brevi note*. Qui l'A., dopo aver chiarito immediatamente la differenza fra IA e IA generativa, si sofferma sulla disamina dei vantaggi e dei rischi connessi a quest'ultima, riflettendo in chiave critica sulle possibili violazioni dei diritti e guardando, con particolare attenzione, soprattutto alla più recente regolamentazione introdotta a seguito dell'approvazione dell'*AI Act.*

Segue il lavoro di Massimo Farina (*L'arte di non degenerare: ecosistemi tracciabili dell'IA tra degrado informativo e rischio*

creativo), che mette in luce il (non remoto) pericolo di degenerazione dell'IA. Trattasi, come spiega l'A., del rovescio della medaglia della complessità tecnica degli attuali sistemi autonomi; una conseguenza con la quale, forse molto presto, saremo costretti a fare i conti.

Lucio Franzese (*Sulla insufficienza dell'Intelligenza Artificiale e della legge*), parte da alcune considerazioni generali sull'uso dell'intelligenza artificiale e ne mette in luce le criticità nella sfera giuridica e sanitaria, come pure le tante difficoltà che si incontrano nell'elaborazione di una *governance* efficace.

Muove da premesse epistemiche il capitolo di Piero Marra, *Un'idea simbiotica della giustizia digitale*, che, analizzando le somiglianze e le differenze tra ragione calcolatrice e umana declinata nell'ambito giudiziale, riconosce nella simbiosi un inedito paradigma procedurale di costruzione socio-tecnica dell'interazione uomo-computer e uomo-macchina.

In una direzione non dissimile, il saggio di Francesco Romeo (*La produzione culturale dell'artificiale: prodromi di uno studio metaepistemologico su diritto ed informazione*), secondo il quale, oggi, anche la cultura umana appare, a suo modo, del tutto diversa da un tempo. Questo è dovuto al fatto che la produzione della cultura stessa ha iniziato a vedere in prima linea agenti e fattori non-umani. Di qui, l'urgenza di tornare a interrogarsi sulla verità e sul ruolo che l'informazione ha nella costruzione della realtà.

Chiude la sezione il lavoro di Francisco José Santamaría Ramos (*Implicaciones éticas y legales relacionadas con el uso creciente de la IA*), che sottolinea con forza l'importanza di un approccio responsabile ed etico all'IA. In modo particolare, guardando alla Dichiarazione di Montreal e alla Raccomandazione dell'UNESCO, Ramos riflette sulla regolamentazione e sul ruolo che l'etica dovrebbe avere nella progettazione e nello sviluppo dei sistemi di IA.

ii) La seconda sezione – *Nuovi diritti e ricadute processuali* – inizia con il contributo di Wanda D'Avanzo (*Intelligenza artificiale e sanità digitale. Il diritto* by design *e il ruolo del* legal design *per la protezione dei dati personali*) che tocca un tema

fra i più noti e significativi: l'interazione fra IA e sfera sanitaria e la necessaria tutela dei dati personali. In questo contesto, l'A. si sofferma sul consenso al trattamento, sul c.d. *legal design*, come pure sulle disposizioni più significative del GDPR.

L'indagine continua con Vincenzo Bruno Muscatiello (Ex Machina. *Il diritto penale, all'improvviso*), che prospetta un'interessante riflessione sul rapporto fra IA e istanze punitive, mettendo in evidenza taluni aspetti della realtà artificiale nei confronti dei quali il diritto penale "tradizionale" potrebbe risultare del tutto impreparato.

Nel medesimo solco tematico, anche Lorenzo Pulito (*Previsioni algoritmiche e valutazioni prognostiche per la sostituzione della pena nella prospettiva della* symbiotic artificial intelligence) che – dinnanzi all'ambizione dell'IA di prevedere il futuro comportamento del reo, supportando, così, l'attività del giudice – smorza l'entusiasmo per il nuovo che avanza, mettendo in guardia nei confronti del possibile asservimento dell'essere umano al "regime algoritmico". L'idea che si prospetta (anche nel contesto delle nuove pene alternative e delle procedure recentemente introdotte dalla riforma Cartabia) è quella di un paradigma di interazione tra "intelligenze" (umana/artificiale) ispirato alle relazioni simbiotiche che si rintracciano in natura.

Segue lo studio di Paolo Sommaggio e Shan Ali (*Cognitive Liberty, Neurolaw, and Digital Governance: Integrating Neuroscience and Blockchain in the Era of Neuro-Rights*), che, guardando alle neuroscienze e al neurodiritto, ipotizza la necessità di prevedere una sorta di "libertà cognitiva" (come nuova forma di tutela) da includersi in una "Dichiarazione dei Neurodiritti Umani". Ciò, a partire dalla consapevolezza che la libertà cognitiva può rappresentare la nuova "arma concettuale" per difendere i diritti umani individuali dai neuro-paradigmi che, di fatto, li svuotano e li riducono a nulla.

Chiude la sezione il saggio di Radosław Strugała (*Liability for Damage Caused to the Data Subject in the Event of GDPR Violation "by Artificial Intelligence": Between Legal Discrepancies and Policy Questions*) che affronta l'importante questione della responsabilità per violazione dei dati da parte dell'IA, ipotizzando, fra l'altro, una possibile interazione fra l'applicazione

dell'articolo 82 del GDPR e un regime speciale di responsabilità pensato per far fronte ai danni causati dall'IA.

iii) Alla terza sezione – *Neuroscienze, s/oggettività e modelli relazionali* – si riferisce il saggio di Salvatore Amato (*Macchinazioni. Soggettività, agentività ed entità esotiche*), che, riprendendo il problema dell'opacità dell'IA, offre al lettore un'indagine sul confine tra la "meccanizzazione della mente" e "l'umanizzazione della macchina", giungendo a domandarsi se, forse, il richiamo – oggi sempre più frequente e diffuso – ad un'etica del *design*, ad un'info-etica, ad un'etica algoritmica, non faccia altro che reintrodurre (con variazioni e adattamenti) i principi e i valori della tradizione del diritto naturale.

Sempre in questa sezione, il contributo di Sveva D'Agosto (*La costruzione della figura femminile fra neuroni specchio e intelligenza artificiale simbiotica*) che, guardando alla costruzione del sistema simbiotico (intelligenza umana/artificiale), ragiona sull'implementazione delle rappresentazioni inautentiche, che sembrerebbero avere delle considerevoli ripercussioni sul funzionamento dei neuroni specchio. In questo contesto, ad emergere è una pericolosa deriva verso l'ingiustizia epistemica nella raffigurazione del femminile *online*, che concorre a generare dinamiche discriminatorie e un diffuso allontanamento dal riconoscimento e dall'affermazione dei diritti fondamentali delle donne.

Teresa María García-Berrio in *Ética de la virtud para una IA compasiva: Ontología, subjetividad y condición humana en la era digital*, sviluppa delle considerazioni etiche circa l'impatto che l'IA ha sul tessuto sociale delle comunità contemporanee e sul benessere umano, richiamando l'attenzione sulla necessità di recuperare (e di riaffermare) valori come la gentilezza, l'empatia e la compassione.

Il percorso di analisi continua con l'articolo di Olimpia G. Loddo (*L'agire sociale ai tempi dell'intelligenza artificiale. Il concetto di "istituzione sociale-digitale"*). Qui, il concetto di istituzioni socio-digitali – inizialmente proposto da Teubner e Backers – viene visto come una possibile risorsa concettuale per affrontare le questioni poste dall'IA con riferimento alle dina-

miche sociali. In particolare, lo studio guarda al ruolo dell'istituzione digitale nei sistemi sociali e a come la comunicazione mediata (o prodotta dall'IA) possa concorrere alla sua ridefinizione.

Ad arricchire l'indagine, il saggio di Paolo Moro (*Persona elettronica. Una finzione giuridica per l'intelligenza artificiale*), nel quale – dopo aver proceduto all'esame della soggettività robotica (e dei parallelismi epistemologici tra la macchina calcolatrice e la coscienza umana) e dopo aver criticato la tendenza ad umanizzare le macchine (che verte su una concezione estremamente riduttiva della soggettività umana) – ci si interroga sulla possibilità di prevedere una nuova "personalità elettronica".

Con il contributo di Guido Saraceni (*Intelligenza Artificiale e benessere psicologico*) la ricerca si apre, inoltre, alla speculazione psicologica e, in modo particolare, alla disamina sulle conseguenze che i modelli di IA possono avere sul benessere psicologico dell'uomo. Lo scopo dello studio è quello di fare luce sull'intricata relazione tra IA e salute mentale, mettendone in evidenza tanto i potenziali vantaggi, quanto i possibili limiti etici, sociali e tecnici.

La sezione termina con il lavoro di Maurizio Sozio (*Il deficit fronetico dello* "Jus ex machina"), in cui – preso atto dell'avvenuta realizzazione della nota profezia di Leibniz (a detta del quale, un giorno, di fronte a una controversia, le parti avrebbero potuto sedersi e affidarsi alla risoluzione tramite un calcolo) – ci si chiede se un'entità calcolatrice, un'intelligenza artificiale, possa davvero sostituire l'uomo e il suo giudizio.

I saggi contenuti in questo volume, rivisti ed aggiornati, sono apparsi per la prima volta sulla rivista "L'Ircocervo. Prima rivista elettronica italiana di Metodologia giuridica, Teoria generale del diritto e Dottrina dello Stato", nel fascicolo 1/2024 intitolato *Artificial Intelligence and Neuro-cognitive Sciences in Law: From Symbiosis to Substitution*. Il Comitato di Direzione della rivista, a cui va il nostro ringraziamento, ne ha concesso la riedizione in un nuovo progetto editoriale. Siamo anche grati al Prof. Donato Malerba, responsabile scientifico dello *Spoke 6 (Symbiotic AI) del progetto FAIR – Future AI Research (PE00000013), nell'ambito del programma MUR del PNRR finanziato dal NextGenera-*

tionEU, per avere creduto in quest'iniziativa, nella comune idea che la rivoluzione digitale necessita di un dialogo profondo ed ancora fondazionale tra diversi saperi, per aprire la strada ad una rielaborazione adattiva o ad un'inedita elaborazione concettuale delle metamorfosi che ci attendono. In questo volume si offrono più angolature del medesimo fenomeno nella consapevolezza, per richiamare Nietzsche, che non esistono fatti, ma solo interpretazioni. E forse anche questa è un'interpretazione.

Roma/Bari, gennaio 2026

Maria Novella Campagnoli e Piero Marra

I.
IA E PENSIERO GIURIDICO:
SCENARI E RISCHI ETICO-GIURIDICI

Agata C. Amato Mangiameli

Intelligenza artificiale generativa: brevi note

Abstract: **Generative Artificial Intelligence: Short Notes.** Generative Artificial Intelligence is posing various dilemmas to all of us – and to jurists in particular. On the one hand, great opportunities are opening up. On the other hand, dangers are advancing. This underscores the imperative to implement robust regulatory measures to protect human rights.

Keywords: Generative Artificial Intelligence, AI Act, Bias, Rights.

Sommario: 1. Strumenti di IA per i più vari compiti: capaci di scrivere, produrre musica, creare opere d'arte digitali – 2. Intelligenza artificiale *vs.* Intelligenza artificiale generativa. I vantaggi – 3. *Segue*: qualche rischio – 4. Conclusioni minime.

1. Strumenti di IA per i più vari compiti: capaci di scrivere, produrre musica, creare opere d'arte digitali

Tra le nuove tecnologie, un ruolo di primo piano, e al contempo una svolta significativa, rivestono gli studi e le applicazioni dell'intelligenza artificiale in pressoché tutti i campi e in particolare in quello giuridico. La ragione è evidente. Una volta la tecnologia serviva per lo più allo svolgimento di quelle attività che erano (e sono) di *routine* o sostanzialmente ripetitive, per le quali le possibili scelte/decisioni erano (e sono) in buona parte già note oppure potevano (e possono) essere determinate con un alto livello di affidabilità. Adesso, invece, l'IA può essere usata per compiere attività che coinvolgono il processo decisionale e mostra di possedere il potenziale per svolgere un lavoro di natura cognitiva che richieda un *know how* anche molto sofisticato.

Basti qui pensare all'impiego dell'IA nella creazione di contenuti e alla capacità del tutto autonoma di produrre opere, tale da sfidare il diritto d'autore, per un verso, e per l'altro, da sollevare questioni riguardanti la protezione delle creazioni medesime. Musica, letteratura, arte, sono sempre più coinvolte: qualche

volta, vengono generate da algoritmi, talaltra, sono prodotte da algoritmi e successivamente perfezionate dall'intelligenza e dalla creatività umane come risultato di un'interessante connubio/ibridazione di naturale e artificiale. E in questo andirivieni di intelligenze (umane e artificiali) viene da sé che debbano essere anzitutto riformulate le nozioni tradizionali di opera d'arte, così come di artista, di originalità e di autorialità, e che, di pari passo, siano prospettate normative consone rispetto ai nuovi strumenti (e cioè che siano non solo in grado di regolamentarne l'utilizzo, ma che sappiano anche scongiurare la totale perdita di senso di certe categorie giuridiche ed economiche classiche, laddove siano da considerare essenziali e determinanti nei rispettivi ambiti).

Del resto, così come in tema di responsabilità ci si interroga su chi debba essere chiamato in causa allorché una IA provochi un incidente, allo stesso modo quando parliamo di opere dell'ingegno non è chiaro se l'autore (l'autrice), il "creatore" (la "creatrice") debba essere considerata l'IA che ha generato l'opera, oppure se sia necessario rivolgersi al programmatore (alla programmatrice), che ha scritto l'algoritmo, piuttosto che all'umano, che, a suo modo e per scopi diversi, se ne è servito. Noto è il recente caso della scrittrice Rie Kudan[1] che ha ammesso di avere usato *ChatGPT*[2] per scrivere alcune parti del romanzo intitolato *Tokio-to-Doyo-to*, opera che peraltro ha anche ricevuto uno dei più prestigiosi premi letterari giapponesi. Allo stesso modo non è chiaro se si possa parlare di diritti degli algoritmi e, quindi, della brevettabilità delle invenzioni scaturite da sistemi di IA, senza che vi sia stato un intervento umano diretto. Significativo, per i suoi diversi risvolti e per le differenti decisioni, è proprio il caso *Dabus*[3], nel quale il suo inventore in varie domande presentate agli uffici preposti (degli Stati Uniti, dell'Unione Europea,

[1] Nota scrittrice giapponese, vincitrice della diciassettesima edizione del *Akutagawa Prize*, importante riconoscimento per la letteratura.

[2] Fra i giornali italiani a darne la notizia, cfr.: "Rie Kudan, la scrittrice vince un prestigioso premio letterario e poi confessa: 'Libro scritto con ChatGpt'", in *Corriere della Sera*, 20/01/2024.

[3] Per un'agile ricognizione della vicenda, e delle diverse questioni che ne sono derivate, rinvio, fra i vari, a B. Marone, G. Pinotti, A. Santosuosso, "L'AI può essere autore o inventore? Tutti gli interrogativi sollevati dalle decisioni Thaler/DABUS", in *Agenda Digitale*, 11/04/2023.

del Regno Unito, del Sud Africa, dell'Australia) ha chiesto il riconoscimento della qualità di inventore in capo alla macchina, pur non essendo un soggetto dotato di capacità giuridica come, invece, sarebbe richiesto dalle diverse norme nazionali sulla brevettabilità. Di qui l'interrogativo: è forse giunto il momento di prevedere una nuova e specifica categoria di inventori nella quale potere includere le stesse intelligenze artificiali?

Tanti gli esempi che possono essere richiamati, tutti estremamente significativi per le loro molteplici implicazioni. Intanto *ChatGPT*, strumento ormai virale, e poi i tantissimi altri strumenti di IA capaci di scrivere (basti pensare a: *Jasper*[4], *Amazon Lex*[5], *AI-Writer*[6], *Writer*[7]), generare immagini (tra i più noti: *Midjourney*[8], *Stable Diffusion*[9], *Dall-E*[10]), produrre musica (come ad esempio: *AIVA*[11], *Soundful*[12], *Boomy*[13], *Am-*

[4] La piattaforma ideata da *Jasper Chat*, pensata per le aziende, che utilizza il linguaggio naturale per consentire agli utenti di richiedere varie attività, come, ad esempio, scrivere un post sul blog dell'azienda stessa, prospettare degli annunci e/o, eventualmente, correggerne lo stile in modo da renderlo più adeguato e funzionale allo scopo commerciale perseguito.

[5] Si tratta di un servizio AWS che permette la creazione di interfacce di comunicazione – mediante voce e testo – in qualsiasi applicazione.

[6] Di cui è disponibile anche una versione di *Microsoft* che, stando a quanto si legge nel sito della multinazionale statunitense, promette di portare la creazione di contenuti a un livello superiore, in termini di efficacia e qualità, con una notevole riduzione delle energie e delle tempistiche richieste.

[7] Che è un rilevatore di contenuti generati dall'IA.

[8] Programma di IA capace di generare immagini a partire da un testo e che, nel giugno del 2022, la rivista inglese *The Economist* ha utilizzato per progettare la copertina di un suo numero.

[9] Utilizzato principalmente per generare immagini dettagliate a partire da descrizioni di testo, ma applicabile anche ad altre attività, fra cui, la pittura e la generazione di traduzioni da immagine.

[10] Algoritmo sviluppato da *OpenAI* e presentato al pubblico nel gennaio del 2021.

[11] Lanciata già nel febbraio del 2016, *AIVA* è il primo compositore virtuale – riconosciuto dalla società musicale SACEM – specializzato nella creazione di musica classica e sinfonica.

[12] Piattaforma musicale che consente agli artisti di trovare ispirazione e di creare nuove melodie. È possibile programmare l'IA affinché generi musica in base a dei criteri specificatamente individuati.

[13] Che consente agli utenti, non soltanto di creare brani partendo da degli stili predefiniti, ma anche di pubblicare la musica così generata sulle

per[14], *Dadabots*[15], *MuseNet*[16]), creare opere d'arte digitali, uniche e NFT (*Non-fungible Token*[17]). Si pensi, a tal proposito, a *Botto*[18], IA il cui processo creativo si basa sulla generazione di migliaia di immagini e al quale partecipa una comunità di utenti che influenza le scelte dell'algoritmo attraverso un sistema di voto. Tale meccanismo – risultato di un'intensa interazione e di una non-convenzionale cooperazione nella creazione – ha già realizzato diverse opere d'arte, dalla cui vendita ha ricavato un notevole profitto (che si aggira intorno ai 1,3 milioni di dollari).

2. Intelligenza artificiale *vs.* Intelligenza artificiale generativa. I vantaggi

Più che di IA *tout court*, si parla ormai sempre più spesso della sua forma avanzata, ovvero dell'Intelligenza Artificiale generativa[19], la cui caratteristica prima è la capacità di produrre

piattaforme di streaming, attraverso le quali, gli stessi potranno poi vedersi corrisposti i diritti d'autore.

[14] Innovativo strumento di composizione musicale che agevola i creatori di contenuti, permettendo loro di creare senza alcuno sforzo brani musicali originali.

[15] Black Metal Band non umana.

[16] Progettata da OpenAI, MuseNet è in grado di creare musica combinando stili e generi diversi, garantendo un notevole grado di originalità e di personalizzazione.

[17] Cfr. F. Macioce, voce "NFT", in A.C. Amato Mangiameli, G. Saraceni (a cura di), *Cento e una voce di Informatica giuridica*, Giappichelli, Torino, 2023, pp. 336-338.

[18] Che sfrutta – simultaneamente – gli algoritmi di machine learning VQGAN (*Vector Quantized Generative Adversarial Network*), che permettono di generare immagini somiglianti ad altre, e CLIP (*Contrastive Language-Image Pre-training*), che garantisce la corrispondenza dell'immagine generata all'*input* di testo inizialmente dato.

[19] Sotto l'egida della più generale formula "IA generativa" sono riconducibili innumerevoli modelli di IA, che vanno, da quelli più risalenti ma tuttora in uso (è il caso delle *RNN* [reti neurali ricorrenti] e delle *CNN* [reti neurali convoluzionali]), sino ai c.d. *transformer models* in grado di rappresentare le sequenze in modo ancor più flessibile e potente rispetto alle prime, motivo per cui *ChatGPT* è capace di rispondere velocemente e bene a richieste di conversazioni. Ci sono inoltre: *i)* gli *autocodificatori variazionali*, costituiti da reti di *encoder* (che sono in grado di apprendere le caratteristiche

(per l'appunto, generare) contenuti in modo totalmente autonomo. Diversamente dall'algoritmo dell'IA classica, che processa un'alta quantità di dati di esempio e, con questi, auto-apprende, riuscendo così a ricondurre un nuovo esempio a una delle casistiche di addestramento, quello di *GAI* (*Generative AI*) – pur essendo basato sempre sulla statistica – è capace di creare contenuti simili a quelli di addestramento, ma del tutto nuovi. L'intelligenza artificiale generativa ha, dalla sua, una straordinaria capacità di produrre testi scritti di qualunque genere (in forma ad esempio di racconto, poesia, lettera, romanzo) e con qualsiasi stile (sia imitando, su richiesta, la prosa di Manzoni oppure la poesia di D'Annunzio, e sia ideando uno stile tutto suo). Ha inoltre una particolare abilità nel realizzare immagini, audio e video, simili a quelli prodotti dall'uomo e, soprattutto, estremamente veritieri e credibili. Ma è soprattutto l'intelligenza artificiale conversazionale *text-oriented* che ha acceso l'immaginazione e che d'altra parte può sollevare qualche apprensione. Com'è noto, il primo consumer chatbot di IA generativa è stato rilasciato al pubblico nell'autunno del 2022, un chatbot questo basato sul modello di rete neurale GPT(-3.5 di OpenAI) che sta per *generative pretrained transformer*[20].

In realtà già molto tempo prima, fece la sua comparsa *Eliza*[21] un agente conversazionale del Massachusetts Institute of Technology, che fu prodotto a metà degli anni '60, e le cui risposte

importanti di un'immagine, di comprimere tali informazioni e di archiviarle come una rappresentazione in memoria) e da reti di *decoder* (che utilizzano quelle medesime informazioni compresse per cercare di ricreare l'originale); *ii)* i *generative adversarial networks*, costituiti da due reti neurali concorrenti (vale a dire, il generatore che crea un'immagine e il discriminatore che decide se l'immagine è reale o generata); *iii)* i *modelli di diffusione* che imparano comprimendo i dati (aggiungendo rumore di sottofondo e tentando di rigenerare l'originale).

[20] Per un approfondimento di taglio tecnico, cfr.: G. Yenduri *et al.*, "GPT (Generative Pre-Trained Transformer). A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions", in *IEEE Access*, 12 (2024), pp. 54608-54649.

[21] J. Weizenbaum, "ELIZA. A computer program for the study of natural language communication between man and machine", in *Communications of the ACM*, 9 (1966), n. 1, pp. 36-45; disponibile in rete al seguente url: https://dl.acm.org/doi/pdf/10.1145/365153.365168.

si limitavano a seguire un insieme di regole e di modelli predefiniti. Al contrario, le attuali intelligenze artificiali generative non seguono precise regole e neppure modelli predefiniti. È, infatti, attraverso l'addestramento sui dati del mondo reale che si sviluppa e procede autonomamente l'intelligenza artificiale generativa, la cui peculiarità risiede nel fatto di prospettare nuovi contenuti in risposta ai *prompt*[22]. Fra l'altro, è assai interessante osservare che neppure gli esperti di IA sanno esattamente come vengano generati i contenuti in risposta, giacché gli algoritmi sono auto-sviluppati e ottimizzati via via che il sistema è addestrato. In altre parole, rispetto all'IA tradizionale (si pensi ai sistemi di *machine learning*[23]), basata su dati etichettati utilizzando tecniche di apprendimento supervisionato e generalmente progettata per eseguire attività specifiche (ad es. rilevare frodi, guidare mezzi di trasporto, ecc.), l'intelligenza artificiale generativa è addestrata su *set* di dati grandi e diversi (a volte, ottimizzati su volumi di dati molto più piccoli correlati a una funzione specifica) e si avvale, quantomeno inizialmente, dell'apprendimento non supervisionato (nel quale in principio i dati non sono etichettati e al software IA non viene fornita alcuna guida esplicita).

Le differenze sin qui brevemente richiamate fra IA tradizionale e *GAI* avranno delle ricadute notevoli in termini di progresso in molteplici settori. Secondo le previsioni contenute in un recente *report* della McKinsey & Company[24], ad esempio, entro il 2025 oltre il 30% dei nuovi farmaci e dei materiali sarà scoperto utilizzando tecniche di intelligenza artificiale generativa; ancora, grazie agli strumenti di IA è (e sarà) possibile comprendere gli effetti dei farmaci su grandi fasce della popolazione.

[22] Si tratta di testi in linguaggio naturale che richiedono all'IA generativa di eseguire una precisa attività.

[23] Formula con la quale si designano gli algoritmi di apprendimento (poi ulteriormente distinguibili in sistemi di *supervised learning*, *unsupervised learning*, *reinforcement* learning). Per un'agile definizione, si veda G. Talamo, voce "Machine Learning", in A.C. Amato Mangiameli, G. Saraceni (a cura di), *Cento e una voce di informatica giuridica*, cit., pp. 310-313.

[24] Il *report* è disponibile e scaricabile online: https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/the-state-of-organizations-2023.

Alcuni ricercatori prevedono inoltre che nei prossimi decenni l'intelligenza artificiale supererà le prestazioni umane in molti compiti, tra cui guidare un camion (entro il 2027), lavorare nella vendita al dettaglio (entro il 2031), scrivere bestseller (entro il 2049), eseguire interventi chirurgici (entro il 2053), sostituire tutti i lavori umani (entro 122 anni)[25]. In generale l'IA generativa sarà in grado di colmare le varie lacune, grazie alle sue interfacce utente semplici e basate su chat, sarà in condizione di verificare rapidamente la presenza di errori e saprà migliorare la comunicazione, potendo, tra l'altro, tradurre il testo in lingue diverse, riuscendo ad automatizzare attività complesse e a risolvere i problemi di codice, perfezionando quest'ultimo in modo più rapido, efficiente, efficace e affidabile.

Di qui, i molteplici vantaggi dell'IA generativa, in prima battuta, nella direzione di una maggiore produttività. Basti pensare che un team di programmatori potrebbe dedicare ore e ore a esaminare un codice difettoso per tentare di risolvere ciò che è andato storto, quando invece uno strumento di intelligenza artificiale generativa potrebbe essere in grado di trovare qualunque falla o errore in pochissimo tempo, segnalandolo in tempo reale e insieme suggerendo le necessarie correzioni. Altri vantaggi sono, ovviamente, rappresentati dai costi ridotti, dalla migliore soddisfazione dei clienti, da processi decisionali più informati, dal controllo di qualità dei prodotti, e, tutto ciò, grazie alla stessa velocità dell'IA generativa e ad approcci sofisticati nei più diversi settori.

3. *Segue*: qualche rischio

Insieme alle sue notevoli prospettive in ordine alla produttività, l'intelligenza artificiale generativa porta con sé anche nuovi potenziali rischi, come l'imprecisione, le c.d. allucinazioni su informazioni false o errate, le violazioni della privacy, l'esposizione della proprietà intellettuale, e, più in generale, la capacità di provocare trasformazioni economiche e sociali non proprio (e non sempre) desiderabili. Ad esempio, è assai improbabile che

[25] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, O. Evans, "When will AI exceed Human Performance. Evidence from AI Expertes", in *Journal of Artificial Intelligence Research*, 62 (2018), pp. 729-754.

i benefici in termini di produttività dell'intelligenza artificiale generativa si realizzino senza sostanziali sforzi di riqualificazione dei lavoratori e, anche in questo caso, molti di questi saranno costretti a cambiare impiego e altri per diversi motivi non troveranno alcuna nuova occupazione.

Si tratta, allora, di interrogarsi anzitutto sull'adozione di *best practice* e di normative in grado di mitigare i rischi dell'IA generativa, nella consapevolezza che – almeno attualmente – perfino coloro che hanno progettato il processo di addestramento di ogni modello non sanno esattamente (né sono in grado di spiegare) come i modelli di IA generativa *fanno ciò che fanno*. Ed infatti, ciò che sappiamo dell'IA generativa è che essa funziona allo stesso modo del cervello umano: prevede quel che accadrà e impara dalle differenze tra le sue previsioni e la realtà dei fatti verificatisi successivamente[26]; sappiamo, inoltre, che i modelli di IA generativa muovono da una rete neurale artificiale codificata nel software e che i ricercatori studiano i modelli in base al numero di connessioni tra i neuroni (celle); sappiamo, altresì, che per effettuare previsioni il modello inserisce parole chiamate *token* ad un livello inferiore, che poi elabora e che, fatto ciò, passa il suo *output* al livello successivo, sino a quando l'*output finale* emerge dalla parte superiore della pila di neuroni artificiali. Ora nelle prime fasi di addestramento, le previsioni del modello non risultano molto accurate, ma una volta che il modello ripete tale processo per bilioni di *token* di testo, si perfeziona sempre più divenendo molto bravo nel prevedere il *token* o la parola successivi. Sappiamo tutto questo, e tuttavia come nota Thompson:

> C'è un enorme 'non lo sappiamo' nel mezzo della mia spiegazione. Quello che sappiamo è che prende tutta la domanda come una sequenza di token, e al primo livello li elabora tutti contemporaneamente. E sappiamo che poi elabora gli output da quel primo livello nel livello successivo e così via fino a quando arriva in cima allo *stack*. E poi sappiamo che usa quel livello superiore per prevedere, cioè produrre un primo token, e quel primo token è rappresentato come un dato in tutto quel sistema per produrre il token successivo,

[26] J. Hawkins, S. Blakeslee, *On intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*, Macmillan, New York, 2004.

> e così via. […] A rigor di logica, la domanda successiva è: a cosa ha pensato in tutta questa elaborazione e come lo ha fatto? Cosa hanno fatto tutti quei livelli? La risposta è: non lo sappiamo. Noi [...] non [...] lo [...] sappiamo. Puoi studiarlo. Puoi osservarlo. Ma la sua complessità va oltre la nostra capacità di analisi. È proprio come una F-MRI [risonanza magnetica funzionale] sul cervello delle persone. È la bozza più rozza di ciò che il modello ha effettivamente fatto. Non lo sappiamo[27].

A tale difficoltà si aggiunga che i modelli di intelligenza artificiale generativa possono introdurre informazioni false o fuorvianti, spesso con un tono così dettagliato e autorevole che anche gli esperti possono esserne ingannati. Allo stesso modo, i loro *output* possono contenere un linguaggio distorto o offensivo appreso dal *set* di dati con cui il modello è stato addestrato, o, ancora, i loro *output* possono contenere delle affermazioni così imprecise da avere "allucinazioni". Allo stesso modo, molti modelli sono addestrati con dati decisamente obsoleti e si rifanno, generalmente, a informazioni pubblicate fino a una certa data, tali da non essere più rilevanti o utili, molti altri, poi, nel raccogliere informazioni personali sensibili aumentano la probabilità di esposizione degli IP protetti e dei dati riservati.

D'altra parte, un *social engineering potenziato* offre anche l'occasione per attacchi informatici mirati, profittando della difficoltà di capire se si stia parlando ad esempio con un bot o con un essere umano online; un'intelligenza artificiale generativa può semplificare e velocizzare le diverse invenzioni-creazioni, ma di per sé non garantisce una qualità superiore, perché, anzi,

[27] Non lo sappiamo e anche per questo non possiamo non condividere la sorpresa che un ricercatore dell'IA riporta nel podcast *This American Life*, quando ha chiesto a GPT-4: "Dammi una ricetta per biscotti con gocce di cioccolato", scrivendo nello stile di una persona molto depressa, e l'agente conversazionale ha risposto all'incirca così sugli ingredienti necessari: una tazza di burro ammorbidito "se si può anche trovare l'energia per ammorbidirlo", un cucchiaino di estratto di vaniglia, "il falso sapore artificiale della felicità", una tazza di gocce di cioccolato semi-dolce, "piccole gioie che alla fine si scioglieranno" (tanto il brano di Thompson, quanto la domanda del ricercatore e la risposta di GPT-4, sono ripresi dall'articolo di G. Pavlik, *Che cos'è l'intelligenza artificiale generativa? Come funziona?*, apparso il 15/09/2023 su oracle.com).

modelli di IA senza una significativa collaborazione umana possono portare a prodotti che diventano standardizzati e difettano di creatività e i modelli addestrati su dati distorti (lacune, pregiudizi, contenuti dannosi) rispecchieranno nel loro *output* il *bias* che sin dall'inizio era presente, perpetuandolo di continuo. Sotto questo e altri profili, gli esseri umani rimangono gli attori principali perché si eviti che *output* difettosi si diffondano e raggiungano i clienti o persino influenzino certe politiche. E sempre agli esseri umani spetta il compito di superare i tanti possibili rischi, anche per non cadere in quel peculiare fenomeno che è chiamato dai ricercatori dell'IA "collasso del modello". Un fenomeno, questo, che potrebbe rendere i modelli di intelligenza artificiale generativa meno utili nel corso del tempo. E infatti, via via che i contenuti generati dall'intelligenza artificiale proliferano, i modelli vengono addestrati su dati sintetici, che, inevitabilmente, conterranno degli errori, ragion per cui alla fine i modelli saranno indotti a dimenticare le caratteristiche dei dati generati dall'uomo su cui sono stati originariamente addestrati. Si tratta di un problema notevole che potrebbe portare ad un vero e proprio punto di rottura nel momento in cui Internet si popolerà sempre più dei contenuti di IA, creando una sorta di ciclo di *feedback* destinato, di fatto, a degradare il modello.

4. Conclusioni minime

L'IA generativa solleva svariate questioni etiche e giuridiche. Intanto, come già detto, è alquanto problematico l'impatto dell'IA generativa sui lavoratori e, in maniera particolare, sulle loro prospettive di lavoro a lungo termine. È intuitivo, inoltre, che l'IA generativa possa avere diversi effetti in altri ambiti, tali da sollecitare interrogativi del tipo: come possiamo eliminare il potenziale *bias*? In che modo possiamo superare gli innumerevoli usi potenziali di IA, che si traducono non di rado in atti vietati[28]? A chi appartiene il lavoro generato da IA, visto che i

[28] È sufficiente pensare, a titolo d'esempio, ai video *Deepfake* che utilizzano la voce e le sembianze di qualcuno, oppure agli strumenti di *hacking* usati per migliorare gli attacchi informatici. E si consideri che la disinformazione diffusa, come pure le campagne di *social engineering*, sono solo alcune delle modalità con le quali i criminali informatici possono fare uso delle

modelli di IA generativa sono addestrati su grandi quantità di dati esterni?

La domanda, qui, da ultimo proposta, richiama i casi in cui un'azienda ottimizzi un modello con propri dati e al contempo l'output del modello includa elementi del lavoro di altre organizzazioni, sollevando questioni quali la violazione del *copyright*, il plagio, ecc. Richiama, altresì, i casi in cui i modelli di IA producano immagini, e anche per questo ormai molti artisti impegnati nei più diversi campi creativi stanno studiando il modo per impedire che il loro lavoro venga utilizzato (e pubblicato) come proprio dai sistemi di IA.

Come è chiaro, quanto più i sistemi di IA generativa si diffondono e si sviluppano in vari modi[29], tanto più le questioni si moltiplicano. Tra queste quelle maggiormente importanti sono legate all'utilizzo dei dati, immessi per l'ulteriore addestramento dell'intelligenza artificiale, perché non si sa se questi stessi dati debbano essere considerati vulnerabili e se rischino di violare il diritto alla *privacy*, quello di *copyright* e i diritti in genere.

risorse dell'intelligenza artificiale generativa. Certo, allo stato, molti modelli hanno salvaguardie, ma il punto è che queste stesse barriere non possono essere considerate perfette. Le aziende che implementano i propri modelli, infatti, devono anzitutto capire di cosa sono capaci i loro sistemi e adottare conseguentemente misure atte a garantirne un uso responsabile.

[29] Solo alcuni esempi. *Snap Inc*., l'azienda responsabile di *Snapchat*, ha implementato un chatbot chiamato *My AI*, che è basato su una versione della tecnologia GPT di *OpenAI*. Personalizzato per adattarsi al tono e allo stile di *Snapchat*, *My AI* è programmato per essere amichevole e gradevole. Gli utenti possono personalizzare il suo aspetto con avatar, sfondi e nomi e possono utilizzarlo per chattare da soli oppure in gruppo con più utenti, simulando il modo tipico in cui gli utenti di Snapchat comunicano con i loro amici. Anche *Bloomberg* ha annunciato *BloombergGPT*, un chatbot addestrato, per metà, con dati generali e, per l'altra metà, con dati proprietari di *Bloomberg* o con dati finanziari puliti. *Oracle*, invece, ha stretto una partnership con lo sviluppatore di *IA Cohere* per aiutare le aziende a creare modelli interni ottimizzati con dati aziendali privati. Più nel dettaglio, *Oracle* prevede di integrare i servizi di intelligenza artificiale generativa nelle piattaforme aziendali per aumentare la produttività e l'efficienza in tutti i processi esistenti, evitando che molte aziende debbano creare e addestrare i propri modelli da zero. Infine, si può ricordare *Slack:* ha rilasciato un chatbot che mira ad aiutare i lavoratori del *customer service* a raccogliere consigli dal *corpus* di conoscenza che risiede nei canali *Slack* istituzionali di ciascun cliente.

Non a caso l'*IA Act*[30] – che mira a promuovere la diffusione di un'intelligenza artificiale antropocentrica e affidabile, garantendo un elevato grado di protezione dei diritti fondamentali[31], e che, proprio per questa ragione, si contraddistingue per la presenza di un *risk based approach*[32], nonché per l'esplicita individuazione di un catalogo di pratiche di IA vietate[33] e ritenute ad "alto rischio"[34] – nella sua versione finale non manca certo di riservare un'attenzione particolare ai modelli di IA generativa[35].

[30] Approvato nel marzo del 2024 (e il cui testo finale, in italiano, è disponibile al seguente url https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/it/pdf).

[31] Secondo quanto dichiarato al considerando n. 1.

[32] Nella consapevolezza che l'IA – a seconda dei suoi diversi impieghi e alla luce dei suoi continui sviluppi – può comportare sugli interessi pubblici e sui diritti fondamentali innumerevoli pregiudizi materiali e immateriali (cfr., in particolare, i considerando nn. 5, 6, 7, 8).

[33] Vd. Capo II (*Pratiche di IA vietate*), art. 5.

[34] Vd. Capo III (*Sistemi ad alto rischio*), sezione 1 (*Classificazione dei sistemi di IA come "ad alto rischio"*) artt. 6 e 7, sezione 2 (*Requisiti per i sistemi di IA ad alto rischio*) artt. 8 ss., sezione 3 (*Obblighi dei fornitori e dei deployer dei sistemi di IA ad alto rischio e di altre parti*) art. 16 ss.

[35] Come si evince dalla lettura del combinato disposto dei considerando n. 99 e n. 105, nei quali si chiarisce che i modelli di IA generativa rientrano nella più ampia categoria dei modelli di IA per finalità generale, rappresentando, al contempo, un'opportunità e un rischio. Così il testo del considerando 99: "I grandi modelli di IA generativi sono un tipico esempio di modello di IA per finalità generali, dato che consentono una generazione flessibile di contenuti, ad esempio sotto forma di testo, audio, immagini o video, che possono prontamente rispondere a un'ampia gamma di compiti distinti". E così, il testo del considerando n. 105: "I modelli di IA per finalità generali, in particolare i grandi modelli di IA generativa, in grado di generare testo, immagini e altri contenuti, presentano opportunità di innovazione uniche, ma anche sfide per artisti, autori e altri creatori e per le modalità con cui i loro contenuti creativi sono creati, distribuiti, utilizzati e fruiti. Lo sviluppo e l'addestramento di tali modelli richiedono l'accesso a grandi quantità di testo, immagini, video e altri dati. Le tecniche di estrazione di testo e di dati possono essere ampiamente utilizzate in tale contesto per il reperimento e l'analisi di tali contenuti, che possono essere protetti dal diritto d'autore e da diritti connessi. Qualsiasi utilizzo di contenuti protetti dal diritto d'autore richiede l'autorizzazione del titolare dei diritti interessato, salvo se si applicano eccezioni e limitazioni pertinenti al diritto d'autore. La direttiva (UE) 2019/790 ha introdotto eccezioni e limi-

Species riconducibile al più ampio *genus* dei modelli di IA per finalità generali[36], l'IA generativa si caratterizza per la presenza di quello che viene definito come un "rischio sistemico", secondo quanto può leggersi nell'articolo 51:

> 1) Un modello di IA per finalità generali è classificato come modello di IA per finalità generali con rischio sistemico se soddisfa una delle condizioni seguenti: *a)* presenta capacità di impatto elevato valutate sulla base di strumenti tecnici e metodologie adeguati, compresi indicatori e parametri di riferimento; *b)* sulla base di una decisione della Commissione, *ex officio* o a seguito di una segnalazione qualificata del gruppo di esperti scientifici, presenta capacità o un impatto equivalenti a quelli di cui alla lettera a) […]. 2) Si presume che un modello di IA per finalità generali abbia capacità di impatto elevato a norma del paragrafo 1, *lettera a)*, quando la quantità cumulativa di calcolo utilizzata per il suo addestramento misurata in operazioni in virgola mobile è superiore a 10^{25}. […].

Condizione, questa, che implica in capo ai fornitori particolari e ben dettagliati oneri di informativa nei confronti della Commissione, che sono specificati nell'articolo 52.

Nel dettaglio, il fornitore deve informare la Commissione – senza ritardo e in ogni caso entro due settimane – dal momento in cui venga a conoscenza del fatto che il modello di IA presenta

tazioni che consentono, a determinate condizioni, riproduzioni ed estrazioni effettuate da opere o altri materiali ai fini dell'estrazione di testo e di dati. In base a tali norme, i titolari dei diritti hanno la facoltà di scegliere che l'utilizzo delle loro opere e di altri materiali sia da essi riservato per evitare l'estrazione di testo e di dati, salvo a fini di ricerca scientifica. Qualora il diritto di sottrarsi sia stato espressamente riservato in modo appropriato, i fornitori di modelli di IA per finalità generali devono ottenere un'autorizzazione dai titolari dei diritti, qualora intendano compiere l'estrazione di testo e di dati su tali opere".

[36] Come definiti dall'articolo 3 dell'*IA Act* al punto n. 63: "'modello di IA per finalità generali': un modello di IA, anche laddove tale modello di IA sia addestrato con grandi quantità di dati utilizzando l'autosupervisione su larga scala, che sia caratterizzato una generalità significativa e sia in grado di svolgere con competenza un'ampia gamma di compiti distinti, indipendentemente dalle modalità con cui il modello è immesso sul mercato, e che può essere integrato in una varietà di sistemi o applicazioni a valle, ad eccezione dei modelli di IA utilizzati per attività di ricerca, sviluppo o prototipazione prima di essere immessi sul mercato".

determinate caratteristiche che lo rendono a rischio sistemico, avendo cura di accompagnare la notifica con tutte le informazioni necessarie a dimostrarne lo *status*. È, inoltre, importante sottolineare che il fornitore di un modello di IA per finalità generali che soddisfi la condizione di cui all'articolo 51, paragrafo 1, *lettera a)*, unitamente alla notifica, ha facoltà di presentare argomentazioni sufficientemente fondate a dimostrare che – nel caso di specie e in via del tutto eccezionale – il modello di IA non dovrebbe essere classificato fra quelli a rischio sistemico. Resta il fatto che la Commissione – nel momento in cui non ritenga sufficientemente fondate tali argomentazioni – potrà comunque respingerle, provvedendo ad includere il modello di IA fra quelli a rischio sistemico. Vi è, poi, anche la possibilità che sia la Commissione stessa, *ex officio*, ad annoverare un certo modello di IA fra quelli a rischi sistemici. Ad evidenziare ulteriormente il particolare ruolo (di garanzia-monitoraggio) affidato alla Commissione, va ricordato che essa garantisce la pubblicazione e l'aggiornamento periodico di un elenco di modelli di IA per finalità generali con rischio sistemico.

Previsioni e disposizioni, queste qui richiamate, che (assieme a quelle contenute negli articoli 53, 54 e 55) testimoniano l'accresciuto impegno – dell'Unione europea e degli Stati membri – nella duplice direzione, dell'implementazione tecnologica e della tutela dei diritti e delle libertà fondamentali dei cittadini europei rispetto ai possibili pericoli insiti nei sistemi e nelle applicazioni di IA.

Massimo Farina

L'arte di non degenerare: ecosistemi tracciabili dell'IA tra degrado informativo e rischio creativo*

Abstract: **The Art of Not Degenerating: Traceable AI Ecosystems Between Information Degradation and Creative Risk.** The essay investigates the tension between "degeneration" and "creative risk" in systems of artificial intelligence for content generation, setting the philosophical notion of "strong artificial intelligence" in dialogue with the legal categories now in force. Starting from a joint reading of data quality and data provenance, it shows how recursive training cycles can lead to model collapse, with loss of informational diversity and increased distortion. From this follows the need for a "lawful diet of data" grounded in traceability of sources, lawfulness of uses, and verifiability of curatorial choices. Along this axis, the article reconstructs the principal obligations introduced by the European regulation on artificial intelligence – documentation of training data, transparency measures for synthetic content, and post-market monitoring – and aligns them with the interventions adopted in the Italian legal order in 2025: the amendment of Article 1 of the Italian Copyright Law, the introduction of Article 70-septies on text and data mining, and new criminal offences concerning deceptive alterations of image and voice. It then proposes an architecture of governance that combines technical practices (content credentials, dataset and model information sheets, registers of transformations) with organisational practices (management systems and risk-assessment procedures), so that responsibility can be demonstrated across the entire chain. The outcome is an integrated model of general safeguards and complementary safeguards capable of containing degenerations without suppressing creative initiative, thereby preserving a space for controlled exploration in research and cultural production.

Keywords: Degeneration, Artificial Intelligence, Accountability, Ethics, Risk.

Sommario: 1. Intelligenza Artificiale e risvolti degenerativi. Spunti introduttivi – 2. Dall'attitudine "dato-centrica" al pericolo di "degenerazione tecnica" – 3. Per uno sguardo alla sfera informativa... – 4. ...e alla sfera tecnico-strutturale – 5. Come scongiurare il rischio di degenerazioni morali e giuridiche? – 6. Conclusioni: verso una governance duale dell'IA fra prevenzione generale e tutela complementare.

* Il presente contributo è stato realizzato nell'ambito del Progetto "Start Up" denominato «Holmes» (*Harmonizing Ownership and Legal Measures for Ethical AI Systems*), realizzato presso il Dipartimento di Ingegneria Elettrica e Elettronica dell'Università degli Studi di Cagliari, allo scopo di ap-

1. Intelligenza Artificiale e risvolti degenerativi. Spunti introduttivi

La società contemporanea si contraddistingue – rispetto al passato – per la presenza di trasformazioni continue e rapidissimi mutamenti socioeconomici. Il risultato è uno scenario assai fluido e in continuo divenire, nella cui cornice la tecnologia riveste un ruolo decisamente primario. Del resto, la Quarta Rivoluzione Industriale[1] è marcatamente segnata dall'uso massiccio delle "macchine"[2].

In un simile contesto, l'Intelligenza Artificiale (in seguito IA) si manifesta come un orizzonte spiccatamente caleidoscopico[3], le cui strabilianti frontiere evolutive si declinano e si ritraducono in moltissimi ambiti[4] grazie all'intrinseca attitudine di modificare sia il nostro *modus vivendi*[5], sia la nostra *ars pensandi*[6].

profondire il rapporto tra intelligenza artificiale generativa e diritto d'autore, binomio che coinvolge non soltanto aspetti giuridici, ma anche molteplici questioni tecnologiche ed etiche.

[1] Cfr., J. Barrat, *La nostra invenzione finale. L'intelligenza artificiale e la fine dell'età dell'uomo*, Nutrimenti, Roma, 2019, p. 9.

[2] Per un maggior approfondimento s.v., A.C. Amato Mangiameli, "Algoritmi e big data. Dalla carta alla robotica", in *Rivista di Filosofia del diritto*, VIII (2019), n. 1, p. 108; G. Pasceri, *Intelligenza artificiale, algoritmo e machine learning*, Giuffrè, Milano, 2021, p. 11, in cui l'A. sottolinea come "l'intelligenza artificiale è figlia del naturale sviluppo dell'innovazione tecnologica come conseguenza ordinaria della crescita scientifica, tecnica e culturale dell'uomo. L'errore, in cui spesso si incorre, è quello di identificarla, diversamente, come un processo tecnologico moderno frutto della capacità di calcolo e dell'informatizzazione dei processi".

[3] L'analisi condotta nel presente contributo ha ad oggetto l'"intelligenza artificiale forte" (di cui è un'esplicazione l'intelligenza artificiale generativa), così definita per la prima volta da John R. Searle nel lontano 1980: J.R. Searle, "Minds, Brains and Programs", in *Behavioral and Brain Sciences*, 3 (1980), n. 3, pp. 417-424. Sugli stadi evolutivi dell'intelligenza artificiale si veda M. Farina, "Brevi riflessioni sullo status delle 'persone elettroniche'", in *L'Ircocervo*, 20 (2021), n. 2, pp. 106-126.

[4] Per una panoramica, si v. M.L. Montagnani, "Governance societaria e governance dell'intelligenza artificiale", in *Mercato, concorrenza, regole*, (2022), n. 2, pp. 271-290.

[5] A. Gehlen, *L'uomo nell'era della tecnica. Problemi socio-psicologici della civiltà industriale*, SugarCo, Milano, 1967, p. 12.

[6] Cfr., H. Jonas, *Dalla fede antica all'uomo tecnologico* (1974), trad. it., il Mulino, Bologna, 1991, p. 9; A. Longo, G. Scorza, *Intelligenza artificiale.*

Di qui, come è intuitivo, l'urgenza che il suo sviluppo e la sua implementazione siano accompagnate da un'attenta e costante riflessione di ordine etico-giuridico[7]. Va detto subito che – diversamente dai precedenti sistemi esperti che erano basati su *set* predefiniti di regole "*if-then*"[8] – le nuove forme di IA tentano di ovviare ai limiti operativi e di apprendimento, per supportare in modo ancor più performante le attività decisionali umane[9],

L'impatto sulle nostre vite, diritti e libertà, Mondadori, Milano, 2020, pp. 57-58.

[7] G. Li, X. Deng, Z. Gao, F. Chen, "Analysis on Ethical Problems of Artificial Intelligence Technology", in *Proceedings of the 2019 International Conference on Modern Educational Technology*, (2019), pp. 101-105. Cfr. anche A.C. Amato Mangiameli, *Corpi docili Corpi gloriosi*, Giappichelli, Torino, 2007; C. Bottari (a cura di), *La salute del futuro. Prospettive e nuove sfide del diritto sanitario*, BUP, Bologna, 2020. Per una valutazione dell'impatto dell'integrazione dell'IA in campo medico si rinvia al primo Rapporto dell'Organizzazione Mondiale della Sanità (OMS) che illustra i benefici dell'IA, riducendo al minimo i suoi rischi ed evitando le sue insidie. V. World Health Organization, *Ethics and governance of artificial intelligence for health: WHO guidance*, 2021. Si veda, *ex multis*, A. Pajno, F. Donati, A. Perrucci (a cura di), *Intelligenza artificiale e diritto: una rivoluzione? Diritti fondamentali, dati personali e regolazione*, I, il Mulino, Bologna, 2022; L. D'Avack, "La rivoluzione tecnologica e la nuova era digitale: problemi etici", in U. Ruffolo (a cura di), *Intelligenza artificiale. Il diritto, i diritti, l'etica*, Giuffrè, Milano, 2020, p. 19.

[8] Sul punto cfr. T. Davenport, R. Kalakota, "The Potential for Artificial Intelligence in Healthcare", in *Future Healthcare Journal*, 6 (2019), n. 2, pp. 94-98. Come precisato dagli Autori (p. 95) "Expert systems based on collections of 'if-then' rules were the dominant technology for AI in the 1980s and were widely used commercially in that and later periods. In healthcare, they were widely employed for 'clinical decision support' purposes over the last couple of decades 5 and are still in wide use today. Many electronic health record (EHR) providers furnish a set of rules with their systems today. Expert systems require human experts and knowledge engineers to construct a series of rules in a particular knowledge domain. They work well up to a point and are easy to understand. However, when the number of rules is large (usually over several thousand) and the rules begin to conflict with each other, they tend to break down. Moreover, if the knowledge domain changes, changing the rules can be difficult and time-consuming. They are slowly being replaced in healthcare by more approaches based on data and machine learning algorithms".

[9] A. Vespigani, *L'algoritmo e l'oracolo*, Il Saggiatore, Milano, 2020.

ricorrendo altresì a una presentazione agile e fruibile delle informazioni ritenute maggiormente rilevanti[10].

Vien da sé che il dibattito sull'impiego dell'IA si sia incrementato notevolmente proprio nel corso dell'ultimo ventennio[11], in relazione alle sue rinnovate potenzialità operative rispetto all'attività umana[12], nonché al possibile rapporto/raffronto tra il pensiero "umano" e l'influenza/ingerenza della tecnica[13] dando il via ad una querelle tra "apocalittici" e "integrati"[14].

[10] In tal senso, Report del Ministero della salute, Consiglio Superiore di Sanità, *I sistemi di intelligenza artificiale come strumento di supporto alla diagnostica*, del 9 novembre 2021, p. 7.

[11] Cfr. R. Trezza, *Diritto e Intelligenza artificiale. Etica. Privacy. Responsabilità. Decisione*, Pacini Giuridica, Pisa, 2020; U. Pagallo (a cura di), *XXVI lezioni di Diritto dell'Intelligenza Artificiale*, Giappichelli, Torino, 2021.

[12] Si v., tra i vari, D. Parisi, "Dodici differenze tra l'intelligenza artificiale e la vita artificiale", in *Sistemi intelligenti*, XVII (2005), n. 1, pp. 155-157; nonché P. Becchi, "Homo sapiens, homo cyber, postorganico. Derive o approdi?", in *Materiali per una storia della cultura giuridica*, (2015), n. 2, pp. 587-596.

[13] Cfr. P.E. Agre, *Computation and Human Experience*, Cambridge University Press, Cambridge, 1997; J. Copeland, *Artificial Intelligence, a Philosophical Introduction*, Black-well, London, 1993. Giova rammentare, a tal proposito, le riflessioni di Natalino Irti, secondo il quale "questa intrinseca connessione fra pensare umano e tecnica fu già intuita da Goethe in una delle prime scene del Faust" e, nel soffermarsi sul rapporto tra diritto e tecnica, ritiene che "il diritto appartiene al pensiero calcolante, che richiede oggettività e impersonalità. Questa secolare tensione, questo sforzo di trascendere le oscure incognite del soggettivismo, è ora approdato alla funzionalità tecnica del robot, che non è il super-uomo né l'anti-uomo, ma appartiene, anch'esso, alla storia dell'uomo": N. Irti, "Il tessitore di Goethe (Per la decisione robotica)", in A. Carleo (a cura di), *Decisione robotica*, Bologna, il Mulino, 2019, pp. 17 ss. Per approfondimenti, si rimanda a R. Borruso, S. Russo, C. Tiberi, *L'informatica per il giurista. Dal Bit a internet*, Giuffrè, Milano, 2009; D. Casalegno, *Uomini e computer. Storia delle macchine che hanno cambiato il mondo*, Hoepli, Milano, 2010; L. Gamberini, L. Chittaro, F. Paternò (a cura di), *Human-computer interaction. Fondamenti teorici e metodologici per lo studio dell'interazione tra persone e tecnologie*, Pearson, Milano, 2012. Da ultimo, J. Barrat, *op. cit.*; E. Brynjolfsson, A. McAfee, *La macchina e la folla. Come dominare il nostro futuro digitale*, Feltrinelli, Milano, 2020. Si v., altresì, E. Perucchietti, *Cyberuomo. Dall'intelligenza artificiale all'ibrido uomo-macchina*, Arianna, Bologna, 2019.

[14] Il richiamo va direttamente a U. Eco, *Apocalittici e integrati*, Bompiani, Milano, 1964; Cfr. anche F. Borgia, *L'uomo senza immagine*, Mimesis, Milano, 2006, p. 129.

Per un verso, l'ascesa dell'IA ha dato l'abbrivio a una nuova era del progresso tecnologico, nella quale i sistemi intelligenti possono analizzare dati complessi, automatizzare processi e assistere (o anche sostituire) gli esseri umani in una vasta ed eterogenea gamma di attività e di compiti[15]. In questo senso, e dal punto di vista prettamente funzionale, l'IA si propone come un utile strumento di facilitazione per l'interazione uomo-macchina (cosiddetta intelligenza artificiale "relazionale"), tuttavia, l'impiego sempre più crescente di tali sistemi ha comportato un uso talvolta improprio della tecnologia sollecitando "quesiti che l'uomo non si era mai posto"[16].

Per un altro verso, il confronto tra l'intelligenza umana e quella artificiale si è intensificato, fino a voler comprendere come quest'ultima possa essere "integrata" nel tessuto legale e sociale, contribuendo al bene comune senza però compromettere i diritti individuali[17]. Così, ecco che alle promesse di innovazione ed efficienza si affiancano nuove sfide e rischi[18] inediti, tra cui spiccano quelli connessi all'impiego dell'intelligenza artificiale "generativa", densa di criticità[19] se si guarda al suo innegabile risvolto "degenerativo". Il fenomeno è certamente articolato, multisfaccettato e tuttora in fase di prima indagine, ma può

[15] S.J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, London, 2016.

[16] In tal senso U. Galimberti, *L'uomo nell'età della tecnica*, Alboversorio, Milano, 2011.

[17] Cfr. A.C. Amato Mangiameli, "IA e diritto. In luogo di una introduzione", in *Journal of Ethics and Legal Technologies*, 5 (2023), n. 2, pp. 81-93.

[18] Sulla percezione del rischio tecnologico, cfr. M. Ferrari, "Progresso tecnologico, macchine intelligenti e autonomia robotica: la 'percezione' del rischio a fondamento delle tutele assicurative e di sicurezza sociale", in *Il Foro it.*, (2021), n. 5, pp. 263 ss. Cfr. S. Quintarelli, F. Corea, F. Fossa, A. Loreggia, S. Sapienza, "AI, profili etici. Una prospettiva etica sull'Intelligenza artificiale: principi, diritti e raccomandazioni", in *BioLaw Journal – Rivista di Biodiritto*, (2019), n. 3, pp. 183-204.

[19] Sebbene il fenomeno dell'IA Generativa sia in costante evoluzione, il dibattito scientifico si manifesta costante da oltre un quinquennio. Si v., *ex multis*, U. Ruffolo, "Piattaforme, A.I. generativa e libertà di (formazione e) manifestazione del pensiero. Il caso ChatGPT", in *Giurisprudenza italiana*, (2024), n. 2, pp. 472-480; M. Ferrari, "Intelligenza artificiale e titolarità dei diritti d'autore: il problema del 'tasso di creatività'", in *Il Foro it.*, (2023), n. 11, pp. 373-379.

essere ricondotto alla tendenza dei modelli di IA a manifestare risposte/comportamenti indesiderati dal punto di vista tecnico e/o giuridico, che ne rischiano di alterare e di compromettere progressivamente l'affidabilità e l'efficienza[20] stessa.

Sotto il profilo tecnico, la degenerazione può derivare da molteplici fattori, tra cui l'accumulo di distorsioni nei dati di addestramento, la complessità eccessiva del modello computazionale adottato, nonché la fragilità dei componenti *hardware* e *software*. Esattamente come accade alle capacità cognitive umane anche quelle di un sistema di IA tendono a deteriorarsi e possono arrivare a degenerarsi nel tempo a causa del combinarsi di tutta una serie di fattori e di fenomeni progressivi. Basti pensare al fatto che l'"anomalia", dovuta alla possibile autoreplicazione di errori o di malfunzionamenti, si ripercuote inevitabilmente sul risultato, comportando a sua volta il rischio di aprire il varco a decisioni che non possono considerarsi conformi agli obiettivi prefissati, o che, addirittura, possono rivelarsi pregiudizievoli.

Sussiste, altresì, una degenerazione di natura giuridica che consiste nell'utilizzo di sistemi di IA in violazione di disposizioni normative esistenti, poste a presidio di posizioni giuridiche soggettive: le più note riguardano la violazione del diritto d'autore e dei dati personali.

Da qui, sorge – in particolare per il filosofo del diritto e per l'informatico giuridico – l'esigenza di interrogarsi su quale possa essere la "cura" da somministrare all'IA per prevenire fenomeni degenerativi (disfunzionali e controproducenti) e, al contempo, per garantire che i sistemi intelligenti operino in modo lecito, etico e responsabile, rispettando i criteri di giustizia e non andando a ledere in alcun modo i diritti umani. Per affrontare questa sfida, è fondamentale, innanzitutto, identificare le potenziali cause di degenerazione dell'IA. Questo ci permetterà di comprendere le principali ricadute etico-giuridiche del suo utilizzo e, conseguentemente, di pensare e di sviluppare adeguate misure preventive.

[20] G. Marcus, "Deep Learning: A Critical Appraisal", in *arXiv*, 2018 (arXiv:1801.00631).

2. Dall'attitudine "dato-centrica" al pericolo di "degenerazione tecnica"

Le criticità connesse allo sviluppo e all'utilizzo dell'IA, quantunque discendano dalle caratteristiche intrinseche alla tecnologia stessa, meritano opportuna considerazione per le motivazioni già esposte. Si badi: i risvolti critici della diffusione e dell'uso sempre più massiccio e pervasivo dell'IA – ancorché siano un precipitato della stessa struttura tecnica – invocano una sensibilità e un'attenzione particolare da parte del giurista, che, se, da un lato, non può non accogliere con favore le innumerevoli facilitazioni apportate dall'IA, dall'altro lato, non può restare silente (o peggio del tutto indifferente) rispetto al profilarsi di pericoli in termini di autonomia decisionale e di effettivo e pieno godimento dei diritti.

È per queste ragioni che, oggi, il fenomeno della degenerazione dell'IA è tra quelli maggiormente preoccupanti ed è espressione di tutte quelle possibili manifestazioni patologiche (di natura tecnica o giuridica) che i sistemi di IA possono presentare. Ora, se è vero che la degenerazione tecnica dell'IA richiama l'idea generale di un declino progressivo – molto simile a quello osservato nelle malattie neurodegenerative che colpiscono gli esseri umani (e gli altri esseri viventi) – e costituisce una conseguenza, per così dire, fisiologica della struttura stessa del sistema, è altrettanto vero che un discorso ben diverso deve essere fatto per la degenerazione giuridica. Ciò, in modo particolare, se si guarda all'attitudine "patologica" tipica della dialettica giuridica, che, coinvolgendo un assetto valoriale ben preciso, ne implica il necessario rispetto.

Procedendo, dunque, lungo la prima direttrice (ovvero quella propriamente tecnica), si osserva che i sistemi di IA si connotano per una latente e sempre maggiore vocazione "dato-centrica"[21].

[21] Sull'origine del concetto di IA "dato-centrica" si veda C. Anderson (2008), "The end of theory: The data deluge makes the scientific method obsolete", in *Wired magazine*, 16 (07). Recuperato da www.wired.com. L'autore, in tempi non sospetti, anticipava che l'analisi massiccia di dati avrebbe superato ogni teoria riguardante la previsione e la distribuzione dei comportamenti e delle tendenze umane. Seguendo quella scia, ben cinque anni dopo (esattamente il 4 febbraio 2013), David Brooks, per la prima volta, in un articolo del New York Times, intitolato *The Philosophy of Data*, utilizzò il termine "*data-i-*

Anzi, più propriamente, l'esigenza di apprendere (e muovere da) una notevole massa critica di dati è prodromica alla funzionalità della tecnologia stessa, che implica una vera e propria "dipendenza dal" dato, anziché una mera "vocazione al" dato.

In altre parole, non si può non riconoscere che, nella pratica, i sistemi di IA sono fortemente dipendenti sia dai dati usati nella fase iniziale – di addestramento del sistema (c.d. *training*) –, sia da quelli che il sistema stesso raccoglie, in maniera autonoma, nella fase di interazione con l'esterno. Ecco, dunque, che le radici genetico-tecniche della degenerazione artificiale si collegano indiscutibilmente ad un ulteriore problema di carattere epistemologico[22], che, ovviamente, trova proprio nei dati il suo punto di massima esplicitazione[23].

Dal punto di vista teleologico, invece, assicurare la qualità e la sicurezza dei dati rappresenta un momento cruciale per il funzionamento dell'IA. Del resto, ogniqualvolta i dati siano difettosi, insufficienti o persino mancanti, oppure vi sia un errore nella comunicazione degli stessi, vien da sé che il funzionamento dell'IA risulterà inevitabilmente alterato e/o del tutto compromesso.

3. Per uno sguardo alla sfera informativa...

Se si considera il ben noto trinomio dato-informazione-conoscenza[24], vien da sé che la degenerazione dei sistemi di IA

sm", poi ripreso nel 2015 da S. Lohr, *Data-ism: Inside the big data revolution*, Oneworld Publications, London, 2015 e da Y.N. Harari, *Homo Deus: A Brief History of Tommorrow*, Harper, New York, 2015. Cfr. D. Brooks (2013), "The Philosophy of Data", in New York Times. Recuperato da www.nytimes.com.

[22] E. Amaturo, B. Aragona, "Per un'epistemologia del digitale: note sull'uso di big data e computazione nella ricerca sociale", in *Quaderni di Sociologia*, (2019) 81-LXIII, pp. 71-90; A. Di Prospero, "Intelligenza Artificiale e inferenza non-monotona: modelli culturali e questioni epistemologiche", in *Scienza&Filosofia*, (2022), n. 27, pp. 179-196.

[23] Ulteriori ed altrettanto rilevanti sono le questioni semantiche di cosa debba intendersi per "intelligenza artificiale", e, soprattutto, per "intelligenza". Sul punto, celebre è il pensiero dello psicologo statunitense Howard Gardner, ideatore della cosiddetta "teoria delle intelligenze multiple", per il quale non esisterebbe proprio nemmeno una facoltà precisa chiamata intelligenza: H. Gardner, *Formae Mentis. Saggio sulla pluralità dell'intelligenza*, trad. it., Feltrinelli, Milano, 2013.

[24] Cfr. A.C. Amato Mangiameli, M.N. Campagnoli, *Strategie digitali. #diritto_educazione_tecnologie*, Giappichelli, Torino, 2020, p. 3, partecipa "alla creazione di nuovi scopi e raggiunge lo stadio retorico, che, in quanto

assuma anzitutto una declinazione informativa[25], che a partire dal presupposto – e parallelamente dal processo – di elaborazione e di diffusione dell'informazione, si ritraduce nel segno della "disinformazione"[26].

In questo senso, una delle principali cause di degenerazione dell'IA risiede nell'utilizzo di dati di addestramento contaminati o non sufficientemente rappresentativi, che possono dar vita a *bias* e distorsioni nei modelli risultanti[27]. Sono, inoltre, sempre più frequenti i casi di addestramento malevolo volontario[28] –

tale, schiude il virtuale come mondo autonomo e favorisce differenti modalità di conoscenza, con propri stili, criteri di valutazione e valori". Si v. anche, M. Palmirani, "*Big Data* e conoscenza", in *Rivista di filosofia del diritto*, IX (2020), n. 1, pp. 73-91.

[25] Cfr. C. Casonato, "Per una intelligenza artificiale costituzionalmente orientata", in A. D'Aloia (a cura di), *Intelligenza artificiale e diritto. Come regolare un metodo nuovo*, FrancoAngeli, Milano, 2020, pp. 131-166. L'A. rileva che le tecnologie dell'informazione della comunicazione (ICT) danno prova che il consenso informato mostra tutte le sue debolezze; "debolezze che lo hanno condotto ad una trasformazione di senso, tanto da essersi convertito da strumento a tutela della riservatezza della persona a mezzo attraverso il quale quotidianamente esponiamo i nostri dati a chiunque ci fornisca un servizio attraverso Internet". Si v. anche M. Capparelli, "Disinformazione online, intelligenza artificiale e ruolo dell'autoregolamentazione", in *Giurisprudenza italiana*, (2024), n. 2, pp. 480-483.

[26] Secondo E. Esposito, *Artificial Communication: How Algorithms Produce Social Intelligence*, MIT Press, Cambridge Ma, 2022, dovrebbe parlarsi non in termini di "intelligenza" artificiale, ma come "comunicazione" artificiale. Si veda, altresì, L. Bittman, "The Use of Disinformation by Democracies", in *International Journal of Intelligence and CounterIntelligence*, 4 (1990), n. 2, pp. 243-261; R. Andriani, "Libertà di espressione e disinformazione. Un conflitto del passato riapparso nel contesto digitale", in *Rivista di scienze della comunicazione e di argomentazione giuridica*, 13 (2021), n. 2, pp. 91-102.

[27] T. Bolukbasi, K.W. Chang, J.Y. Zou, V. Saligrama, A. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings", in *Advances in neural information processing systems*, (2016), n. 26, pp. 4349-4357.

[28] Tra le tecniche di attacco recentemente scoperte per ottenere risultati sbagliati, vi è quella denominata "*prompt-specific poisoning attack*". Il fenomeno è stato analizzato e spiegato da un gruppo di ricercatori dell'Università di Chicago: S. Shan, W. Ding, J. Passananti, H. Zheng, B.Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models", in *arXiv*, 2024 (arXiv:2310.13828).

veri e propri attacchi volti ad "avvelenare"[29] i dati (*data poisoning*) in ingresso e, di conseguenza, in uscita – consistenti nella somministrazione all'algoritmo di dati fallati perché non corretti o non completi.

Ma, si osservi, il fenomeno disinformativo presenta anche un'ulteriore origine che risiede nella necessità di un continuo aumento del volume di dati disponibili per proseguire il processo di addestramento stesso nell'intento di proporre modelli sempre più performanti (ovviamente a patto che i dati immessi siano di qualità). Come è *ictu oculi* evidente, questa crescente ricerca di dati, che tende pressoché all'infinito, fa affiorare un dilemma fondamentale: arriveremo al momento in cui le risorse informative a disposizione dei modelli non saranno più sufficienti per garantire il loro sviluppo ottimale? Il ritmo della raccolta dei dati con cui avanzano i sistemi di IA, infatti, è di gran lunga superiore alla produzione di informazioni accessibili sul *web*. Ma se è così, non ci si può non avvedere che ciò comporterà l'inevitabile rallentamento del ritmo di addestramento dei modelli stessi e la conseguente limitazione delle potenzialità dell'IA nel suo complesso. Di qui, la necessità di escogitare soluzioni innovative di approvvigionamento da fonti alternative di informazioni per garantire che l'IA continui a progredire non perdendo (ma, anzi, incrementando via via) la sua capacità di rispondere alle crescenti esigenze delle applicazioni reali.

Una delle proposte e dei tentativi più frequenti di soluzione è quello di ricorrere all'utilizzo dei cosiddetti dati sintetici[30], ossia ai dati generati dal modello di IA sulla scorta dei dati originali.

[29] Per una disamina ad ampio spettro sull'avvelenamento dei dati, non soltanto in ambito IA, e sulle sue conseguenze, si veda G. Ziccardi, *Dati avvelenati. Truffe, virus informatici e falso online*, Raffaello Cortina, Milano, 2024.

[30] Una definizione di "dati sintetici" è rinvenibile nel *Glossary Of Statistical Terms*, emanato dall'OCSE nel 2007, ove si precisa come il termine sia espressivo di un approccio "to confidentiality where instead of disseminating real data, synthetic data that have been generated from one or more population models are released" (p. 768). Per un approfondimento tecnico sulla creazione e l'utilizzo di dati sintetici si rimanda a S.I. Nikolenko, *Synthetic Data for Deep Learning*, Springer, Cham, 2021, pp. 139 ss. e pp. 269 ss. Sul punto si v. anche M.G. Peluso, "Intelligenza Artificiale e dati di qualità: la tecnologia come valido alleato", in *MediaLaws*, (2022), n. 2, pp. 322-337.

Certo, l'uso di dati sintetici è destinato ad ampliare il *dataset*, ma, trattandosi di informazioni non reali, questi stessi dati potrebbero anche determinare a loro volta delle disfunzioni informative. Un recente studio[31], effettuato da un gruppo di ricercatori della Rice University in collaborazione con alcuni studiosi di Stanford, ha constatato che è sempre più frequente l'impiego di *dataset* di addestramento composti da dati sintetici generati da altri modelli generativi: in pratica i modelli generativi consumano sempre più i risultati di altri modelli generativi. Il fenomeno è stato definito MAD (*Model Autophagy Disorder*) e viene descritto come un ciclo autofagico o semplicemente di autoconsumo. Il gruppo di ricercatori, evocando l'analogia con la malattia della mucca pazza, ha osservato che, dopo un certo numero di cicli di autoconsumo, diversi da modello a modello, il sistema impazzisce, o, meglio, comincia a restituire *output* errati, talvolta persino discriminatori, finendo per disattendere del tutto il risultato pratico inizialmente perseguito.

Vien da sé che si tratta di un problema che si acuisce con entità e implicazioni diverse a seconda del settore preso in considerazione. Nell'ambito della diagnostica medica, ad esempio, la manipolazione dei dati reali mediante tecniche di IA potrebbe comportare il concreto rischio di perdita di alcuni elementi indispensabili per l'indagine medica, rendendo i dati così "creati in vitro" non più idonei (né utili) allo scopo per il quale erano stati inizialmente raccolti.

Va chiarito: quanto detto non ha ovviamente l'obbiettivo di "mettere al bando" e condannare l'impiego dei dati sintetici. Al contrario, difatti, in alcuni casi, la loro produzione può essere persino necessaria per soddisfare una maggiore tutela degli individui (i cosiddetti "interessati", secondo la locuzione del GDPR) a cui si riferiscono i dati personali reali utilizzati in fase di addestramento. I dati sintetici, difatti, proprio perché sono fittizi (ma comunque realistici), non possono essere classificati come dati personali (perché non identificano o non rendono identificabile, direttamente o indirettamente, una persona fisica) di modo che – grazie al loro impiego – il rischio di illecito trattamento

[31] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, *et al.*, "Self-consuming generative models go MAD", in *arXiv*, 2023 (arXiv:2307.01850).

diviene minimo anche negli usi secondari che vengono fatti dei dati raccolti. Nondimeno, resta il fatto che la scarsità (qualitativa e quantitativa) dei *dataset* può causare un eccessivo adattamento del modello con conseguente perdita della capacità di generalizzare su nuovi *set* di dati, cioè dati non visti durante la fase di addestramento (*overfitting*)[32]. Per meglio intendere, si consideri che il vero punto di forza dei sistemi di IA risiede proprio nella capacità di generalizzare, ossia di operare correttamente su dati nuovi e mai visti prima. Essi non possiedono una vera comprensione semantica e anche quando producono informazioni apparentemente intelligenti, ciò è dovuto principalmente alla loro capacità di elaborare e manipolare dati esterni. Sono impermeabili alla memorizzazione passiva della sintassi, della semantica o della pragmatica dei dati di addestramento. Il loro apprendimento consiste nell'acquisizione di modelli e nell'estrazione di regole per generare un risultato coerente dal punto di vista sintattico, semantico e pragmatico. Al contrario, le attività di memorizzazione non presentano lo stesso grado di estro intelligente e, quando diventano prevalenti, il sistema è colpito da una "luce troppo forte, [che] offende la sua vista e lo rende quasi cieco, [ossia] incapace di vedere gli oggetti che generano le ombre a lui più famigliari"[33].

In situazioni diametralmente opposte, quando il modello non impara abbastanza dai dati di addestramento, può determinarsi una scarsa performance sia sul *dataset* di addestramento che su quello di test (*underfitting*)[34].

[32] Diverse possibili cause di *overfitting* sono state segnalate nella letteratura, tra cui: la complessità eccessiva del modello, che lo porta a modellarsi troppo strettamente ai dati di addestramento; il *dataset* di addestramento limitato, che può portare il modello a non avere una rappresentazione adeguata della variabilità complessiva del fenomeno che si vuole modellare; la durata eccessiva dell'addestramento, che lo induce a memorizzare piuttosto che imparare a generalizzare. Per approfondimenti, si veda C. Howard, "Choices Matter When Training Machine Learning Models For Return Prediction", in *Financial Analysts Journal*, Forthcoming, recuperabile da SSRN, https://ssrn.com/abstract=4497739 o da http://dx.doi.org/10.2139/ssrn.4497739.

[33] Il riferimento è al mito della caverna di Platone.

[34] Tra le principali cause dell'*underfitting* si riscontrano: eccessiva semplicità del modello rispetto alla complessità dei dati; insufficienti o inappropriate caratteristiche informative, che impediscono di fare previsioni accurate; addestramento insufficiente (ad esempio, pochi cicli di addestramento), che impedisce di apprendere adeguatamente dai dati. Per approfondimenti,

Si badi, i due aspetti patologici qui brevemente descritti si collocano all'origine del cosiddetto "stato allucinatorio dell'intelligenza artificiale" (*AI Hallucination State*)[35], caratterizzato dalla produzione di *output* inesatti o, più in generale, non corrispondenti al *set* di dati di addestramento o alle previsioni del modello impiegato[36]. Ciò compromette la validità dello strumento tecnologico rendendolo occulto fautore di disinformazione, con evidenti ricadute etiche.

Le "allucinazioni" (*hallucination*), talvolta simili ai deliri paranoici umani[37], si manifestano in diverse varianti e declinazioni. Paradigmatico è ciò che può accadere nei sistemi di riconoscimento di immagini – largamente adoperati nel settore sanitario – che potrebbero identificare elementi e oggetti non presenti nella scena o anche suggerire la generazione di oggetti incoerenti con le caratteristiche reali.

4. ...e alla sfera tecnico-strutturale

Va detto che il fenomeno degenerativo qui in esame può essere frutto anche di aspetti strutturali dell'intero processo di estrazione ed elaborazione dell'informazione. Si pensi ai cosid-

si veda H. Zhang, L. Zhang, Y. Jiang, "Overfitting and underfitting analysis for deep learning based end-to-end communication systems", in *2019 11th international conference on wireless communications and signal processing (WCSP)*, 2019, pp. 1-6.

[35] In argomento, si vedano, tra i tanti: Y. Bang, *et al.*, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity", in *arXiv*, 2023 (arXiv:2302.04023); N. Dziri, S. Milton, M. Yu, O. Zaiane, S. Reddy, "On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?", in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 5271-5285; G. Eysenbach, "The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers", in *JMIR Medical Education*, 9 (2023), n. 1 (https://doi.org/10.2196/46885).

[36] C. Cilardo (2023), "Ma perché i chatbot hanno (così tante) allucinazioni", in *AgendaDigitale*. Recuperato da www.agendadigitale.eu.

[37] B.A. Huberman, S. Mukherjee (2023), "Hallucinations and Emergence in Large Language Models", in *SSRN*, September 18, 2023. Recuperato da www.ssrn.com.

detti scenari *black box*, in cui si inseriscono i dati per ottenere un risultato specifico senza alcuna consapevolezza del processo che lo determina[38]. Con le tecniche di *deep learning*[39], infatti, le macchine sono in grado di sviluppare un apprendimento "automatico" (più che "autonomo"), la cui opacità e scarsa trasparenza di funzionamento non permette di verificare appieno la correttezza dei risultati ottenuti.

Questa situazione dà vita, tra gli altri, anche alcuni limiti di natura applicativa[40]. La maggior parte dei timori legati all'applicazione di tali tecnologie sono connessi alla scarsa prevedibilità dei risultati che, in virtù dell'autonomia[41] delle macchine, pos-

[38] Per una prima panoramica dei problemi sollevati dall'opacità dei sistemi di intelligenza artificiale, v. Y. Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation", in *Harvard Journal of Law & Technology*, 31 (2018), n. 2, pp. 889-938. L'espressione *black box* è mutuata da F. Pasquale, *The Black Box Society*, Harvard University Press, Cambridge (MA), 2015.

[39] I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, MIT press, Cambridge (MA), 2016.

[40] Per maggiori dettagli, si vedano D.E. Goldberg, J.H. Holland, "Genetic algorithms and machine learning", in *Machine learning*, 2 (1988), pp. 95-99; P. Čerka, J. Grigienė, G. Sirbikytė, "Liability for damages caused by artificial intelligence", in *Computer Law & Security Review*, 31 (2015), n. 3, pp. 376-389.

[41] A tal proposito, guardando non soltanto allo stato dell'arte ma anche alle possibili evoluzioni future dell'IA, è utile ricordare la distinzione tra: *Artificial Narrow Intelligence* (ANI); *Artificial General Intelligence* (AGI); e *Artificial Super Intelligence* (ASI). Le intelligenze del primo tipo (cosiddette limitate), le uniche attualmente disponibili, sono dotate di capacità computazionale per eseguire, in modo efficiente compiti unici (non sono versatili), come il tracciamento delle pagine, il gioco degli scacchi, il riconoscimento dei numeri scritti a mano, etc. Le intelligenze generali (secondo tipo), invece, ripropongono il concetto originario di intelligenza, traducendolo in algoritmi con prestazioni equivalenti a quelle di un essere umano e caratterizzati da una competenza volutamente programmata in un unico dominio ristretto. Semplificando, si tratta di intelligenze artificiali in grado di fare ogni cosa a livello umano. Le super intelligenze (terzo tipo), infine, superano (anzi, supereranno) di gran lunga le prestazioni cognitive degli esseri umani praticamente in ogni campo di interesse. Come si è detto, nel contesto tecnologico contemporaneo si rilevano soltanto intelligenze di tipo ANI. Per quanto oggi si sostenga che le intelligenze AGI e ASI siano in fase di sviluppo attraverso le tecniche di *machine learning* e *deep learning*, una parte ottimista della letteratura scientifica stima che le prime saranno disponibili solo nel 2029,

sono suggerire decisioni non pienamente governate dal controllo umano[42]. In questo caso, l'opacità decisionale della macchina neutralizza l'intelligibilità dell'intero processo di estrazione e generazione dell'informazione.

Sul piano tecnico, in particolare, sussiste una concreta difficoltà nella ricostruzione a posteriori dell'*iter* logico della predizione e dei vari passaggi che conducono alla decisione finale (si tratta del c.d. problema dell'intelligibilità).

A seconda del modello matematico posto alla base dell'algoritmo – per quanto sia dotato di estrema precisione e ritenuto di assoluta efficacia – spesso risulta impossibile ripercorrere il processo decisionale che ha condotto all'estrazione di un determinato *output*. E come è intuitivo, questa sorta di "opacità" degli

mentre le seconde nel 2045. Secondo un orientamento prevalente, invece, si individua l'anno 2100 per l'intelligenza generale e il 2130 per la super intelligenza. Per maggiori approfondimenti sulle tematiche appena citate, tra i tanti, si vedano A. Signorelli, *Rivoluzione artificiale: l'uomo nell'epoca delle macchine intelligenti*, Ledizioni, Milano, 2019; N. Bostrom, *Superintelligenza. Tendenze, pericoli, strategie,* trad. it., Bollati Boringhieri, Torino, 2018; N. Bostrom, "Ethical Issues in Advanced Artificial Intelligence", in *Review of Contemporary Philosophy*, (2006), n. 1-2, pp. 66-73, consultabile in www.nickbostrom.com; N. Bostrom, E. Yudkowsky, "The ethics of artificial intelligence", in W. Ramsey, K. Frankish, *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, Cambridge, 2011, pp. 316-334.

[42] Tali preoccupazioni, in particolare, trovano conferma all'interno di vari studi e ricerche internazionali. In un documento, elaborato nel 2020, dalla Commissione europea, infatti, si legge: "the specific characteristics of many AI technologies, including opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour, may make it hard to verify compliance with, and may hamper the effective enforcement of rules of existing EU law meant to protect fundamental right" (European commission, *White Paper On Artificial Intelligence: A European approach to excellence and trust*, 19 febbraio 2020, p. 12). Anche le linee guida (2019) a cura dell'High-Level Expert Group on Artificial Intelligence (il gruppo di esperti nominati dalla Commissione europea per lo studio di una strategia sull'intelligenza artificiale) suggeriscono la necessaria preservazione della dimensione "umano-centrica" (*human-centric*) delle nuove tecnologie. Infine, anche l'European Group on Ethics in Science and New Technologies, nel documento pubblicato il 9 marzo 2018, intitolato *Artificial Intelligence, Robotics and 'Autonomous' Systems* evidenzia: "that humans and not computers and their algorithms – should ultimately remain in control, and thus be morally responsible".

algoritmi e dei sistemi di IA si accresce in misura proporzionale rispetto al numero delle variabili e dei dati che sono oggetto di analisi[43].

La veloce disamina delle più note forme degenerative tecniche dell'IA – qui sinteticamente proposta – facilita la comprensione della già accennata degenerazione giuridica, nell'intento di individuare quelli che possono essere i rimedi come pure le strategie e le soluzioni di natura preventiva.

5. Come scongiurare il rischio di degenerazioni morali e giuridiche?

Chiaramente, la degenerazione dell'IA – intesa come riduzione e declino delle sue prestazioni – comporta significative implicazioni anche sul piano etico e giuridico[44].

Basti pensare che qualsiasi uso e/o sviluppo dei sistemi di IA, che comporti un potenziale *vulnus* ai diritti umani, si inscrive comunque nel novero della degenerazione.

In quest'ottica non possono non menzionarsi le tante criticità connesse al fenomeno della c.d. discriminazione algoritmica. Situazioni patologiche, che risultano alquanto ricorrenti oltreché

[43] Un recente algoritmo di IA, sviluppato da Google, ha ricevuto notevoli attenzioni per le sue promettenti prestazioni, in quanto, pur essendo applicato a popolazioni diverse rispetto a quelle utilizzate per la raccolta dei dati di addestramento, si è dimostrato in grado di migliorare la velocità e la qualità dello *screening* del tumore al seno, con un livello di precisione superiore a quello degli specialisti radiologi (Cfr. S.M. McKinney, *et al.*, "International evaluation of an AI system for breast cancer screening", in *Nature*, (2020), n. 577, pp. 89-94). Tuttavia, l'algoritmo è stato fortemente criticato, da parte della comunità scientifica, in merito alla sua scarsa trasparenza, poiché è stato presentato senza riportare nel dettaglio le specifiche tecniche sulle modalità di costruzione e funzionamento (Cfr. K. Wiggers, 2020), "Google's breast cancer-predicting AI research is useless without transparency, critics say", in *VentureBeat*. Recuperato da www.venturebeat.com; R. Williams (2020), "Lack of transparency in AI breast cancer screening study 'could lead to harmful clinical trials', scientists say", in *Inews*, recuperato da www.inews.co.uk; B. Haibe-Kains, G.A. Adam, A. Hosny, *et al.*, "Transparency and reproducibility in artificial intelligence", in *Nature*, 586 (2020), n. 7829, pp. 14-16).

[44] L. Floridi, *The Cambridge handbook of information and computer ethics*, Cambridge University Press, Cambridge, 2019.

trasversali e pervasive, dando vita a riflessioni sul degrado nella *performance* dell'IA e sulle tante conseguenze – individuali e collettive – generate da risultanze decisionali errate (il dibattito sul punto, non a caso, si spinge sino a verificare in che modo tali discriminazioni possano arrecare pregiudizio nei confronti di determinate popolazioni, disattendendo i principi di equità e di giustizia).

Altrettante questioni sorgono in ordine alla violazione (o all'uso non corretto) dei dati personali e con riguardo alle conseguenze, che derivano dalla perdita di controllo delle informazioni di carattere personale, sia per l'individuo e sia per la collettività (di qui, fenomeni collaterali come la sorveglianza abusiva e la possibile compressione delle libertà civili).

La complessità del fenomeno in esame, insieme con l'indistricabile – ed evidente – connessione tecnico-etica-giuridica, richiede un approccio olistico, che sia in grado di mettere in luce le connessioni e le interazioni tra i principali fattori che generano e che alimentano la degenerazione.

Per ovviare al rischio di una visione limitata e poco funzionale, è doveroso adottare una strategia che – certo – si muova in ottica "preventiva", ma che sia anche integrata – coniugando cybersicurezza, gestione dei rischi, consapevolezza e governance – e costantemente orientata a ponderare e a verificare tutti quei risvolti etici che, se non valutati attentamente in ogni fase dell'*iter* tecnico, potrebbero dar luogo all'attenuazione (se non addirittura alla neutralizzazione) degli effetti positivi dell'impiego dell'IA nei più svariati settori[45].

Sulla scorta di tali presupposti, si ritiene che l'approccio preventivo al fenomeno degenerativo debba essere orientato alla promozione dei principi etici fondamentali e, specularmente, al

[45] L'intelligenza artificiale, specialmente se applicata a settori particolarmente delicati, come quello sanitario, mette in luce nuove sfide, poiché consente alle macchine di "imparare", compiere delle scelte ed eseguirle senza l'intervento umano. Quando le capacità di apprendimento degli algoritmi si determinano su dati incompleti e, dunque, inaffidabili, oppure alterati in seguito ad attacchi informatici o condizionati da fattori imprevedibili o semplicemente errati, le decisioni che ne conseguono potrebbero assumere forme e direzioni contrarie alla dignità umana, ai principi fondamentali posti a tutela dell'individuo e ai precetti normativi che governano un determinato settore.

benessere generale degli esseri umani[46] in ossequio ad una prospettiva antropocentrica, tesa, per l'appunto, a porre la persona – e la tutela dei suoi diritti – in primo piano.

Indubbiamente, un'azione preventiva efficace si connette intimamente con interventi regolatori (aventi anch'essi carattere "generale") che, definendo e perimetrando i confini di utilizzo dell'IA mirano alla tutela effettiva dei diritti e alla sicurezza dei cittadini, *a fortiori* in un contesto – quale è quello attuale – iper-tecnologico.

A tuto ciò si aggiunga che l'adozione di testi normativi – nazionali e sovranazionali[47] – dedicati all'IA non è di per sé affatto scevra di sfide. Da un lato, è necessario dotare lo strumento regolatorio di adeguata flessibilità, per adattarsi ai rapidi progressi tecnologici, evitando di soffocare l'innovazione; dall'altro, è fondamentale codificare principi etici chiari, che guidino lo sviluppo e l'utilizzo dell'IA, enfatizzando la trasparenza, l'equità, la responsabilità e la non discriminazione. Se così, in pratica, le

[46] In questi termini si è sviluppata la strategia europea per lo sviluppo dell'IA: Comunicazione della Commissione al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni, *Creare fiducia nell'intelligenza artificiale antropocentrica*, 8 aprile 2019, COM(2019) 168 final, consultabile nel sito ufficiale dell'Unione Europea https://eur-lex.europa.eu. La Commissione europea evidenzia che (p. 1) "Per affrontare queste sfide e sfruttare al massimo le opportunità offerte dall'IA, nell'aprile 2018 la Commissione ha pubblicato una strategia europea che pone l'essere umano al centro dello sviluppo dell'IA – un'IA antropocentrica".

[47] Oggi la cornice regolatoria, a livello europeo, è principalmente rappresentata dal Regolamento (UE) 2024/1689 (AI Act), che introduce obblighi graduati per pratiche vietate, sistemi ad alto rischio e modelli di IA di carattere generale, con un calendario di applicazione scaglionato (art. 113). Tra gli obblighi spiccano, per i modelli GPAI, la policy sul diritto d'autore e la tracciabilità dei dataset (art. 53). In parallelo, l'ordinamento italiano ha approvato la Legge 23 settembre 2025, n. 132 che, tra l'altro, designa AgID e ACN come autorità nazionali, rafforza la tutela d'autore per opere create con ausilio di IA (novellando la L. 633/1941) e introduce specifiche fattispecie penali per abusi di contenuti sintetici (*deepfake*) (artt. 20, 25, 26). Questo doppio livello, europeo e nazionale, incide direttamente sul tema del "rischio creativo" qui discusso, perché sposta sul piano organizzativo e documentale la responsabilità di dimostrare scelte di *design*, alimentazione dei modelli e controllo degli effetti degenerativi. Cfr. M. Farina, "Degenerazione e rischio creativo dell'Intelligenza Artificiale "forte": forme di prevenzione e tutela complementare", in *L'Ircocervo*, 23 (2024) n. 1, pp. 122-138.

azioni preventive, oltre ad essere "generali" ed "anteriori", rispetto al concreto uso dell'IA, dovrebbero risultare, altresì, "concomitanti" rispetto all'effettivo utilizzo di tali sistemi. Aspirazioni, queste, tutt'altro semplici da realizzarsi nella pratica.

In tal senso, occorrerebbe anzitutto procedere all'implementazione di *policy* adeguate alla manutenzione regolare e all'aggiornamento dei sistemi IA: ad esempio, sarebbe utile la predisposizione di protocolli standardizzati, per garantire la loro "perdurante" affidabilità e sicurezza a lungo termine.

Nella medesima direzione, si dovrebbero promuovere revisioni periodiche da parte di terzi indipendenti, per assicurare che i sistemi di IA rispettino i principi etici e non violino i diritti umani.

Infine, importanti misure preventive dovrebbero agire anche sul fattore culturale, con lo sviluppo di programmi di formazione per gli operatori di sistemi IA, volti ad assicurare un uso responsabile e informato delle tecnologie.

L'approccio proposto mira non soltanto a prevenire la degenerazione[48], come fenomeno in sé considerato, ma anche a promuovere un utilizzo dell'IA che sia etico e conforme ai criteri di giustizia e rispettoso dei diritti umani[49].

La recente diffusione delle IA generative, oltre a stimolare molteplici riflessioni di carattere etico-giuridico, offre l'occasione per richiamare, sul piano metodologico, annose questioni sul rapporto tra attività umana e tecnica, che inducono a riflettere anche sull'effettiva completezza di misure di prevenzione generali.

Com'è noto, invero, una delle più dibattute questioni in materia si ricollega al conflitto etico-giuridico tra l'autore di un'opera dell'ingegno umano[50] e il pubblico fruitore della stessa.

[48] A. D'Aloia (a cura di), *op. cit.*; G. Sartor, *L'intelligenza artificiale e il diritto*, Giappichelli, Torino, 2022; A. Pajno, F. Donati, A. Perrucci (a cura di), *op. cit.*

[49] Cfr. E. Grassi, *Etica e intelligenza artificiale. Questioni aperte*, Aracne, Roma, 2020; C. Mannelli, *Etica e Intelligenza artificiale. Il caso sanitario*, Donzelli editore, Roma, 2022; G. Tamburrini, *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Carocci, Roma, 2020.

[50] L'art. 25 della Legge 23 settembre 2025, n. 132 è intervenuto sul primo comma dell'art. 1 della legge sul diritto d'autore (L. 22 aprile 1941, n. 633) con un duplice ritocco: ha esplicitato la natura umana dell'ingegno,

Il dibattito, in particolare, insiste sul risultato dell'elaborazione compiuta dal sistema di IA generativa, per valutare se il diritto d'autore di quel prodotto ricada in capo all'utilizzatore, al creatore, a entrambi (congiuntamente) ovvero se possa ipotizzarsi una titolarità in capo alla stessa IA[51]. Tralasciando l'ultima ipotesi, impraticabile perché negherebbe il concetto stesso di opera dell'ingegno (umano), la soluzione (oggi non unanime) a favore delle altre si basa tutta sulla riconducibilità del risultato all'intelletto di uno, o più di uno, dei protagonisti citati.

Si ripropongono, seppur in termini tecnologici più complessi, le medesime questioni giuridiche che si posero per la titolarità del *software* e la riconducibilità (anche in termini di prevedibilità) delle azioni compiute dall'uomo per mezzo del programma[52].

Rispetto al dibattito generatosi sulla tutela del *software*, sistemi di IA richiedono puntuali riflessioni su ulteriori e stringenti fattori di complessità – sul piano tecnico, etico e giuridico –, che ricadono sulle antecedenti (ma fondamentali) attività di costitu-

inserendo dopo le parole "opere dell'ingegno" il termine "umano"; ha aggiunto, subito dopo "qualunque ne sia il modo o la forma di espressione", la clausola: "anche laddove create con l'ausilio di strumenti di intelligenza artificiale, purché costituenti risultato del lavoro intellettuale dell'autore". Il secondo comma dell'art. 1 (relativo alla protezione di programmi per elaboratore e banche dati) non è stato modificato, ma va ora riletto in coerenza con il principio, reso espresso, della paternità creativa umana. Per completezza, lo stesso art. 25 introduce anche il nuovo art. 70-*septies* L.d.A. in tema di estrazione di testo e dati tramite IA.

[51] Sul punto, *ex multis*, M. Ferrari, "Intelligenza artificiale e titolarità dei diritti d'autore", cit., pp. 373-379; S. Lavagnini, "Intelligenza artificiale e proprietà intellettuale: proteggibilità delle opere e titolarità dei diritti", in *Il Diritto di autore*, (2018), n. 3, pp. 360-375; M. Franzosi, "'Copyright': chi è l'autore delle opere generate a computer?", in *Il Diritto di autore*, (2018), n. 2, pp. 168-171; G. Sanseverino, "'Ex machina'. La novità e l'originalità dell'invenzione 'prodotta' dall'IA", in *AIDA*, (2018), pp. 3-22; A. Ramalho, "Originality redux: an analysis of the originality requirement in AI-generated works", in *AIDA*, (2018), pp. 23-41; S. Guizzardi, "La protezione d'autore dell'opera dell'ingegno creata dall'Intelligenza Artificiale", in *AIDA*, (2018), pp. 42-68; G. Noto La Diega, "Artificial Intelligence and databases in the age of big machine data", in *AIDA*, (2018), pp. 93-149.

[52] G. Sartor, "L'intenzionalità degli agenti software e la loro disciplina giuridica", in *Riv. Trim. Dir. Proc. Civ.*, (2003), n. 1, pp. 23-51; M.A. Biasiotti, F. Romano, M.T. Sagri, "La responsabilità degli agenti software per i danni prodotti a terzi", in *Informatica e diritto*, 28 (2002), n. 2, pp. 155-164.

zione del *dataset* per nutrire e istruire l'IA, imponenti l'utilizzo lecito delle *fonti* di addestramento.

In passato l'avvento di Internet e l'immediata disponibilità di opere dell'ingegno (testi, foto, filmati, *software*, etc.) in rete, nel silenzio dell'autore, ha spinto molti a valutare tale inerzia quale libera utilizzazione, nonché a risolvere tali quesiti ricorrendo ai principi generali – comuni alle tradizioni giuridiche della quasi totalità degli ordinamenti – in materia di diritto d'autore[53].

All'indomani dell'avvento dell'IA generativa, invece, il problema si ripropone in misura più pressante, in relazione alla scelta, nonché alle modalità di creazione e popolamento dei *datasets* di addestramento: in particolare, si discute se il libero "rastrellamento" di fonti informative dalla Rete, sia rispettoso della posizione autorale (tanto morale, quanto patrimoniale).

Da più parti, peraltro, si lamenta l'inadeguatezza della disciplina positiva e l'anacronismo degli accordi internazionali (sorti in un contesto pre-tecnologico, nel quale le opere dell'ingegno digitali non esistevano) e si invocano riforme adeguate all'attuale contesto[54].

Per quanto qui di interesse, tenuto conto che il diritto d'autore non dipende dalla natura del supporto informativo, si ritiene che le sollecitazioni richiamate manifestino una ben più rilevante criticità. Infatti, i *deficit* di tutela non dipendono dalla carente completezza delle misure di prevenzione generali; piuttosto, il fattore critico risiede proprio nel prevedere azioni preventive unicamente incentrate sullo strumento legale.

In altri termini, per ovviare alle criticità connesse alla composizione e all'addestramento dei *datasets*, occorre prendere le distanze da qualunque approccio che si limiti a fornire una soluzione basata esclusivamente sulla tutela "legale" del diritto d'autore.

[53] Per una panoramica sull'argomento, si veda, S. Dell'Arte, *Fondamenti di diritto d'autore nell'era digitale*, ed. 2, Key Editore, Milano, 2023.

[54] Di particolare interesse è il Considerando n. 105 del già citato AI ACT, che compie espresso riferimento alla necessità di accesso, da parte dei modelli di IA generativa, "a grandi quantità di testo, immagini, video e altri dati" per il loro addestramento e all'"autorizzazione del titolare dei diritti interessato" salve le eccezioni introdotte dalla Direttiva (UE) 2019/790 (che difficilmente potranno estendersi ai modelli realizzati dai privati per finalità commerciali).

Se è pur vero, infatti, che un buon impianto normativo, di equo bilanciamento tra gli interessi sottesi, può essere utile per tutelare gli autori delle opere utilizzate in addestramento, è altrettanto vero che si rischia – come tuttora accade in Rete – che la tutela approntata venga confinata in una dimensione totalmente teorica. Il maggiore fattore di *alea* è, dunque, che tali previsioni risultino, nella sostanza, incapaci di fronteggiare le concrete esigenze di garanzia autorale. In quest'ottica, in particolare, è utile ricordare che le misure preventive generali incontrano un inossidabile limite esterno, consistente nel rapporto di autonomia delle parti e nella propria libertà negoziale.

Per tali ragioni, si ritiene che accanto alle misure di tutela legale – a carattere "generale" – si debba sviluppare una direttrice "complementare" che coglie la dimensione negoziale del rapporto intersoggettivo.

Tale combinazione di approcci e garanzie potrebbe assicurare quella tutela dell'efficienza pratica di cui si necessita e, considerato che i rimedi complementari[55], non si sostituiscono e non prevalgono sulla tutela legale, si scongiurerebbe il pericolo di esporsi alle consequenziali critiche – astrattamente condivisibili, se avulse dalla visione insiemistica proposta – sull'inerzia (e, spesso, incapacità) degli autori a predisporre chiare licenze con le quali regolamentare la circolazione delle proprie opere digitali.

A differenza del passato, tuttavia, l'avanzamento tecnologico può indubbiamente costituire un fattore di vantaggio, per asservirne gli usi agli obiettivi di tutela.

Più precisamente, la tecnologia può efficacemente intervenire nell'ambito della tutela complementare, per almeno due ordini di ragioni. In primo luogo, versando in un ambiente di natura tipicamente negoziale, l'autodeterminazione delle parti ben può indirizzarsi, altresì, sull'uso degli strumenti – anche a carattere tecnologico – più adeguati, per il perseguimento e soddisfacimento dei propri interessi; in questo senso, la tecnologia potrebbe fare la differenza, attraverso un sistema di auto-esecuzione della volontà autoriale. Si potrebbero, infatti, integrare le opere, con appositi *metatag* leggibili dall'IA, che si ciberebbe

[55] Sulla tutela complementare, si veda M. Farina, *Elementi di Diritto dell'Informatica*, Cedam-Wolters Kluwer Italia, Milano, 2019, pp. 15-17.

lecitamente soltanto di quelle licenziate e, viceversa, scarterebbe quelle prive di autorizzazione[56]. Siffatta soluzione potrebbe dar vita ad una nuova era delle IA generative di "sana alimentazione" e far tramontare quelle "malnutrite".

In secondo luogo, trattandosi di una forma di tutela "complementare", che si aggiunge – senza sovrapporsi – a quella legale di carattere generale, non implicherebbe alcun ridimensionamento o ripudio di quest'ultima, la quale, al contrario, troverebbe sempre piena applicazione – espandendosi e prevalendo sulla dimensione negoziale – in presenza di violazione di regole imperative. La prevenzione del fenomeno degenerativo implica, pertanto, un approccio olistico – tanto sui profili tecnologici, quanto, (e congiuntamente) su quelli etico-giuridici – e duale: simultaneamente e sincronicamente a carattere generale e complementare.

6. Conclusioni: verso una *governance* duale dell'IA fra prevenzione generale e tutela complementare

Le considerazioni svolte mostrano come la degenerazione dei sistemi di intelligenza artificiale non sia un fattore marginale, ma un rischio strutturale che accompagna l'adozione di tecnologie capaci di incidere in profondità sull'ordito delle relazioni economiche e sociali. L'obiettivo non è soltanto preservare l'efficienza degli strumenti, bensì salvaguardare, insieme, la qualità

[56] La soluzione, mediante *watermark*, è stata proposta, limitatamente allo scarto dei dati sintetici, dal gruppo di ricercatori statunitensi già sopra citati: S. Alemohammad, J. Casco-Rodriguez, L. Luzi, *op. cit.* La soluzione, in quel caso, è volta a risolvere una delle cause di degenerazione tecnica. I metadati che integrano la licenza nell'opera digitale, invece, se rigidamente applicati, potrebbero risolvere ogni causa di degenerazione giuridica connessa alla violazione dei diritti, morali e patrimoniali, dell'autore. Una soluzione simile, per lo meno nell'idea di base, è stata già sperimentata nel panorama italiano, con il Provvedimento del Garante per la protezione dei dati personali, recante "Linee guida in materia di trattamento di dati personali contenuti anche in atti e documenti amministrativi, effettuato da soggetti pubblici per finalità di pubblicazione e diffusione sul *web*" – 2 marzo 2011", (doc-web n. 1793203), consultabile in www.garanteprivacy.it. In quell'occasione, allo scopo di evitare che i motori di ricerca indicizzino i contenuti pubblicati nei siti *web* delle PA si consigliava l'inserimento di "*metatag noindex* e *noarchive* nelle intestazioni delle pagine *web*".

degli esiti e i presupposti etico-giuridici che fondano la convivenza civile.

Ne discende la necessità di un'impostazione interdisciplinare e multilivello in cui la filosofia del diritto, le scienze dell'informazione, l'informatica giuridica e le neuroscienze cognitive devono convergere in una cornice di governo responsabile dell'IA, capace di tenere insieme affidabilità tecnica e rispetto dei valori che informano i nostri ordinamenti. Solo una tale integrazione consente di tradurre in prassi i principi e di evitare che l'innovazione si risolva in nuove forme di asimmetria.

Sul versante tecnico, la fenomenologia degenerativa riconduce a fragilità note, come la dipendenza dal dato, le dinamiche disinformative, le allucinazioni e l'opacità dei processi decisionali. Per questo, il controllo umano non è un presidio esterno e occasionale, ma un elemento strutturale della catena di trattamento. L'intervento umano qualificato orienta l'uso dei modelli, rileva precocemente derive indesiderate e stabilizza la qualità dei risultati.

Dal punto di vista etico-giuridico, l'attenzione si concentra sui rischi discriminatori, sulle diseguaglianze informative e sulla trasparenza dei processi. Ridurre tali rischi implica trattare la degenerazione anche come questione di giustizia per evitare nuove esclusioni, rendere verificabili i passaggi di generazione dell'informazione e garantire spiegabilità sufficiente delle decisioni automatizzate.

Su questo sfondo si giustifica un modello di "governance duale".

Da un lato, operano le misure generali di fonte pubblica, che fissano obblighi e standard. Così, nella dimensione europea, il Regolamento 2024/1689 impone requisiti graduati lungo l'intero ciclo di vita dei sistemi, con specifici oneri documentali e di tracciabilità per i modelli di carattere generale. Mentre, sul piano nazionale, la legge 23 settembre 2025, n. 132 coordina funzioni di vigilanza, rafforza la protezione autorale nelle opere realizzate con ausilio di sistemi intelligenti e introduce specifiche fattispecie sanzionatorie per abusi legati a contenuti sintetici.

Dall'altro lato, opera la tutela "complementare", affidata all'autonomia privata e a strumenti tecnico-organizzativi (me-

tadati di licenza leggibili dai sistemi, *standard* di *content provenance*, clausole contrattuali sull'addestramento), che calano i principi nel concreto flusso delle transazioni informative. La coesistenza dei due binari – pubblico/regolatorio e negoziale/tecnico – riduce il divario tra regole e pratiche, trasformando i principi in procedure ripetibili e verificabili.

In termini operativi, ciò si traduce in alcune priorità: protocolli interni di manutenzione e aggiornamento che diano continuità alla qualità del dato e del modello; verifiche periodiche indipendenti sull'aderenza ai principi etici e sulla non lesione di diritti fondamentali; programmi di formazione continua per progettisti e utilizzatori; documentazione *ex ante* delle scelte di *design* e delle fonti di addestramento, per rendere tracciabili le responsabilità lungo la filiera. Queste leve organizzative, se congiunte a presìdi tecnici e normativi, permettono di fronteggiare la degenerazione come rischio prevedibile e gestibile.

In definitiva, l'uscita dal "paradosso" di una tecnologia insieme potente e fragile passa per l'assestamento di un'ecologia dell'IA dotata di regole generali chiare, di strumenti complementari efficaci, di controllo umano competente e di misurabilità degli effetti. Soltanto così può preservarsi il "rischio creativo" come motore di progresso – non come fattore di vulnerazione – e si ricompone la dialettica tra tecnica e diritto in una direzione pienamente antropocentrica.

Lucio Franzese

Sulla insufficienza dell'Intelligenza Artificiale e della legge

Abstract: **On the Inadequacy of Artificial Intelligence and the Law.** The article intends to highlight the problems that the use of Artificial Intelligence reveals in legal matters and the healthcare sector, starting from some considerations of the use of this technique.

Keywords: Artificial Intelligence, Right, Compas, AI Act, Digital Governance.

Sommario: 1. Premessa – 2. Un giudice algoritmico? – 3. Un medico algoritmico? – 4. *AI Act* e *Governance digitale*.

1. Premessa

Occorre imparare a convivere con l'Intelligenza Artificiale poiché lo sviluppo e la sua affermazione appaiono inarrestabili. È necessario comprendere come governare questo cambiamento epocale per non esserne travolti. Il problema infatti non risiede nel "se" ma nel "come" farlo, poiché non sembra più possibile tornare a un mondo senza l'IA[1].

Gli studi intorno all'IA nascono a metà del secolo scorso dalla insoddisfazione dell'uomo nei confronti dell'informatica e nel tentativo di costruire sistemi che simulino il funzionamento della mente umana, con l'intento dichiarato di sostituire il più possibile le persone nello svolgimento di vari compiti, rendendo le attività più veloci in modo da massimizzarne l'efficienza economica.

Durante gli anni Ottanta – ma come è forse possibile constatare anche ai nostri giorni – le aspettative sulle "macchine intelligenti" erano esageratamente alte, trainate soprattutto dai media e da grandiose promesse che a loro volta trainavano ingenti investimenti pubblici. L'entusiasmo e gli investimenti si alimentavano reciprocamente e concorrevano a far sperare nella migliore

[1] Cfr. N. Cristianini, *La scorciatoia*, il Mulino, Bologna, 2022, p. 187.

riuscita di questa impresa. Le promesse però non riuscirono a tenere il passo con il clamore suscitato e ben presto iniziarono a diffondersi umori meno ottimistici che portarono a un netto calo degli investimenti e dell'interesse per questa disciplina, sia da parte del mondo scientifico che da quello del grande pubblico. All'epoca le limitazioni di carattere oggettivo, le fragilità e gli eccessivi costi dei sistemi che fino ad allora erano stati sviluppati, stavano rendendo sempre meno probabile la realizzazione di una macchina in grado di comprendere una lingua, riconoscere oggetti e ragionare come una persona. Durante questo periodo di "inverno" per l'IA il settore risultò meno attraente, ma ciò nonostante iniziarono a imporsi delle idee teoriche coltivate in nicchie specialistiche per anni e con pochissimi finanziamenti. Queste idee basate su metodi statistici e sulle reti neurali posero le basi per la "lingua franca" dell'IA dei nostri giorni, grazie anche al perfetto connubio con la creazione del *World Wide Web* lanciato nel 1994. La ricerca sull'Intelligenza Artificiale ne venne completamente trasformata[2].

Com'è noto, nel novembre del 2022 OpenAI ha presentato ChatGPT, chatbot capace di rispondere in modo più o meno coerente alle domande che gli si pongono su tutto lo scibile. Questo evento ha dato notorietà mondiale al fenomeno dell'IA e ha alimentato un dibattito che coinvolge tutta la vita sociale e in primo luogo il settore giuridico e quello sanitario.

Tuttavia non si può nascondere che quanto più l'IA sembra progredire in potenza e capacità, tanto più desta legittime preoccupazioni. Basti pensare che durante il *World Governments Summit* (WGS), Sam Altman, fondatore di Open AI e creatore di ChatGPT, accennando, con soddisfazione ed entusiasmo, alle grandi potenzialità legate alla prossima evoluzione del nuovo modello di GPT, ovvero GPT-5, ha confessato di "non dormire la notte" pensando alle possibili conseguenze dell'IA in mancanza di un quadro di regole efficaci.

L'aspetto rilevante è che i dibattiti non sono polarizzati su posizioni antitetiche, IA sì, IA no, e molto spesso, a conclusioni fatte, emerge forte il richiamo, soprattutto e sorprendentemente, da parte del mondo scientifico di introdurre norme che ci

[2] Cfr. *ivi*, pp. 33-35.

consentano di convivere in sicurezza con queste tecnologie. Invero un regolamento, come quello approvato dall'Unione Europea (*AI Act*) può rappresentare un aiuto. Ciononostante senza il supporto anche di una *pedagogia* all'uso di questi strumenti e soprattutto senza il recupero nell'ambito del discorso giuridico della concreta esperienza dei rapporti umani, ogni tentativo di regolare una materia così complessa può risultare, in ultima istanza, insufficiente.

Al giorno d'oggi è possibile ipotizzare che, in un futuro non troppo lontano, l'IA riuscirà a fare molte cose di quelle che oggi sono ancora prerogative dell'uomo. Nello stesso tempo però è necessario chiedersi se ci sono dei limiti a quello che possiamo lasciar fare all'IA, particolarmente in materia giuridica e in ambito sanitario.

Sistemi algoritmici, più o meno complessi, fanno già parte dell'esperienza giuridica e medica. È possibile pertanto proporre qualche spunto di riflessione sul fenomeno proprio a partire da alcuni esempi relativi all'applicazione di questi strumenti.

2. Un giudice algoritmico?

Nel sistema giudiziario americano viene utilizzato un software, il "Compas" (*Correctional offender management profiling for alternative sanctions*) per "calcolare" la pericolosità sociale dell'imputato.

Esso è stato portato alla ribalta dal "caso Loomis"[3], riguardante un cittadino americano che non si era fermato ad un controllo di polizia mentre era alla guida di un'auto poi risultata coinvolta in un conflitto a fuoco. Il tribunale circondariale di La Crosse nel determinare la pena da comminare a Loomis ha tenuto conto anche del punteggio assegnato da Compas che, sulla base di un questionario di 137 domande stabilisce il grado di rischio che l'imputato commetta nuovamente un reato. Il cittadino americano è stato condannato alla reclusione di sei anni perché recidivo e perché ritenuta alta la possibilità che in futuro commetta nuovamente dei reati. Dato che il funzionamento dell'algoritmo è segreto perché di proprietà di una azienda privata (la Northpointe,

[3] Cfr. S. Carrer, "Se l'*amicus curiae* è un algoritmo: il chiacchierato caso Loomis alla Corte Suprema del Wisconsin", in *Giurisprudenza Penale*, 4 (2019).

ora Equivant), non si sa su quali basi si è giunti a ritenere Loomis soggetto ad alto rischio di recidiva – condizione decisa da Compas ma sottratta al contraddittorio tra le parti – e alla conseguente applicazione della pena. La cosa ancora più sorprendente è che su richiesta di revisione della sentenza da parte dell'imputato, il quale lamentava che la decisione, presa sulla base dei risultati del Compas, aveva violato il proprio diritto ad un equo processo, la Corte Suprema del Wisconsin ne ha confermato la validità. Ha infatti espressamente stabilito che l'uso del Compas può essere legittimo nell'ambito dei giudizi di determinazione della pena se vengono rispettati, da parte degli organi giudicanti, alcuni limiti e alcune cautele indicate dalla stessa Corte. Resta però il fatto che non è possibile conoscere come i dati vengano processati dal Compas in quanto la società proprietaria ne mantiene legittimamente il segreto industriale e il brevetto.

Al fine di una prima riflessione si può notare che un giudizio automatizzato sembra, in definitiva, riproporre la figura del giudice quale *bocca che pronuncia le parole della legge*[4], per tal modo eludendosi tutta la problematica dell'interpretazione del diritto e della composizione delle liti intersoggettive, che è il fine cui tende il processo e che non è la mera applicazione della legge. Le parti in giudizio intendono ottenere il riconoscimento del loro proprio che non può ridursi in un "rigido formalismo" secondo cui la rilevanza giuridica dell'agire "si manifesta propriamente nella sua conformità, nel suo lasciarsi individuare in base a criteri figurativi posti dal legislatore"[5]. Il "giudice algoritmico" preconizzato dallo straordinario sviluppo dell'IA sembra dare compimento alla volontà di de-soggettivazione del giudizio teorizzato e auspicato dall'illuminismo giuridico. Inoltre sembra dominare "l'assenza di un'intersoggettività reale, creativa, luogo di istituzione del diritto"[6].

[4] Cfr. V. Ferrari, "Note socio-giuridiche introduttive per una discussione su diritto, intelligenza artificiale e big data", in *Sociologia del diritto*, (2021), n. 3, pp. 9-31.

[5] Cfr. N. Irti, *Studi sul formalismo negoziale*, Cedam, Padova, 1997, pp. VII-VIII.

[6] L. Avitabile, "Il diritto davanti all'algoritmo", in *Rivista Italiana per le Scienze Giuridiche*, 8 (2017), p. 321.

Si deve riflettere anche sul fatto che l'uso del Compas mostra un altro aspetto controverso, ovvero la presenza di *bias*. È emerso infatti che i dati di cui si avvale per calcolare statisticamente la pericolosità sociale degli imputati sono intrisi di pregiudizi razziali. La proporzione di imputati neri soggetti impropriamente a detenzione preventiva, poiché dichiarati successivamente innocenti, tende ad essere maggiore dell'analoga proporzione di imputati bianchi risultati anch'essi innocenti, così come per la quantificazione della pena si registra disparità di trattamento a secondo dell'etnia del condannato[7].

L'applicazione dell'IA al fenomeno giuridico non potrà che produrre l'uniformazione del giudizio escludendosi per tal modo quell'equità che rappresenta la giustizia del caso concreto, il cui riconoscimento resta una prerogativa del giudice umano[8]. Francesco Gentile ci ricorda a riguardo che "il riconoscimento della diversità di ognuno costituisce un'autentica garanzia per il rispetto dell'essenza comune a tutti"[9].

L'applicazione automatizzata della legge sarebbe un'operazione che non abbisogna di una previa interpretazione della norma e l'idea stessa di processo, incentrata com'è sul contraddittorio tra le parti, guidato e stimolato dal giudice, che lo rende refrattario ad ogni automatismo decisionale, a ben vedere viene messa da parte[10]. C'è poi il rischio che le decisioni basate su dati storici, finiscano in sostanza per replicare all'infinito sentenze passate, senza tener conto della concreta esperienza giuridica[11]. E in ultima istanza la produzione (e il termine non è a caso) di automatismi impedisce all'imputato di vedersi riconosciuto il *suo proprio* che consiste nel "diritto alla pena", come icasticamente

[7] Sul punto cfr. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A Survey on Bias and Fairness in Machine Learning", in *ACM Computing Surveys*, 54 (2021), n. 6, pp. 1-35 e A.Z. Huq, "Artificial Intelligence and the Rule of Law", in *Public Law and Legal Theory Working Paper*, (2021), n. 764, p. 4.

[8] Cfr. G. Zaccaria, *La responsabilità del giudice e l'algoritmo*, Mucchi Editore, Modena, 2023.

[9] F. Gentile, *Politica aut/et statistica*, Giuffrè, Milano, 2003, p. 185.

[10] Cfr. A. Garapon, J. Lassègue, *La giustizia digitale. Determinismo tecnologico e libertà*, trad. it., di M.R. Ferrarese, il Mulino, Bologna, 2021.

[11] Sempre capitale il rinvio al concetto di esperienza giuridica così come formulato da Giuseppe Capograssi.

affermato da Giuseppe Bettiol, per il quale il compito dell'avvocato non è quello di fare ottenere al suo assistito la pena minima, ma quella idonea per riacquistarlo al consorzio sociale dopo che esso se ne è collocato al di fuori con la commissione del reato[12].

Tali automatismi si fondano sul mero dato statistico e non sul principio dello *stare decisis* per il quale un orientamento vincola perché ha dato buona prova di sé, cioè ha contribuito a risolvere una lite riattivando la comunicazione intersoggettiva che era stata interrotta dalla commissione dell'illecito.

Soprattutto si deve accennare che questi mezzi possono rappresentare forme di controllo sociale finalizzate al dominio dell'uomo sull'uomo, mentre il diritto classicamente mira allo sviluppo della persona umana, alla sua piena realizzazione mediante la promozione della libertà individuale[13].

3. Un medico algoritmico?

Potremmo chiederci in ambito medico cosa implica l'uso di sistemi di Intelligenza Artificiale e possiamo farlo ponendoci due domande: "È possibile sostituire il medico con un algoritmo?" E in tal caso: "Chi si farebbe curare da un algoritmo?"

Va subito detto che per le diagnosi, la scelta della cura e la fase post-ospedaliera, la relazione tra medico e paziente resta fondamentale. Occorre valorizzare la relazione di cura e di fiducia tra medico e paziente nel quale si incontrano l'affidamento del paziente e la competenza e la responsabilità del medico. Le persone hanno diritto di conoscere le proprie condizioni di salute e di essere edotte sulla diagnosi e quindi sui benefici, rischi e sulle eventuali alternative in modo chiaro, completo e aggiornato, nonché devono essere informati sulle conseguenze di un eventuale rifiuto del trattamento sanitario e dell'accertamento diagnostico[14].

[12] Per un approfondimento cfr. G. Marini, *Giuseppe Bettiol. Diritto penale come filosofia*, ESI, Napoli, 1992.

[13] Esemplare la lezione di Feliciano Benvenuti nel diritto pubblico-amministrativo tradizionalmente incentrato sulla nozione di autorità: cfr. L. Franzese, *Feliciano Benvenuti, Il diritto come scienza umana*, ESI, Napoli, 1999.

[14] Cfr. l'art. 1 della Legge 219/2017.

La prima considerazione è che un algoritmo non può fare altro che considerare la persona come un insieme di dati clinici e quindi, in sostanza, riduce la stessa alla sua cartella clinica, e non può fare altrimenti: guarda alla malattia e non alla persona. Le persone non sono però delle "macchine" da aggiustare.

Per un medico il rapporto con il paziente non si basa solo sui dati clinici, né tanto meno solo sulla dimensione cognitiva dello stesso. Egli osserva il corpo del paziente, le posture, lo sguardo; ascolta come si esprime e che linguaggio utilizza; ritiene indispensabile conoscere anche la storia personale e familiare, in sostanza guarda alla totalità della persona.

L'Intelligenza Artificiale non ha l'obiettivo di comprendere ma solo di operare, cioè di padroneggiare un determinato evento. Attualmente le "macchine intelligenti" si avvalgono del modello statistico e quindi della raccolta di un'enorme mole di dati per inferirne le invarianti statistiche, che rappresentano la soluzione proposta dalla macchina al problema che le è stato sottoposto[15].

> Il potere degli algoritmi persegue scopi, concepiti e voluti non da un qualche algoritmo, ma da esseri umani orientati ad incontrare gli altri come semplici esecutori di un calcolo e non come titolari del diritto di esistere per una formazione originale della loro personalità[16].

Si pensi a quanto stanno facendo alcuni ospedali in Italia che inseriscono ed elaborano in un unico archivio digitale le cartelle cliniche dei loro pazienti, al fine di fissare dei protocolli per le diagnosi e per le cure, mettendo tra parentesi le differenti espressioni della personalità di ciascuno che inevitabilmente ne condiziona anche il percorso terapeutico.

Il rischio è che così facendo il medico si limiti a consultare questi protocolli, rinunciando ad esercitare l'arte medica. Invero il medico potrebbe chiedersi: "Ma chi sono io rispetto alla mole di informazioni di cui dispone la macchina?". Senza considerare l'impatto che questi strumenti possono avere sull'operosità dei medici che potrebbero anche perdere la motivazione a cimentarsi personalmente con i casi loro sottoposti. Fino alla possibile

[15] Cfr. N. Cristianini, *op. cit*.

[16] B. Romano, *Algoritmi al potere. Calcolo Giudizio Pensiero*, Giappichelli, Torino, 2018, p. 2.

perdita della stessa capacità di effettuare una diagnosi. Rinunciando, per tal modo, alla propria libertà e responsabilità.

Infatti una procedura o un dispositivo tecnico non possono essere chiamati a rispondere delle proprie azioni poiché la responsabilità è un elemento imprescindibile dell'impegno personale che implica la disponibilità a comprendere le aspettative e le preoccupazioni dell'altro per cercare un terreno di condivisione[17].

È opportuno richiamare l'intervista a uno dei chirurghi italiani di fama mondiale, Luigi Bonavina, che alla domanda: "Qual è la medicina del futuro", ha risposto: "È quella di precisione che cura la persona caso per caso; il *modus operandi* deve avere come target un solo paziente con le sue specifiche caratteristiche anatomiche e funzionali". Di qui il suo richiamo al pensiero di Ippocrate, secondo cui: "Non può saper di medicina chi non sa che cosa è l'uomo; questo si deve conoscere per curare correttamente gli uomini". Un pensiero che Bonavina ritiene

> più che mai attuale oggi che l'innovazione tecnologica ha rivoluzionato in meglio cure e terapie, senza trascurare la fase post operatoria, forse la più delicata, e dove l'empatia con il paziente fa la grande differenza. Un approccio, questo, che però – non può non osservare il chirurgo – spesso oggi stride con la prassi di interventi veloci e in batteria, in cui si deve massimizzare tutto[18].

La medicina è molto più che raccogliere dati e mettere insieme le informazioni come può fare un sistema di Intelligenza Artificiale. Resta fondamentale l'esperienza dell'incontro empatico tra due persone, il medico e il paziente che insieme partecipano, ciascuno a modo suo, al cammino terapeutico: il paziente con la volontà di guarire e nutrendo fiducia nell'operato del medico; e il medico mettendosi a fianco dell'ammalato con-

[17] Cfr. G. Gorgoni, R. Gianni, "Responsibility, Technology, and Innovation", in W. Reijers, A. Romele, M. Coeckelbergh (eds.), *Interpreting Technology: Ricoeur on Questions Concerning Ethics and Philosophy of Technology*, Rowman & Littlefield, Lanham, 2021, p. 171.

[18] L. Bonavina (2024), "La medicina del futuro? É quella di precisione che cura la persona caso per caso", in *Il Sole 24 Ore*. Recuperato da https://www.ilsole24ore.com/art/la-medicina-futuro-e-quella-precisione-che-cura-persona-caso-caso-AFqzkFmC?refresh_ce=1.

dividendone le aspettative di guarigione. In definitiva, se anche il medico non può fare a meno dell'IA vale allora il monito di papa Francesco, secondo cui "il cuore che manca alle macchine, lo deve mettere l'uomo".

4. *AI Act* e *Governance digitale*

L'IA sta avendo uno sviluppo strabiliante. Ciò tuttavia non significa che essa sia la panacea di ogni male né, all'opposto, che si può demonizzarla e dire che è fonte di ogni male.

Nei suoi confronti bisogna adottare l'atteggiamento che Platone e Aristotele consigliano nei confronti del farmaco, che ha sì un effetto abilitante ma anche distruttivo. Occorre quindi esercitare una pedagogia dell'Intelligenza Artificiale in ogni suo ambito di applicazione, a partire dal campo medico e da quello giuridico, che sono settori in cui l'IA sta avendo più successo e anche i più delicati perché involgono direttamente la persona umana e le relazioni intersoggettive. Bisogna problematizzare il ricorso all'IA sottolineando, in particolare, che essa non può sostituirsi all'uomo nell'individuazione dei fini che ne orientano l'azione, potendo avere una funzione soltanto strumentale, riguardante cioè l'organizzazione automatizzata di quanto è necessario per portare ad effetto le decisioni adottate dall'uomo. Anche nei più recenti congegni l'IA solo apparentemente infatti individua i fini della propria azione che in realtà sono le mere risultanze delle statistiche che vengono individuate sulla base dell'elaborazione dei dati somministrati alla macchina[19].

Un cenno va fatto al Regolamento sull'IA, il primo al mondo, che l'Unione Europea ha sentito la necessità di adottare per disciplinare i sistemi di Intelligenza Artificiale, ovvero stabilire definizioni e linee d'azione comuni, vietare pratiche non conformi ai principi europei, tra cui quelli legati alla tutela della privacy, e stabilire regole sulla base dei diversi livelli di rischio rappresentati dall'Intelligenza Artificiale per la tutela della persona umana[20].

[19] Cfr. B. Romano, *op. cit.*, p. 2.

[20] Cfr. A. Simoncini, "L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà", in *BioLaw Journal - Rivista di BioDiritto*, (2019), n. 1, pp. 411-417.

È molto opportuno che l'Unione Europea abbia adottato una regolamentazione giuridica che, come è avvenuto con il Regolamento per la protezione dei dati personali (GDPR 679/2016), costituisce un precedente a livello mondiale. Gli stessi Stati Uniti e la Cina, tradizionalmente impegnati soltanto nello studio e nello sviluppo dei sistemi di IA, perché vedevano nella regolamentazione un intralcio alla diffusione del suo sviluppo, stanno adottando delle loro discipline. L'Unione Europea ha interesse a stimolare le imprese a fare ricerca e a immettere sul mercato sistemi di Intelligenza Artificiale, volendo allo stesso tempo tutelare e salvaguardare l'integrale sviluppo della persona umana nei confronti delle tecniche irrispettose della stessa. Invero il messaggio che passa dal Regolamento, noto anche come *AI Act*, è che per l'Unione Europea *non tutto ciò che è tecnicamente possibile è anche meritevole di essere realizzato e utilizzato*, per cui ci sono dei sistemi di IA che sono vietati o circoscritti da tante cautele perché ritenuti lesivi della persona umana e dei suoi diritti, quali ad esempio la libertà, la dignità, la privacy. Per tal modo si evita il rischio che questi strumenti da utili diventino dannosi per lo sviluppo integrale della persona umana.

I problemi legati allo sviluppo e utilizzo dell'IA mostrano forse con maggiore evidenza i limiti che caratterizzano le "geometrie" della scienza giuridica moderna, che riduce il diritto alla legge intesa come manifestazione di volontà del titolare del potere; rivelandoci come sia invece necessario concepire il diritto quale fatto sociale e nella sua dimensione sussidiaria rispetto alla capacità di autoregolamentazione delle persone nel consorzio umano.

> L'incapacità d'intendere l'esigenza d'autocontrollo da parte del singolo, per non far intendere l'esigenza del controllo da parte della società, per neutralizzare le reciproche violenze degli individui, costituiscono la ragione dell'ordinamento politico, inteso come dominio dell'uomo sull'uomo[21].

Per cui "il soggetto umano in quanto tale, cioè in quanto eticamente orientato, libero e responsabile, responsabile perché libero e libero perché eticamente orientato"[22], non può rispecchiarsi in una concezione del diritto che fonda la sua legittimazione uni-

[21] Cfr. F. Gentile, *op. cit.*, p. 57.
[22] Cfr. *ivi*, p. 99.

camente sul controllo che riesce a esercitare sulle persone assunte come atomi sociali e sensibili soltanto al proprio *particulare* e dunque bisognose di essere eterodirette "dalla culla alla bara".

Se guardiamo senza preconcetti al consorzio umano si può notare come il singolo riesca a esprimersi in una dimensione solidaristica, anche senza l'intervento legislativo diretto a imporre comportamenti di apertura sociale. Le persone, infatti, non possono essere intese come delle monadi chiuse alle esigenze del vivere insieme: difficilmente riuscirebbero a sopravvivere.

Per questo la dimensione sussidiaria del diritto consente di cogliere l'idea secondo cui regolamentare una materia così complessa, come quella dell'IA, è possibile solo mediante una vera e propria "*governance digitale*"[23] che ci permetta appunto di "governare" lo sviluppo tecnologico, evitando di lasciarsi travolgere, per renderlo funzionale alle persone e ancorarlo a una base antropologica con al centro le persone e i loro diritti fondamentali.

Questa *governance* presuppone una efficace dialettica tra le dinamiche di autoregolamentazione, la possibile *coregulation*, l'utilizzo di raccomandazioni e *soft regulation* (tra cui le linee guida), fino ad arrivare in ultima istanza a una regolazione vera e propria (*hard regulation*). Un buon esempio, a riguardo, è rappresentato dal Regolamento generale sulla protezione dei dati personali (GDPR 679/2016)[24] e, a ben vedere, dallo stesso Regolamento europeo sull'IA. L'*AI Act* prevede infatti diverse esigenze di regolazione in base ai diversi livelli di rischio, classificati in quattro gruppi: rischio inaccettabile; rischio elevato; basso rischio e rischio minimo. Che corrispondono sinteticamente a: tecnologia di IA vietata; tecnologia di IA che prima di essere immessa sul mercato europeo deve soddisfare determinati requisiti; tecnologie di IA soggette al principio di trasparenza[25].

[23] Cfr. L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, ed. it. a cura di M. Durante, Raffaello Cortina Editore, Milano, 2022, pp. 106-108.

[24] *Ibidem*.

[25] Per un approfondimento critico si veda J. Schuett, "Risk Management in the Artificial Intelligence Act", in *European Journal of Risk Regulation*, (2023), pp. 1-19.

Inoltre il Regolamento incentiva ad adottare e applicare su base volontaria dei codici di condotta[26]. Le buone pratiche dimostrano che un'etica nello sviluppo e utilizzo di questi strumenti[27] può risultare altrettanto efficace e sicuramente di supporto alla sua regolamentazione e alla più ampia e promettente *governance digitale*[28].

Rischiosa resta la via di chi, tra gli sviluppatori di sistemi di IA, non si pone limiti e indifferente alle implicazioni etiche mira ad ogni costo a massimizzare il profitto anche se questo comporta dei pericoli, aspettando poi che siano le istituzioni a porre rimedio ai danni eventualmente causati all'intera società.

[26] L'art. 69 del Regolamento europeo sull'IA prevede la possibilità di adottare volontariamente, anche per i sistemi di IA diversi da quelli ad alto rischio, un Codice di Condotta nel rispetto dei requisiti previsti dal Regolamento per i sistemi di IA ad alto rischio. Inoltre il 30 ottobre 2023 i leader del G7 hanno trovato un accordo sui Principi Guida Internazionali sull'IA e sul Codice di Condotta nel contesto dell'"*Hiroshima AI Process*", che riguarderà gli sviluppatori di IA.

[27] Sul punto cfr. P. Benanti, *Le macchine sapienti. Intelligenze artificiali e decisioni umane*, Marietti, Bologna, 2018, cap. 5.

[28] Cfr. L. Floridi, *op.cit.*, pp. 106-108.

Piero Marra

Un'idea simbiotica della giustizia digitale*

Abstract: **A Symbiotic Idea of Digital Justice.** Starting from an interpretation of Joseph Carl Robnett Licklider's paper "Man-Computer Symbiosis" (1960), the essay considers the possibilities of achieving human-computer/machine symbiosis for judicial decision-making from a procedural point of view. There are numerous limitations and risks associated with the use of artificial intelligence in the judicial process, but there can also be advantages, provided that the conditions of sense and meaning that make such a decision conceivable and possible are determined—conditions that can be synthetically found in the procedural nature of legal experience.

Keywords: Symbiotic Artificial Intelligence, Judicial Decision, Procedural Conditions, Significant Human Contribution, Socio-digital Reason.

Sommario: 1. L'evoluzione simbiotica dell'intelligenza artificiale – 2. Somiglianze – 3. Differenze – 4. Dalla sostituzione… alla simbiosi tra procedure e istituzioni.

1. L'evoluzione simbiotica dell'intelligenza artificiale

In un saggio intitolato *Man-Computer Symbiosis* (1960)[1], l'informatico e psicologo statunitense Joseph Carl Robnett Licklider ci introduceva la simbiosi tra l'uomo e il computer, distinguendola dai meccanismi di estensione delle capacità umane di cui aveva parlato, nel 1954, lo storico della scienza John David North[2]. Secondo Licklider, l'evoluzione simbiotica dell'intelligenza artificiale avrebbe operato in modo differente dalle tecnologie semi-automatiche che invece rimanevano strettamente condizionate dalla direzione, integrazione e pre-determinazione dell'uomo.

* Questo lavoro è stato parzialmente sostenuto dal progetto FAIR – Future AI Research (PE00000013), nell'ambito del programma MUR del PNRR finanziato dal NextGenerationEU.

[1] J.C.R. Licklider, "Man-computer symbiosis", in *IRE Transactions on Human Factors in Electronics*, (1960), n. 1, pp. 4-11.

[2] J.D. North, *The rational behaviour of mechanically extended man*, Boulton Paul Aircraft Ltd., Wolverhampton, 1954.

Secondo tale inedita visione, il punto focale dell'interazione uomo-computer non stava tanto nelle *risposte* che un sistema artificiale poteva dare, ma erano le *domande* di partenza a condizionare gli *outputs* del sistema[3]. Veniva così tracciata la possibilità che entità simbiotiche potessero generare decisioni su questioni complesse attraverso la "problematizzazione in tempo reale" di *casi* e combinando l'euristica umana con processi computazionali[4].

Il senso della simbiosi è abbastanza intuitivo: due entità *dissimili* co-esistono intimamente *associate* e traggono reciproci benefici. Un po' come accade tra il *Ficus* e il *Blastophaga grossorum*, il più noto tra gli aganoidi impollinatori. È un'assimilazione che ancora una volta antropomorfizza[5] la tecnica, similmente all'evocativa "trovata"di un gruppo di giovani ricercatori che, in occasione della conferenza di Dartmouth del 1956[6], definirono "intelligenti" macchine solo capaci di calcolare. Qui, però, il termine 'simbiosi' è usato in senso *informazionale*: la cooperazione uomo-computer veniva risolta mediante l'uso di interfacce (e dunque, ancora, algoritmi) per accorciare la separazione tra il linguaggio naturale dell'uomo e quello simbolico

[3] J.C.R. Licklider, *op. cit.*, p. 5.

[4] *Ivi*, p. 6: "[…] men are noisy, narrow-band devices, but their nervous systems have very many parallel and simultaneously active channels. Relative to men, computing machines are very fast and very accurate, but they are constrained to perform only one or a few elementary operations at a time. Men are flexible, capable of 'programming themselves contingently' on the basis of newly received information. Computing machines are single-minded, constrained by their "pre-programming.' Men naturally speak redundant languages organized around unitary objects and coherent actions and employing 20 to 60 elementary symbols. Computers 'naturally' speak nonredundant languages, usually with only two elementary symbols and no inherent appreciation either of unitary objects or of coherent actions".

[5] Sugli equivoci ingenerati da un uso antropomorfico dei termini in tema di *IA* che parlano di intelligenza, oracoli, allucinazioni, incantesimi, e non di sistemi, esiti delle ricerche, errori, difetti o malfunzionamenti, cfr. G. Finocchiaro, *Intelligenza artificiale. Quali regole?*, il Mulino, Bologna, 2024, p. 25. Si veda anche S. Zorzetto, "La metafora della IA: una giungla lessicale e foresta simbolica", in *Notizie di Politeia*, 151 (2023), pp. 179-185, ma anche A. Rossetti, "Dare regole alle tecnologie", in S. Ricci, A. Rossetti (a cura di), *Aspetti giuridici della società dell'informazione*/Legal Issue in the Information Society, Giuffrè Francis Lefebvre, Milano, 2025, pp. 1-14.

[6] Si tratta della proposta di J. McCarthy, M.L. Minsky, N. Rochester, C.E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 31/09/1955.

dell'elaboratore elettronico. Ne costituivano i prerequisiti: la condivisione in tempo reale delle informazioni; grandi capacità di memorizzazione ed organizzazione dei dati; inter-azione "paritetica" in cui l'uomo fissa gli obiettivi e il calcolatore traccia il percorso[7]. In effetti, il mondo ha intrapreso una direzione non diversa da quella appena accennata. Ve n'è traccia in alcuni passaggi decisivi della recente rivoluzione industriale[8]: la geometria cede il proprio spazio all'aritmetica ed alla fisica del *cloud computing*[9]; il *software* è *prius* rispetto all'*hardware* e la decisione informazionale si impone sull'operatività macchinica.

Ben oltre questi momenti, gli esiti divergenti delle tecniche di apprendimento automatico [*machine learning*] e di apprendimento profondo [*deep learning*] lasciano presagire un'imminente sostituzione dell'uomo in molte – e potenzialmente in tutte – le sue attività. A ben vedere, tuttavia, i recenti ritrovati della tecnica alla base dei modelli fondativi [*foundation model*] ora definiti come *general purpose AI model* (GPAI) e cioè sistemi che esibiscono "originali" capacità generali senza comportamenti deterministici[10], pur presentando notevoli rischi *generici*

[7] Nella sostanza, tuttavia, l'uomo e i suoi bisogni restano primari, mentre macchina e uomo contribuiscono entrambi ad implementare il sistema: da "ciascuno" secondo le proprie capacità (rispettivamente intenzionali e computazionali), all'uomo secondo i suoi bisogni. In questo modo l'inter-azione è solo in apparenza paritetica. D'altronde si tratta di entità onticamente differenti. Lo rilevano A. Beckers, G. Teubner, *Three Liability Regimes for Artificial Intelligence. Algorithmic Actants, Hybrids, Crowds*, Hart Publishing, Oxford, 2021; ne dà un'ampia trattazione O.G. Loddo, "L'agire sociale ai tempi dell'intelligenza artificiale. Il concetto di 'istituzione sociale-digitale'", in *L'Ircocervo*, 23 (2024), n. 1, pp. 359-371, in part. p. 361.

[8] L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Raffaello Cortina, Milano, 2017.

[9] Sulla despazializzazione del diritto e della giustizia è dedicato il libro di A. Garapon, *La despazializzazione della giustizia*, trad. it., Mimesis, Milano, 2021; per la logica del *cloud computing* si rinvia a P. Sommaggio, "Dalla scrivania alla nuvola e ritorno. Riflessioni filosofico-giuridiche sul *cloud computing*", in P. Moro, C. Sarra (a cura di), *Tecnodiritto. Temi e problemi di informatica e robotica giuridica*, FrancoAngeli, Milano, 2017, pp. 179-218; ed ancora, per i temi collegati alle applicazioni amministrative, A.C. Amato Mangiameli, "Connettività pubblica e rete internazionale", in Id., *Informatica giuridica*, Giappichelli, Torino, 2015, pp. 289-294.

[10] Su rischi e opportunità di questo modello, si rinvia all'ampio contributo di R. Bommasani *et al*., "On the opportunities and risks of foundation models", in *arXiv*, 2022 (arXiv:2108.07258); ed anche A.C. Amato

e *sistemici*, costituiscono al contempo l'occasione per ripensare la stessa interazione uomo-computer [*human-computer-interaction*] e uomo-macchina [*human-machine-interaction*].

Al centro della sfida tecnologica, dunque, tornano il momento decisionale e il ruolo della simbiosi *informazionale*. Si tratta di una visione non limitata alle tecnologie incorporate di tipo "simbiontico", che pure aprono a molti interrogativi e a rilevanti questioni di compatibilità etica e giuridica[11]. Non si può infatti ritenere che la simbiosi si realizzi solo per quei sistemi artificiali indossabili [*wearable*] o impiantabili [*implantable*]. V'è in altre parole un livello di simbiosi digitale differente che riguarda in particolare la decisione critica a cui l'uomo, quale agente libero e responsabile, è naturalmente chiamato a prendervi parte, non come ricettore passivo di dati, ma come "elaboratore" attivo di informazioni.

In generale, l'idea che la simbiosi possa *problematizzare* gli esiti della decisione digitale accorda i differenti approcci alla materia, suddivisi tra chi identifica l'intelligenza dell'automa con il pensiero (rappresentazione della conoscenza e ragionamento), e quanti la considerano solo come percezione e interazione con l'ambiente; tra la posizione di chi ritiene che l'intelligenza debba riprodurre fedelmente le capacità dell'uomo (con

Mangiameli, "Intelligenza artificiale generativa: brevi note", in *L'Ircocervo*, 23 (2024), n. 1, pp. 101-111 e per le definizioni normative, *ivi*, p. 109 nt. 35, ora pubblicato in questo volume.

[11] Sono ormai note le sperimentazioni della società *Neuralink* e l'uso di microchip neurali nel cervello umano per curare alcune patologie neurologiche. Cfr. A. Agnihotri, S. Bhattacharya, *Neuralink. Invasive neurotechnology for human welfare*, Sage publications, Los Angeles, 2023. Per un'introduzione al tema ed ai rischi connessi, con particolare riferimento alla tutela della *privacy* e agli strumenti *by design* di controllo e comprensione umana si rinvia a W. D'Avanzo, "Intelligenza artificiale e sanità digitale. Il diritto *by design* e il ruolo del *legal design* per la protezione dei dati personali", in *L'Ircocervo*, 23 (2024), n. 1, pp. 421-423, ora in questo volume. In generale, offre un quadro dei principî etici in tema di intelligenza artificiale F.J. Santamaría Ramos, "Implicaciones éticas y legales relacionadas con el uso creciente de la", in *L'Ircocervo*, 23 (2024), n. 1, pp. 139-174 e T. Garcia Berrio, "Ética de la virtud para una IA compasiva", in *L'Ircocervo*, 23 (2024), n. 1, pp. 318-340, entrambi pubblicati, con aggiornamenti, in questo volume. Sul fronte inedito dei "neurodiritti", rinvio all'ampia trattazione di A. D'Aloia, M.C. Errigo (eds.), *Neuroscience and Law: Complicated Crossing and New Perspectives*, Springer, Chaim, 2020.

tutti i limiti), e quella opposta di coloro che vogliono realizzare sistemi solamente razionali (sfruttando la correlazione delle informazioni per realizzare scelte ottimali)[12].

Restano, però, da capire le condizioni di possibilità e pensabilità della decisione simbiotica all'interno di una certa "istituzione sociale-digitale"[13]. Non è una questione irrilevante se si pensi che la dimensione informazionale dei sistemi digitali, oltre a incidere sulle scelte umane, costituisce essa stessa nuove realtà[14]. Basti ricordare che all'"inconscio digitale" è seguita la "morte digitale". La stessa solitudine o la demenza diventano stati digitali, e così ogni concetto tradizionalmente legato alla spazialità vissuta ha il suo analogo informazionale[15]. Due mondi incomunicabili, ma comunque inter-dipendenti. Il tema è tra l'altro particolarmente sensibile per il discorso giudiziale che è luogo privilegiato della discussione problematica di casi (è questo il tratto distintivo del *case method*) spesso risolti nella disposizione unica e irripetibile della decisione del giudice. Anche questa, in effetti e in senso ampio, è l'esito di processi informazionali che si condensano in un linguaggio dal chiaro senso performativo. A tale campo di esperienza umana sono tra l'altro dedicate le minime riflessioni di questo capitolo.

L'originale suggestione di Licklider, almeno nell'interpretazione che è stata qui tratteggiata, è invece il nostro *incipit*. Discuterò

[12] S.J. Russel, P. Norvig, *Artificial Intelligence. A Modern Approach*, ed. 2, Prentice Hall, Upper Saddle River, 2010, p. 2.

[13] Il concetto di "istituzione sociale-digitale" è spiegato da O.G. Loddo, *op. cit.*, p. 362.

[14] Su questo aspetto e sui rilievi epistemologici, v. F. Romeo, "La produzione culturale dell'artificiale: prodromi di uno studio metaepistemologico su diritto ed informazione", in *L'Ircocervo*, 23 (2024), n. 1, pp. 51-69, ora in questo volume.

[15] D. de Kerckhove, "Conclusioni. L'inconscio digitale", in A. Buffardi, D. de Kerckhove (a cura di), *Il sapere digitale. Pensiero ipertestuale e conoscenza connettiva*, Liguori editore, Napoli, 2011, pp. 65 ss.; sul secondo concetto rinvio all'ampia trattazione di G. Resta, "La 'morte' digitale", in *Il diritto dell'informazione e dell'informatica*, XXIX (2014), n. 6, pp. 891-920; quanto agli *status* psicologici, si veda ampiamente G. Saraceni, "La IA(D). Intelligenza Artificiale e benessere psicologico", in *L'Ircocervo*, 23 (2024), n. 1, in part. p. 394 e nt. 12, ma anche in questo volume Id. "Intelligenza Artificiale e benessere psicologico".

dialetticamente alcune questioni riguardanti il controverso rapporto tra decisione umana e artificiale nel campo giudiziale, per verificare se sia possibile, e in che termini, pensare alla simbiosi come condizione procedurale necessaria del *tecnodiritto*, e della dimensione tecno-informatica dell'esperienza giuridica in generale[16].

2. Somiglianze

Non v'è dubbio che la correlazione tra i dati operata da sistemi di intelligenza artificiale funzioni con meccanismi non del tutto dissimili dalla costruzione del ragionamento umano in campo giudiziale. Ma c'è di più. Se da un lato le recenti applicazioni di intelligenza artificiale dimostrano abilità sintattiche, dall'altro *mostrano* anche autonome e altamente sofisticate capacità di interazione con l'ambiente circostante. E per tale ragione che non sono pochi ad auspicare l'avvento di sistemi di giustizia automatizzata, più efficienti e senza alcuno spazio per orpelli ed incertezze. Almeno nelle linee programmatiche. È una possibilità, tra le molte, che può anche darsi nella costruzione di un certo tipo di giustizia "digitale". Le questioni investono, in questo modo, i punti di contatto tra la ragione umana e quella computazionale. Vediamone i tratti essenziali.

In astratto (e solo in astratto) la giurisdizione viene pensata come proiezione ortogonale della legislazione. Sta per garantire che i tanti divieti, diritti, permessi, obblighi vengano effettivamente applicati. Al diritto serve la giustizia (in senso legale, per dirla con Norberto Bobbio[17]) e questa al diritto[18]. Il discorso ri-

[16] Sui risvolti teorici, fenomenologici e pratici del *tecnodiritto*, rinvio all'interessante volume di P. Moro, C. Sarra (a cura di), *op. cit.*, in part. pp. 7-10 e di recente P. Moro, "Intelligenza artificiale e tecnodiritto. Fondamenti etici ed innovazione legislativa", in Id. (a cura di), *Etica, diritto e tecnologia. Percorsi dell'informatica giuridica contemporanea*, FrancoAngeli, Milano, pp. 7-24, nonché a A.C. Amato Mangiameli, "Tecno-diritto e tecno-regolazione. Spunti di riflessione", in *Rivista di filosofia del diritto*, 6 (2017), pp. 87-112. Altro contributo alla tematica è di recente apparso anche nel fascicolo n. 1 (2025) della Rivista di filosofia del diritto, *special issue* dedicato a "Intelligenza artificiale e diritto" a cura di L. Palazzani, pp. 7-80.

[17] N. Bobbio, voce "Ragionamento giuridico", in *Contributi ad un dizionario giuridico*, Giappichelli, Torino, 1994, p. 284.

[18] G. Zagrebelsky, *La legge e la sua giustizia. Tre capitoli di giustizia costituzionale*, il Mulino, Bologna, 2008, p. 161.

mane lo stesso a voler considerare non il codice, ma il vincolo giudiziale del precedente proprio dei sistemi di *Common Law* e in parte anche del vecchio Continente se si guarda alla formazione essenzialmente casistica del diritto dell'Unione Europea[19]. Anche il precedente, infatti, è formalizzabile in una regola che procede per imitazione. In entrambi i casi, il diritto dei giudici [*Richterrecht*] ben potrebbe ridursi alla logica deduttiva in senso stretto che fa derivare da tutta la conoscenza giuridica (rappresentabile mediante proposizioni o regole) conclusioni logiche fondate su operazioni sintattiche che pongono (è il *modus ponendo ponens*) o tolgono (è il *modus tollendo tollens*) la verità dell'asserzione conclusiva. È questa la logica proposizionale, di derivazione aristotelica, definibile come *analitica* o *formale* e che pure ha contrassegnato molti passaggi della nostra storia. Era tra l'altro l'ideale seguito dalla "scuola dell'esegesi" (la francese "*Ecole de l'exégese*") che ha pervaso tutto il XIX secolo, per realizzare l'esigenza di calcolabilità del diritto. Stando a questo modello, l'attività del giudice consisterebbe nel trarre "diritto" *dal* "diritto che si ha" attraverso operazioni prima sintattiche e poi semantiche. *Sintatticamente* il passaggio è da una norma generale ad una individuale; *semanticamente* il passaggio viene assicurato dalla possibilità di far rientrare oggettivamente il significato di una proposizione particolare in quello di una proposizione generale.

Anche i modelli dialettico o retorico e quello *fuzzy* possono essere assimilati alle logiche proposizionali proprie della calcolabilità tecnica che subisce gli effetti di discorsi apofantici e cioè dei discorsi inerenti ai giudizi di verità o falsità. Da un lato, il modello *dialettico* o *retorico* tratta le proposizioni possibilmente vere e possibilmente false[20]. A contare sono gli argomenti, così

[19] Per un approfondimento sul ruolo del precedente rinvio al volume di A. Carleo (a cura di), *Il vincolo giudiziale del passato. I precedenti*, il Mulino, 2018. Per uno studio applicativo del precedente G. De Martino, P. Marra, A. D'Aversa, L. Pulito, A. Pellicani, G. Pio, M. Ceci, "Leveraging textual content, citational aspects and dissenting opinions through a multi-view contrastive learning methodology for legal precedent analysis", in *Computer Law & Security Review*, 60 (2026), pp. 1-17 (https://doi.org/10.1016/j.clsr.2025.106257).

[20] Segnalo sul tema A. Incampo, *Metafisica del processo. Idee per una critica della ragione giuridica*, ed. 2, Cacucci, Bari, 2016 e G.A. Ferrari, M.

come le tecniche retoriche per persuadere chi ascolta, dal momento che la persuasione è, a richiamare ancora Aristotele, un concetto di relazione[21]. Il modello *fuzzy*, dall'altro lato, rappresenta la logica della verità per *approssimazione* o secondo una scala di valori. Questa volta sono interessate le proposizioni "più o meno vere", "non del tutto vere" e "non del tutto false"[22]. Anzi, secondo i sostenitori di questo modello, la decisione giudiziale è sempre l'effetto della *fuzziness* e cioè della vaghezza e/o indeterminatezza e/o ambiguità del linguaggio normativo. Lo si vede, in particolare, in tutti i casi in cui occorre risolvere questioni giuridiche aperte o per determinare le stesse conseguenze sanzionatorie dell'illecito.

Non ragionano diversamente i sistemi artificiali esperti che usano le inferenze per passare da certe premesse alle conclusioni. Entrambi questi schemi sono formalizzabili ed automatizzabili con strutture sintattiche che richiamano i simboli della matematica, organizzano i dati con la programmazione logica e, infine, stabiliscono correlazioni[23]. Queste *rappresentazioni*

Manzin, *La retorica fra scienza e professione legale. Questioni di metodo*, Giuffrè, Milano, 2004.

[21] Ce lo ricorda Aristotele, *Ars Rhetorica*, 1356b, [recensuit et brevique adnotatione critica instruxit W. D. Ross], 1959, trad. it. di M. Dorati, *Retorica*, Mondadori, Milano, 1996, p. 17.

Da questo punto di vista, non sono da confondere le tecniche artificiali di convincimento con quelle retoriche di persuasione che dipendono dalla relazione con l'uditore: è il modello della nuova retorica di C. Perelman, *Trattato dell'argomentazione. La nuova retorica*, trad. it con una prefazione di N. Bobbio, Einaudi, Torino, 1966.

[22] Alcune caratterizzazioni della logica *fuzzy* sono formulate da L.A. Zadeh, "Fuzzy Logic and Approximate Reasoning", in *Synthese*, 30 (1975), pp. 407-428. Per la teoria della logica *fuzzy* nei modelli esplicativi delle decisioni giudiziali si rinvia a T. Mazzarese, *Forme di razionalità delle decisioni giudiziali*, Giappichelli, Torino, 1996.

[23] La proposta di usare i connettivi della logica formale per esprimere i rapporti tra proposizioni nei testi normativi e nei contratti è in L.E. Allen, "Toward a Normalized Language to Clarify the Structure of Legal Discourse", in A.A. Martino (ed.), *Deontic Logic, Computational Linguistics, and Legal Informations Systems*, North-Holland Publishing Company (Elsevier), North Holland, 1982, pp. 349-407. La programmazione logica consiste nell'uso della logica simbolica per la rappresentazione esplicita di problemi e delle basi di conoscenza a questi associate, assieme all'uso di inferenze logiche controllate

della conoscenza sono combinate con le istruzioni (gli algoritmi) per compiere diversi tipi di inferenze (deduttive, presuntive, induttive, probabilistiche, argomentative, basate su casi, ecc.). In generale, tali tecniche di programmazione postulano l'idea che l'intelligenza dipenda unicamente dalla conoscenza. Potrebbe pertanto ritenersi sufficiente fornire al sistema tutta quella necessaria (*knowledge-based system*) e tradurla mediante regole (*rule-based system*) secondo due schemi: (i) *conclusivo* (è il mondo della deduzione analitica in senso stretto) e (ii) *presuntivo* (è il mondo di ciò che avviene normalmente ed ammette eccezioni)[24].

Negli ultimi anni, tuttavia, grazie all'uso computazionale di grandi quantità di dati (*big data*), le tecniche di apprendimento automatico hanno superato i limiti dei primi sistemi. Con questi metodi le macchine possono estrarre informazioni rilevanti dai dati di *input* e così imparare meglio a svolgere le loro funzioni. Per questa ragione, come aveva anticipato Alan Turing negli anni '50, una macchina capace di apprendere può operare in modi che non siano stati previsti dai suoi creatori e con meccanismi di funzionamento oscuri. A volte, oltre i compiti assegnati, le stesse macchine costruiscono il proprio ingegno con procedure incontrollabili e risultati inattesi (*black box effect*)[25]. Sono state così realizzate numerose applicazioni che generano esse stesse il modello di conoscenza su cui si fondano le proprie prestazioni inferenziali, senza trascurare gli esiti dell'apprendi-

per la soluzione effettiva di tali problemi: così la definisce C.J. Hogger, R.A. Kowalski, "Logic Programming", in S.C. Shapiro (eds.), *Encyclopedia of Artificial Intelligence*, Wiley, Hoboken (NJ), 1987, pp. 544-558.

[24] Sul tema del rapporto tra normalità e norme si veda G. Siniscalchi, *Normalità di norme*, Cacucci, Bari, 2007.

[25] Così A.M. Turing, "Computer Machinery and Intelligence", in *Mind*, 59 (1950), n. 236, pp. 433-460. Sul problema del *black box effect* e sulle sue ragioni e implicazioni, S. Amato, "Macchinazioni. Soggettività, agentività ed entità esotiche", in *L'Ircocervo*, 23 (2024), n. 1, pp. 289-306, ora in questo volume. Sui rischi collegati ai modelli *general purpose*, oltre ad A.C. Amato Mangiameli, "Intelligenza artificiale generativa", cit., si veda anche M. Farina, "Degenerazione e rischio creativo dell'Intelligenza Artificiale "forte": forme di prevenzione e tutela complementare", in *L'Ircocervo*, 23 (2024), n. 1, pp. 122-138 e, in questo volume, Id. "L'arte di non degenerare: ecosistemi tracciabili dell'IA tra degrado informativo e rischio creativo".

mento profondo che sfrutta addirittura reti neurali per imitare la complessità dell'organizzazione del cervello umano. Infine, vi sono i sistemi *General Purpose* che costantemente danno risposte a tutte le questioni, apparendo originali[26].

La realtà ha così travalicato l'immaginazione, ben oltre l'idea aristotelica della meccanica come tecnica finalizzata al superamento dei limiti naturali dell'uomo[27]. Il *Machine Learning* (ML), il *Deep Learning* (DL) e la combinazione di *Artificial Neural Network* e *Foundation Model* sono effettivamente capaci di pervadere ogni processo decisionale umano. Si parla, non a caso, di tecnologie *disruptive*[28]. Esistono, ad esempio, sistemi che rilevano e raccolgono dati sulle emozioni e che, catturando le espressioni facciali, la direzione dello sguardo, le intonazioni della voce e altri segnali, svolgono operazioni assimilabili a quelle proprie del comprendere. È il campo dell'Internet dei corpi [*Internet of Bodies*] e di una sua branca, l'Internet dei comportamenti [*Internet of Behaviours*] che sfocia negli studi sinergici della interazione uomo-macchina[29]. Si tratta di tecniche che,

[26] P. Thagard, *Cervelli a confronto. Perché l'intelligenza umana è diversa da quella degli animali e dei robot*, trad. it., FrancoAngeli, Milano, 2021, p. 34.

[27] Aristotele, *Meccanica*, trad. it., Bompiani, Milano, 2010.

[28] La *disruption* indica una trasformazione permanente e non un momento accidentale. Lo sottolineano A. Garapon, J. Lassègue, *La giustizia digitale. Determinismo tecnologico e libertà*, trad. it., il Mulino, Bologna, 2021, pp. 234-237. D'altro canto, l'*AI* è "una delle forze più trasformative del nostro tempo […] destinata a modificare il tessuto della società" (così *Hight-Level Expert Group on Artificial Intelligence*, AI HLEG 2019, nominato dalla Commissione europea per la redazione dei principî etici per una AI affidabile). Lo ricordano L. Floridi, F. Cabitza, *Intelligenza artificiale. L'uso delle nuove macchine*, Bompiani, Firenze, 2021.

[29] Sulla rilevanza dello studio dei comportamenti nel campo sanitario si v. L.A. Marsch, S.E. Lord, J. Dallery, *Behavioral healthcare and technology*, Oxford University Press, Oxford, 2015, p. 83. Sono anche interessanti gli studi sulle applicazioni che mostrano le somiglianze tra l'agire umano e quello artificiale. In questo senso, mi permetto di segnalare, sempre per il settore sanitario: M.A. Bochicchio, S. Corciulo, "Symbiosis and Synesthesia in Proactive Conversational Agents for Healthy Ageing", in A. Dix, M. Roach, T. Turchi, A. Malizia, B. Wilson (eds.), *Proceedings of the 1st International Workshop on Designing and Building Hybrid Human-AI Systems co-located with 17th International Conference on Advanced Visual Interfaces (AVI 2024)*, CEUR, Vol. 3701, 2024 (https://ceur-ws.org/Vol-3701/paper10.

sfruttando i sistemi generici e l'*imaging* di risonanza magnetica funzionale (fMRI), consentono all'automa di leggere la mente del soggetto umano che interagisce fino a mostrarsi emotivamente sensibili[30]. E questo solo per citare alcuni esempi.

Appena, dunque, si cerca di vedere le differenze emergono le somiglianze. Molte caratteristiche dell'intelligenza umana le si rinvengono, in definitiva, nell'automa: percezione, pianificazione, decisione, comprensione, apprendimento, astrazione, correlazione, creazione, ragionamento, comunicazione, e finanche la mimica dei sentimenti, sono solo alcuni dei modi di funzionamento delle intelligenze artificiali[31].

Ma ciò può essere sufficiente per sostenere la sostituzione? È il tema del prossimo paragrafo.

3. Differenze

Accanto alle analogie, numerose sono le differenze con riferimento all'ambito giudiziale. Se ne possono individuare almeno tre, corrispondenti ad altrettanti piani di indagine.

pdf); B. Wilson, D. Scott, M. Roach, E. Nielsen, B. Müller, "Designing for situated AI-human decision making: Lessons learned from a primary care deployment", in A. Dix, M. Roach, T. Turchi, A. Malizia, B. Wilson (eds.), *op. cit.* (https://ceur-ws.org/Vol-3701/paper11.pdf). In generale, per gli studi che non si limitino al campo medico, rimando ai recenti lavori di B. Fu, P. Gatsby, A. Ryan Soriano, K. Chu, N. Guardado, "Towards Intelligent Flight Deck – A Preliminary Study of Applied Eye Tracking in The Predictions of Pilot Success and Failure During Simulated Flight Takeoff", in A. Dix, M. Roach, T. Turchi, A. Malizia, B. Wilson (eds.), *op. cit.* (https://ceur-ws.org/Vol-3701/paper12.pdf).

[30] Molti sono gli interrogativi che sorgono a partire alle scoperte ingegneristiche che vanno oltre le macchine. Alcune sono riportate in S. Amato, *op. cit.*, pp. 297-298.

[31] P. Thagard, *op. cit.*, pp. 30 ss. Sulle distinzioni tra il modo tipicamente umano di stare al mondo e le intelligenze artificiali, P. Moro, "Persona elettronica. Una finzione giuridica per l'intelligenza artificiale", in *L'Ircocervo*, 23 (2024), n. 1, pp. 372-389, ora in questo volume.

(i) *Differenze epistemiche*. Toccano innanzitutto la grammatica della decisione umana e di quella artificiale. Si guarda innanzitutto allo statuto epistemico del numero. La rivoluzione digitale è, infatti, una rivoluzione grafica che assegna al numero e alla combinazione binaria un ruolo primario: la scrittura digitale codifica, sotto forma di numeri, fatti ed eventi del mondo fisico[32]. Ma i numeri non hanno carattere deontico, così come non costituiscono *status* deontici i dati che sfruttano la loro correlazione con i segni algebrici[33]. In questo senso, non sono pochi i programmatori che confondono il piano dell'*essere* con quello del *dover essere*, entrambi appartenenti all'esperienza giuridica[34]. Non a caso i fatti decisivi per il diritto e per il discorso giudiziale non sono semplici fatti (o fatti bruti, nell'accezione data da John Searle)[35], ma fatti che dipendono dalle istituzioni in cui trovano senso e significato. Da un polo all'altro c'è uno scarto che richiede innanzitutto un processo di costruzione e ricostruzione del significato del fatto da parte del soggetto. E ciò postula quantomeno che il fatto da conoscere sia *in sé* e *per sé* conoscibile. A rilevare in questo passaggio non è tanto l'efficacia della decisione giudiziale, quanto la sua esistenza. In questo modo, il numerico rimane un "feticcio" e non è, dunque, giuridicamente rilevante. Tuttavia, scomparendo il fatto, non c'è nemmeno il diritto. *Da mihi factum, dabo tibi ius*.

D'altronde, la stessa relazione tra i numeri in base alle regole dei segni algebrici, "regole che fissano una precisa sintassi del-

[32] Lo evidenziano con efficacia A. Garapon, J. Lassègue, *op. cit*., pp. 37-48.

[33] Solo in parte coincidente con quanto afferma V. Zeno-Zencovich, "'Big data' ed epistemologia giuridica", in. S. Faro, T.E. Frosini, G. Peruginelli (a cura di), *Dati e algoritmi. Diritto e diritti nella società digitale*, il Mulino, Bologna, 2020, pp. 13-24. Il dato, infatti, è una traduzione informazionale del numero, come tale non idoneo in sé a costituire gli *status* deontici che, invece, hanno bisogno di ulteriori procedure di trasformazione e convalida *in e per* un ordinamento normativo. Semplifico di molto le tesi di A.G. Conte, di recente ampiamente riportate in P. Di Lucia, L. Passerini Glazel, "La norma come *status* deontico. Una svolta ontologica nella semiotica del normativo", in *L'Ircocervo*, 21 (2022), n. 1, pp. 77-109, in part. pp. 105-106.

[34] P. Marra, *Per una moralità procedurale del diritto. Considerazioni attuali a partire da Lon L. Fuller*, Cacucci, Bari, 2022.

[35] J.R. Searle, *Making the Social World. The Structure of Human Civilization*, Oxford University Press, Oxford, 2010.

le operazioni tra i numeri", fa emergere l'intreccio tra intelletto e ragione inferenziale. I numeri ordinano, ma non manipolano. Basti pensare che le relazioni tra numeri (relazioni di quantità) sono segnate profondamente dal gioco dei segni algebrici (e cioè relazioni di qualità); a loro volta, queste ultime consentono al numero di dare risultati in una procedura di calcolo solo grazie all'intelletto[36]. Per altra parte, la sovrapposizione del piano fattuale con quello assiologico richiede la mediazione umana di tipo *intenzionale* in grado di comprendere il carattere finalizzato e propositivo dell'esperienza umana e di quella giuridica in quanto tale[37]. Si tratta di un'attività creativa, seppure vincolata alla responsabilità del giudice, di un nuovo diritto anche oltre quello esistente. Dinanzi al giudicante si palesano perciò i "casi" ossia dei fatti dotati di senso in cui la distinzione tra ciò che *è* e ciò che *deve essere* appare fortemente attenuata[38]. Al diritto non bastano i numeri, né le relazioni di quantità.

(ii) *Differenze retoriche*. Questa volta si guarda alla possibilità dei fenomeni che dipendono strettamente dalla volontà particolare di chi decide. Vi rientra, per esempio, il carattere dialettico della costruzione della verità nel processo con riferimento sia all'interpretazione delle disposizioni normative (*quaestio iuris*), sia alla prova dei fatti o dei giudizi sui fatti (*quaestio facti*). La dimensione retorica dell'intero processo è la dimostrazione

[36] Lo rileva P. Savarese, "Diritto, sapere e intelligenza artificiale: una traccia epistemica", in *L'Ircocervo*, 23 (2024), n. 1, pp. 72-79 il quale ricostruisce la distinzione di Platone, *Repubblica*, 509d-510a tra *nous* (intelletto o intelligenza, capacità di comprendere l'essenza delle cose) e *dianoia* (ragione argomentativa, ossia capacità di calcolare o inferire conseguenze a partire da verità colte con intelligenza). Una distinzione in parte differente da quella tracciata da Aristotele in *Etica Nicomachea*.

[37] Lon L. Fuller, "Reason and Fiat", in *Harvard Law Review*, 59 (1946), p. 376. Sul ruolo della intenzionalità del soggetto per la costruzione giuridica, rinvio a P. Marra, "'Intenzionalità' del diritto in Lon L. Fuller", in *L'Ircocervo*, 20 (2021), n. 2, pp. 186-202.

[38] Tra l'altro è questa la tesi di G. Zagrebelsky, *op. cit.*, pp. 179-198. In queste pagine, il riferimento al "caso vivo" più che al "fatto morto" per l'interpretazione giuridica, consentirebbe, secondo Zagrebelsky, di superare la *Grande Divisione* tra *essere* e *dover essere* alla base della *legge di Hume* poiché l'attività del giurista si rivolgerebbe e seguirebbe le leggi di essenza.

più chiara della precarietà della verità che si alimenta del sapere *doxastico* proprio delle opinioni[39].

Ne sono prova il ruolo decisivo del giudizio di equità che fa proprio il linguaggio della morale[40]; l'integrazione valutativa delle clausole generali e le tecniche di formazione giudiziale necessarie per la costruzione di concetti indeterminati od elastici (quali "inadempimento grave", licenziamento "senza giustificato motivo o giusta causa"…) o di quelli normativi ("diritto altrui", "bene demaniale", "danno ingiusto"…)[41]; le numerose tecniche di retorica che sfruttano l'elaborazione empatica e umoristica dell'argomentazione, puntando alla *brevitas* e persino all'uso accorto del silenzio che in effetti non sembra confacente a sistemi sempre alla ricerca di risposte computabili[42]; per non tacere del ruolo della misericordia (e fors'anche della carità) nella determinazione della pena la cui effettività può essere condizionata dagli orientamenti culturali[43].

[39] S. Fuselli, *La verità nel processo. Percorsi di logica ed epistemologia*, FrancoAngeli, Milano, 2022

[40] L'espressione è di R.M. Hare, *Il linguaggio della morale*, trad. it., Ubaldini, Roma, 1968.

[41] La letteratura sul tema è particolarmente vasta, mi permetto di segnalare, senza che possano esaurire la trattazione, C. Luzzati, *La vaghezza delle norme. Un'analisi del linguaggio giuridico*, Giuffrè, Milano, 1990 (sul tema delle clausole generali, in particolare pp. 302-303), V. Velluzzi, *Le clausole generali. Semantica e politica del diritto*, Giuffré, Milano, 2006, in particolare pp. 10-11, M. Barcellona, *Clausole generali e giustizia contrattuale. Equità e buona fede tra codice civile e diritto europeo*, Giappichelli, Torino, 2006, S. Patti, *Ragionevolezza e clausole generali*, Giuffré, Milano, 2016, p. 53. Per il legame di tali aspetti con l'informatica giuridica, L. Lombardi Vallauri, "Norme vaghe e teoria generale del diritto", in *Ars Interpretandi*, (1998), pp. 153-163, in part. pp. 160-161.

[42] Su questo profilo rinvio all'interessante studio di A. Arnese, Advocati fides *e strategie difensive nelle Lettere di Plinio*, Cacucci, Bari, 2023. Sul ruolo della virtù nelle professioni, I. Trujillo, "Virtù e professioni giuridiche: i limiti della deontologia, in *Ars Interpretandi*, XI (2022), n. 1, pp. 87-102; dal lato del giudicare, L. Corso, "La virtù del giudice fra emozioni, giustizia particolare e ruolo istituzionale. Un abbozzo a partire da Aristotele", in *Ars Interpretandi*, XI (2022), n. 1, p. 53.

[43] T. Snarski, "'Merciful Justice' and the Challenges of Posthumanism in Legal Philosophy: Some Comments on the Criterion for Evaluating Technological Progress in the Criminal Justice System", in *L'Ircocervo*, 23 (2024), n. 1, pp. 273-288, ma anche H.E. May, "Virtue Jurisprudence and the

E che dire dei metodi di interpretazione? Sono illuminanti alcune considerazioni di Luigi Lombardi Vallauri nel suo *Corso di filosofia del diritto* (1981). I vari tipi di interpretazione possono essere distribuite in coppie antitetiche. Poniamo: 1) l'interpretazione "oggettiva" e "soggettiva"; 2) l'interpretazione "storica" ed "evolutiva"; 3) l'interpretazione "letterale" e "fondamentale"; 4) l'interpretazione fondamentale "concettuale" e fondamentale "teleologica"; 5) l'interpretazione "settoriale" e "sistematica"; 6) l'interpretazione "sistematica parziale" e "sistematica totale". La cosa interessante è che le sei coppie non sono mai esclusive, ma si possono combinare e, tolte le combinazioni prive di senso e applicando per ogni metodo l'argomento *a simili* e quello *a contrario*, al giurista si apre un ventaglio di almeno 144 usi possibili di ogni proposizione normativa[44]. Ora, pur non essendo un risultato assoluto (comunque minimo), non è tanto il rapporto sintattico tra questi metodi a rilevare, quanto lo spazio ermeneutico in cui è chiamato a muoversi il soggetto per valutare sia le incognite del divenire (di ciò che non è prevedibile), sia le incognite dell'individuale (di ciò che, pur essendo astrattamente prevedibile, non tiene conto delle singolarità e, dunque, torna a non essere prevedibile)[45].

Non solo. Nel confondere generalità e particolarità, vi è il giudizio sui fatti. Ad essere impossibile per il processo è questa volta l'oggettività dei fatti umani. Non è sufficiente nemmeno la previsione *ex ante* di una "generalizzazione empirica" espressa in massime di esperienza o in leggi scientifiche. Queste, in effetti, non sono in grado di schematizzare tutte le circostanze o condizioni singolari, comprese le anomalie delle serie causali in un fatto passato *qua* passato. D'altronde, a mancare è proprio la

Function of Punishment in International Criminal Law", in A. Incampo, W. Żełaniec (eds.), *Universality of Punishment*, Cacucci, Bari, 2015, pp. 187-213. Sulla rilevanza della cultura per la percezione della pena, O. Höffe, *Globalizzazione e diritto penale*, trad. it., Comunità, Torino, 2001.

[44] Le riflessioni sono di L. Lombardi Vallauri, *Corso di filosofia del diritto*, ed. 2, Cedam, Padova, 1981, pp. 55-58.

[45] Sulla libertà metodologica e sulla distinzione tra "interpretare" e la semplice ricognizione dei tanti risultati possibili dei canoni di interpretazione, cfr. L. Lombardi Vallauri, "Informatica e criteri 'politici' o valutativi della decisione giuridica", in *Jus*, (1982), n. 3, pp. 303-315.

simultaneità tra l'accadimento del fatto e la sua conoscenza. Sicché, il fatto non può mai essere accertato *in sé* e *per sé.* In questo senso, i risultati della statistica (comprensiva dell'apporto neuroscientifico) sono sempre insoddisfacenti nella misura in cui si riducono alla fallacia del formalismo interpretativo che aspira a catturare e/o misurare *simul et idem* la fatticità del reale. Difatti, la verifica scientifica di un "antecedente" nei sistemi complessi porta in genere a stabilire una condizione soltanto *sufficiente* e non anche *necessaria* di un accadimento[46].

Restano pure insufficienti i risultati della statistica alla base dei modelli di apprendimento automatico e profondo, oltre a quello *general purpose*[47]. Il rischio è quello di far entrare nel giudizio il concetto di *causabilità* piuttosto che di causa di un fatto particolare. I problemi, tuttavia, sono almeno due: il primo incide su consolidati paradigmi epistemologici, in particolare nel giudizio penale ove vige il principio di tassatività e determinatezza della fattispecie; il secondo rimane l'insufficienza del calcolo di probabilità nell'effettivo accertamento causale tra fatto ed evento. In pratica, quando il dato si confronta con la realtà e con i casi unici ed irripetibili della vita, mostra tutta la propria inconsistenza.

[46] Antonio Incampo rileva che la difficoltà di giungere ad una conclusione certa induce i giudici da un lato ad affermare che non è sufficiente una prova fondata su un criterio di sola probabilità, dall'altro a trasformare i reati di danno in reati di pericolo. Lo fanno tutte le volte che accertano il nesso di causa non come ciò che *ha* effettivamente determinato un certo evento, bensì come ciò che *potrebbe* astrattamente avere proprio quegli esiti. Con l'effetto paradossale ed in sé contraddittorio di ragionare sempre in astratto, ma con la logica della probabilità. Cfr. A. Incampo, "*Propter iustos*. Probabilità e ragioni del dubbio nella giustizia penale", in A. Incampo, A. Scalfati (a cura di), *Giudizio penale e ragionevole dubbio*, Cacucci, Bari, 2017, pp. 69-72. Per il diritto civile, ricordo che per la Corte di Cassazione, Sez. lav., 3 febbraio 2021, n. 2474 (ma v. pure Sez. IV^ penale, 24 febbraio 2021, n. 16843), il criterio di ragionevole probabilità scientifica ispirato al principio del "più probabile che non", deve ancorarsi non esclusivamente alla determinazione quantitativo-statistica delle frequenze di classe di eventi (cd. probabilità quantitativa), ma riconducendone il grado di fondatezza all'ambito degli elementi di conferma disponibili nel caso concreto.

[47] Sono condivisibili le conclusioni di S. Bauzon, "La place de l'Intelligence Artificielle (IA) dans le monde du droit", in *L'Ircocervo*, 23 (2024), n. 1, pp. 112-121.

(iii) *Differenze logiche*. Non solo la verità nel processo non è dicibile, ma è lo stesso concetto di "informazione" a non spiegare adeguatamente quello di "verità" per il giudizio[48].

Si è detto che il numero è una rappresentazione simbolica della realtà, ma non è ancora realtà. È in questo senso, un feticcio. Il numero come il dato descrive il mondo attraverso la combinazione di 0 ed 1 e cioè dei corrispondenti giudizi di verità o falsità sugli stati di cose. Seguono la stessa idea anche le rappresentazioni della statistica che procede per verità *doxastiche* in base a ciò che accade per la massima parte o per lo più (si tratta degli *endoxa*[49]).

E tuttavia le relazioni di quantità tra i numeri rimangono indipendenti dagli usi che se ne possono fare. Le stesse proposizioni possono, infatti, avere usi differenti. In questo senso, il *discorso cognitivo* si stacca da quello *normativo* almeno sotto il profilo *pragmatico*. La funzione pragmatica identifica esattamente l'uso che si fa di un certo enunciato, e si distingue dal relativo valore semantico che coincide con il significato delle parole. Dal punto di vista pragmatico, dunque, lo stesso enunciato può avere usi differenti. Ad esempio, l'enunciato "Tutti i ladri sono puniti" può valere come: (i) un'asserzione sull'esistenza di una norma che prevede una punizione per i ladri; (ii) un'ipotesi statistica sulle conseguenze del furto per chi ne fosse artefice; (iii) una norma che prescrive la punizione per tutti i ladri[50]. I tanti usi di una proposizione non possono ignorare la circostanza che un enunciato con una stessa grammatica possa avere usi e significati molto diversi tra loro. In definitiva, anche a voler immaginare un'intelligenza artificiale che operi sul piano della semantica del discorso (il piano dei diversi significati di un enunciato), resterebbe il problema della capacità di *astrarre* il normativo in un dato spazio pragmatico. Con quali implicazioni?

[48] È solo in parte l'interrogativo da cui muove il saggio di S. D'Agosto, "La costruzione della figura femminile fra neuroni specchio e intelligenza artificiale simbiotica", in *L'Ircocervo*, 23 (2024), n. 1, pp. 307-317, ora anche in questo volume.

[49] Aristotele, *Topici*, trad. it., Laterza, Roma/Bari, 1985, p. 17.

[50] Traggo spunto da A. Incampo, *Metafisica del processo*, cit., pp. 40-42.

Le logiche proposizionali ed i ragionamenti stocastici a cui si è fatto cenno presuppongono un discorso essenzialmente apofantico, un discorso, cioè, la cui verità è dettata dalla corrispondenza tra ciò che si dice e lo stato di cose descritto. La logica del discorso giudiziale è, invece, la logica della validità piuttosto che della verità[51]. Le sentenze non sono vere o false, ma valide o invalide. In tal modo, lo stesso concetto di "verità processuale" è ambiguo. Il processo culmina, infatti, con un'azione, piuttosto che con una verità. È, appunto, la decisione del giudice, e la decisione solo del giudice condizionata da un atto di volontà [*Willensakt*], per riprendere un tema fondamentale della *Reine Rechtslehre* (1934) di Kelsen[52]. È questo un fatto che non può essere trascurato se si vuole ricostruire una logica rigorosa del discorso giudiziale.

4. Dalla sostituzione… alla simbiosi tra procedure e istituzioni

Come conciliare le differenze dal punto di vista del diritto? Abbiamo visto che a rendersi inconsistente è la sostituzione, quantomeno per il discorso giudiziale. Pesano soprattutto le questioni epistemiche, retoriche e logiche. Non basta dunque la conoscenza, effettivamente disponibile ai sistemi informazionali[53]. Alla decisione giudiziale (la norma individuale contenuta in una sentenza) serve più che altro la volontà del soggetto

[51] Sulla distinzione tra verità semantica (verità *de dicto*) e ontica (verità *de re*) rinvio ad A. Incampo, *ivi*, pp. 39-44; Id., *Filosofia del dovere giuridico*, ed. 3, Cacucci, Bari, 2019, pp. 77-89. La distinzione è particolarmente rilevante per la caratterizzazione della logica giudiziale come logica della validità.

[52] H. Kelsen, *Reine Rechtslehre. Einleitung in die rechtswissenschaftliche Problematik*, Franz Deuticke, Wien, 1934, trad. it. *La dottrina pura del diritto*, Einaudi, Torino, 1966, p. 29. Peraltro, Kelsen affermerà che "il giudizio del giudice non è un giudizio nel senso logico del termine, così come non lo è la legge che mediante tale giudizio viene applicata; esso è una norma". Cfr. Al riguardo H. Kelsen, *Allgemeine Theorie der Normen*, herausgegeben von Kurt Ringhofer und Robert Walter, Manzsche Verlags- und Universitätsbuchhandlung, Wien, 1979, p. 185, trad. it. *Teoria generale delle norme*, Einaudi, Torino 1985, p. 387.

[53] B. Romano, *Algoritmi al potere. Calcolo giudizio pensiero*, Giappichelli, Torino, 2018.

agente-decidente. C'è, tuttavia, da chiedersi se l'idea kelseniana prima accennata rappresenti tutto il senso dell'esperienza giuridica[54]. Il rischio è che richiamare la sola volontà non basti, e che la stessa volontà si tramuti in arbitrio, facendo dipendere la giustizia da "emotività devianti", da filtri digitali o da fattori esogeni, e persino da *input* pescati a caso tra gli infiniti dati (o, peggio, tra *bad data*), un po' come con il lancio dei dadi del giudice Bridoye, uno dei personaggi narrati da François Rabelais[55]. È tra l'altro il pericolo che la giustizia predittiva vorrebbe scongiurare coerentemente con l'ideale di certezza del diritto[56].

Invero, anche la giustizia predittiva, nel suo significato estremo, è ideologia. Sfrutta la neutralità del numero, ma lo fa mediante rappresentazioni simboliche nient'affatto neutrali e finisce così per condizionare i giudizi umani attraverso filtri informazionali sacrificando il libero convincimento del giudice e la libertà metodologica del giurista critico. È, infatti, ovvio che sia più facile calcolare la decisione del giudice di "partito" che si conforma alle elaborazioni di dati, rispetto a quella del legalista-logicista tradizionale[57]. Tuttavia, in questo modo, si sottovaluta che molte situazioni (e potenzialmente tutte) richiedano la ridiscussione critica di procedimenti consolidati. Per il diritto, si può richiamare il ruolo della *ratio decidendi* di una sentenza alla base di tesi opposte (c.d. *distinguishing*) per prospettare un cam-

[54] Sono in questo senso rilevanti le riflessioni di L. Franzese, "Sulla insufficienza dell'Intelligenza Artificiale e della legge", in *L'Ircocervo*, 23 (2024), n. 1, pp. 27-35, ora in questo volume, e F. Recanati, "L'Intelligenza Artificiale e il senso del diritto", in *L'Ircocervo*, 23 (2024), n. 1, pp. 36-50.

[55] F. Rabelais, *Gargantua e Pantagruele*, II, ed. 6, trad. it., Rizzoli, Milano, 2004 e in particolare *Il terzo libro dei fatti e detti eroici del nobile Pantagruele*. È noto che Bridoye decideva i casi lanciando i dadi. Tra i tanti, V. Marinelli, "I dadi del giudice Bridoye", in *Materiali per una storia della cultura giuridica*", 32 (2002), pp. 529-532. Sul ruolo delle emozioni, F. Corona, "L'evoluzione del processo decisionale tra giudizio automatico e sostituzione giudiziaria", in *L'Ircocervo*, 23 (2024), n. 1, pp. 175-193.

[56] Si veda l'ampia trattazione in A. Carleo (a cura di), *Decisione robotica*, il Mulino, Bologna, 2019.

[57] L. Lombardi Vallauri, *Corso di filosofia del diritto*, cit., p. 16. Sul tema anche S. Crafa, "Dalle competenze alla consapevolezza digitale: capire la complessità e la non neutralità del software", in P. Moro (a cura di), *Etica, diritto e tecnologia*, cit., pp. 63-78.

bio di direzione (*overruling*) quando si ritiene che i casi definiti sulla base del vecchio precedente siano stati decisi *ab iniuria*, contro il diritto[58]; oppure, su un versante estremo, all'invalidità *ex tunc* della stessa legge quando è odiosamente ingiusta e contraria al "diritto sovralegale" [*übergesetzliches Recht*], nella nota formula coniata da Gustav Radbruch in *Gesetzliches Unrecht und übergesetzliches Recht* (1946)[59].

Si tratta comunque di attività vincolate alle strutture che le rendono possibili. È questo il connotato procedurale dell'esperienza giuridica. Le procedure danno forma alle azioni e alle relazioni che nascono, e sono al contempo condizioni significativamente *deontiche* di ogni attività umana[60]. D'altronde, hanno natura procedurale anche i meccanismi di traduzione del numero in linguaggio naturale[61]. In che senso?

A metà degli anni '30, Alan Turing stava lavorando al problema della decisione in logica [*Entscheidungsproblem*], posto nel 1928 dal pensiero finitario di David Hilbert i cui esiti vennero codificati dai noti teoremi di incompletezza di Kurt Gödel[62]. La

[58] Richiamo l'interessante contributo di A. Di Porto, "Intelligenza artificiale e professione forense. Il robot al servizio dell'avvocato", in P. Perlingieri, S. Giova, I. Prisco (a cura di), *Rapporti civilistici e intelligenze artificiali: attività e responsabilità. Atti del 15° Convegno Nazionale*, ESI, Napoli, 2020, p. 221.

[59] G. Radbruch, "Gesetzliches Unrecht und übergesetzliches Recht", in *Süddeutsche Juristen-Zeitung*", 5 (1946), pp. 105-108, nella edizione a cura di E. Fittipaldi, il titolo è: "Ingiustizia legale e diritto sovralegale", in A.G. Conte, P. Di Lucia, L. Ferrajoli, M. Jori (a cura di), *Filosofia del diritto*, Raffaello Cortina, Milano, 2002, pp. 149-163. Per un'analisi approfondita, si rinvia a A.M. Campanale, *Il diritto oltre la legge in G. Radbruch e R. Dworkin*, Adriatica, Bari, 1997 e G. Vassalli, *Formula di Radbruch e diritto penale. Note sulla punizione dei "delitti di Stato" nella Germania postnazista e nella Germania postcomunista*, Giuffrè, Milano, 2001. Di recente, L. Avitabile, *Legalità e giustizia. I Feuerbach e Radbruch. Un'interpretazione con Giuliano Vassalli*, Giappichelli, Torino, 2021.

[60] È una idea che può rinvenirsi nel pensiero di Lon L. Fuller, *The Morality of Law*, ed. 2, Yale University Press, New Haven, 1969 e che ho approfondito in P. Marra, *Per una moralità procedurale del diritto*, cit.

[61] In questa stessa direzione, l'analisi epistemologica di F. Romeo, *op. cit.*, pp. 51-69

[62] I testi fondamentali di riferimento sono, nell'ordine di trattazione, A.M. Turing, "On Computable Numbers with an Application to the Entscheidungsproblem", in *Proceedings of the London Mathematica Society*, 42

questione era se all'interno di un sistema assiomatico del primo ordine si potesse decidere attraverso una procedura meccanica, per qualsiasi formula di quell'ordine, e se tale formula fosse un teorema del primo ordine oppure no. Già i teoremi di Gödel avevano dimostrato l'impossibilità dei sistemi assiomatici di dedurre tutte le formule o proposizioni vere che vi appartengono. Un sistema di formalizzazione dell'aritmetica elementare è in grado di comprendere insieme, in base a regole di trasformazione del sistema, sia proposizioni aritmetiche, sia "proposizioni formalmente indecidibili"[63]. C'è in estrema sintesi un livello di comprensione che sfugge a qualsiasi formalizzazione e il cui risultato è *indecidibile* dal punto di vista del calcolo. La potenza di calcolo di una macchina universale, dunque, non implica anche la capacità di fondare tutte le proprietà contenute nei propri assiomi. Su queste basi, Ludwig Wittgenstein sostiene che la fondazione della normatività non può che avvenire in una certa "forma di vita" [*Lebensform*] e cioè in base a giudizi di convalida che poggiano sul modo tipicamente umano di stare al mondo[64]. Ecco perché, quando l'algoritmo entra nella forma

(1937), n. 1, pp. 230-265; D. Hilbert, "Über das Unendliche", in *Mathematische Annalen*, 95 (1926), n. 1, pp. 161-190 e Id. "Die logischen Grundlagen der Mathematik", in *Mathematische Annalen*, 88 (1923), nn. 1-2, pp. 151-157; K. Gödel, "Über formal unentscheidbare Sätze del Principia Mathematica und verwandter Systeme I", in *Monatshefte für Mathematik*, 38 (1931), n. 1, pp. 173-198.

[63] Sulla rilevanza dei teoremi per una fondazione dei sistemi giuridici, si veda A. Incampo, *Metafisica del processo*, cit., pp. 221-223 e prima Id., *Sul fondamento della validità deontica. Identità non-contraddizione*, Edizioni di Giuseppe Laterza, Bari, 1996, pp. 330-351. Per alcune applicazioni dei teoremi di Gödel alla costruzione dei sistemi giuridici, cfr. S. Agrusti, "Incompletezza gödeliana dei sistemi giuridici", in *Annali del Dipartimento Jonico in Sistemi Giuridici ed Economici del Mediterraneo: Società, Ambiente, Culture*, I (2013), pp. 1-17.

[64] L. Wittgenstein, *Ricerche filosofiche*, trad. it., Einaudi, Torino, 1995, part. 25. Sul rapporto tra Turing e Wittgenstein v. J. Floyd, "Chains of Life: Turing, *Lebensform*, and the Emergence of Wittgenstein's Later Style", in *Nordic Wittgenstein Review*, 5 (2016), n. 2, pp. 7-89 ed ancora Id., "*Lebensformen*: Living Logic", in C. Martin (ed.), *Language, Form(s) of Life and Logic. Investigations after Wittgenstein*, De Gruyter, Berlin, 2018, pp. 59-92. Spunti interessanti si rinvengono anche in G. De Anna, "Automi,

di vita, perde la sua neutralità perché non resta indipendente, ad esempio, rispetto alle traduzioni della programmazione[65].

È analogo il meccanismo di validità di una decisione giudiziale. Essa conta per la nostra forma di vita se motivata. Nessun destinatario, infatti, prenderebbe in considerazione una qualsiasi decisione priva di giustificazione. Tra l'altro, per l'ordinamento italiano l'obbligo è costituzionalmente imposto dall'art. 111, comma 6, della Costituzione. Ma è prima ancora un obbligo *praeter legem*, al di là della legge. La motivazione è una condizione che risale all'idea stessa di sentenza e, in generale, di ogni processo decisorio. Una sentenza immotivata, frutto soltanto del capriccio per dirla con Oliver W. Holmes, è un atto privo di senso. Se stiamo all'uso della decisione giudiziale, e cioè alla sua pragmatica, la motivazione ha un contenuto *praxeologico* che, cioè, deriva dal concetto primo di sentenza. Senza la motivazione, non vi sarebbe una vera decisione giudiziale.

Sul versante inedito della decisione digitale, le cose non possono essere diverse. La decisione del o con il sistema artificiale non può restare altrettanto priva di motivazione se vuole essere effettiva in una *forma di vita*. Questo perché la sentenza del giudice, pur adoperando meccanismi informazionali, non può fare a meno di formalizzarsi. Ha dunque un modo particolare di "fare cose con parole" perché non tutte le parole sono capaci di costituire *status* deontici, ma solo quelle che rispettano "condizioni di felicità" [*felicity conditions*] procedurali in un ambiente necessariamente umano in cui sono assicurati stabili ed effettivi canali di comunicazione[66]. Non sono, tuttavia, poche le incognite legate principalmente alle incapacità dell'uomo di spiegare i risultanti delle infinite connessioni tra dati che richiedono altrettante capacità computazionali. Non basta il principio di trasparenza che rischia di rappresentare un falso mito se non si collega

responsabilità e diritto", in *Rivista di filosofia del diritto*, VIII (2019), n. 1, pp. 125-142.

[65] O.G. Loddo, *op. cit.*, p. 363.

[66] Di questa idea è in parte anche L. Lombardi Vallauri, "Informatica e criteri 'politici' o valutativi della decisione giuridica", cit., p. 308 il quale già nel 1982 si chiedeva se fosse possibile "fare un automa conversazionale che ponga al giudice la domanda giusta […] traendone poi velocemente le conseguenze giuste".

alla dimensione trascendentalmente pubblica delle regole[67]. La trasparenza senza immediata intelligibilità e comprensibilità dei processi decisionali digitali rischia di essere in partenza cieca. Di qui le esigenze di costruire la spiegazione non solo su basi algoritmiche (è il paradigma della *Explainable Artificial Intelligence* – XAI), ma procedurali se si vuole condurre la decisione giudiziale nell'*habitat* umano. Se è praxeologica la natura della motivazione, procedurale è dunque la costruzione di una motivazione sufficientemente comprensibile per l'uomo. Occorre però stabilire i mezzi. Ricavo solo qualche conclusiva considerazione dal dibattito sul tema.

Un controllo umano supervisionato [*supervisory human control* – SHC], e cioè l'esito di un approccio predominante usato a partire dai sistemi autonomi militari, dimostra tutti i suoi limiti proprio perché riduce l'interazione uomo-macchina alle sole interfacce, sacrificando la "consapevolezza situazionale" da parte dell'uomo[68]. In un contesto associativo, invece, la cooperazione uomo-macchina [*human-machine teaming* – HMT] richiede la partecipazione attiva degli agenti (artificiali e umani) in sistemi multi-verso. Con questo approccio, il controllo umano non è basi-

[67] L'idea che la forma possa indicare un senso fondamentale del diritto [*Grundwert*] non è dissimile dalla "formula trascendentale del diritto pubblico" [*transzendentale Formel des öffentlichen*] di I. Kant, "Zum ewigen Frieden. Ein philosophischer Entwurf", in *Kleinere Schriften zur Geschichtsphilosophie, Ethik und Politik*, 1913: trad. it. di Vincenzo Cicero, *Pace perpetua*, Rusconi, Milano, 1997. Al tema della Öffentlichkeit e delle sue implicazioni per l'ideazione di un *Grundwert* si veda A. Incampo, *Filosofia del dovere giuridico*, cit., pp. 161-183 e Id., *Metafisica del processo*, cit., pp. 303-335. Si occupa della centralità del principio di trasparenza, comunque alla base del diritto alla spiegazione, il saggio di M. Martoni, "Spiegare le pieghe. Alcune osservazioni su trasparenza e spiegabilità", in *L'Ircocervo*, 23 (2024), n. 1, pp. 212-227, nonché P.B. Helzel, M. Leonetti, "*Machine learning* e depotenziamento dell'umano: quale futuro?", in *L'Ircocervo*, 23 (2024), n. 1, pp. 341-358.

[68] È il motivo da cui si muove lo studio di A. Tsamados, L. Floridi, M. Taddeo, "Human control of AI systems: from supervision to teaming", in *AI and Ethics*, (2024), pp. 1-14, part. pp. 7-8 (https://doi.org/10.1007/s43681-024-00489-4). Nel saggio che presento in questo numero, tuttavia, accenno ad una critica.

lare quanto il principio associativo[69]. È una idea che abbiamo già visto all'inizio. Tuttavia, per essere effettiva la cooperazione non può che fondarsi sulle capacità comunicative non riducibili alla semplice trasmissione di contenuti, del tutto inutile se questi non vengono processati ed effettivamente compresi dai destinatari[70]. È un aspetto che si tende a sottovalutare. Perciò, affidare l'interazione uomo-macchina unicamente alle interfacce (e cioè agli algoritmi e alla combinazione di dati), e tramite queste veicolare gli *input* per le decisioni digitali, non basta a rendere queste ultime neutrali, certe ed effettive; migliori, in poche parole, delle decisioni umane. Sono, infatti, le stesse interfacce a filtrare, condizionare, manipolare, predeterminare la volontà del soggetto/utente inconsapevole; d'altra parte, i dati non possono spiegare altri dati[71].

Nell'ibridazione dei processi decisionali, a metà tra l'informazionale del digitale e la formalizzazione del reale, sarà, dunque, ancora rilevante dotarsi di procedure? È forse la vera sfida tecnologica che attende l'umanità. Non c'è dubbio che ogni processo di interazione o decisione, anche in un contesto socio-digitale, avrà bisogno di condizioni minime di con-valida da parte dei destinatari. Questo vale anche per la decisione digitale nel processo che, in altri termini, dovrà mantenere un legame stretto con il destinatario attraverso una motivazione sufficiente che possa spiegare ogni momento del processo cooperativo (o simbiotico) uomo-computer/macchina che ne è alla base, problematizzando, ancora con Licklider, tutte le questioni presupposte, anche a carattere tecnico. Ed è solo in un contesto associativo che la decisione resta imputabile[72].

[69] T. O'Neill, N. McNeese, A. Barron, B. Schelble, "Human–autonomy teaming: a review and analysis of the empirical literature", in *Human Factors*, 64 (2020), n. 5, pp. 904-938 (https://doi.org/10.1177/0018720820960865). Qui si legge: "at least one human working cooperatively with at least one autonomous agent, where an autonomous agent is a computer entity with a partial or high degree of self-governance with respect to decision-making, adaptation, and communication".

[70] Cfr. O.G. Loddo, *op. cit.*, pp. 367-368, che richiama la teoria dei sistemi di Niklas Luhmann.

[71] G. D'Acquisto, *Decisioni algoritmiche. Equità, causalità, trasparenza*, Giappichelli, Torino, 2022, pp. 12 e 23.

[72] La necessità di un controllo umano significativo nel processo decisionale è embrionalmente introdotto in P. Marra, L. Pulito, A. Carnevale, A.

Oltre a delimitare il perimetro di ciò che è valido, e a definire un contesto responsabile secondo criteri di imputazione che hanno natura simbiotica, le procedure restano centrali anche per la concettualizzazione di nuovi diritti indispensabili nei contesti inediti che la realtà digitale offre. Nuove categorie di diritti che non potranno a loro volta fare a meno delle tecniche digitali (si pensi alla versatilità della *blockchain* e agli stessi sistemi GPAI), sfruttandone le potenzialità e cogliendone pure i limiti[73].

La giustizia digitale non potendo contare sulla pura tecnica per non essere fagocitata da ciò che non può comprendere, né spiegare, avrà dunque bisogno di "istituzioni sociali-digitali", e cioè di procedure e nuove ritualità che forniscano quantomeno una cornice simbolica (non geometricamente delimitata) in cui il mondo immateriale dell'infosfera e l'a-spazialità aritmetica del digitale possano ancora trovare una propria determinazione sul fronte della produzione di regole e della loro applicazione a partire sempre dall'uomo come fonte e scopo del diritto. Quando a mancare sono le procedure è, infatti, facile imbattersi nei

Lombardi, A. Dyoub, F.A. Lisi, "A Procedural Idea of Decision-making in the Context of Symbiotic AI", in A. Dix, M. Roach, T. Turchi, A. Malizia, B. Wilson (eds.), *op. cit.* (https://ceur-ws.org/Vol-3701/paper9.pdf). Spunti interessanti si rinvengono nel *leading case* Loomis (sentenza *Wisconsin Supreme Court*, *State v. Loomis* 881N.W. 2d 749 (Wis. 2016) in merito all'applicazione del software COMPAS - *Correctional Offender Management Profiling for Alternative Sanctions*. Per i saggi di questo numero speciale, rinvio a C. Grieco, "Tra simbiosi e condizionamento IA-centrico. Il divorzio fra legittimazione tecnica e *agency* umana", in *L'Ircocervo*, 23 (2024), n. 1, pp. 194-211, V.B. Muscatiello, "*Ex Machina*. Il diritto penale, all'improvviso", in *L'Ircocervo*, 23 (2024), n. 1, pp. 228-251, anche in questo volume, e L. Pulito, "Algoritmi predittivi e valutazioni prognostiche per la sostituzione della pena nella prospettiva della *symbiotic artificial intelligence*", in *L'Ircocervo*, 23 (2024), n. 1, pp. 252-272, anche in questo volume, con nuovo titolo.

[73] È la prospettiva che emerge in particolare dai saggi di P. Sommaggio, S. Ali, "The Role of Blockchain in Managing Neuroscientific Evidence and Its Impact on Cognitive Liberty", in *L'Ircocervo*, 23 (2024), n. 1, pp. 457-474, in questo volume Idd., "Cognitive Liberty, Neurolaw, and Digital Governance: Integrating Neuroscience and Blockchain in the Era of Neuro-Rights", nonché R. Strugała, "Liability for Damage Caused to the Data Subject in the Event of GDPR Violation 'by Artificial Intelligence': Between Legal Discrepancies and Policy Questions", in *L'Ircocervo*, 23 (2024), n. 1, pp. 475-488, anche in questo volume.

fallimenti socio-tecnici, fallimenti che si potrebbero vedere, sul piano etico, per la mancanza di adesione alle condizioni intrinseche ad ogni forma di interazione. Proprio per questo, tra l'altro, il diritto non può essere pensato solo su basi efficientistiche.

Licklider immaginava il futuro dell'intelligenza artificiale e vedeva la simbiosi come una fase intermedia dell'evoluzione tecnica dei sistemi artificiali[74]. Prima della *singolarità tecnologica*[75]. Non è, tuttavia, una fase irrilevante dell'evoluzione umana. Anzi. La simbiosi è l'opposto della sostituzione e della delega in bianco alle macchine nei campi sensibili dell'agire propositivo in cui è, invece, richiesto un *apporto umano significativo* irriducibile a semplici questioni di interfacce. Lo stesso è per la decisione giudiziale. Non vi è dubbio che quella simbiotica debba essere pensata a partire dalle condizioni minime della sua esistenza. La motivazione è indubbiamente una di queste. Vale per la decisione giudiziale, ma anche tutte le volte in cui l'uomo è chiamato responsabilmente ad agire.

Lo sviluppo, in ultima analisi, di una *Symbiotic AI* come terzo livello socio-digitale non può fare a meno di un'idea procedurale dell'esperienza. A patto di specificarne le condizioni per non rimanere spettatori inermi di mutamenti barocchi già in atto[76]. Prima ancora della singolarità e di inedite e multiverse realtà ancora da esplorare.

[74] J.C.R. Licklider, *op. cit.*, p. 2.

[75] Per una trattazione, M. Sozio, "Il deficit fronetico dello '*Jus ex machina*'", in *L'Ircocervo*, 23 (2024), n. 1, pp. 408-410, ora in questo volume, e P. Moro, "Intelligenza artificiale e tecnodiritto", cit., pp. 9-11.

[76] Uso un'immagine suggestiva di G. Siniscalchi, *Barocco giuridico. Osservatori, osservanti, spettatori*, FrancoAngeli, Milano, 2017. Da questo punto di vista, l'approccio *risk based* espresso dal recente regolamento europeo in tema di intelligenza artificiale è analizzato in chiave comparatistica da L. Fabiano, "*Algorithmic Recommendation* ed *Artificial Intelligence* nell'esperienza statunitense ed europea fra esigenze strategiche e tutela dei diritti fondamentali", in *L'Ircocervo*, 23 (2024), n. 1, pp. 440-456. Seppure rappresenti un primo passo, nella prospettiva procedurale, la regolamentazione europea non è sufficiente. Il diritto ha anche un carattere promozionale. Una regolamentazione ulteriore potrebbe predisporre regole di base per costituire su basi procedurali una istituzione socio-digitale di tipo simbiotico per un ulteriore sviluppo tecnologico.

Francesco Romeo

La produzione culturale dell'artificiale: prodromi di uno studio metaepistemologico su diritto ed informazione

Abstract: **The Cultural Production of the Artificial: A Metaepistemological Study on Law and Information.** Human culture is not only a matter of human beings because the production of culture is no longer an exclusive skill of *them*. It is a major change: culture is beginning to evolve in a different way. It is necessary to repropose the questions about truth and reality, as the two leading players perceive and conceptualize differently but coexist in the same cultural environment. The research focuses on the debate between realism, correspondence theory, and new realism regarding legal sciences.

Keywords: Legal Metaepistemology, Information, Meaning, Reality, Correspondence Theory of Truth.

Sommario: 1. Come definire la realtà ed il significato – 2. Realtà, classificazione ed informazione – 3. La proposta: informazione che genera significati.

1. Come definire la realtà ed il significato

"Quando non riesci a guardare in faccia il tuo avversario,
devi porre molte più domande a te stesso".[1]

La cultura umana, oggi, non è più solo umana perché la produzione della cultura, che regola la vita del *sapiens*, non è più una sua esclusiva abilità, esistono altre fonti di produzione che umane non sono, ma che producono cultura destinata a regolare la vita, in particolare quella sociale, dell'essere umano. Questi sistemi cognitivi artificiali producono cultura ma con una rappresentazione della realtà diversa dalla nostra, con concetti, riferimenti e necessità diversi dai nostri.

[1] Commento di uno spettatore anonimo nella prima partita tra il campione Lee Se-dol e AlphaGo, https://www.youtube.com/watch?v=WXuK-6gekU1Y (36:50 di 1:30:27), da me tradotto. AlphaGo vinse 4 a 1 e determinò il ritiro del campione dalle competizioni: "Anche se diventassi il numero uno, c'è un'entità che non può essere battuta" fu il suo commento alla decisione di ritirarsi, dichiarazione al Yonhap News Agency a Seoul, https://en.yna.co.kr/view/AEN20191127004800315.

È un cambiamento di prospettiva epocale, segna un confine: la cultura sta iniziando ad evolvere in modo diverso da quello fino ad ora percorso, con nuove entità non riducibili alle precedenti, l'albero evolutivo si biforca, è come se una mutazione avesse cambiato il soggetto agente. In questa trasformazione è necessario riproporre le usate domande sul vero e sulla realtà, visto che i protagonisti, i sistemi artificiali, percepiscono e concettualizzano in modo diverso, ma convivono nel medesimo ambiente culturale. Occorre situarsi a livello gnoseologico o metaepistemologico per comprendere quale sia la realtà per la quale dobbiamo trovare un corretto o adeguato modo di rappresentarla e di agirvi.

Chi percepisce fa parte della realtà ed oggi si cerca di riprodurre artificialmente questa capacità di percezione e riflessione sul mondo, per questo, è necessario accettare solo ipotesi che siano empiricamente verificabili anche sull'uomo e sulla sua mente. Il metodo speculativo non è più fonte di spiegazioni accettabili scientificamente, ma solo di ipotesi da verificare. La spiegazione dell'uomo, compresa la sua mente e la sua cultura si deve accompagnare con la sua riproduzione artificiale, con la verifica empirica della riproducibilità.

Il livello di analisi metaepistemologico si occupa della adeguatezza delle regole per la conoscenza, è, generalmente, il luogo in cui troviamo risposte metodologiche sulla scienza o sul sapere attorno alla realtà e alla verità. A questo livello, però, le domande da porci cambiano, giacché a noi interessano ora anche le regole per la corretta riproduzione dell'umano nei sistemi artificiali. Da questo cambiamento di orizzonti si evidenziano o scaturiscono tre conseguenze che esamineremo nell'articolo.

1. Le rappresentazioni della realtà sono state create dall'uomo con i suoi strumenti percettivi e adeguate ai suoi bisogni (non rappresentazione della realtà vera, ma umana). Il metodo classificatorio aristotelico permette questo, ma non permette di superare questo ancoraggio. Questa parzialità della rappresentazione appare solo ora in pieno, a confronto con rappresentazioni che umane non sono, anche se la dipendenza delle nostre rappresentazioni scientifiche dagli strumenti di osservazione era oramai nota.

2. Le domande gnoseologiche e metaepistemologiche non riguardano più solo la gnoseologia scientifica, ma l'intera cultura, i suoi – necessari all'origine – presupposti umani, che hanno permesso l'evoluzione delle società attuali. C'è una normatività invisibile, sottesa e sottointesa in tutta la cultura, che si costruisce sulle necessità umane, direi sulla sua natura, meglio sulla sua genetica e sulla sua biologia. La nostra cultura ha rispecchiato questa normatività, oltre alla realtà percepita, e la ha costruita in una realtà rappresentata a sua misura. I sistemi artificiali rendono evidente questa invisibile normatività, obbligando a porre la domanda se essa vada mantenuta, aprendo anche alla possibile risposta negativa.

3. I due punti precedenti riguardano anche il diritto. Gli ordinamenti giuridici sono evoluti all'interno delle società umane con una loro precisa teleologia (ed assiologia): risolvere i conflitti all'interno di esse. Questa *ratio* immanente al diritto lo informa in ogni sua parte. Il vero ed il giusto giuridici, così come l'efficienza degli ordinamenti giuridici e delle norme, si devono misurare con la lente di questa *ratio*. La risposta che deve dare la metaepistemologia giuridica è forse diversa da quella della fisica o della matematica: per il giurista, la realtà e il vero comprendono il come essi vengono costruiti e condivisi all'interno di una cultura, nella sua rappresentazione di senso comune. Per il giurista non è possibile pensare al suo oggetto, il diritto, senza comprendervi colui che lo pensa o agisce.

Una risposta a questi cambiamenti ed alle domande che da essi sorgono richiede non solo una riproposizione delle domande sul metodo di indagine, ma, più in profondità, sulla realtà stessa, su come si formino le nostre rappresentazioni e quanto esse rispecchino la realtà, ed in queste che cosa appartenga alla realtà esterna all'uomo e cosa appartenga all'uomo.

Tento, in queste pagine, di proporre una definizione ed una descrizione nuova.

Per facilitarmi il cammino del tentativo di chiarificazione mi riallaccio ad un esempio di chi già si è posto questa domanda, trovando risposte che forse vanno verificate. Mi riferisco al *Ma-*

nifesto del nuovo realismo di Maurizio Ferraris ed in particolare all'esperimento mentale della ciabatta[2].

Se un uomo dice ad una altro di portargli una ciabatta che si trova sul tappeto, normalmente il riconoscimento dell'oggetto e l'azione hanno luogo senza difficoltà. Nonostante le differenze biologiche, cerebrali, culturali, di esperienze vissute dei due soggetti, essi si capiscono. Qui – secondo Ferraris – occorre distinguere le differenze di opinione, ad esempio su concezioni mentali quali libertà, o spirito o trinità, dall'identificazione di una cosa separata ed esterna rispetto a noi. La ciabatta è sul tappeto indipendentemente dalla nostra sensazione, o percezione o opinione o struttura mentale. Che equivale a dire:

La realtà esiste indipendentemente dalla nostra sensazione [1]

oppure

La realtà è indipendente dalle nostre rappresentazioni della realtà [2]

Su questi due asserti concordo, ma la ciabatta in quanto rappresentazione mentale non è il significato della ciabatta e neppure il suo senso[3], se così fosse la [2] verrebbe negata. La ciabatta esiste come oggetto reale, ed essa è il riferimento o denotazione della parola ciabatta. Secondo l'impostazione di Frege, ma anche successivamente a lui, la ciabatta è il significato, *Bedeutung*. L'elaborazione successiva a Frege, in particolare da parte di Russell e Carnap creò i lemmi denotazione e riferimento per evitare equivocità. La ciabatta, però, ottiene la sua rappresentazione differentemente, a seconda che venga considerata dal punto di vista dell'uomo che la utilizza o da quello della scienza che la inserisce nelle sue classificazioni, o da quello del cane che la porta al padrone, o da quello dell'edera che la avviluppa. Qui va evidenziata

[2] M. Ferraris, *Manifesto del nuovo realismo*, Laterza, Roma-Bari, 2012, pp. 39 ss.

[3] Frege distingue tra senso, significato e rappresentazione: Sinn, Bedeutung, Vorstellung. Vedi G. Frege, "Über Sinn und Bedeutung", in *Zeitschrift für Philosophie und philosophische Kritik*, 100 (1892), pp. 25-50.

un'ambiguità semantica di notevole portata. La speculazione logica ha definito senso e significato in modo stringente e formale, lontano dal significato del linguaggio naturale. In Frege, ad esempio, *Sinn* significa senso, ovverossia il mondo dei pensieri oggettivi (*Gedanken*), cioè quella parte separabile del pensiero (*Denken*), che ha possibilità di essere oggettivata[4]. Mentre *Bedeutung* significa l'oggetto materiale al quale ci si riferisce, il riferimento. In Frege esiste però una terza categoria mentale, quella delle *Vorstellungen*[5], le rappresentazioni soggettive del mondo. Rientrano nelle *Vorstellungen* anche gli stati emozionali oltre alle descrizioni soggettive. Il significato soggettivo, quello che nella cultura va a costituire il mondo dei significati comuni per l'uomo, nella loro pienezza, ed a determinarne l'azione, non si ferma alle coppie estensione-intensione o riferimento-intensione o senso-significato. Il significato della realtà si forma nel soggetto senziente e percipiente nella triade fregeana *Vorstellung*, *Sinn* e *Bedeutung*[6].

[4] Il senso (*Sinn*) costruisce per Frege il terzo regno dell'oggettività, contenente il "pensiero puro", i *Gedanken*. Ma questo non costituisce, per Frege, l'intero pensiero umano. Frege non tratta, ma non per questo non riconosce, altre parti, come le rappresentazioni e tutto il pensiero emozionale, perché, per Frege, queste parti sono esprimibili solo in termini di linguaggio naturale, con la ineliminabile caratteristica della equivocità, dovuta alla ineliminabile mescolanza di *Sinn* con *Vorstellungen* ed *Anschauungen*. Per questo Frege separerà il mondo del *Sinn* nel *Begriffsschrift*, in un linguaggio interamente formalizzato e non esprimibile in termini di linguaggio naturale.

[5] Traduco *Vorstellung* con rappresentazione e non con idea, per diversi motivi. Come prima considerazione noto che il lemma tedesco ammette le due traduzioni ed i due campi semantici. Secondariamente, il lemma idea è stato usato, in lingua italiana, a torto o a ragione, non ne disquisisco qui, per tradurre l'importante lemma *Begriff,* che nella logica di Frege costituisce il contenuto di quel che lui chiama il terzo regno del vero e che verrà esaminato e, direi, portato alla luce nel *Begriffschrift*, il saggio del neokantiano Frege che verrà tradotto, con malaccorto ammiccamento platonico, come *Ideografia*. Il terzo importante motivo è dato dall'uso dello stesso lemma in altri autori, contemporanei o di poco anteriori, ed in contesti di significato vicini, come, ad esempio, in Franz Brentano. Il lemma esprime, nella sua etimologia, esattamente l'attività mentale della quale stiamo parlando: il tentativo di riproduzione a specchio della realtà nella mente.

[6] "Bisogna distinguere il significato e il senso di un segno dalla rappresentazione ad esso associata. Se il significato di un segno è un oggetto percepibile con i sensi, la mia rappresentazione di esso è un'immagine interiore che si forma dai ricordi delle impressioni sensoriali che ho avuto e delle

Ho approfondito in altro luogo quanto espresso da Frege, il suo fraintendimento nella filosofia analitica del Novecento e le semplificazioni del "significato di significato" fatte proprie nel Novecento da alcune speculazioni analitiche[7]. Il significato di ciabatta, per l'essere umano, si accompagna alla rappresentazione, diversa per ciascun diverso individuo, delle sensazioni che prova indossandola ed al fatto stesso che, invece di mangiarla, la indossa. Il lemma ciabatta è accompagnato da un campo semantico complesso che nel suo insieme costituisce la rappresentazione per Frege, esprimibile solo attraverso il linguaggio naturale, che costituisce, in parte, per noi quel che chiamiamo significato e per Frege *Vorstellungen*, ma che non si identifica con il riferimento o

attività, sia interiori che esteriori, che ho svolto. Questa immagine è spesso imbevuta di emozioni; la chiarezza delle sue singole parti varia e fluttua. Non sempre la stessa rappresentazione è collegata allo stesso senso, anche nella stessa persona. La rappresentazione è soggettiva: la rappresentazione di una persona non è quella di un'altra. Perciò ci sono di per sé molteplici differenze nelle rappresentazioni associate allo stesso senso. [...] La rappresentazione si differenzia quindi in modo sostanziale dal senso di un segno, che può essere proprietà comune di molti, e non è quindi una parte o una modalità del singolo individuo; non si può infatti negare che l'umanità abbia un tesoro comune di pensieri, che trasmette da una generazione all'altra. Se, dunque, non si esita a parlare di senso direttamente, bisogna aggiungere, nel caso della rappresentazione, a chi appartiene e in quale momento. Si potrebbe forse dire: così come alla stessa parola una persona associa una rappresentazione, un'altra ne associa un'altra, così una persona può associare un senso ed un'altra un altro senso. Allora la differenza consiste solo nel tipo di collegamento. Nulla impedisce che entrambi assumano lo stesso senso, ma non possono avere la stessa rappresentazione. *Si duo idem faciunt, non est idem*. Se due si rappresentano la stessa cosa, ognuno ha, ciononostante, la propria rappresentazione. Anche se a volte è possibile stabilire delle differenze nelle rappresentazioni, persino nelle percezioni, di persone diverse, un confronto esatto non è tuttavia possibile perché non possiamo avere queste rappresentazioni insieme nella stessa coscienza". G. Frege, "Über Sinn und Bedeutung", cit., p. 29 s.

[7] F. Romeo, "Senso e significato in Kelsen e nei suoi interpreti", in F. Riccobono, F. Romeo (a cura di), *Dieci obiezioni a Kelsen. Discutendo con Luigi Ferrajoli*, Editoriale Scientifica, Napoli, 2018, pp. 45 ss.; Id., "Vorstellungen, Anschauungen und Rechtsentscheidung: ein Denkpfad von Frege nach Yoshino", in E. Schweighofer, M. Araszkiewicz, F. Lachmayer, M. Pavčnik (eds.), *Formalising Jurisprudence, Festschrift für Hajime Yoshino*, Edition Weblaw, Bern, 2019, p. 259 ss.

la denotazione o la estensione o l'intensione logicamente intesi[8]. Ciononostante l'equivoco resta culturalmente consolidato ed alle concezioni neorealiste della realtà si congiungono, spesso, ulteriori nascoste premesse corrispondentiste che identificano, nel significato, la corrispondenza tra le nostre rappresentazioni ed il riferimento mettendoli in corrispondenza biunivoca. Si perde così tutta la soggettività delle *Vorstellungen* fregeane, e del pensiero che realmente avviene, che sono quelle che costruiscono linguisticamente il significato nella sua interezza. Ma si perdono anche le peculiarità dell'oggetto reale, che può esser descritto o rappresentato in molti diversi modi, ma anche sentito e percepito in molti diversi modi. I nostri sensi non hanno possibilità di sentire la realtà per quello che è, sono strumenti parziali di raccolta di informazioni sull'ambiente che ci circonda. La realtà indagata con strumenti di rilevamento più precisi è ben diversa. Per questo, le teorie realiste-corrispondentiste negano la realtà, sono autocontraddittorie. L'errore in esse si incardina sulla confusione tra realtà e rappresentazione. Questa svolta del Novecento segue Carnap in alcune correnti di analisi del linguaggio ma non Frege. Che Frege invece avesse ragione nella sua triade è dimostrato oggi proprio dalla storia della Intelligenza Artificiale.

Alcune abilità umane sono state facili da replicare artificialmente, così è stato per il calcolo o l'elaborazione logica di dati. In queste i sistemi artificiali hanno immediatamente mostrato di poter raggiungere abilità superiori a quelle dell'essere umano. Per altre abilità, quelle ritenute meno specificamente umane, condivise in diversa misura da tutti gli animali, come la capacità di riconoscere oggetti o di apprendere comportamenti dall'esperienza, si sono riscontrate difficoltà sia fattuali che teoriche impreviste. La cosiddetta *pattern recognition* è una attività tutt'altro che banale ed immediata. L'ipotesi della Intelligenza Artificiale classica consistente nella descrizione-definizione della realtà

[8] Occorre ricordare che Frege non è stato il fondatore dell'analisi del linguaggio. Anzi, nel *Begriffsschrift* abbandona il linguaggio naturale per formalizzare il "terzo regno del vero", cosa che gli sarebbe stata impossibile all'interno di esso. Frege si esprime ripetutamente in senso negativo nei riguardi della possibilità di formalizzare il linguaggio naturale, v. E. Mayer, *Der Wert der Gedanken*, Peter Lang, Frankfurt a.M., 1986-89, pp. 142 ss.; F. Romeo, "Vorstellungen, Anschauungen und Rechtsentscheidung", cit., p. 264.

nelle caratteristiche essenziali per permettere il riconoscimento di questa e la corretta classificazione, operando quindi con sistemi logici chiusi, ha fallito il suo intento. I sistemi programmati così semplicemente non sono in grado di riconoscere la realtà, di differenziare gli oggetti in autonomia. La *pattern recognition* ha, invece, compiuto un avanzamento rivoluzionario quando sono stati utilizzati programmi di elaborazione dati basati su presupposti diversi, nei quali il sistema autonomamente poteva formare le sue classificazioni, a dipendenza di moltissimi fattori. In questi programmi, le reti neurali, la classificazione si forma in un grafo, in una rete di connessioni tra unità di calcolo. I caratteri definitori non sono presenti o assenti, ma possono essere quantificati in un continuo, ed inoltre non si tratta di veri caratteri definitori di classi o specie, perché la loro importanza ai fini del corretto riconoscimento viene stabilita nell'apprendimento della rete e può anche essere nulla. Insomma, il riconoscimento di oggetti o realtà è una attività cerebrale e mentale estremamente complessa, che coinvolge molto più sapere o conoscenze di quelle strettamente attinenti alla realtà da riconoscere. È necessario ripensare alcune categorie in materia con particolare riguardo al diritto: esso è un mondo di significati espressi linguisticamente, per la maggior parte, circa comportamenti umani. Per questo la domanda gnoseologica o metaepistemologica si completa in ambito giuridico domandandosi quale sia una concezione di realtà adeguata ai sistemi giuridici: la risposta deve comprendere la normatività e quindi non basta la mera rappresentazione scientifica della realtà; è, forse, necessario comprendere la ponderazione umana della stessa secondo l'assiologia giuridica e le possibilità umane di operare tale ponderazione.

2. Realtà, classificazione ed informazione

> "I thought AlphaGo was based on probability calculation and that it was merely a machine, but, when I saw this move I changed my mind. Surely AlphaGo is creative. The move was really creative and beautiful"[9]

[9] Lee Se-dol (52:13 di 1:30:27): https://www.youtube.com/watch?v=WXuK6gekU1Y.

Oggi non si comprenderebbe il faticoso darsi da fare dello *homo sapiens* se non lo si inquadrasse al suo posto, nella casella che gli riserva la spiegazione evoluzionista della cultura umana. É – questo lavorìo continuo – derivazione immediata della spinta genetica tesa a migliorare la posizione di utilità genetica e, in subordine, individuale e sociale. Sia l'immediata apprensione delle ricchezze sia l'agire all'apparenza ultroneo rispetto all'utile, caratteristico del *sapiens*, diventano, tuttavia, significativi nella spiegazione evoluzionista[10]. L'intera cultura appare, in quest'ottica, un enorme strumento di massimizzazione della *fitness* individuale anche se tramite l'aumento del benessere sociale, direttamente, e di quello individuale solo derivatamente[11]. In questa ipotesi, sono

[10] Un ruolo centrale nell'evoluzione della cultura del *sapiens* lo ha avuto la cooperazione, vedi C. Boehm, *Hierarchy in the Forest. The Evolution of Egalitarian Behavior*, Harvard University Press, Cambridge-Londra, 2001; C. Boehm, "The natural selection of altruistic traits", in *Human Nature*, 10 (1999), n. 3, pp. 205-252; D.S. Wilson, E.O. Wilson, "Rethinking the theoretical foundation of sociobiology", in *The Quarterly Review of Biology*, 82 (2007), n. 4, pp. 327 ss., a p. 343: "Le nostre capacità di pensiero simbolico e di trasmissione sociale delle informazioni sono attività fondamentalmente comunitarie che probabilmente hanno richiesto uno spostamento dell'equilibrio tra i livelli di selezione prima di potersi evolvere. Solo quando potevamo fidarci dei nostri partner sociali per lavorare verso obiettivi condivisi, potevamo contare su di loro per condividere informazioni significative". Una fiducia sorvegliata, che permette la cooperazione negli animali prosociali è un bias che imposta anche l'azione sociale umana D. Cesarini, C.T. Dawes, J.H. Fowler, M. Johannesson, P. Lichtenstein, B. Wallace, "Heritability of cooperative behavior in the trust game", in *PNAS*, 105 (2008), n. 10, p. 3721 ss. (http://www.pnas.org/content/105/10/3721.full); B. Wallace, D. Cesarini, P. Lichtenstein, M. Johannesson, "Heritability of ultimatum game responder behavior", in *PNAS*, 104 (2007), n. 40, p. 15631 ss. (http://www.pnas.org/content/104/40/15631.full).

[11] L'evoluzione della cultura, basata sullo scambio di informazioni è il punto nodale della discussione evoluzionista dagli anni Settanta dello scorso secolo ad oggi. In questa evoluzione il sorgere di relazioni di tipo giuridico, rivolte alla soluzione dei conflitti che potevano facilmente rompere la cooperazione dell'intero gruppo, ha svolto un ruolo cardine; sul punto ho avanzato una mia ipotesi, vedi F. Romeo, *Antropologia giuridica, un percorso evoluzionista verso l'origine della relazione giuridica*, Giappichelli, Torino, 2012. Dal sorgere della cultura in poi, l'evoluzione genetica è andata di pari passo con quella culturale, vedi L.L. Cavalli Sforza, *L'evoluzione della cultura*, Codice, Torino, 2004; P.J. Richerson, R. Boyd, *Non di soli geni. Come la cultura ha trasformato l'evoluzione umana*, trad. it., Codice, Torino, 2006.

strumenti artificiali – ma nell'ottica evoluzionista sarebbe meglio dire naturali – anche le rappresentazioni concettuali della realtà, esse, come ogni altra parte della cultura umana, condividono la stessa natura dei primissimi strumenti materiali, frecce e oggetti contundenti, che anche erano utili a massimizzare la *fitness* dello *habilis* e delle sue prime società. Lo sono i linguaggi, ma anche la matematica e la logica, prodotti culturali al pari delle lance e delle frecce. La cultura, sin dall'origine, svolge una funzione adattiva e se gli strumenti materiali permettono di intervenire sulla realtà, così come percepita dai nostri sensi, quelli concettuali permettono di andare oltre le limitazioni sensoriali e percepire o rappresentare la realtà in modi molto più soddisfacenti e aderenti a quello che essa veramente è e che resta, nella sua completezza, largamente sconosciuto. Possiamo pensare agli strumenti concettuali come a degli "apriscatole" per la realtà, permettono di penetrare e "vedere", almeno parzialmente, ciò che è chiuso in una scatola e quindi inesistente per i nostri sensi[12]. La realtà trovata dagli strumenti concettuali è una realtà non vissuta, non percepita, inabitata dai *sapiens*; sono strumenti, quelli concettuali, che ci portano oltre i confini della nostra biologia e dei nostri presupposti genetici, ci precipitano in un universo a noi sconosciuto e non agibile con i nostri soli sensi, ma solo descrivibile concettualmente; essi ci permettono di superare i nostri limiti sensoriali e di avvicinarci ad una rappresentazione della realtà che favorisca la nostra azione in essa, certa delle conseguenze. Tutta la cultura dell'uomo si è costruita sulla natura che lo costituisce, essendo un mezzo amplificativo delle possibilità di azione del *sapiens* sulla realtà e, prima di lui, dello *habilis*.

La lingua tedesca, in concordanza con le due radici, germanica e latina, della lingua, ha due lemmi per significare realtà: *Wirklichkeit* e *Realität.* Qui sarebbe possibile una differenziazione

[12] A mio avviso, le congetture culturali su quello che la realtà sia non mettono mai in dubbio la sua esistenza, indipendentemente dal soggetto che la concettualizza o dalla congettura. Ciò che funge da prova di verità delle rappresentazioni e delle congetture sulla realtà è l'adeguatezza delle stesse, il permettere una ripetibilità dell'azione dell'individuo o una migliore definizione degli obiettivi che si possono raggiungere con le rappresentazioni elaborate o utilizzate nella realtà. Ma conoscere la vera realtà è impossibile, perché anche il soggetto conoscente appartiene ad essa e nel momento in cui conosce cambia la realtà.

tra realtà rappresentata e realtà oggettiva, ma la storia delle idee ha visto usare il lemma in modo contrastante nel significato, fino ad annullare le differenze e ad intenderli oggi come sinonimi. Le differenze di significato tra i due lemmi restano però, con notevoli diversità, nella filosofia degli autori che li hanno differenziati.

In italiano il lemma realtà è accompagnato dalla sostantivazione dell'aggettivo "reale". In questo studio uso il lemma realtà per ciò che esiste oggettivamente con una sua materialità, in sé, tutta la realtà è oggettiva e materiale. Uso, inoltre, il lemma reale ad individuare la realtà rappresentata in un sistema cognitivo; tutte le rappresentazioni della realtà, e i suoi tentativi, ivi comprese le forme immaginarie della realtà, appartengono al reale. Nel mio uso, quindi, cervo ed ircocervo sono ugualmente due reali, ma il primo può essere realtà in ogni sua forma, l'altro è realtà se individua questa rivista, altrimenti è una ricomposizione immaginaria di forme rappresentative, reale ma non esistente nella realtà[13]. Non sempre questa differenza viene tenuta in adeguato conto anche da chi si professa realista[14], ma la rappresentazione della realtà nel reale porta con sé una trasformazione e quindi un'inevitabile perdita di informazione, anche se aggiunge significato[15].

[13] Dell'ircocervo si è parlato forse troppo nella storia del pensiero; tuttavia, le discussioni non sono ancor sopite e le trattazioni in merito raggiungono livelli di estrema acutezza e affilatura concettuale. Sul *certamen* filosofico ancora in corso vedi C. Nitsch, "Dell'ircocervo, o di una storia fantastica della filosofia del diritto", in *Rivista di filosofia del diritto*, (2021), n. 2, pp. 428 ss. L'autore dà conto che, per quel che riguarda la parola greca *tragélaphos* significante ircocervo, essa designa oramai una realtà, non un mero reale, cioè un genere di antilopi che in effetti, in alcune specie, assomigliano a quanto descritto dai filosofi. Il *tragélaphos* aristotelico si è materializzato nel XIX secolo ad opera di Henri Marie Ducrotay de Blainville, che ha creato il *taxon*, latinizzato in *tragelaphus*, donando così, con raro *sense of humor*, materialità all'incongruenza e al mito, nonché, mi si permetta l'ironia, affermando pragmaticamente la forza creatrice della cultura sulla realtà: *fiat tragélaphus!*

[14] Cadono alle volte nel fraintendimento i neorealisti contemporanei che adottano un punto di vista corrispondentista tra realtà e reale. Ma l'errore è assai diffuso, quando, ad esempio, si parla di realtà univocamente individuata dal segno o dal riferimento, significando non che si individua una realtà non descritta o percepita per quello che è, ma che quel che si percepisce e descrive corrisponde esattamente alla porzione di realtà descritta. In queste concezioni realtà, reale e significato coincidono.

[15] La rappresentazione della realtà fa perdere informazioni relative all'oggetto che si rappresenta, ma aggiunge il significato, inserendo quelle

Il reale è simile alla realtà, mai uguale [3]

Questo divario tra reale e realtà si compone nel soggetto senziente, che è l'unico punto in cui i due mondi si incontrano, in lui la realtà acquista significato; i dati provenienti dalla realtà giungono al soggetto attraverso i sensi e vengono elaborati dal soggetto senziente in informazioni e poi ordinati secondo le sue possibilità mentali e necessità biologiche in significati. Tali significati sono inestricabilmente connessi al soggetto che elabora l'informazione[16], ne determinano la volontà e l'azione.

Il soggetto senziente elabora il significato [4]

La realtà è un complesso di dati naturali strutturati in determinate forme, seppur mutevoli, si può quindi pensare come informazione[17]. Il reale, invece, porta con sé non le strutture naturali rappresentate esattamente per come sono, bensì queste adattate al soggetto senziente ed alla cultura. La realtà rappresentata, il reale, può essere concepito come un regno dei significati, che si originano nei processi culturali individualmente elaborati, secondo le modalità di questi. Di ciò che compone la cultura, tutto può essere concepito come informazione strutturata in significati condivisi, elaborati attorno o sull'essere umano, secondo sue necessità e bisogni: essa dà forma a rappresentazioni dell'universo, sempre più aderenti alla realtà, cangianti con lo svolgersi dell'evoluzione culturale storicamente realizzata. Ho percorso in altro scritto la storia del lemma informazione[18], evidenzian-

informazioni, seppur limitate, nel sistema cognitivo che le percepisce e le valuta nell'ottica delle proprie necessità, mentali sì, ma, in genere, biologiche.

[16] A partire dalla Teoria della *Gestalt*, le scienze cognitive sono andate in questa direzione, senza tentennamento alcuno, in quanto maggiormente esplicativa. Sul punto vedi F. Romeo, *Lezioni di logica ed informatica giuridica*, Giappichelli, Torino, 2012, pp. 32 ss., nonché, per un quadro degli sviluppi recenti, molto ben approfondito v. B. Lotto, *Percezioni. Come il cervello costruisce il mondo*, Bollati Boringhieri, Torino, 2017, l'intero libro è imperniato su questo punto.

[17] È questa l'ipotesi cibernetica: se si sapesse tutto dell'uomo sarebbe faxabile.

[18] F. Romeo, "In-formatio, la formazione della realtà", in *Sociologia*, LV (2021), n. 2, pp. 112 ss.

do come, fino a tutto il Medio Evo, esso venisse utilizzato nel significato di in-formare, dare forma al mondo, non cercando di individuare in modo neutrale e non modificativo una realtà, ma ponendo in essere un'attività umana ordinatrice della realtà, formandola in rappresentazioni e classificazioni: suddividendo, quindi, e raggruppando secondo differenza e somiglianza, presenza o assenza di determinate caratteristiche, dette essenziali, creando classi di oggetti e creando i nomi a designazione delle classi. È, questa, un'attività descrittiva ma anche creatrice, che vede il soggetto creatore della cultura alla quale spetta il ruolo di rappresentare il reale[19]. Ricordo, a titolo di esempio, il grande lavoro classificatorio e descrittivo compiuto nell'antichità a riguardo del mondo naturale: piante, funghi, minerali ed animali vennero classificati, secondo la metodologia aristotelica del genere prossimo e della differenza specifica[20], ed utilizzati poi, negli usi della vita, secondo le proprietà della classe. Questa metodologia permette di rendere identici individui, all'apparenza assai diversi, ma tutti accomunati dalla condivisione delle caratteristiche essenziali, rendendo trascurabili quelle accidentali. Questa operazione di riduzione di ogni singolo individuo all'essenza definitoria trasforma la semplice somiglianza in identità: o di numero o di genere o di specie[21]. Il metodo

[19] Di questo uso del lemma resta traccia anche nella lingua contemporanea. Ad esempio, l'aggettivo e participio passato del verbo informare "informato" significa sia una persona a conoscenza di fatti o notizie in genere, ma anche qualche cosa conformata in un determinato modo o un'azione improntata a determinati principi.

[20] Aristotele, *Topici*, I(A), 8 103b, 11-19.

[21] Aristotele, *Topici*, I(A), 7 103a, 1-39. Il problema fu affrontato differentemente da Tommaso d'Aquino, per il quale può esistere, tra gli individui della stessa classe, non identità ma somiglianza e la classificazione del reale può avvenire tramite ragionamenti analogici, anche quantificabili, basati sulla somiglianza tra individui non sulla identità. "3. L'uno non è che l'ente indiviso e quindi la specie indifferenziata non è altro che la specie unica. Ora, a una cosa si può attribuire l'unità, non solo secondo il numero, la specie, o il genere, ma anche secondo una certa analogia e proporzione, e questa è l'unità o comunanza esistente tra la creatura e Dio. Le parole poi: 'fatta per uguagliarne un'altra', si riferiscono all'immagine perfetta". Tommaso d'Aquino, Somma Teologica [32360] Iª q. 93 a. 1 ad 3, reperibile in: http://www.carimo.it/somma-teologica/somma.htm. L'analogia è, per Tommaso, il *remedium aequivocationis*, che permette di passare dal molteplice all'uno, dal diver-

classificatorio aristotelico permette di descrivere accuratamente il mondo percepibile ai nostri sensi, e con gran risparmio di energie; esso permette di trarre dalla realtà l'utile necessario per la massimizzazione della *fitness* individuale, ci permette di riconoscere tutte le proprietà essenziali dell'individuo, almeno, quelle ritenute tali, per la sola attribuzione alla classe, anche in mancanza della diretta percezione di esse.

La realtà è fatta di massa ed energia, gli oggetti reali sono tutti diversi l'uno dall'altro; invece, quello che forma individui identici è l'operazione di individuazione e classificazione; i singoli individui, di per sé, sono ognuno diverso da ogni altro, l'in-formare, il dare forma al mondo, ordina e trasforma il mondo, che non sarà più composto da individui singoli o da materia, ma da classi. Ciò che viene detto di un appartenente alla classe è detto di tutti, senza necessità di testare ex novo ogni singolo: l'universo da infinito diviene finito e quindi calcolabile. Questa è la cultura.

Così il *sapiens* ha creato una realtà rappresentata in classi di oggetti, con proprietà definite, ma non un duplicato della realtà. È di tutta evidenza che l'informare il mondo secondo la metodologia aristotelica è compiuto a partire dai cinque sensi: non è l'attribuzione agli elementi della realtà di un valore o qualificazione assoluti, ma è il calcolo di un rapporto tra loro e l'informante. La realtà viene misurata e classificata in rapporto ai sensi dell'uomo. Conoscere la realtà secondo i sensi, o secondo la percezione, non equivale a conoscere la realtà per come essa è, per se stessa.

La realtà oggettiva non è conoscibile nella sua piena oggettività [5]

so al simile e dal simile all'uguale. È una impostazione epistemologica che sarà molto feconda, basti pensare che darà i suoi frutti più maturi con Franz Brentano, che segna il passaggio delle indagini sull'intelletto dalla filosofia alla psicologia. Brentano era cattolico e fu istruito su Tommaso d'Aquino già in giovane età. Questa peculiarità gli permise di innovare rispetto alla tradizione culturale dominante in Germania. Sarà, nella sua impostazione ed in molteplici innovazioni concettuali, il lontano progenitore delle scienze cognitive. Vedi M. Antonelli, F. Boccaccini, *Franz Brentano. Mente, coscienza, realtà*, Carocci editore, Roma, 2021.

Questa forma di conoscenza può determinare nulla più di una corrispondenza tra la realtà e l'essere senziente[22]. La cultura umana si è costruita, nei millenni, su questa corrispondenza, perché serviva all'agire dell'uomo, che, a sua volta, è organizzato su determinati sensi. Sarebbe diversa se altri fossero i sensi, se, ad esempio, l'essere umano possedesse un apparato sonar, come i pipistrelli, oppure una vista adatta a penetrare nel mondo degli atomi. Sarebbe anche diversa se, invece di sederci avessimo bisogno di appenderci a testa in giù, per riposare, sempre come i pipistrelli[23]. Ma sarebbe diversa la cultura e sarebbero diverse le nostre rappresentazioni della realtà, il reale, non la realtà.

Questo aiuta anche a comprendere perché i nostri sensi si sono evoluti per rappresentare la realtà nel nostro modo, con i nostri cinque sensi e non con altri. Questa rappresentazione della realtà ha permesso alla cultura di evolvere trasformando il *sapiens* e la sua costituzione biologica, agendo anche sulle predisposizioni genetiche e trasformando e costruendo a poco a poco la sua socialità. La cultura umana è inestricabilmente connessa con la storia evolutiva dell'uomo e della cultura stessa, la realtà rappresentata nella cultura è il risultato di questa evoluzione.

Il reale è un prodotto culturale dipendente dalla storia evolutiva dello *homo sapiens* [6]

Con il consolidarsi ed evolvere della cultura, le diverse rappresentazioni della realtà si assolutizzano andando a sostituire la realtà stessa. La cultura diviene, essa da sola, prova della realtà, si sostituisce ad essa: il perfezionamento delle rappresentazioni porta a supporre che tra esse e la realtà esista un rapporto di isomorfia, una corrispondenza isomorfa, e che per ogni concetto esista nella realtà un oggetto esattamente corrispondente a quello descritto nel concetto. Questa è forse l'utopia, ma anche l'errore, o l'approssimazione, più ingombrante nella cultura contempora-

[22] È, a questo punto, legittima la domanda se sia possibile una rappresentazione neutrale del reale, o se qualsiasi rappresentazione sia possibile solo come misurazione o parametrizzazione ad un punto di riferimento più o meno arbitrario.

[23] F. Romeo, *Il diritto artificiale*, Giappichelli, Torino, 2002, pp. 21 ss.

nea ed anche la più pericolosa. La metodologia aristotelica permette sì di creare un insieme di concetti-arnese utili all'agire del *sapiens*, ma contemporaneamente crea una forte separazione tra *sapiens* e realtà, tra i suoi geni e la natura dalla quale provengono. La cultura funge da vera e propria ipòstasi della realtà: allo stesso tempo, la rappresenta, nella sua replica-finzione, ma la nasconde nella sua esistenza-verità. La prova empirica, sempre presente nella metodologia scientifica, verifica costantemente la correttezza dell'isomorfia, ma tanta parte del pensiero filosofico o storico se ne distacca, costruendo immagini di mondi inesistenti, il più delle volte impossibili[24], o, peggio, inadatti agli esseri umani, reali ma non esatte rappresentazioni di essa, ircocervi non cervi. Quest'ultimo caso riguarda senz'altro le scienze normative, come tutte le scienze giuridiche, ma anche quelle economiche, alcune discipline ingegneristiche, l'informatica e l'intelligenza artificiale, insieme a molte filosofie. Sono esposte al pericolo soprattutto quelle impostazioni scientifiche e filosofiche che non accettano la prova empirica come ultima *ratio decidendi*, oppure la sostituiscono con una prova logica. Quanto

[24] Franz Brentano fu tra i più critici nei confronti dell'idealismo suo contemporaneo. È nota la sua posizione di critica radicale all'idealismo tedesco, accusato di non aver letto correttamente Aristotele, di averlo, anzi, usato maldestramente. Brentano era un anti-idealista, ma tutta la sua contemporaneità filosofica viveva la crisi dei grandi sistemi idealisti. Già negli anni Sessanta Brentano elabora la sua teoria delle quattro fasi della filosofia. "La teoria brentaniana individua nella storia della filosofia una regolarità ben precisa, che ne scandisce internamente lo sviluppo. In ciascuno dei tre grandi periodi in cui essa si articola – antico, medioevale e moderno – Brentano rileva la presenza di fasi alterne, di ascesa e di decadenza, che ritornano ciclicamente. A una prima fase ascendente, caratterizzata da un genuino interesse teoretico e dall'adozione di un metodo 'conforme a natura' (*naturgemäß*, è il caso di Aristotele nell'età antica, Tommaso d'Aquino in quella medievale, Cartesio, Locke e Leibniz in quella moderna), fanno seguito tre fasi di decadenza, espressione di un progressivo indebolimento di quell'interesse e di quell'ideale metodologico. All'interesse teoretico si sostituisce progressivamente quello pratico (stoici, epicurei – tarda scolastica – illuminismo), che sfocia in una tendenza scettica (pirronismo, tarda accademia – nominalismo – Hume), cui subentra, per reazione, una forma di misticismo privo di ogni fondamento gnoseologico (neoplatonismo – mistica medievale, Raimondo Lullo, Nicola Cusano – Kant, Fichte, Schelling, Hegel)". M. Antonelli, F. Boccaccini, *op. cit.*, pp. 49-50.

detto consegue dal fatto che il controllo della corrispondenza tra la realtà e le nostre rappresentazioni non può dare come risultato la verità, o non può essere indice di vera riproduzione della realtà, della realtà come essa è, ma solo come da noi percepita. La nostra rappresentazione si chiude nella ipòstasi della realtà, non nella realtà stessa, l'ipòstasi nasconde la vera realtà. La realtà dietro alla cultura potrà presentarsi, in futuro, in forme lontane da quelle rappresentate, impreviste ed impossibili da prevedere nell'ipòstasi culturale, e questo non può che essere accettato, mentre non può essere accettato che di questa modalità di costruzione della cultura e di queste limitazioni inerenti ai nostri percetti e concetti non se ne tenga conto. L'irrompere devastante ed irrazionale della realtà nelle nostre vite è sempre lì, minaccioso, limitabile ma non eliminabile completamente, ma quello che possiamo controllare è l'adeguatezza della nostra azione all'ipòstasi culturale: essa tiene presente non solo la realtà, ma anche la nostra particolare forma di accesso ad essa.

L'ipòstasi culturale, il reale, permette controlli ulteriori oltre alla semplice corrispondenza del concetto con l'oggetto: il percetto contiene in sé la misura dell'adeguatezza all'umano. La cultura si è evoluta come misura della realtà, ma a misura dell'uomo, è l'essere umano l'unità di misura. Tanto è vero quanto detto, che è assai facile trovarne la dimostrazione proprio nella storia della cultura e nei suoi cambiamenti nella descrizione della realtà. In botanica, dopo la lezione aristotelica i suoi allievi ne approfondirono le possibilità e si iniziò a classificare l'intero mondo delle piante. Teofrasto creò la tassonomia botanica, che portò a distinguere, e perciò conoscere, le piante secondo caratteristiche essenziali comuni. Fu un passo fondamentale nella conoscenza, che aprì le porte all'uso di queste piante. Dall'uso a fini alimentari alla elaborazione di farmaci, dallo sfruttamento delle proprietà tintorie a quelle allucinatorie o, anche, venefiche[25]. La tassonomia nell'antichità si basava su caratteristiche morfologiche macroscopiche, quali, ad esempio, il colore del

[25] Non di soli pugnali e aspidi perivano gli imperatori romani, ma anche di funghi velenosi, come ci narrano molti storiografi romani. Il fungo prediletto, all'epoca, è oggi chiamato Amanita Cesarea, amanita dei cesari, in loro onore. Ma il simile Amanita Muscaria, venefico mortale in grandi quantità o Amanita Phalloides, venefico in piccolissime quantità sono con esso confon-

gambo o del cappello, la presenza di lamelle o lamellule o tubuli o idni o qualsiasi altra caratteristica percepibile scelta ai fini classificatori. Sempre su queste basi Linneo trasformò il metodo tassonomico in sistematica, dove una caratteristica veniva presa principalmente in considerazione: la forma degli organi sessuali. Oggi la tassonomia botanica è ulteriormente cambiata e si basa su caratteristiche genetiche o biologiche, relegando quelle morfologiche a criterio valido *prima facie* ma inaffidabile. Quindi, da Aristotele ad oggi la descrizione del mondo dei funghi o delle piante è cambiata, anche se la realtà di riferimento è rimasta la stessa. Non sono semplicemente cambiati i nomi di famiglie, generi e specie, ma anche il metodo di classificazione, i caratteri definitori, la comprensività della classe, la sua estensione. Infatti, individui che nelle precedenti classificazioni si trovavano nella stessa classe, possono anche trovarsi, oggi, in classi diverse e questa è una differenza non meramente di *nomen* ma di proprietà o caratteristiche[26]. Questo non perché sia cambiata la realtà, cambiando la classificazione, bensì perché le precedenti caratteristiche definitorie non permettevano di differenziare le due specie, è cambiato il reale. Concordo pienamente con la critica di Ferraris[27] al postmodernismo: la realtà non cambia a dipendenza delle nostre rappresentazioni di essa, ma le nostre rappresentazioni sono migliorate e ci permettono di differenziare e quindi di mangiare alcuni tipi di funghi, senza avvelenarci, che con la classificazione precedente non potevamo mangiare perché non erano distinti dai simili velenosi. La realtà rappresentata, che ci permette di agire adeguatamente nella realtà oggettiva, si è popolata di nuove specie, ora le conosciamo e riconosciamo andando nel bosco, hanno un nome e delle proprietà diverse. È

dibili allo stato di ovulo; confondibili, ovviamente, per chi non conosceva il metodo classificatorio. L'imperatore Claudio morì così.

[26] Tra le quali la velenosità non è la meno importante.

[27] Mi riferisco a: M. Ferraris, *Manifesto del nuovo realismo*, cit., *passim*, il libro è impostato quasi interamente a critica del postmodernismo ma anche dei seguaci di Heidegger. Le risposte, a volte smodatamente aggressive, certo oltre le righe scientifiche, non sono mancate e hanno spesso varcato la soglia emozionale, questo è un indice dell'importanza dell'argomento e della sua attualità vedi D. Di Cesare, C. Ocone, S. Regazzoni, *Il nuovo realismo è un populismo*, Il Melangolo, Genova, 2013.

quindi la rappresentazione della realtà, il reale, che è cambiato e che ci permette un'azione più adatta all'ambiente, non la realtà.

L'esempio serve a veicolare il significato prima esposto, che ferma la rappresentazione della differenza tra reale e realtà:

Realtà rappresentata e realtà oggettiva sono due diverse entità [7]

La domanda sulla esistenza di una realtà oggettiva oltre o fuori di noi è forse la domanda prima della filosofia, certamente del razionalismo e dell'empirismo. Il rapporto tra realtà e rappresentazioni è alla base delle speculazioni filosofiche dall'antichità ad oggi; sarebbe iperbolico affermare che le risposte siano tante quanti sono i filosofi, ma le classificazioni che cerchino di raggrupparle appaiono in genere errate. Ad esempio, affermare l'antirealismo di filosofi che negano la corrispondenza tra rappresentazioni e realtà oggettiva è un errore comune: all'interno di quella vasta categoria ci sono filosofi che ammettono l'esistenza di una realtà oggettiva, ma negano la possibilità di rappresentarla con esatta corrispondenza e negano che la rappresentazione possa modificare, senza altra azione, la realtà, come sostenuto in questo scritto[28]. Definire questi filosofi antirealisti è un errore. Così, non può essere definito antirealista chi fa del metodo empirico scientifico il suo metodo, anche se è un relativista e non ammette l'univocità logica o la mancanza di errore come possibili. Invece dà prova di antirealismo chi suppone l'esatta corrispondenza tra l'oggetto ed il concetto, tra la realtà e le nostre rappresentazioni di essa. Questa posizione nega la realtà sotto un duplice profilo. In prima istanza, in quanto nega l'esistenza di tutte quelle caratteristiche che i nostri sensi non possono percepire ma che tuttavia sono parte dell'oggetto. In seconda istanza, perché attribuisce alla realtà caratteristiche di valore (positive o negative) che derivano dal soggetto percipiente e che sarebbero diverse cambiando il soggetto. Noi percepiamo una pantofola nel suo essere pantofola, con caratteristiche fisiche legate alla funzione da essa svolta per noi o, comunque, avente

[28] Ricordo, *en passant*, che la fisica quantistica afferma una posizione vicina a questa nel principio di indeterminazione di Heisenberg e nell'equazione di Schrödinger.

una forma fisica percepibile ai nostri sensi, non la percepiamo come equazione d'onda né nel suo aspetto molecolare, quindi, come esattamente sia la pantofola – che percepiamo e rappresentiamo grossolanamente in un concetto di comodità che adatta il percetto alle nostre necessità – non lo sappiamo. Il mondo delle pantofole è un mondo di significati – non di sole informazioni – propri del sistema cognitivo che le percepisce, è il mondo della cultura umana, del reale, non è la realtà.

3. La proposta: informazione che genera significati

"I couldn't celebrate. It was fantastic that we had won,
but there was such a big part of me that saw this man
trying so hard and being so disappointed"[29]
"I'm happy when I win not like when my colleague wins.
But this time it felt like my win"[30]

Quanto prima esposto, riassunto nelle sette tesi evidenziate, sembra portare ad una situazione di stallo.

Le tesi [1] [2] [5] e [7] mostrano una spaccatura tra realtà e reale, che però si compone nel soggetto, che è, egli stesso, parte della realtà e creatore del reale che lo conforma.

Le tesi [3] [6] e [4] mostrano il significato di significato: questo si forma nell'essere umano attraverso il chiasma, offerto dalla sensazione e dalla percezione, e permette all'uomo non di rappresentare la vera realtà, ma di rappresentare un reale che gli permetta un'azione adeguata in una realtà nella quale, altrimenti, sarebbe cieco.

Quindi arriviamo a concludere che:

La realtà esiste ma è inconoscibile in sé [8]

[29] Maddy componente del gruppo di AlphaGo; avrebbe dovuto essere completamente dalla parte dell'intelligenza artificiale, alla cui vittoria ha contribuito, ma... (1:01:25 di 1:30:27): https://www.youtube.com/watch?v=WXuK6gekU1Y.

[30] Hajin Lee, Secretary General, International Go Federation (1:13:15 di 1:30:27); https://www.youtube.com/watch?v=WXuK6gekU1Y, a proposito dell'unica vittoria su 5 partite di Lee Se-dol.

Il significato si forma all'interno della nostra mente, o meglio cervello, e modifica o adatta alle nostre necessità il risultato della sensazione, in questo consiste la percezione. I significati costruiscono il mondo del reale. Non abbiamo quindi accesso diretto e veritiero alla realtà, il vero si costruisce nel reale. Possiamo a questo punto riesaminare alcune posizioni. La nota tesi nietzschiana "non ci sono fatti, ma solo interpretazioni" diventa: fatti e interpretazioni sono due mondi separati: uno oggettivo, l'altro immaginato; la realtà è indipendente dal pensiero ma non dalla nostra azione su di essa, viceversa il pensiero è indipendente dalla realtà ma non dalle conseguenze della nostra azione su di essa. Qualsiasi pensiero avviene in una scatola isolata dalla realtà, ma qualsiasi esternazione dello stesso diviene informazione, cioè parte della realtà. Qualsiasi fatto fa parte della realtà ed è informazione, quindi modifica il reale, tant'è che i fatti possono essere interpretati e comunicati come racconti su di essi. Sia il fatto percepito che quello comunicato possono dar luogo a rappresentazioni inadeguate per la nostra azione e possono non corrispondere in maggior o minor misura alla realtà, ma anche una rappresentazione inadeguata, se cagiona un'azione, modifica la realtà. È il reale, questo mondo immaginato che determina la nostra azione nella realtà. In questo senso si dirigono anche le ipotesi provenienti da alcuni neuroscienziati interessati a problemi epistemologici. Così Beau Lotto afferma:

> Quando apriamo gli occhi vediamo il mondo come è realmente? Vediamo la realtà? […] La risposta è che non vediamo la realtà.
>
> Il mondo possiede una sua oggettività solo che noi non la vediamo. La nostra esperienza del mondo non coincide con ciò che quest'ultimo è realmente *perché il nostro cervello non si è evoluto per fare questo*[31].

La metaepistemologia scientifica, sia dalla parte della meccanica quantistica, che da quella delle neuroscienze, quindi sia dagli studi sulla materia che da quelli sull'uomo, dà risposte concordi circa la nostra impossibilità di rappresentare con verità la realtà.

[31] B. Lotto, *op. cit*., p. 13.

A noi interessa ora introdurre un'ulteriore distinzione. La loro duplice natura di realtà e reale la mostrano i cosiddetti "oggetti sociali" introdotti come punto speculativo da Searle[32], tra questi rientrano senz'altro i frutti dell'operare giuridico. La loro materialità ed il loro far parte della realtà, come atti cartacei e quant'altro, non va confuso con il loro far parte del reale. Quegli atti contengono molecole e atomi per la realtà. Nel reale essi costituiscono significati nei soggetti che li leggono. Nella mia ipotesi, i fatti sociali hanno sempre una duplice natura che però si differenzia e diverge a dipendenza del sistema cognitivo che li vede. Infatti la natura atomica potrà interessare il fisico o il chimico, mentre quella reale, cioè sociale o culturale interesserà l'individuo nella sua costruzione di significati, che non sono presenti nell'atto cartaceo in sé; su di questo si possono trovare solo segni, ma sono o divengono nel cervello di chi li vede dei significati. Parlare di fatti sociali alla stregua degli accadimenti naturali è una ulteriore fonte di equivoci.

L'evoluzione della cultura a misura del *sapiens*, delle sue possibilità cognitive e delle sue necessità, seppur possa essere considerata una limitazione, nelle metodologie scientifiche, per la corretta o vera rappresentazione del proprio oggetto, tuttavia, può essere un vantaggio nelle teorie normative sulla socialità umana, perché esse regolano il comportamento dell'uomo nel mondo percepibile dai suoi sensi e rappresentabile con gli strumenti logici e di ragionamento in lui evoluti. Ciò comporta che non ogni realtà possa essere presa in considerazione, come tale, da esse (scienze normative) e non ogni verità. La verità scientifica non costituisce l'assiologia giuridica, può solo essere un mezzo, uno dei tanti, per raggiungere tale assiologia o realizzarla, nulla più. Parimenti, la validità argomentativa della logica classica non è il necessario presupposto di una verità normativa, a garanzia della conservazione della stessa[33]. Non ogni argomento valido deve essere un argomento accettato in una scienza

[32] J.R. Searle, *Creare la realtà sociale*, trad. it., Raffaello Cortina Editore, Milano, 2010.

[33] Il principio di conservazione del valore di verità non riguarda la decisione che invece crea il valore di verità, lo costituisce.

normativa[34], così come il principio di non contraddizione non è un principio insuperabile nella decisione giuridica. È nota la posizione di Hans Kelsen sulla non applicabilità della logica classica in genere e del principio di contraddizione esclusa in particolare alla decisione giuridica. Se anche la soluzione adottata dal giudice nella decisione di una controversia non si riferisce ad una norma valida presente nell'ordinamento, la disapplica, purtuttavia la disapplicazione non costituisce contraddizione. Questa parte è la più emendata, tralasciata, mormorata con sottaciuta disapprovazione, negli scritti dell'autore praghese, ma Kelsen ha ragione e con lui anche Frege, una decisione non è una conclusione logica, anzi, il decisore può decidere anche di applicare il principio di non contraddizione, o di non applicarlo, senza commettere un errore logico. Questo è vero in ogni sistema normativo, e così anche negli ordinamenti giuridici, in essi è il decisore che sceglie quali inferenze logiche utilizzare, la sua decisione è costitutiva delle regole inferenziali da utilizzare nell'analisi del caso per la sentenza o per la decisione. Prima della decisione il giudice si trova di fronte ad una contraddizione di partenza, nella realtà rappresentata, che deve sciogliere nella decisione. Per farlo deve ammettere nel suo ragionamento come vera e, contemporaneamente, come non vera una determinata qualificazione giuridica.

Prendiamo l'esempio più lampante di questa contraddizione di fondo nel ragionamento del giudice:

L'imputato è 'presunto' innocente e 'sospettato' colpevole [9]

L'imputato è contemporaneamente innocente e colpevole, fino alla decisione, attraverso la quale, dopo le inferenze scelte dal giudice, viene esclusa la contraddizione e l'imputato diviene o innocente o colpevole. Si noti che l'imputato non è innocente o colpevole a dipendenza della commissione o meno del fatto considerato reato, ma a dipendenza della decisione del giudice. La dimostrazione viene data dalla necessità di una nuova sentenza, in caso di emersione di nuove prove, per il cambiamento di sta-

[34] Ad esempio la fallacia dell'argomento *ad personam* è accettata per la valutazione della attendibilità della testimonianza.

tus. La decisione del giudice "imputa" le conseguenze previste dalla norma al soggetto "come se" fosse colpevole o viceversa, in base ad una ponderazione di probabilità.

Gli aggettivi *presunto* e *sospettato* della [9] si riferiscono ad un tipo di ragionamento che deve portare alla trasformazione della contraddizione attuale, inammissible nella logica classica, ma come può avvenire?

All'epoca in cui Kelsen scriveva non esistevano molte possibilità di ricorso a logiche alternative rispetto a quella classica, ma oggi ve ne sono in abbondanza ed il processo di decisione può essere descritto altrimenti.

Nella logica classica, la contraddizione nelle premesse rende vera qualsiasi conseguenza, dando luogo ad una esplosione di possibili conseguenze tutte parimenti vere e false, ma la rappresentazione della contraddizione tramite le logiche paraconsistenti, ad esempio, può fornire risultati coerenti ed adeguati a ciò che avviene nel diritto. In effetti, il giudice si trova di fronte ad una contraddizione che genera un'infinità di conclusioni possibili, tra le quali però il giudice è tenuto ad escludere e scegliere. Le ipotesi ricostruttive, presentate dalle parti, possono essere consistenti al loro interno, ma contraddittorie tra di loro, anzi, quasi sempre lo sono. Per il giudice esse hanno lo stesso valore di verità e devono essere prese in considerazione tutte, contemporaneamente, sono quindi contraddittorie. Le logiche paraconsistenti si definiscono proprio sulla possibilità di limitare le conseguenze che derivano da due o più premesse contraddittorie, falsificando il principio "*ex contradictione quodlibet*" e non impedendo l'utilizzazione di premesse contraddittorie. Un'altra possibilità è fornita dall'uso di logiche probabilistiche o statistiche o altre logiche non standard.

La normatività della decisione implica che essa sia espressione delle rappresentazioni della realtà e ad esse appartenga, non alla realtà: la decisione del giudice, la sentenza, é un fatto sociale ed appartiene al reale. L'esistenza di una contraddizione nelle rappresentazioni della realtà è non solo possibile, ma anche frequente. Eliminare la contraddizione serve a adeguare la rappresentazione alla realtà, non a pensare. Non c'è contraddizione impossibile nel pensiero fantastico o immaginativo, tutto ha pari

possibilità ed un uguale valore di verità, tutto coesiste. Invece nel pensiero ricostruttivo di una realtà si apre la caccia alle contraddizioni, esse esprimono, nella modalità della rappresentazione, le impossibilità, sono gli occhiali di Aristotele che ci permettono di vedere meglio la realtà. Sono occhiali, però, che manifestano i propri limiti nel confronto con i sistemi artificiali. In questi il processo che va dalla sensazione e raccolta delle informazioni alla creazione del significato è radicalmente diverso rispetto al nostro. Il loro reale è diverso dal nostro anche se condividiamo la realtà nella quale viviamo, così anche i significati da essi proposti[35] e che si formano nelle loro rappresentazioni della realtà. I sistemi artificiali hanno capacità di elaborazione che eccedono di gran lunga quelle a noi disponibili tramite la logica classica. Essi possono possedere, inoltre, possibilità di percezione più ampie di quelle che si sono in noi evolute. La loro possibilità di avvicinarsi alla realtà e di rappresentarla è superiore alla nostra. Tuttavia il loro reale si discosta dal nostro e non sarà mai possibile renderlo uguale. Il significato di pantofola che è in noi sarà sempre diverso da quello che si troverà in uno di questi sistemi, che non condivide la nostra storia evolutiva.

Coesistono ora due reali ontologicamente diversi, due insiemi di significati, riferentisi alla medesima realtà.

Questa conclusione può portarci ad un tentativo di avvicinamento, se non altro teorico, rendendo comuni i concetti utilizzati per descrivere la loro cognizione e la nostra, ciò che costruisce i due diversi reali.

Ho già discusso il significato di significato, ma occorre seguire il processo di formazione del significato nel reale, partendo dalla realtà, in un modo che sia compatibile sia con i sistemi artificiali che con gli esseri umani[36]. La mia proposta si basa sui seguenti momenti che riassumo in punti:

1. la realtà è informazione;

[35] Ad esempio nella IA generativa dove si creano nuovi contenuti senz'altro dotati di originalità e di creatività.

[36] Vedi F. Romeo, "In-formatio", cit., pp. 112 ss.; Id., *Lezioni di logica ed informatica giuridica*, cit., pp. 32 ss.

2. l'informazione, sotto forma di dati, stimola i nostri sensi o quelli dei sistemi artificiali ed entra nel sistema cognitivo, artificiale o naturale;
3. il sistema cognitivo elabora i dati ricevuti unitamente alle informazioni su se stesso che già possiede, comprese quelle biologiche o genetiche, accessibili o non accessibili;
4. il sistema cognitivo forma il significato;
5. il significato determina la volontà e permette l'azione del sistema cognitivo sulla realtà, adattandola a se stesso.

Già quel poco che ho esposto fin qui lascia intravedere alle spalle questioni filosofiche di grande rilievo, la concezione si distacca dalla discussione filosofica e si avvicina a quella fornita dalle scienze cognitive nel loro vasto complesso. In questo modo, però, i due mondi di significati diversi, l'uno già consolidato e l'altro che si sta creando, vengono a riunirsi in un nuovo chiasma, quello della comune comunicazione tramite l'informazione.

Francisco José Santamaría Ramos

Implicaciones éticas y legales relacionadas con el uso creciente de la IA

Abstract: **Related Ethical and Legal Implications with the Increasing Use of AI.** This paper examines the ethical implications of artificial intelligence, emphasizing the importance of a responsible and ethical approach in its development and application. Through the analysis of the Montreal Declaration and the UNESCO Recommendation, reflections are made on the regulation and the role ethics should play in the design and development of artificial intelligence, bearing in mind that the core objective of this technology is to be a benefit to society while respecting fundamental human values and rights.

Keywords: Artificial Intelligence, Morality, Ethics, Human Rights.

Resumen: 1. Introducción – 2. Definiendo la inteligencia artificial – 3. La Declaración de Montreal: Análisis y Crítica – 3.1. Principio de bienestar – 3.2. Principio de respeto a la autonomía – 3.3. Principio de protección de la privacidad y la intimidad – 3.4. Principio de solidaridad – 3.5. Principio de participación democrática – 3.6. Principio de equidad – 3.7. Principio de inclusión de la diversidad – 3.8. Principio de prudencia – 3.9. Principio de responsabilidad – 3.10. Principio de desarrollo sostenible – 4. UNESCO: Recomendación sobre la ética de la inteligencia artificial – 4.1. Introducción – 4.2. Evaluación del impacto ético – 4.3. Gobernanza y administración éticas – 4.4. Política de datos – 4.5. Desarrollo y cooperación internacional – 4.6. Medio ambiente y ecosistemas – 4.7. Género – 4.8. Cultura – 4.9. Educación e investigación – 4.10. Comunicación e información – 4.11. Economía y trabajo – 4.12. Salud y bienestar social – 5. Conclusiones.

1. Introducción

La sociedad actual asiste expectante a una revolución de carácter tecnológico que no sólo promete unos grandes beneficios al ser humano, sino que también plantea una serie de preguntas sobre cómo se van a implementar esos cambios y, sobre todo, cuáles van a ser las consecuencias de estos.

No puedo sino coincidir con el economista e ingeniero alemán Klaus Schwab que asistimos a: "La cuarta revolución industrial donde se plasma que el ser humano asiste a una nueva

transformación de la humanidad que está cambiando por completo la forma de vivir, trabajar y relacionarse"[1].

La presente revolución es consecuencia necesaria de "la confluencia de numerosos avances tecnológicos en diversos campos: impresión 3D, vehículos autónomos, biotecnología, computación cuántica, etc"[2]. Desde el punto de vista que más nos interesa, para el presente trabajo, el avance tecnológico de importancia es la Inteligencia Artificial.

Es más, en la actualidad podemos hablar de las denominadas tecnologías: "NBIC (nano-bio-info-cogno): nanotecnologías, biotecnologías, tecnologías de la información y ciencias cognitivas (IA, ciencias de datos, robótica, interfaces cerebro-máquina, biología sintética, nanotecnología)"[3].

Atrás queda octubre de 1950 cuando Alan Turing se hizo la pregunta que cambió radicalmente el mundo de la informática: "¿Pueden pensar las máquinas?"[4]. En menos de un siglo hemos pasado de pensar que el Test de Turing era una mera fábula a *democratizar* el uso de la inteligencia artificial a través de sistemas como *Chat GPT o Midjourney*, entre otros.

Hablar sobre cómo la inteligencia artificial ha penetrado a nivel social, supone examinar el impacto de la presente tecnología en diversos aspectos de la vida humana y, al mismo tiempo, supone hablar de los desafíos y oportunidades de una tecnología que, a pesar de que cada vez nos sorprende más, no podemos olvidar que estamos en presencia de una tecnología emergente.

Como el lector puede suponer, el presente trabajo no puede examinar de forma detallada todas y cada una de las áreas sociales en las que la inteligencia artificial va a impactar. Sin embargo, lo que sí puedo hacer es enumerar las que, en mi opinión, considero más importantes, aunque no por ello las únicas:

[1] K. Schwab, *La cuarta revolución industrial*, Penguin Random House Grupo Editorial, Barcelona, 2016.

[2] F.J. Santamaría Ramos, "Robótica, Inteligencia Artificial y Desempleo", in J.P. Pampillo Baliño y S. Botero Gómez (coordinadores), *Justicial Social global: Perspectivas, reflexiones y propuestas desde Iberoamérica*, Tirant Lo Blanch, Ciudad de México, 2022, pp. 425-458.

[3] T. Ausín, "¿Por qué ética para la Inteligencia Artificial? Lo viejo, lo nuevo y lo espurio", in *Sociología y Tecnología*, 11 (2021), p. 2.

[4] A.M. Turing, "Computing machinery and intelligence", in *Mind*, (1950), n. 49, pp. 433-460.

- *Empleo*: Sin duda, el uso de la inteligencia artificial en los mercados laborales va a marcar un antes y un después en la historia del ser humano y en la sociedad del siglo XXI. Hoy en día, la pérdida de empleos en algunos sectores es una realidad. Incluso, sin ser absolutamente pesimista, incluso sin pensar que el ser humano se puede ver desprovisto de su activo laboral, no podemos negar que la inteligencia artificial trae consigo desplazamiento laboral e incluso la necesidad de adaptación y capacitación del empleado para obtener un *trabajador del siglo XXI* acorde a las expectativas del mercado laboral.

- *Brechas*: Un uso desigual de la inteligencia artificial puede generar brechas tanto sociales como económicas. Hablar de brechas supone hablar de una desigualdad desde el punto de vista del acceso a las tecnologías y, sobre todo, a las oportunidades. Todo ello, nos conduciría a hablar de la posibilidad real de que se produzca una concentración tanto de pobreza como de poder en manos de aquellos que controlen sistemas basados en la inteligencia artificial.

- *Cambios sociales y culturales*: Las interacciones sociales e incluso la cultura también son sectores en donde la penetración de la inteligencia artificial va a penetrar en mayor o menor medida. Sin duda alguna, su impacto va a poder ser observado en cuestiones como el ámbito laboral, el ocio y el entretenimiento así como la forma en la que los seres humanos nos relacionamos con los demás.

- *Formación y educación*: Hablamos de cómo la penetración de la inteligencia artificial va a modificar la forma en la que se adquieren habilidades y conocimientos así como cuestiones relativas a la accesibilidad y equidad en materia formativa y/o educativa.

- *Ética, moral y derechos humanos*: La inteligencia artificial es una tecnología que impacta de lleno, debido a cuestiones como la toma de decisiones y la responsabilidad, en dilemas de carácter moral, ético, así como en una cuestión esencial: los derechos humanos.

Es justo en este último punto donde vamos a focalizar el objeto de discurso del presente trabajo. Vamos a tratar de analizar las cuestiones éticas y morales que implica la penetración de la

inteligencia artificial en la sociedad. Asimismo, trataremos de analizar el impacto de esta tecnología en los derechos humanos.

A medida que la sociedad empieza a convivir con los sistemas de inteligencia artificial, el ser humano se empieza a hacer preguntas e incluso a cuestionar el avance de la presente tecnología. ¿Debemos limitar el desarrollo de la inteligencia artificial? ¿Las máquinas harán todo el trabajo por nosotros? ¿Hasta qué punto debemos dejar que las inteligencias artificiales tomen decisiones sobre nosotros? La sociedad, los seres humanos, nos hacemos este tipo de preguntas. Y son preguntas totalmente lícitas. Al fin y al cabo "todo el mundo está interesado en la ética. Todos tenemos nuestras propias ideas sobre lo que está bien y lo que está mal y cómo podemos diferenciarlo"[5].

Si bien la Real Academia de la Lengua Española define la ética como "aquel conjunto de normas morales que rigen la conducta de la persona en cualquier ámbito de la vida" yo prefiero personalmente entender la ética como "aquella dimensión de la filosofía que reflexiona sobre la moralidad"[6].

Esta conceptualización de la ética me parece mucho más apropiada, en tanto en cuanto, la ética se encarga de estudiar la moral. ¿Y qué es la moral? Podemos decir que la moral es "una dimensión del mundo vital compuesta de valoraciones, actitudes, normas y costumbres que orientan o regulan el obrar humano"[7].

Como podemos observar, el campo de la ética cobra importancia en un ámbito tecnológico de tanta trascendencia como es la inteligencia artificial. Al fin y al cabo, hemos de tener en cuenta que las personas que diseñan y desarrollan un sistema de inteligencia artificial se encuentran sometidas al ámbito de la moral y, por tanto, en cierto modo, se encuentran sometidas al campo de la ética.

A lo largo de las presentes líneas trataremos de analizar cómo la ética puede ser aplicada al campo de la inteligencia artificial. Para ello procederemos a analizar dos de los grandes textos in-

[5] D. Robinson, C. Garrat, *Ética: una guía ilustrada*, Editorial Tecnos, Madrid, 2021, p. 11.

[6] A. Cortina, *Ética sin moral*, Editorial Tecnos, Madrid, 2000, p. 29.

[7] J. de Zan, *La ética, los derechos y la justicia*, Fundación Konrad Adenauer, Montevideo, 2004, p. 19.

ternacionales en la materia: la Declaración de Montreal para un desarrollo responsable de la inteligencia artificial y la Recomendación de la UNESCO, sobre la ética de la inteligencia artificial.

Sin embargo, antes de embarcarnos en el análisis de los citados textos, debemos comenzar por definir qué es la inteligencia artificial.

2. Definiendo la inteligencia artificial

Debemos comenzar diciendo que no existe una definición *pacífica* sobre qué es la inteligencia artificial. Los propios avances tecnológicos sobre la materia hacen que, de alguna forma, estemos en constante revisión del término o del concepto.

En 2017 el Parlamento Europeo, realiza una aproximación al conceto de inteligencia artificial estableciendo las características que debe tener un "robot inteligente":

- Capacidad de adquirir autonomía mediante sensores y/o mediante el intercambio de datos con su entorno (interconectividad) y el análisis de dichos datos.
- Capacidad de aprender a través de la experiencia y la interacción.
- Forma del soporte físico del robot.
- Capacidad de adaptar su comportamiento y acciones al entorno[8].

En 2018, la Declaración de Montreal para un desarrollo responsable de la inteligencia artificial define el concepto como:

Una serie de técnicas que permiten que una máquina imite el aprendizaje humano, en concreto para aprender, hacer predicciones, tomar decisiones y percibir lo que la rodea. En el caso de un sistema informático, la inteligencia artificial se aplica a los datos digitales[9].

[8] Parlamento Europeo (2017), Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL)). Recuperado de https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_ES.html#title2.

[9] Centro de Investigación en Ética (CRE), Universidad de Montreal (2018), Declaración de Montreal para un desarrollo responsable de la inteligencia artificial. Recuperado de https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/01/ES-UdeM_Decl-IA-Resp_LA-Declaration_v4.pdf.

En 2020, la Comisión Europea definía la Inteligencia Artificial como: "una combinación de tecnologías que agrupa datos, algoritmos y capacidad informática"[10].

En 2021, la UNESCO, en su Recomendación sobre la ética de la inteligencia artificial ya es consciente de que no es posible proporcionar una única definición del concepto y que incluso la definición debe ser actualizada en función de los avances tecnológicos, considera que la inteligencia artificial puede ser definida como "sistemas capaces de procesar datos e información de una manera que se asemeja a un comportamiento inteligente, y abarca generalmente aspectos de razonamiento, aprendizaje, percepción, predicción, planificación o control"[11].

En el mismo año que la UNESCO, la Comisión Europea lanzaba una propuesta de normas armonizadas, comúnmente denominada "Ley de Inteligencia Artificial" que definía el concepto de "Sistema de inteligencia artificial" como:

> El software que se desarrolla empleando una o varias de las técnicas y estrategias que figuran en el anexo I[12] y que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa[13].

[10] Comisión Europea (2020), *Libro blanco sobre la inteligencia artificial – un enfoque europeo orientado a la excelencia y la confianza*. Recuperado de https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_es?filename=commission-white-paper-artificial-intelligence-feb2020_es.pdf).

[11] UNESCO (2021), *Recomendación sobre la ética de la inteligencia artificial*. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa.

[12] Las técnicas y estrategias de dicho anexo son las siguientes:

"Estrategias de aprendizaje automático, incluidos el aprendizaje supervisado, el no supervisado y el realizado por refuerzo, que emplean una amplia variedad de métodos, entre ellos el aprendizaje profundo.

Estrategias basadas en la lógica y el conocimiento, especialmente la representación del conocimiento, la programación (lógica) inductiva, las bases de conocimiento, los motores de inferencia y deducción, los sistemas expertos y de razonamiento (simbólico).

Estrategias estadísticas, estimación bayesiana, métodos de búsqueda y optimización".

[13] Comisión Europea (2021), Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en ma-

Recientemente, la presente definición ha sido modificada para alinearla con el trabajo de organizaciones internacionales que trabajan en la materia. Asimismo, se ha encomendado a la Comisión Europea que trabaje en aras a desarrollar directrices sobre cómo debe aplicarse la definición de "sistema de inteligencia artificial".

En todo caso, el proyecto final establecido por la Unión Europea, el 26 de enero de 2024, define como "sistema de inteligencia artificial":

> Un sistema basado en una máquina diseñado para funcionar con distintos niveles de autonomía y que pueden mostrar adaptabilidad después del despliegue y que, por razones explícitas, u objetivos implícitos, infiere, a partir de los insumos que recibe, como generar resultados tales como predicciones, contenidos, recomendaciones o decisiones que puedan influir físicamente o en entornos virtuales[14].

Definición que finalmente, en el Reglamento 2024/1689 quedó de la siguiente forma:

> Un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales[15].

teria de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión. Recuperado de https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52021PC0206.

[14] Comisión Europea (2024), Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión. Recuperado de https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

[15] Parlamento y Consejo de la Unión Europea (2024), Reglamento (UE) 2024/1689 del Parlamento Europeo y del consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.o 300/2008, (UE) n.o 167/2013, (UE) n.o 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE)

Desde el punto de vista doctrinal, también observamos que lograr una definición *pacífica* del concepto inteligencia artificial no es fácil.

Algunos autores consideran que puede definirse como:

> la ciencia y la ingeniería para fabricar máquinas inteligentes, especialmente programas informáticos inteligentes. Está relacionada con la tarea similar de usar computadoras para comprender la inteligencia humana, pero la IA no tiene que limitarse a métodos que son biológicamente observables[16].

Otros autores incluso hablan de distintos tipos de definiciones en función del pensamiento y la racionalidad ante la actuación:

- Sistemas que piensan como los humanos.
- Sistemas que actúan como los humanos.
- Sistemas que piensan racionalmente.
- Sistemas que actúan racionalmente[17].

Otros autores prefieren utilizar el concepto *agente*:

> Un agente es algo capaz de comportamiento inteligente. Puede ser un robot o un programa de ordenador. Los agentes físicos, como los robots, tienen una interpretación clara. Toman cuerpo como aparatos físicos que interactúan con un entorno físico. La mayoría de las investigaciones de IA, sin embargo, se centran en los agentes *virtuales* o de *software* que existen como modelos que ocupan un entorno virtual dentro de un ordenador[18].

La presente definición creo que es bastante acertada para definir el estado actual de la técnica. Lo que hoy en día podemos denominar como inteligencia artificial no es ni más ni menos que un sistema que tiene la capacidad de comportarse de una forma inteligente y que puede tener corporeidad y, por tanto,

2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial)

[16] J. McCarthy (2007), "What is Artificial Intelligence?", in *Universidad de Stanford*. Recuperado de http://jmc.stanford.edu/articles/whatisai/whatisai.pdf.

[17] S. Russel, P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education, Hoboken, 2003.

[18] H. Brighton, H. Selina, *Inteligencia Artificial: Una guía ilustrada*, Editorial Tecnos, Madrid, 2019.

capacidad de interactuar con su entorno físico o bien, pueden ser incorpóreas y, por tanto, ser únicamente un software con comportamiento inteligente.

No obstante, también nos podemos encontrar con definiciones que diferencian no por la cuestión corpórea sino por la consciencia. Así encontramos definiciones como la siguiente:

> […] Podemos denominar inteligencia artificial blanda (agentes artificiales blandos) o robótica a aquel tipo de agentes que, aunque son capaces de tomar decisiones por sí mismas y, además, son capaces de interactuar con su entorno, aún no son capaces de tomar conciencia de sí mismas. […] Hablar de un agente artificial fuerte, supone hablar de un agente que, en potencia, podría tener todas las habilidades y talentos del ser humano y no sólo eso, sino que también podrían encontrarse en posesión de otros talentos que no se encontrasen en el ser humano. Es decir, en potencia, un agente artificial fuerte, podría llegar a superar incluso al ser humano. […][19].

Quizá sea demasiado prematuro poder hablar de la posibilidad de que un agente artificial adquiera autoconciencia y, por tanto, que una inteligencia artificial sea capaz no sólo de detectar todo lo que sucede a su alrededor, sino que también sea capaz de descubrir aquello que sucede dentro de ella misma. A pesar de ello es una definición que probablemente tengamos que aplicar. Y seguramente no tardemos demasiado en hacerlo.

No obstante, considero que con las presentes definiciones y a pesar de que puedan existir muchas más, el lector puede tomar conciencia de lo difícil que es lograr definir el concepto de inteligencia artificial.

Creo que, en este sentido, podemos tomar como definición de aproximación la siguiente: una inteligencia artificial es un agente con capacidad de comportamiento inteligente, es decir, puede comportarse y razonar como un ser humano, con independencia de si es un agente virtual o un agente con corporeidad.

[19] F.J. Santamaría Ramos, "Agentes Artificiales", in F.J. Santamaría Ramos (coordinador), *Tecnoretos del Derecho*, Tirant Lo Blanch, Valencia, 2021, pp. 421-440.

3. La Declaración de Montreal: Análisis y Crítica

En el año 2017 la Universidad de Montreal comenzó a desarrollar y redactar una Declaración para el desarrollo responsable de la inteligencia artificial, conocida comúnmente como Declaración de Montreal y que vio la luz finalmente en el año 2018.

Los objetivos esenciales de la presente declaración son los siguientes:

- Crear un marco ético para el desarrollo y la implementación de la inteligencia artificial.
- Guiar la transición digital para que todos puedan beneficiarse de esta revolución tecnológica.
- Abrir un foro de discusión nacional e internacional a fin de lograr conjuntamente un desarrollo equitativo, inclusivo y ecológicamente sostenible dela inteligencia artificial[20].

A lo largo de las presentes líneas vamos a tratar de analizar y, en su caso, criticar los principios y valores éticos insertos en la presente declaración para tratar de dilucidar si en el campo de la inteligencia artificial es posible canalizar una serie de valores éticos esenciales que realmente respeten los intereses de la humanidad.

Comencemos, por tanto, enumerando los principios de la presente declaración:

- Principio de bienestar.
- Principio de respeto a la autonomía.
- Principio de protección de la privacidad y la intimidad.
- Principio de solidaridad.
- Principio de participación democrática.
- Principio de equidad.
- Principio de inclusión de la diversidad.
- Principio de prudencia.
- Principio de responsabilidad.
- Principio de desarrollo sostenible.

Pasemos a analizar detalladamente cada uno de estos principios.

[20] Universidad de Montreal (2018), *Declaración de Montreal para un desarrollo responsable de la inteligencia artificial*. Recuperado de https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/01/ES-UdeM_Decl-IA-Resp_LA-Declaration_v4.pdf.

3.1. Principio de bienestar

Sin entrar en su contenido completo, podemos decir que el presente principio se puede resumir en la siguiente frase: "El desarrollo y el uso de los sistemas de inteligencia artificial deben permitir que se cultive el bienestar de todos los seres conscientes".

Parto de la base, y creo que es necesario dejarlo claro, de que todos y cada uno de los principios establecidos en la presente declaración están pensados y diseñados como mecánicas para que las inteligencias artificiales puedan ser utilizadas siempre en beneficio de la humanidad.

Dicho esto, existen una serie de preguntas o de retos, dentro del presente sector, que considero que es necesario plantearse sino replantearse dado que cuando uno lee los principios, si bien los considera loables, no deja de tener la sensación de que les falta definición, de que les falta contenido para poder ser aplicados de una forma clara, concisa y contundente.

Por ejemplo, en el segundo punto del presente principio se establece que una inteligencia artificial "debe permitir que las personas vayan tras sus deseos, siempre y cuando no causen daño a otros seres conscientes". Asimismo, su cuarto apartado establece que una inteligencia artificial "no debe generar malestar, a menos que nos permitan alcanzar un bienestar superior".

Analicemos detenidamente estos dos puntos. En primer lugar, el concepto de daño es un concepto extremadamente abstracto. ¿Qué es causar daño a otro ser o a otra persona? Sinceramente, no creo que a la hora de construir el presente principio se estuviese pensando el en principio jurídico de no causar daño a otro (*alterum non laedere*) sino que más bien se está pensando en otro tipo de daño, más no estrictamente el de carácter jurídico. Debemos entender, por tanto, que hablamos de daño es su acepción más común o cotidiana.

Por tanto, considero que debemos entender por daño, cualquier cuestión que pueda afectar de forma negativa a un ser consciente. En este sentido, podríamos hablar de cualquier acción física, psicológica, económica o incluso discriminatoria y tanto de forma intencional como accidental.

El contexto, como podemos observar, es muy amplio e incluso puede dar lugar a situaciones contradictorias que puedan

impedir que la propia inteligencia artificial sea incapaz de tomar una decisión. Un ejemplo, podría ser la recomendación, por parte de la inteligencia artificial, de un comportamiento imprudente. Evidentemente un comportamiento imprudente, por ejemplo, conducir por encima del límite permitido, no debiera ser recomendado por una inteligencia artificial. Sin embargo, ¿y si esa recomendación se realiza porque en el vehículo va otra persona que necesita una intervención médica urgente? En este caso, existen dos daños al mismo tiempo, el del pasajero que necesita atención médica y los posibles daños a terceras personas como consecuencia de una conducción imprudente.

Idéntico razonamiento me produce el punto cuarto donde se habla de malestar y bienestar. Malestar es un término demasiado genérico y abstracto. Definirlo no es fácil… podemos decir que es cualquier situación que altera la tranquilidad o el confort de una persona desde un punto de vista negativo. Por tanto, el bienestar también afecta a la tranquilidad o el confort, pero desde un punto de vista positivo.

Mi primera impresión es que una inteligencia artificial, al igual que una persona, tiene que valorar, equilibrar, cuando va a tomar una decisión que va a generar un malestar, a pesar de que con esa decisión se va a alcanzar un bienestar superior que no se podría lograr de otra forma. Dado que estamos en presencia de conceptos demasiado genéricos, se me hace difícil tanto para las personas como para las inteligencias artificiales el calibrar la presente situación. Máxime cuando cada persona es singular. Lo que a mí me genera malestar no tiene porqué generárselo a otra persona. ¿Cómo puede una inteligencia artificial tomar una decisión sin conocer a la persona? No sería capaz de determinar si la situación le va a provocar malestar a esa persona. Incluso, algo muy común en sociedad. Un comentario inapropiado puede generar malestar en otra persona y muchas veces, ni siquiera nos hace saber dicho malestar.

El presente principio me parece poco práctico. Es incluso poco práctico para el propio ser humano. Por supuesto, desde un punto de vista bienintencionado o incluso optimista, pensamos que todas las personas buscamos cultivar el bienestar de otras personas o de todos los seres conscientes. Sin embargo, el ser

humano sigue entrando en guerra y causando daño y dolor e incluso, causamos daño y malestar a otros seres conscientes. Por ejemplo, seguimos matando animales para poder nutrirnos.

En definitiva, no creo que sea un principio que pueda ser aplicado a todos los sistemas de inteligencia artificial. Existen cierto tipo de recomendaciones o de decisiones que, de forma inevitable, van a causar daño o malestar en terceras personas. Al igual que sucede con el ser humano, en muchas ocasiones, la inteligencia artificial deberá asumir la toma de decisiones aún a sabiendas de que se va a causar un cierto daño o malestar. Lo importante será, en este caso, el equilibro o *beneficio* logrado.

3.2. Principio de respeto a la autonomía

El presente principio establece que las inteligencias artificiales "se deben desarrollar y utilizar respetando la autonomía de las personas, a fin de que tengan un mayor control sobre sus vidas y lo que las rodea".

Sobre el presente principio me interesa entrar a analizar sus apartados segundo, quinto y sexto.

El segundo apartado es tremendamente categórico:

> No se deben desarrollar ni utilizar las inteligencias artificiales para imponer un estilo de vida específico a las personas, ya sea de forma directa o indirecta, mediante la implementación de mecanismos opresivos de vigilancia y evaluación o de incentivos.

Tengo serias dudas sobre la aplicación práctica del presente apartado en el día a día de cualquier sociedad. Puedo lograr entender el sentido e incluso la dirección que se quiere tomar con el presente apartado y, sin embargo, creo que carece de sentido. ¿Al fin y al cabo, no se impone un estilo de vida específico a las personas en cualquier sociedad? Incluso en las sociedades más democráticas y tolerantes existe un estilo de vida perfectamente marcado y tasado. Aquel estilo de vida que marca el Derecho.

> La sociedad implica comportamientos: colectivos e individuales y dichos comportamientos deben ser sometidos a unas reglas del juego, a unas normas reguladoras, en definitiva, a

> unas mecánicas que nos permitan resolver las posibles controversias que surgen, de forma inevitable, en la sociedad[21].

Entiendo que este apartado se refiere a la posibilidad de que un régimen autoritario utilice sistemas de inteligencia artificial para oprimir, vigilar o evaluar a sus ciudadanos, pero seamos sinceros, los regímenes autoritarios han existido a lo largo de la historia y no han necesitado de la inteligencia artificial para lograr sus fines.

Además, en el supuesto de que eso pasase… ¿Quién va a hacer frente a esa situación? Baste ver cómo los diversos Estados giran la cabeza ante una guerra o un ataque frontal a un país que no sea el suyo. Sí, surgen sanciones internacionales, discursos de buenas intenciones y un largo etcétera que, luego, a la postre, no acaba con la guerra… Dudo mucho que ningún Estado haga poco más que criticar el uso de la inteligencia artificial con fines autoritarios…

El quinto apartado también tiene un buen sentido y dirección, máxime si nos atenemos a las propias características del siglo XXI: "No se deben desarrollar inteligencias artificiales para difundir información no confiable, mentiras o propaganda; se los debe diseñar con el objetivo de contener su difusión".

Las noticias falsas o *fakes news* son una de las lacras más importantes del siglo XXI. La llegada de la inteligencia artificial no hace sino agravar la cuestión. La presente cuestión es tan importante que la Unión Europea presentó el 16 de junio de 2022 un Código de buenas prácticas en materia de desinformación[22]. El presente Código suscribe cuestiones relevantes como la desmonetización de la desinformación, la transparencia de la publicidad política, el empoderamiento de los usuarios, las mecánicas de cooperación con los verificadores de datos, etc.

Sin embargo, considero que este apartado también carece de una utilidad práctica. Estoy de acuerdo y comparto el peligro que supone la desinformación en la sociedad globalizada del siglo XXI. Sin embargo, y aunque busquemos contextos de autorre-

[21] F.J. Santamaría Ramos, *Introducción al sistema jurídico español*, Tirant Lo Blanch, Valencia, 2022.

[22] Unión Europea (2022), *Código de buenas prácticas en materia de desinformación*. Recuperado de https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

gulación como son los códigos de buenas prácticas, finalmente necesitamos recurrir al Derecho. El Reglamento Europeo de Servicios Digitales busca lograr un entorno en línea seguro y "aborda la difusión de contenidos ilícitos en línea y los riesgos para la sociedad que puede generar la difusión de desinformación y otros contenidos"[23]. Como podemos observar, si desde el punto de vista social no logramos evitar las mentiras y la desinformación que generan las propias personas, difícilmente unos principios éticos ligados a la inteligencia artificial lo van a lograr.

Nos queda por analizar el sexto apartado del presente principio que establece que en el "desarrollo de la inteligencia artificial, se debe evitar la creación de dependencias mediante técnicas que capten la atención o imiten las características humanas (apariencia, voz, etc.) que pueden causar confusión entre las inteligencias artificiales y los humanos".

El apartado es interesante y se logra ver la finalidad de este. La sociedad tiene miedo a una pérdida de la socialización de las personas entre sí. Al fin y al cabo si una inteligencia artificial no sólo piensa y razona como el ser humano sino que además tiene rasgos similares o idénticos a los del propio ser humano, se puede caer en un alto grado de aislamiento social. Es decir, que las personas prefiramos interactuar con inteligencias artificiales antes que con otros seres humanos.

No obstante, fíjese el lector en que si bien, en otros apartados analizados, la Declaración es tremendamente tajante y categórica, no lo es en este caso. De un "no se deben crear dependencias" hemos pasado a un "se debe evitar". Considero que no es un descuido, sino que es algo que incluso en una Declaración de estas características es evidente… En un futuro, no muy lejano, los desarrolladores dotarán de apariencia, voz y etc. a las inteligencias artificiales para ser mucho más *amigables* para el ser humano.

[23] Parlamento y Consejo Europeo (2022), Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo de 19 de octubre de 2022 relativo a un mercado único de servicios digitales y por el que se modifica la Directiva 2000/31/CE (Reglamento de Servicios Digitales). Recuperado de https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX-:32022R2065#d1e2102-1-1.

3.3. Principio de protección de la privacidad y la intimidad

El presente principio establece la necesidad de proteger la intimidad contra cualquier intromisión de la inteligencia artificial, así como de los sistemas de adquisición de archivos de datos[24].

Sin duda alguna, el presente principio me parece de lo más oportuno y quizá, uno de los principios que, desde el punto de vista jurídico, sea de los más fáciles de implementar dado que aquí estamos tratando un valor social que ya ha tenido una regulación importante a nivel mundial.

No obstante, me gustaría analizar una serie de cuestiones desarrolladas en el presente principio que pueden ser, en cierto modo, debatibles.

El presente principio se desarrolla a través de 8 apartados de los que particularmente me interesa entrar en dos de ellos.

En el primer apartado se estable la protección de la intrusión de la inteligencia artificial así como de los sistemas de adquisición de archivos de datos de aquellos "espacios personales" en los que las personas no se encuentran bajo vigilancia e evaluación digital. Me surgen varias inquietudes al respecto.

La primera inquietud que me surge es qué debemos entender por "*espacio personal*". Y digo esto, porque considero que es un concepto que podemos denominar generacional. Es decir, la percepción de dicho espacio va a depender en gran medida, de la época en la que las personas hallamos nacido y madurado.

Bajo mi punto de vista, quizá *clásico,* en gran medida porque nací y me crie en el siglo XX, no tengo más remedio que considerar que ese "espacio personal" no es ni más ni menos que la propia vida privada en su conjunto, es decir, aquel lugar reservado sólo "a unos pocos, familiares y amigos"[25].

No obstante, parece que este no es el sentir de los redactores de la Declaración de Montreal ya que en su texto dan a entender que es posible diferenciar distintos "espacios personales" y que,

[24] Un sistema de adquisición y archivo de datos se puede definir como "todo sistema informático que pueda recopilar y registrar datos. En última instancia, estos datos se utilizan para entrenar los sistemas de IA como parámetros para la toma de decisiones".

[25] J.A. Soler Martínez, "Protección constitucional de la intimidad y de los datos de carácter personal frente a las nuevas tecnologías", in *Anuario de Derecho Canónico*, 11 (2022), pp. 93-126.

por tanto, los que deben protegerse son los que no se encuentran bajo vigilancia o evaluación digital. Y yo me pregunto, ¿existe algún espacio donde el ser humano ya no se encuentre sometido a vigilancia o evaluación? Sinceramente, considero que no. Alexa, Cortana o Siri ya son asistentes artificiales que están las 24 horas del día "pendientes de nuestros movimientos"[26] y que ya son "uno más de la familia, uno más del hogar familiar".

Nótese que incluso la Ley sobre Inteligencia Artificial de la Unión Europea considera que existen una serie de sistemas de inteligencia artificial que "son componentes de seguridad de productos, o que son en sí mismos productos"[27], que son clasificados como de "*alto riesgo*". Alguno de estos productos son los equipos a presión, aparatos de gas, etc. Sorprende sobremanera que dentro de esta clasificación de alto riesgo se encuentren también los juguetes. O quizá no sorprende tanto, dado que es posible, que el legislador europeo ya esté previendo las posibles vulneraciones sobre la intimidad que pueden tener los sistemas de inteligencia artificial aplicados a los juguetes. Sin duda, me parecen el producto perfecto, para *colarse* en la intimidad de todos los hogares.

Otro de los puntos que me interesa analizar es el segundo punto que establece la protección de los "pensamientos y las emociones contra aquellos usos que puedan ocasionar daños, en especial aquellos usos que imponen juicios morales sobre las personas o sus elecciones de vida".

Comparto plenamente este posicionamiento. Existe un peligro potencial a que los sistemas de inteligencia artificial puedan ser utilizados como *agentes de la moral*. No obstante, existe otro riesgo, que, bajo mi punto de vista, es aún más peligroso. Y es

[26] Diario "El País", "Empleados de Amazon escuchan a diario conversaciones que mantienen los usuarios con Alexa". Recuperado de https://elpais.com/tecnologia/2019/04/11/actualidad/1554992401_521050.html.

[27] Parlamento y Consejo de la Unión Europea (2024), Reglamento (UE) 2024/1689 del Parlamento Europeo y del consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.o 300/2008, (UE) n.o 167/2013, (UE) n.o 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial)

convertirnos en estados sometidos a la "*dictadura tecnológica*". Soy consciente de la dureza de mis palabras pero me sorprende, por ejemplo, que en Estados democráticos, como son los Estados que conforman la Unión Europea nos planteemos cuestiones como las que aparecen en el apartado tercero del artículo 50 de la Ley de Inteligencia Artificial donde se deja entrever que va a ser posible utilizar sistemas de inteligencia artificial enfocados a la "categorización biométrica y el reconocimiento de emociones que hayan sido autorizados por ley para detectar, prevenir e investigar delitos", eso sí, haciendo referencia a que dichos sistemas se sujetarán garantías adecuadas para los derechos y libertades de terceros y de conformidad con el Derecho de la Unión.. Pero fíjense en el matiz, no sólo se habla de la investigación propiamente dicha, hablamos de la detección y prevención, dos conceptos que implican anticipación, es decir, adelantarse a la propia comisión del delito.

Me asaltan las dudas… ¿Dónde queda el aforismo romano "*Cogitationis poenam nemo patitur*"? ¿Acaso una conquista propia de la revolución francesa[28] como el establecimiento de que el pensamiento no delinque ya no tienen vigor en la era del conocimiento? Considero que cometemos un grave error como sociedad si utilizamos los sistemas de inteligencia artificial no sólo como *agentes de la moralidad* sino también como sistemas que indaguen en nuestro pensamiento y en nuestras emociones para la prevención de delitos. Una persona determinada, en un momento concreto puede tener unos sentimientos y emociones muy intensos como pueden ser el odio y la ira y eso no significa que vayan a cometer un delito. Prejuzgar de esta forma a las personas no es propio de ninguna sociedad democrática del siglo XXI.

3.4. Principio de solidaridad

El presente principio establece que "el desarrollo de los sistemas de inteligencia artificial debe intentar preservar los lazos de solidaridad entre las personas y las generaciones".

[28] C.A. Gómez Pavajeau, "Injusto y culpabilidad en los grandes filósofos griegos", in *Revista de Derecho Penal y Criminología*, 31 (2010), n. 91, pp. 137-151.

Sobre el presente principio únicamente me interesa entrar a analizar el apartado tercero que establece que "no se deben implementar los sistemas de inteligencia artificial para remplazar a las personas en tareas que requieren relaciones humanas de calidad, sino que se los debe desarrollar para facilitar dichas relaciones".

Lo primero que me llama la atención es que hablar de relaciones humanas de calidad es un concepto bastante genérico. Quiero entender que se están refiriendo a un tipo específico de interacción que se da entre las personas basado en la empatía, la comunicación efectiva, el respeto mutuo y la confianza y que derivan, por tanto, en un estado saludable y satisfactorio para aquellos que las practican.

Entiendo el sentido y la dirección que implica la presente cuestión. Se trata de minimizar el impacto que los sistemas de inteligencia artificial puedan tener en un futuro a medio plazo, el remplazo de la socialización entre las personas por la socialización entre personas y agentes artificiales. Y no es descabellado pensar que esto pueda lograr. Si es posible que dos personas se enamoren simplemente chateando, no es descabellado pensar que las personas podamos tener afinidades con agentes artificiales y decidamos cultivar esas relaciones antes que cultivar relaciones con otras personas.

No obstante, hemos de pensar que, en algún momento, vamos a tener que implementar sistemas de inteligencia artificial que desarrollen relaciones de calidad, sobre todo en Estados donde la natalidad es realmente baja y no va a ser sencillo encontrar personas para desarrollar determinadas tareas como pueden ser el cuidado de personas mayores. Soy consciente de que mis palabras van contra el antropocentrismo propio de nuestra especie pero, también soy una persona realista y considero que a medio plazo, no existirán tantas facilidades para cultivar las relaciones de calidad en determinados contextos socio-laborales.

3.5. Principio de participación democrática

El presente principio establece que "los sistemas de inteligencia artificial deben cumplir con los criterios de inteligibilidad,

capacidad de justificación y accesibilidad, y se los debe someter a un análisis, un debate y control democráticos".

Sobre este punto poco tengo que objetar, simplemente realizar dos matices de cuestiones que considero que deben ser siempre así, con independencia de encontrarnos o no en un contexto de *participación democrática*.

Los apartados 7 y 9 del presente principio estableces que "en todo momento debemos ser capaces de verificar que los sistemas de inteligencia artificial estén haciendo aquello para lo que se los programó y para lo que se los utiliza" y que, "todo usuario de un servicio que utiliza un bot de chat[29] debe poder identificar con facilidad si está interactuando con un sistema de inteligencia artificial o una persona real".

Me parecen dos cuestiones de lo más razonables. Considero que, al menos con el estado actual de la técnica, es necesario que el ser humano sea capaz de constatar que los agentes artificiales realmente están realizando las tareas para las cuales se han creado. No me refiero a analizar todos y cada uno de sus procesos pero sí, ser conscientes de que se están dedicando a cumplir las finalidades para las cuales queremos utilizarlas y que, por tanto, no están desarrollando tareas secundarias o alternativas o, si lo están haciendo, que sepamos el porqué de dicha acción.

Asimismo, me parece totalmente pertinente que los seres humanos sepamos identificar si nos encontramos en presencia de otro ser humano o si, por el contrario, nuestra interacción se realiza con un agente artificial.

3.6. Principio de equidad

En mi opinión estamos en presencia de uno de los principios de la Declaración más subjetivos y, sobre todo, más ambiciosos. De una forma breve, sus siete apartados se resumen en establecer que "el desarrollo y uso de los sistemas de inteligencia artificial deben contribuir a crear una sociedad justa y equitativa".

Creo que el presente supuesto es demasiado ambicioso y parte de un razonamiento que, a mi juicio, es erróneo. Parte de la base de

[29] Hablamos de "un sistema de inteligencia artificial que pueden conversar con una persona en lenguaje natural".

que la sociedad es local, pero, la sociedad del siglo XXI es global. Voy a tratar de explicarme lo mejor posible, dado que este principio necesita de un desarrollo que no puede abarcar el presente estudio. Y aunque mi análisis será breve no carecerá de rigor.

Todo el presente principio pivota en esperanzas muy loables:

> - Evitar la creación, el refuerzo o la reproducción de patrones de discriminación (social, sexual, étnica, cultural, religiosa …).
> - Tratar de eliminar las relaciones de dominación entre grupos y personas en función de diferencias de poder, riqueza o conocimiento.
> - La reducción de las desigualdades y las vulnerabilidades.
> - Un acceso igualitario a los recursos, conocimientos así como a las herramientas digitales esenciales.
> - La equidad social.

Sin embargo, "el mundo no es igualitario ni puede serlo"[30]. En este sentido coincido plenamente con la siguiente afirmación:

> La equidad social suele referirse a diferentes ámbitos: justicia social y cohesión; de igualdad de oportunidades, derechos, libertades, de trato y de logro/resultados; recibir prestaciones básicas del estado y una distribución de la renta/riqueza "razonable"; evitar situaciones de pobreza o discriminaciones (por etnia, lugar de nacimiento o residencia, capacidad, género, edad, etc.). Es por tanto, un fenómeno histórico y cultural que se da en todas las sociedades, si bien el grado, ámbito (o bien valorado), el sentido… para establecer la desigualdad variará según el país, la cultura, el sistema de creencia o las condiciones del entorno, entre otras[31].

Esta es la cuestión importante, el punto nuclear, el concepto de equidad social es un concepto que se da en todas las socieda-

[30] C. Pacella, "Aparatos estéticos", in *Revista del Centro de Investigaciones de la Facultad de Filosofía y Humanidades, Áreas Letras*, 11 (2020), n. 18, p. 8.

[31] R. Calvo Palomares, J.J. Portet Tiebas, M. Bou Pastor, "La equidad social como elemento esencial para el desarrollo local", in J.A. Márquez Domínguez, J.M. Jurado Almonte, F.J. Pazos García (coordinadores), *Desarrollo local en territorios de fronteras*, Servicio de publicaciones de la Universidad de Málaga, Málaga, 2016.

des pero que varía según el país, es decir, es un fenómeno *local*. Los agentes artificiales tienen una dimensión de carácter global. Una persona en España puede utilizar un sistema de inteligencia artificial desarrollado en China. ¿Qué el ser humano trata de evitar las discriminaciones? Podemos decir que, a nivel general, es uno de los objetivos de la humanidad. ¿Realmente es una realidad? Todos sabemos que no lo es.

Algunos ejemplos:

(i) Según las Naciones Unidas, en 2023 casi 700 millones de personas en todo el mundo subsistirán con menos de 2,15 dólares. Es más, se prevé que no se alcanzará el objetivo mundial de acabar con la pobreza extrema para 2030, ya que casi 600 millones de personas siguen viviendo en la pobreza extrema[32].

(ii) Amnistía Internacional establece que, durante el año 2022, los conflictos globales provocaron graves violaciones de los derechos humanos y del derecho internacional humanitario, con una respuesta internacional insuficiente. La violencia de género persistió a pesar de algunos avances legislativos, y el derecho al aborto experimentó altibajos. La pandemia de COVID-19 exacerbó las crisis económicas, alimentarias y climáticas, afectando desproporcionadamente a los grupos marginados y aumentando la desigualdad. La represión de la disidencia política alimentó los conflictos armados y los golpes de Estado, mientras que la violencia sexual en estos contextos se mantuvo como una cruel realidad[33].

(iii) Naciones Unidas, en 2022 llegó a establecer que,

> En los Estados Unidos de América se habían aprobado o estaban en curso de examen leyes que limitaban los debates en las escuelas sobre la raza y otras características personales,

[32] Naciones Unidas (2023), "Desafíos globales: Acabar con la pobreza". Recuperado de https://www.un.org/es/global-issues/ending-poverty#:~:text=Seg%C3%BAn%20estimaciones%20recientes%2C%20en%202023,menos%20de%202%2C15%20d%C3%B3lares.&text=Se%20prev%C3%A9%20que%20no%20se,viviendo%20en%20la%20pobreza%20extrema.

[33] Amnistía Internacional, *Informe 2022/2023 Amnistía Internacional: La situación de los derechos humanos en el mundo*, Amnesty Internacional Ltd, Reino Unido, 2023.

> y también iniciativas para prohibir libros en las bibliotecas escolares y públicas, en especial los que trataban cuestiones relativas a la raza, así como a otras identidades. En algunos estados se habían aprobado leyes y medidas restrictivas del voto, que podían afectar de forma desproporcionada a los afrodescendientes[34].

A la humanidad le queda por recorrer un largo camino en cuanto a la equidad social se refiere y, sin embargo, el temor a los sistemas de inteligencia artificial es tal, que buscamos que un producto creado por el ser humano sea un espejo en el que desearíamos vernos reflejados. La realidad, en mi opinión, es que al igual que el ser humano no es equitativo socialmente, sus productos, las inteligencias artificiales, tampoco lo serán. Por mucho empeño que pongamos en ello.

3.7. Principio de inclusión de la diversidad

El presente principio se puede resumir estableciendo que "el desarrollo y el uso de los sistemas de inteligencia artificial deben contribuir a preservar la diversidad social y cultural y no debe restringir las posibilidades en lo que respecta a las elecciones de vida o las experiencias personales".

Realmente, al igual que me sucede con el principio anteriormente expuesto, considero que el presente principio necesita de un desarrollo más exhaustivo del que se le puede dar en el presente trabajo. No obstante, trataré de centrarme en alguna de sus cuestiones más importantes.

En su primer apartado establece que los sistemas de inteligencia artificial "no deben conducir a una homogeneización de la sociedad mediante la normalización de los comportamientos y opiniones".

Sin embargo, la homogeneización ya es una realidad, con independencia de la penetración en la sociedad de los agentes artificiales. De hecho:

[34] Asamblea General de Naciones Unidas, *Promoción y protección de los derechos humanos y las libertades fundamentales de los africanos y los afrodescendientes frente al uso excesivo de la fuerza y otras violaciones de los derechos humanos por agentes del orden, mediante un cambio transformador en favor de la justicia y la igualdad raciales*, Nueva York, 2022.

> La estandarización cultural, fruto de la modernización y la creciente homogeneización de la sociedad, la mundialización de las problemáticas, la aldea global, no ha conseguido que desaparezcan los desequilibrios geopolíticos y culturales, que se reflejan en la estructura jerarquizada de las culturas y en las nuevas formas de estratificación social, en consecuencia, nuevas formas de exclusión social, educativa y cultural[35].

No puedo sino suscribir estas palabras, ya estamos asistiendo a un fenómeno complejo y de consecuencias ciertamente imprevisibles para la sociedad, nuestra homogeneización cultural.

Esta cuestión, también se hace patente en el apartado cuarto del presente principio cuando se habla de la posibilidad de que los agentes artificiales puedan "encerrar a las personas en un perfil de usuario o confinarlas en un filtro burbuja"[36].

Es decir, determinados servicios de internet pueden, en cierta forma limitar o censurar nuestra libertad de información. O mejor dicho, nuestras propias experiencias como usuarios pueden hacer que nos encerremos en nosotros mismos y no seamos capaces de localizar otros puntos de vista. Sin duda, como muy bien se dice en la Declaración, esto "limita las posibilidades de desarrollo personal".

Por último, en su último apartado, se establece que "se debe diversificar la oferta de los sistemas de inteligencia artificial para evitar que se formen monopolios de facto y se socaven las libertades individuales".

Me sorprende sobremanera el presente principio cuando desde hace varios años, antes incluso de la llegada de la inteligencia artificial a nuestras vidas, éramos conscientes de que: "Una característica de las plataformas digitales que más éxito tienen es la rapidez con que consiguen una cuota de mercado considera-

[35] B. Mulet Trobat, "Apuntes para la interpretación sociológica de las consecuencias socioculturales y educativas de la Cibercultura en la era de la globalización", in *Estudios sobre las culturas contemporáneas*, 28 (2023), n. 56, pp.169-206.

[36] Hablamos de "una información filtrada que llega a una persona en internet. Diferentes servicios, ofrecen resultados personalizados para los usuarios. Esto puede aislar a las personas porque ya no pueden acceder a la información común".

ble, dando lugar a que dichas plataformas tiendan a convertirse en monopolios"[37].

En este caso, suscribo totalmente la presente afirmación, aunque yo diría aún más. Debemos empezar a ser conscientes de que el gran monopolio no es ni más ni menos que la propia sociedad digital.

3.8. Principio de prudencia

De una forma resumida el presente principio establece que "toda persona que esté involucrada en el desarrollo de la inteligencia artificial debe actuar con prudencia y anticipar, en la medida de lo posible, los efectos adversos del uso de los sistemas de inteligencia artificial y tomar las medidas adecuadas para evitarlos.

Sin duda alguna, este principio me parece de lo más útil, oportuno y, además, coherente. Considero que es un principio que sí debe ser esencial en el desarrollo de los agentes artificiales. Sobre todo, en dos cuestiones importantes que interesa reseñar.

Por un lado, su primer apartado establece que

> es necesario desarrollar mecanismos que tengan en cuenta las posibilidades de uso doble (beneficioso o perjudicial) de la investigación en inteligencia artificial y el desarrollo de los sistemas de inteligencia artificial (ya sean públicos o privados) a fin de limitar los usos perjudiciales.

No puedo estar más de acuerdo, con independencia de que el desarrollo sea público o privado debemos desarrollar protocolos que nos permitan observar aquello que quizá no queramos o no sepamos ver. Cuando se desarrolla una tecnología, sin duda alguna, siempre se ve el beneficio, pero como regla general, nunca observamos el peligro que su uso puede comportar. Y sino, que se lo pregunten a Alfred Nobel que inventó la dinamita para agilizar los trabajos en la construcción… Quizá si hubiese visto el peligro potencial de su invención, nunca se hubiese animado a hacerla pública…

[37] UNCTAD, *Informe sobre la economía digital 2019*, Servicio de publicaciones de las Naciones Unidas, New York, 2019.

Por otro lado, otra cuestión que me parece de lo más oportuno y coherente reside en el tercer punto de este principio:

> Antes de lanzarlos al mercado, sin importar si son gratuitos o pagos, los sistemas de inteligencia artificial deben cumplir requisitos rigurosos de fiabilidad, seguridad e integridad, y se los debe someter a pruebas que no pongan en peligro la vida de las personas, perjudiquen su calidad de vida o afecten de forma negativa su reputación o integridad psicológica.

No obstante, y a pesar de que me parece algo muy oportuno, considero que la redacción podría mejorarse. Yo no hablaría de su lanzamiento al mercado, sino que directamente establecería el lanzamiento a la propia sociedad. Hablar de mercado, bajo mi punto de vista, podría dar a entender que sólo hablamos de aquellos agentes artificiales elaborados por el sector privado, dejando fuera de la ecuación a los agentes artificiales elaborados por el sector público… Creo que eso es un error. Cuando una entidad pública proceda a lanzar a la sociedad un agente artificial debe garantizar todas las cuestiones que acabamos de analizar: fiabilidad, integridad, seguridad, que no pongan en peligro a las personas o perjudiquen su vida.

3.9. Principio de responsabilidad

El presente principio establece, de una forma resumida que "el desarrollo y el uso de los sistemas de inteligencia artificial no deben contribuir a reducir la responsabilidad de los seres humanos a la hora de tomar decisiones".

La idea clave que reside sobre el presente principio es que única y exclusivamente el ser humano es el responsable de las recomendaciones o acciones de un agente artificial.

Si bien entiendo el presente principio dado que, de lo que se trata es de que no existan lagunas desde el punto de vista de la responsabilidad, ya sea de carácter civil, penal, administrativa, etc.

De hecho, existen en este principio una serie de *líneas rojas*[38], fruto del potencial que tienen las inteligencias artificiales. Un agente artificial, por sí sólo nunca puede tomar decisiones que:

[38] Es decir, la decisión final debe ser tomada siempre por el ser humano.

(i) Tengan que ver con la decisión de matar.

(ii) Impliquen la comisión de un crimen o delito. Es decir, sí que podrán tomar este tipo de decisiones, pero la responsabilidad de dichos actos va a recaer directamente en la persona o personas que autorizan a que el agente artificial lo cometa.

Aun así, el presente principio también establece que "cuando un sistema de inteligencia artificial provoca daños o perjuicios y se demuestra que dicho agente es responsable pero se utilizó según lo previsto, no es razonable culpar a las personas involucradas en su desarrollo o uso". Es decir, asistimos a una exoneración de responsabilidad que, a mi juicio, es de lo más coherente.

3.10. Principio de desarrollo sostenible

El presente principio establece que "el desarrollo y el uso de los sistemas de inteligencia artificial se deben llevar a cabo con el objetivo de garantizar una sostenibilidad fuerte del medio ambiente"

Quizá uno de los temas menos comentados cuando se habla de inteligencia artificial. Su desarrollo puede implicar un impacto ambiental importante. De hecho, su impacto ambiental "deriva principalmente de la cantidad ingente de energía requerida por los sistemas de IA para su eficaz desarrollo, entrenamiento y uso"[39].

En algún momento, la presente cuestión deberá ser debatida en foros internacionales ya que se deberán tomar decisiones en tanto en cuanto, a mayor desarrollo de los agentes artificiales, mayo impacto ambiental sufrirá nuestro planeta.

4. UNESCO: Recomendación sobre la ética de la inteligencia artificial

4.1. Introducción

A finales de 2021, la Conferencia General de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultu-

[39] D.E. Araiz Huarte, "La inteligencia artificial como agente contaminante: Concepto jurídico, impacto ambiental y futura regulación", in *Actualidad Jurídica Ambiental*, (2023), n. 130.

ra (UNESCO) en su 41ª reunión adoptó la Recomendación sobre la ética de la inteligencia artificial.

Dentro de su preámbulo, la UNESCO reconoce que "el uso de los agentes artificiales puede llegar a influir en el pensamiento, las interacciones, así como en la adopción de decisiones, por parte de los seres humanos".

Asimismo, la UNESCO, es consciente de la preocupación que existe, a nivel mundial, en relación con las preocupaciones éticas, en particular con cuestiones relativas a la "discriminación, desigualdad, brechas digitales, exclusión social, entre otras". Por supuesto, la presente preocupación lleva asociado la consideración del impacto que estas tecnologías pueden tener sobre "la dignidad humana, los derechos humanos, las libertades fundamentales, la igualdad de género, la democracia, los procesos sociales, económicos, políticos y culturales, las prácticas científicas de ingeniería, el bienestar animal y el medio ambiente y los ecosistemas".

Es por ello que la presente recomendación pretender ser una brújula que los Estados Miembros de la UNESCO, pueden aplicar, de forma voluntaria, a través de medidas, fundamentalmente, de carácter legislativo.

Tal y como se establece en el ámbito de aplicación de la presente Recomendación:

> Aborda la ética de la inteligencia artificial como una reflexión normativa sistemática, basada en un marco integral, global, multicultural y evolutivo de valores, principios y acciones interdependientes, que puede guiar a las sociedades a la hora de afrontar de manera responsable los efectos conocidos y desconocidos de las tecnologías de la inteligencia artificial en los seres humanos, las sociedades y el medio ambiente y los ecosistemas, y les ofrece una base para aceptar o rechazar las tecnologías de la inteligencia artificial. Considera la ética como una base dinámica para la evaluación y la orientación normativas de las tecnologías de la inteligencia artificial, tomando como referencia la dignidad humana, el bienestar y la prevención de daños apoyándose en la ética de la ciencia y la tecnología.

La Recomendación de la UNESCO se basa en una serie de principios y aunque la propia organización reconoce que todos

y cada uno de ellos son deseables y pertinentes, existe la posibilidad de que, en algunas ocasiones, los mismos puedan entrar en contradicción y, por tanto, los Estados Miembros deberán ser capaces de gestionar dichas tensiones basándose en el principio de proporcionalidad, así como en el respeto a los derechos humanos y a las libertades fundamentales. No obstante, y en todo caso "las tecnologías de la inteligencia artificial deben estar al servicio del bien de la humanidad". La UNESCO sin duda, aglutina en sus palabras los deseos de la humanidad. No obstante, me pregunto si todos y cada uno de sus Estados Miembros desean y quieren hacer realidad dichos deseos, máxime cuando países como Sudán, Rusia y Ucrania, Países Miembros, no lo olvidemos, han cogido las armas para dirimir sus disputas.

En todo caso, el texto ofrece una serie de valores y principios que merece la pena enumerar. Los valores de la presente Recomendación son los siguientes:

- Respeto, protección y promoción de los derechos humanos, las libertades fundamentales y la dignidad humana.
- Prosperidad del medio ambiente y los ecosistemas.
- Garantía de la diversidad y la inclusión.
- Vida en sociedades pacíficas, justas e interconectadas.

En relación con los principios, son los siguientes:

- Proporcionalidad e inocuidad.
- Seguridad y protección.
- Equidad y no discriminación.
- Sostenibilidad.
- Derecho de la intimidad y protección de datos.
- Supervisión y decisión humanas.
- Transparencia y explicabilidad.
- Responsabilidad y rendición de cuentas.
- Sensibilización y educación.
- Gobernanza y colaboración adaptativas y de múltiples partes interesadas.

No obstante, quizá la parte más interesante de la presente Recomendación son los ámbitos de actuación que plantea. Vamos a

proceder a su enumeración para, de forma posterior, analizarlos detenidamente:

- Evaluación del impacto ético.
- Gobernanza y administración éticas.
- Política de datos.
- Desarrollo y cooperación internacional.
- Medio ambiente y ecosistemas.
- Género.
- Cultura.
- Educación e investigación.
- Comunicación e información.
- Economía y trabajo.
- Salud y bienestar social.

4.2. Evaluación del impacto ético

El presente ámbito de actuación me parece de sumo interés. En él se plantean varias cuestiones que considero que deben ser necesarias en cualquier regulación sobre inteligencia artificial:

- Evaluaciones de impacto ético.
- Mecanismos de diligencia y supervisión.
- Investigación de los efectos sociológicos y psicológicos de las recomendaciones basadas en la inteligencia artificial sobre los seres humanos en lo relativo a su autonomía de decisión.

Considero totalmente necesarios los presentes mecanismos. Ante una tecnología como la presente es recomendable detenerse a analizar los beneficios, así como los perjuicios y riesgos de los agentes artificiales. Y dicho análisis debe ser realizado de forma previa a su puesta en funcionamiento.

El análisis no sólo debe ser previo sino también durante el funcionamiento de las presentes tecnologías. Debemos implementar mecánicas capaces de "determinar, prevenir y atenuar los riesgos", en caso de que éstos se materialicen. Incluso, como plantea la Recomendación los agentes artificiales deben ser probados, incluso en escenarios reales, antes de su puesta en funcionamiento. Bajo mi punto de vista, esta es la única opción de poder saber, con seguridad, que los agentes artificiales no suponen ningún riesgo para el ser humano.

En este sentido, la Ley de Inteligencia Artificial, elaborado por la Unión Europea[40], contempla la presente cuestión es su artículo 27 que lleva por título "Evaluación del impacto sobre los derechos fundamentales de los sistemas de inteligencia artificial de alto riesgo". En este sentido, se alaba la intención del legislador europeo, aunque también es criticable que la presente evaluación únicamente se plantea para los sistemas de inteligencia artificial catalogados por la Unión Europea como de alto riesgo. No obstante, la Unión Europea es pionera en este sentido y cualquier norma, tanto nacional como internacional debiera incluir regulación en el presente sentido.

4.3. Gobernanza y administración éticas

Considero que el presente ámbito de aplicación tiene una derivada jurídica de interés, en tanto en cuanto, es un ámbito que se puede materializar, siempre y cuando haya interés y voluntad política.

A grandes rasgos, el presente ámbito se centra en las siguientes cuestiones:

- Velar por que los mecanismos de gobernanza de la inteligencia artificial sean inclusivos, transparentes, multidisciplinarios y multilaterales.
- Velar por que se investiguen y reparen los daños causados mediante sistemas de inteligencia artificial, estableciendo mecanismos de aplicación estrictos y medidas correctivas.
- Elaborar estrategias, tanto de carácter nacional como regional sobre inteligencia artificial.
- Elaborar mecanismos de certificación, así como sistemas de reconocimiento mutuo de dichas certificaciones.

[40] Parlamento y Consejo de la Unión Europea (2024), Reglamento (UE) 2024/1689 del Parlamento Europeo y del consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.o 300/2008, (UE) n.o 167/2013, (UE) n.o 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial).

- Realizar autoevaluaciones transparentes de los sistemas de inteligencia artificial, tanto de los ya existentes, como de aquellos que se puedan ir proponiendo.
- Diseñar e implementar una figura de responsable independiente de la ética de la inteligencia artificial, así como crear una red de responsables independientes en los planos nacional, regional e internacional.
- Elaborar legislaciones nacionales respetuosas con los derechos humanos y las libertades fundamentales.
- Vigilar el impacto social y económico que puede tener el uso de inteligencia artificial en ámbitos de carácter sensible tales como los derechos humanos, la aplicación de la ley, la asistencia social, el empleo o la salud, por citar algunos ejemplos.
- Elaborar normas internacionales que describan los niveles de seguridad y transparencia mensurables y comprobables, de modo que se puedan evaluar objetivamente los sistemas y determinar los niveles de cumplimiento.
- Establecer marcos reguladores para alcanzar la rendición de cuentas y la responsabilidad por el contenido y los resultados de los sistemas de inteligencia artificial en las diferentes etapas de su ciclo de vida.
- Establecer requisitos claros de transparencia y explicabilidad de los sistemas de inteligencia artificial para ayudar a garantizar la fiabilidad de dichos sistemas durante todo su ciclo de vida.

Si bien existen algunas cuestiones que considero que pueden ser difíciles de materializar, como puede ser establecer un consenso internacional sobre qué podemos considerar inclusivo y transparente, por ejemplo, sí que coincido en la mayoría de los objetivos aquí analizados.

Se hace totalmente necesario elaborar una legislación homogénea en materia de inteligencia artificial. Es evidente, que un primer paso, debe centrarse en la elaboración de normas nacionales o incluso regionales, como es el caso de la Ley de Inteligencia artificial que se está elaborando por parte de la Unión Europea. No obstante, el paso definitivo y totalmente necesario reside en la elaboración de una norma internacional, que debiera ser firmada y ratificada por todos los Estados, es decir, un texto

con aplicabilidad y, por tanto, con carácter sancionador y coercitivo a nivel mundial. En esta materia, no basta con que algunos estados ratifiquen, la ratificación de un texto sobre la presente materia debe ser global, ya que lo que está en juego es la salvaguarda de la propia humanidad como especie.

En relación a los mecanismos de certificación, por ejemplo, la Unión Europea ya está trabajando en su regulación. La Ley de Inteligencia artificial dedica una sección completa a esta cuestión que lleva por título "Normas, evaluación de la conformidad, certificados y registro". Asimismo, la presente norma tiene todo un capítulo dedicado a la Gobernanza donde se establece la creación del Consejo Europeo de Inteligencia Artificial, así como la creación de un Grupo de expertos científicos independientes y la designación de autoridades nacionales competentes en la materia.

4.4. Política de datos

El presente ámbito de aplicación se centra en una de las cuestiones que más preocupa en relación con los agentes artificiales. Su impacto en nuestro derecho a la intimidad y a la protección de nuestros datos personales.

En este sentido, este ámbito de actuación se centra en las presentes cuestiones:

- Elaborar estrategias que garanticen la evaluación continua de la calidad de los datos de entrenamiento para los sistemas de inteligencia artificial.
- Establecer mecanismos de retroalimentación para aprender de los errores y compartir las mejores prácticas entre todos los actores de la inteligencia artificial.
- Establecer salvaguardias para proteger el derecho a la intimidad, así como a la protección de datos de carácter personal.
- Promover repositorios abiertos de datos y códigos fuente públicos o de financiación pública y fideicomisos de datos, a fin de apoyar el intercambio seguro, equitativo, legal y ético de datos, entre otros.
- Promover y facilitar el uso de conjunto de datos sólidos y de calidad para el entrenamiento, el desarrollo y la utilización

de los sistemas de inteligencia artificial y ejercer vigilancia al supervisar su recopilación y utilización.

- Alentar a las empresas del sector privado a que compartan con todas las partes interesadas, los datos que recopilan, en beneficio de la investigación, la innovación o el interés público.

Sin duda, a medida que avanzamos en los ámbitos de actuación propuestos por la UNESCO, nos damos cuenta de la necesidad de establecer un marco internacional global en la presente materia. Un claro ejemplo es la protección de datos de carácter personal. Si bien, la presente materia está empezando a tener recorrido en las legislaciones nacionales, es evidente que, a nivel mundial, la situación es muy dispar. Aún existen muchos países que o bien no disponen de normativa relativa a la protección de datos, o bien disponen de normativa sectorial sobre la materia.

Asimismo, nótese, que en relación a los datos abiertos (*open data*) pero aplicados al campo de la inteligencia artificial, la propia UNESCO establece que se debe alentar, es decir, animar o estimular a que el sector privado comparta información Sin duda, la UNESCO está cargada de buenas intenciones pero dudo mucho que el sector privado comparta información sino es a través de una normativa que implique la realización de dicho comportamiento.

4.5. Desarrollo y cooperación internacional

Una vez más, en el presente ámbito de actuación, me toca repetirme. No basta con las buenas intenciones en una materia como la presente. Se hace necesario una norma internacional mundial ratificada por todos los Estados y en donde se contemple, entre otras muchas cosas, cómo vamos a lograr el presente ámbito de actuación: el desarrollo y la cooperación internacional que se sustenta sobre las siguientes cuestiones:

- A nivel mundial se debe dar prioridad a la ética de la inteligencia artificial.
- Establecer plataformas de cooperación internacional.
- Procurar promover la colaboración internacional en materia de investigación e innovación en inteligencia artificial.

- Alentar la cooperación y colaboración internacionales en el ámbito de la inteligencia artificial para salvar las divisiones geotecnológicas.

4.6. Medio ambiente y ecosistemas

Otra cuestión de interés es el desarrollo sostenible y la protección del medio ambiente y los ecosistemas. Este tema, quizá no es el más tratado por la doctrina pero considero que todos los Estados, en su conjunto, debieran ir evaluando el impacto ambiental, tanto de corte directo como de corte indirecto que implica el desarrollo y puesta en funcionamiento de los agentes artificiales. Está claro que cuestiones como la huella de carbono, el consumo de energía y, por supuesto, el impacto ambiental debe ser analizado convenientemente. Un análisis exhaustivo de la presente cuestión incluso podría implicar que se desarrollen normas jurídicas que de alguna forma limiten el desarrollo de estos sistemas

4.7. Género

La igualdad de género es una cuestión que tiene un impacto transcendental en la humanidad y por la cual, a nivel nacional, regional e internacional se están dando grandes pasos que aún así, son insuficientes. Siendo conscientes de la importancia y la necesidad de reducir la brecha de género, la UNESCO busca los siguientes objetivos, en lo que a la inteligencia artificial se refiere:

- Lograr la igualdad de género y garantizar que no se conculcan los derechos humanos y las libertades fundamentales de las niñas y las mujeres, ni su seguridad e integridad, en ninguna etapa del ciclo de vida de los sistemas de inteligencia artificial.
- Velar por que las políticas digitales nacionales incluyan un plan de acción en materia de género destinado a apoyar a las niñas y las mujeres para que no queden excluidas de la economía digital impulsada por la Inteligencia artificial.
- Asegurarse de que estas tecnologías no exacerben las ya amplias brechas que existen entre los géneros en varios ámbitos del mundo analógico.

- Velar porque los estereotipos de género y los sesgos discriminatorios no se trasladen a los sistemas de inteligencia artificial.

- Alentar la iniciativa empresarial, la participación y el compromiso de las mujeres en todas las etapas del ciclo de vida de los sistemas de inteligencia artificial.

- Promover la diversidad de género en la investigación en IA en el mundo universitario y la industria.

No me cabe duda de que en esta materia hay mucho trabajo por hacer. Un trabajo que es totalmente necesario. En el siglo XXI la brecha de género debiera ser algo minoritario y en clara recesión. No obstante, permítame el lector que en la presente materia no sea tan optimista. En primer lugar, dada las claras diferencias culturales que existen en el mundo. Tal y como establece el observatorio de la islamofobia en los medios: "[…] es importante reconocer que la brecha de la desigualdad de género es más difícil de reducir en las sociedades donde la religión sigue siendo un referencial fundamental – como es el caso de la mayoría de los países arabo-musulmanes"[41].

La religión ciertamente puede ser una de las cuestiones que impidan reducir la brecha pero que no se me malinterprete, no es la única. Cuestiones de carácter político, ideológico o incluso social pueden ayudar a que la brecha siga estando presente. ONU Mujeres, sin entrar en la cuestión de la inteligencia artificial propiamente dicha, hace argumentaciones ciertamente desalentadoras: "La diferencia salarial de género permanece en el 23% a nivel mundial y, si no se toman medidas contundentes, pasarán otros 68 años antes de lograr la igualdad en esta esfera"[42].

Creo que el punto clave es la *contundencia*. Si queremos lograr avances en la reducción de la brecha de género hay que ser

[41] Observatorio de la islamofobia en los medios (2019), *La igualdad de género en el mundo musulmán: se necesita una visión reformista*. Recuperado de https://www.observatorioislamofobia.org/2019/09/30/la-igualdad-genero-mundo-musulman-se-necesita-una-vision-reformista/.

[42] ONU Mujeres (2018), *Hacer las promesas realidad: La igualdad de género en la agenda 2030 para el desarrollo sostenible*. Recuperado de https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/Library/Publications/2018/SDG-report-Gender-equality-in-the-2030-Agenda-for-Sustainable-Development-2018-es.pdf).

contundentes. En lo que a la inteligencia artificial y esta materia respecta, vuelvo a repetirme, necesitamos una regulación internacional de escala mundial.

4.8. Cultura

Un ámbito de actuación de "*creciente actualidad*" en tanto en cuanto, hoy en día, se están viendo grandes avances en los agentes artificiales enfocados al presente sector. En este sentido, la UNESCO establece las siguientes cuestiones:

- Incorporar agentes artificiales a la preservación, el enriquecimiento, la comprensión, la promoción, la gestión y la accesibilidad del patrimonio cultural, material, documental e inmaterial.
- Se debe examinar y abordar las repercusiones generadas en la cultura por los sistemas de inteligencia artificial.
- Se debe promover la educación en inteligencia artificial de los artistas y profesionales creativos.
- Se debe promover investigaciones que analicen la intersección entre la inteligencia artificial y la propiedad intelectual.
- Alentar a los museos, las galerías, las bibliotecas y los archivos de ámbito nacional a que utilicen sistemas de inteligencia artificial para realizar sus colecciones y enriquecer sus bibliotecas, bases de datos y bases de conocimientos, ofreciendo al mismo tiempo acceso a sus usuarios.

Más allá de los claros beneficios que pueden tener la inteligencia artificial en el desarrollo y progreso de la cultura sí que es cierto que necesitamos resolver algunas cuestiones esenciales. La más importante, en mi opinión es clarificar qué sucede con las creaciones de carácter intelectual desarrolladas por agentes artificiales. ¿Se pueden proteger obras generadas por agentes artificiales en base a las normativas de propiedad intelectual existentes en la actualidad? ¿Se hace necesario replantear la propiedad intelectual y su normativa? Considero que es la primera cuestión que debe ser resuelta.

4.9. Educación e investigación

En el presente ámbito de actuación la UNESCO plantea las siguientes cuestiones:

- Impartición para todo el mundo de conocimientos adecuados en materia de inteligencia artificial.

- Promover programas generales de sensibilización sobre los avances en inteligencia artificial.

- Alentar las iniciativas de investigación sobre la utilización responsable y ética de las tecnologías de la inteligencia artificial en la enseñanza, la formación de docentes y el aprendizaje electrónico.

- Velar porque las tecnologías de la inteligencia artificial empoderen a los educandos y los docentes y mejoren su experiencia, teniendo presente que los aspectos relacionales y sociales y el valor de las formas tradicionales de educación son fundamentales.

- Velar porque los sistemas de inteligencia artificial apoyen el proceso de aprendizaje sin reducir las capacidades cognitivas y sin recabar información sensible.

- Promover la inclusión digital en los programas de educación en materia de inteligencia artificial en todos los niveles.

- Elaborar planes de estudios sobre la ética de la inteligencia artificial para todos los niveles y promover la colaboración cruzada entre la enseñanza de competencias técnicas de inteligencia artificial y los aspectos humanísticos, éticos y sociales de educación en inteligencia artificial.

- Promover y apoyar investigaciones sobre la ética de la inteligencia artificial.

- Promover públicamente las mejores prácticas de los investigadores y las empresas que desarrollan la inteligencia artificial de forma ética y cooperar con ellos.

- Velar por que los investigadores en inteligencia artificial reciban formación en ética de la investigación y exigirles que tengan en cuenta consideraciones éticas en sus concepciones, productos y publicaciones.

- Alentar a las empresas a que faciliten el acceso de la comunidad científica a sus datos para la investigación.

- Velar por que cualquier evolución futura relacionada con las tecnologías de la inteligencia artificial se base en investigaciones científicas rigurosas e independientes.

- Promover la investigación interdisciplinaria en inteligencia artificial mediante la inclusión de disciplinas distintas de la ciencia, la tecnología, la ingeniería y las matemáticas, como pueden ser los estudios culturales, la educación, la ética, las relaciones internacionales, el derecho, la lingüística, la filosofía, las ciencias políticas, la sociología y la psicología.

No me cabe duda de que nos encontramos en un punto de inflexión. El desarrollo de los sistemas de inteligencia artificial puede generar una revolución sin precedentes en la historia del ser humano, mayor si cabe, que la generada con el desarrollo de las tecnologías de la información y de las comunicaciones. Por este motivo tanto la formación como la investigación me parecen dos cuestiones esenciales que deben ser potenciadas.

Por un lado, considero que toda persona debe tener unos conocimientos, al menos esenciales, sobre inteligencia artificial y, por tanto, en la educación reglada, incluso a niveles de primaria, se debería acercar al estudiante al conocimiento relacionado con la inteligencia artificial. Evidentemente, no sólo la educación reglada debe encargarse de hacer llegar el conocimiento. Por este motivo, los Estados creo que tienen no sólo el deber sino también la obligación de generar planes de sensibilización para poder llegar al mayor número de personas.

Por otro lado, aquellas personas que investigan, diseñan y desarrollan sistemas de inteligencia artificial debiera tener una formación multidisciplinar donde se aúnen tanto la ciencia como las humanidades. Los sistemas de inteligencia artificial, por su propia definición y características deben llevar *incorporados* una serie de valores y una ética para el desarrollo de sus actividades. Por tanto, sus creadores, deben ser los encargados de dar esos primeros valores éticos a los sistemas de inteligencia artificial y, por este motivo, su pensamiento debe ser completamente multidisciplinar.

4.10. Comunicación e información

Otra de las cuestiones sobre las que pueden impactar los sistemas de inteligencia artificial es en la libertad de expresión y el acceso a la información. Por este motivo, la UNESCO considera que sobre el presente ámbito de actuación es necesario:

- Utilizar los sistemas de inteligencia artificial para mejorar el acceso a la información y el conocimiento.

- Garantizar que los sistemas de inteligencia artificial respetan y promueven la libertad de expresión y el acceso a la información.

- Lograr que los Estados miembros inviertan en competencias digitales de alfabetización mediática e informacional y promoverlas.

- Que se creen entornos propicios para que los medios de comunicación tengan los derechos y recursos necesarios para informar eficazmente sobre las ventajas e inconvenientes de los sistemas de inteligencia artificial.

- Alentar a los medios de comunicación a que hagan un uso ético de los sistemas de inteligencia artificial en su trabajo.

Sin duda alguna, "las fake news o noticias falsas se han convertido en una constante dentro del Periodismo y en un verdadero problema a la hora de discernir la verisimilitud de una información"[43]. El avance de los sistemas de inteligencia artificial no hace sino agravar la presente cuestión.

4.11. Economía y trabajo

Bajo mi punto de vista, la presente es una de las cuestiones que más nos debe preocupar como sociedad. ¿Qué impacto va a tener el desarrollo de los agentes artificiales sobre el empleo? En este sentido, el presente ámbito de actuación se centra en las presentes cuestiones:

- Se debe evaluar y abordar el impacto de los sistemas de inteligencia artificial en los mercados de trabajo.

- Se debe garantizar una transición equitativa a los empleados en situación de riesgo.

- Se deben explorar programas de protección social para aquellos que no puedan reconvertirse.

[43] M. Alonso González, "Las *fake news* o noticias falsas se han convertido en una constante dentro del Periodismo y en un verdadero problema a la hora de discernir la verisimilitud de una información", in *Revista internacional de comunicación*, (2019), n. 45, pp. 29-52.

- Se deben adoptar medidas adecuadas para garantizar la competitividad de los mercados y la protección de los consumidores con la finalidad de evitar los abusos de posición dominante.

Considero que estos cuatro puntos son imprescindibles y deben ser abordados por todos los Estados. Debemos analizar cuidadosamente cómo se va a ver modificado el mercado de trabajo con la irrupción de los sistemas de inteligencia artificial.

No debemos olvidar que:

> En numerosos estudios relacionados con el potencial impacto de la automatización se encontró que casi la mitad de los empleos en Estados Unidos y el 60% de los trabajos de los países de la Asociación de Naciones del Sudeste Asiático (ASEAN) están en riesgo de ser automatizados[44].

La automatización supondrá la pérdida de muchos puestos de trabajo. Por tanto, la sociedad debe estar preparada para activar programas de reciclaje de los trabajadores, así como sistemas de protección social para aquellas personas que no logren un correcto acomodo en el mercado laboral. De hecho, la Unión Europea, en 2015, ya estableció la posibilidad "de que debería considerarse seriamente la posibilidad de introducir una renta básica universal"[45].

Considero que los Estados deben comenzar a trabajar sobre el presente asunto antes de que sea demasiado tarde. Es totalmente necesario anticiparse, en lo que a la presente cuestión se refiere.

4.12. Salud y bienestar social

El último ámbito de actuación planteado por la UNESCO tiene que ver con la salud y el bienestar social. En este sentido, me parece de interés matizar las siguientes cuestiones:

- Se deben emplear sistemas de inteligencia artificial para mejorar la salud humana y proteger el derecho a la vida.

[44] A.M. López Estupiñán, L. Peña Mesa, "Inteligencia Artificial: El futuro del empleo", in *Revista Lecciones Vitales*, 1 (2023), pp. 1-7.

[45] Parlamento Europeo (2015), *Recomendaciones destinadas a la Comisión sobre normas de Derecho Civil sobre robótica*. Recuperado de https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_ES.html.

- El desarrollo y despliegue de los sistemas de inteligencia artificial relacionados con la salud deben ser seguros, eficaces, eficientes y probados desde el punto de vista científico y médico. Asimismo, deben facilitar la innovación y el progreso médico con base empírica.

- Se debe prestar especial atención a la regulación de las soluciones de predicción, detección y tratamiento médicos en los agentes artificiales, con especial énfasis, entre otras cuestiones, de los sesgos, la privacidad y, sobre todo, el que la decisión final sobre el diagnóstico médico y el tratamiento correspondan siempre a seres humanos.

- Desarrollo de investigaciones sobre los efectos y la regulación de los posibles daños de los sistemas de inteligencia artificial para la salud mental.

- Elaborar directrices sobre las interacciones entre seres humanos y los sistemas de inteligencia artificial y sus repercusiones en las relaciones entre los seres humanos, con especial énfasis, en aquellos colectivos más vulnerables como pueden ser las personas de avanzada edad, las personas con discapacidad o los menores de edad.

- Las cuestiones éticas relativas a los sistemas basados en inteligencia artificial utilizados en las neurotecnologías y las interfaces cerebro-ordenador deberían tenerse en cuenta a fin de preservar la dignidad y la autonomía humanas.

- Velar por que los usuarios puedan determinar fácilmente si interactúan con un ser vivo o con un sistema de inteligencia artificial. Asimismo, deben ser capaces de rechazar eficazmente dicha interacción y solicitar la intervención humana.

- Alentar y promover la investigación colaborativa sobre los efectos de la interacción a largo plazo de las personas con los sistemas de inteligencia artificial.

Quizá una de las mayores preocupaciones, a nivel social tiene que ver con la posibilidad de que se use la inteligencia artificial como un sistema o mecanismo que vaya contra nuestra propia existencia. De ahí la preocupación y especial sensibilidad por el respeto a la salud y a la vida humana. Si bien, es algo que nos debe preocupar, considero que se pueden dar otros usos, que nos pueden pasar más desapercibidos y que puedan tener una mayor

repercusión para el ser humano. Por ejemplo, cuestiones relativas a cómo van a interactuar las inteligencias artificiales con los colectivos más vulnerables, ya que pueden ser mecanismos que pueden utilizarse para modificar comportamientos, actitudes o incluso pensamientos en los presentes colectivos.

Asimismo, también considero que es necesario poner el foco en cómo la interacción con los sistemas de inteligencia artificial nos puede afectar en nuestras relaciones con el resto de los seres humanos. Es posible que pueda darse un efecto de asilamiento social en personas que sólo tengan interés por interactuar con sistemas de inteligencia artificial. Si bien, "la literatura revela que las TIC pueden ser un medio para aliviar el sentimiento de soledad y la percepción de aislamiento social en adultos mayores durante los contextos de crisis"[46], es posible que este efecto pueda darse a la inversa en las interacciones con agentes artificiales, tanto en situaciones de normalidad como en situaciones de crisis.

Por último, también creo que las personas tenemos derecho a decidir. Decidir si deseamos entablar una interacción con un agente artificial o si, por el contrario, deseamos que nuestra interacción se realice con un ser vivo, es decir, con un ser humano.

5. Conclusiones

Nos encontramos inmersos en una revolución tecnológica de dimensiones titánicas y que ya está transformando nuestra forma de vida, nuestro empleo y cómo nos relacionamos. Sin duda uno de los avances más prometedores y que, al mismo tiempo más preocupación nos aporta es el desarrollo de los sistemas de inteligencia artificial.

Para el desarrollo de los sistemas de inteligencia artificial se hace necesario reconocer a la ética como una disciplina de relevancia en la presente materia. Es totalmente necesario, para el correcto desarrollo de la inteligencia artificial, la aplicación de una serie de consideraciones éticas y morales que son esenciales

[46] A. Urbina Carreño, "La relación entre el uso de las TIC y el sentimiento de soledad: una mirada desde los Determinantes Sociales de la Salud de las personas mayores en Chile", in *Revista española de comunicación en salud*, suplemento, 1 (2023), pp. 74-80.

si queremos garantizar un desarrollo responsable y respetuoso de la presente tecnología.

Como hemos tenido oportunidad de ver, definir a la presente tecnología no es sencillo. Es una cuestión compleja y en constante evolución, debido al rápido desarrollo de la materia y a los constantes cambios sobre la percepción de esta disciplina. Aunque no existe un consenso en torno a la definición sí que podemos observar algunas características de los sistemas de inteligencia artificial tales como la capacidad de autonomía, su aprendizaje, su adaptabilidad y su capacidad para la toma de decisiones. También hemos tenido oportunidad de ver que distintas instituciones y organizaciones han realizado importantes esfuerzos para tratar de llegar a una definición de inteligencia artificial e incluso hemos analizado enfoques doctrinales que llegan a considerar aspectos, dentro de los agentes artificiales, tales como el comportamiento inteligente y la conciencia.

A grandes rasgos podemos concluir que los sistemas de inteligencia artificial son sistemas capaces de comportarse de manera inteligente, tienen capacidad para aprender, adaptarse y tomar decisiones y los podemos dividir en dos categorías: agentes virtuales y agentes físicos.

No obstante, creo que debemos tener muy presente que debemos seguir avanzando en la definición de inteligencia artificial. Para ello es necesario seguir revisando y actualizando el concepto de inteligencia artificial a medida que avance la tecnología y vayamos profundizando en el entendimiento de la presente disciplina.

A lo largo del presente trabajo hemos analizado la Declaración de Montreal y hemos tratado de desarrollar una opinión crítica sobre los principios éticos en ella contenidos. Dichos principios tienen una clara misión: guiar el desarrollo y la implementación de la inteligencia artificial de manera que se beneficie a la humanidad. Hemos sido críticos en nuestras reflexiones. Muchos de los principios planteados por la presente Declaración adolecen de definiciones claras y concisas lo cual no beneficia a su aplicación práctica o a su implementación efectiva.

Asimismo, algunos de sus principios plantean desafíos en lo relativo a la prevención del uso de inteligencias artificiales para

incluir en las decisiones de las personas o en torno a que a las personas se les puede imponer un estilo de vida específico gracias a los agentes artificiales.

Surgen muchos interrogantes sobre cómo proteger la intimidad y los datos personales en un mundo cada vez más digitalizado y fiscalizado. Asimismo, debemos enfrentar claros desafíos en lo relativo a cómo van a interactuar los agentes artificiales con los seres humanos y sobre todo, si estamos dispuestos a que las relaciones humanas de calidad puedan ser sustituidas por relaciones con agentes artificiales en un futuro próximo.

Enfrentamos un desafío de importancia en relación a la equidad social dado que el uso y desarrollo de la inteligencia artificial se va a ver claramente condicionado por unas condiciones locales y culturales muy variadas a lo largo del mundo. Nuestras sociedades no tienen las mismas condiciones locales y culturales y, por tanto, esto afecta a una materia como la inteligencia artificial.

Los Estados deben armarse de mecanismos que aborden los posibles usos perjudiciales de la inteligencia artificial, así como de mecánicas para mitigar los riesgos en el uso de la inteligencia artificial y garantizar un uso y desarrollo responsables.

Por otro lado, la responsabilidad humana en la toma de decisiones relacionadas con la inteligencia artificial es fundamental. Por tanto, las personas debemos asumir la responsabilidad final por las acciones de los agentes artificiales de cara a evitar lagunas de responsabilidad y garantizar la rendición de cuentas.

También debemos destacar la importancia de considerar el impacto ambiental en el desarrollo y uso de la inteligencia artificial. Cuestiones como el consumo masivo de energía y el impacto ambiental de estas tecnologías no pueden dejarse en un segundo plano.

Por otro lado, hemos tenido oportunidad de analizar la Recomendación sobre la ética de la inteligencia artificial adoptada por la UNESCO. En el presente texto se reconoce la influencia potencial de la inteligencia artificial en el pensamiento, las interacciones y las decisiones humanas, así como la necesidad de abordar preocupaciones éticas, como la discriminación y la

exclusión social, y sus impactos en la dignidad humana, los derechos humanos y el medio ambiente.

La Recomendación establece una serie de valores y principios que deben guiar el desarrollo y uso de la inteligencia artificial, incluyendo el respeto a los derechos humanos, la promoción de la diversidad y la inclusión, y la garantía de la seguridad y la transparencia. Además, propone áreas de acción específicas, como la evaluación del impacto ético, la gobernanza y administración éticas, y la política de datos, entre otras.

Sin embargo, para que estos principios se conviertan en realidad, es necesario un compromiso global y una cooperación internacional efectiva.

En resumen, el desarrollo y la implementación de la inteligencia artificial plantean una serie de desafíos éticos, sociales y legales que requieren una respuesta integral y colaborativa a nivel internacional. Es necesario garantizar que la inteligencia artificial se utilice de manera responsable y ética para promover el bienestar humano y proteger los derechos fundamentales en todos los ámbitos de la sociedad.

Podemos concluir que los textos analizados en el presente trabajo establecen una base ética para el desarrollo de la inteligencia artificial aunque es evidente que aún quedan muchos aspectos por definir y abordar para garantizar que el desarrollo y uso de la inteligencia artificial sea responsable y en beneficio de la humanidad.

Todos los principios analizados en este trabajo requieren de una atención continua así como de una acción coordinada a nivel global si queremos garantizar un futuro sostenible y equitativo para la humanidad.

II.
NUOVI DIRITTI E RICADUTE PROCESSUALI

Wanda D'Avanzo

Intelligenza artificiale e sanità digitale. Il diritto *by design* e il ruolo del *legal design* per la protezione dei dati personali*

Abstract: **Artificial Intelligence and Digital Health. The Law by Design and the Role of Legal Design for Personal Data Protection.** Artificial intelligence represents the most debated topic in the philosophical and legal IT fields and its exponential development in multiple sectors generates new opportunities, but also great challenges, for the future of humanity. The convergence between artificial intelligence and healthcare, in particular, which begins to hypothesize the incorporation of technology into humans, requires in-depth reflection on the urgent need for a balance between technological innovation and the protection of people's fundamental rights, especially in relation to protection of data. In this regard, an essential role is played by the clarity and quality of the information provided to citizens on the processing of personal data concerning them and on the way in which health data is collected and analysed. Although this aspect still presents several critical issues, a way to encourage better communication on patient health could be found in the so-called legal design, understood as an innovative approach to law, aimed at making legal documents easily understandable for all users. The article aims to analyze the regulatory provisions currently in force contained in the GDPR as well as the new communication techniques, such as the so-called legal design, to be applied to particular sectors such as healthcare, to investigate their potential in the face of the new possibilities offered by artificial intelligence systems.

Keywords: Artificial Intelligence, Neuroscience, GDPR, Law by Design, Legal Design.

Sommario: 1. Introduzione – 2. Intelligenza artificiale e neuroscienze – 3. La protezione dei neurodiritti nel GDPR. Il diritto *by design* – 4. La trasparenza informativa in sanità. Riprogettare la comunicazione attraverso il *legal design* – 5. I *legal design patterns* come nuovo approccio multidisciplinare del diritto – 6. Conclusioni.

1. Introduzione

L'intelligenza artificiale rappresenta il tema più dibattuto in ambito filosofico ed informatico-giuridico e il suo sviluppo

* Questo lavoro è stato parzialmente sostenuto dal progetto FAIR – Future AI Research (PE00000013), nell'ambito del programma MUR del

esponenziale in molteplici settori genera nuove opportunità, ma anche grandi sfide, per il futuro dell'umanità. Questa specifica branca dell'informatica sfrutta il potenziale di calcolo degli algoritmi matematici per dotare i calcolatori elettronici di abilità proprie dell'intelligenza umana.

Nata intorno ai primi anni '60 del Novecento, l'intelligenza artificiale trova oggi, nella enorme disponibilità di dati favorita da Internet, una nuova linfa che ne permette una sempre maggiore implementazione e il perfezionamento di modelli che stanno aprendo la via a nuove forme di interazione tra l'uomo e le macchine.

La convergenza tra questa tecnologia innovativa e la sanità, in specie, sta muovendo i primi passi verso una svolta significativa che sembra voler spingersi verso una incorporazione della tecnologia nell'uomo: un esempio tra tutti è rappresentato dalle recentissime sperimentazioni della società *Neuralink*, che puntano, attraverso l'impianto di microchip neurali nel cervello umano, ad offrire una cura per le patologie neurologiche[1].

A tal proposito, occorre considerare che, già da qualche anno, il mondo dell'informatica applicato alla medicina è proiettato verso la creazione di una nuova generazione di strumenti in grado di raccogliere ed elaborare in tempo reale dati e informazioni, rendendoli di uso comune e diffuso, in grado di integrarsi e di interagire costantemente con il corpo di chi li utilizza.

Questa vasta area di sperimentazione prende il nome di *Internet of Bodies*, o Internet dei corpi, espressione coniata nel 2016 per descrivere l'ampia gamma di dispositivi connessi ad Internet che monitorano il corpo umano, raccolgono dati fisiologici, biometrici o comportamentali, e scambiano informazioni attraverso una rete *wireless*. In sostanza, l'Internet dei corpi si avvale di una serie di strumenti tecnologici che collegano l'essere umano a dei dispositivi artificiali connessi alla rete. Essi monitorano il corpo umano, raccolgono parametri sanitari e altre informazioni personali e trasmettono tali dati grazie alle tecnologie di connettività[2].

PNRR finanziato dal NextGenerationEU.

[1] A. Aghihotri, S. Bhattacharya, *Neuralink. Invasive neurotechnology for human welfare*, Sage publications, Los Angeles, 2023.

[2] A.M. Matwyshyn, "The internet of bodies", in *Wm & Mary Law Review*, 61 (2019), p. 77. Sul punto si veda anche C. Amato, "Internet of Bodies:

Del tutto simile all'Internet delle cose, l'Internet dei corpi opera attraverso un hardware che elabora i dati raccolti da applicazioni comuni presenti ormai su qualsiasi smartphone e, attraverso la rete, li scambia con un *hub* centrale a intervalli prestabiliti o in tempo reale, permettendone al contempo l'archiviazione, l'analisi e la visualizzazione[3].

Gli strumenti di cui si avvale l'Internet dei corpi per funzionare possono essere di due tipologie: non invasivi ed indossabili, oppure invasivi, cioè iniettabili, ingeribili o impiantabili[4].

I dispositivi indossabili hanno trovato fin da subito largo impiego nel settore biomedicale e sono utilizzati per il monitoraggio continuo a distanza delle condizioni fisiche di pazienti a rischio o affetti da particolari patologie[5]. E a seguito del crescente interesse del mercato nei confronti di supporti mobili sempre più versatili, grazie all'abbassamento dei costi, sono stati perfezionati anche dispositivi con una vocazione più spiccatamente commerciale, utilizzabili dagli utenti nella normale vita quotidiana, nel tempo libero oppure come supporto ad attività pratiche di vario genere.

I *fitness tracker*, ad esempio, sono tecnologie esterne al corpo che raccolgono informazioni che vengono analizzate da software proprietari di aziende private. I braccialetti o gli orologi di uso comune connessi allo smartphone possono tenere traccia dei movimenti del corpo, della frequenza cardiaca, degli schemi di sonno e altri dati fisici utili per il monitoraggio.

Digital content directive, and beyond", in *Jipitec*, 12 (2021), pp. 181 ss.; G. Boddington, "The Internet of Bodies – alive, connected and collective: the virtual physical future of our bodies and our senses", in *AI & society*, 38 (2021), n. 5, pp. 1-17.

[3] S. Baker, W. Xiang, I. Atkinson, "Internet of things for smart healthcare. Technologies, challenges, and opportunities", in *IEEE Access*, (2017), n. 5, pp. 26521-26544; J. Calvillo-Arbizu, I. Román-Martínez, J. Reina-Tosina, "Internet of things in health: Requirements, issues and gaps", in *Computer Methods and Programs in Biomedicine*, 208 (2021), n. 4, p. 106231.

[4] Esistono anche una serie di applicazioni del c.d. *Internet of Medical Things*, che creano delle soluzioni ospedaliere intelligenti, come ad esempio i letti connessi con funzionalità di monitoraggio dei segni vitali.

[5] E. Germani, L. Ferola, "Il wearable computing e gli orizzonti futuri della privacy", in *Diritto dell'Informazione e dell'Informatica*, (2014), n. 1, p. 76.

Accanto ai dispositivi indossabili, gli strumenti dell'Internet dei corpi comprendono tecnologie più invasive che, al momento, sono solo in sperimentazione o in via di prima applicazione.

Per citare solo qualche esempio si possono menzionare i pancreas artificiali che potrebbero automatizzare il dosaggio dell'insulina per i soggetti affetti da diabete; i nuovi *pacemaker* cardiaci e defibrillatori impiantabili potrebbero consentire il controllo continuo ed in tempo reale delle fluttuazioni cardiache di un paziente, regolando la frequenza dei battiti.

Una ulteriore tecnologia, approvata già dal 2017 dalla *Food and Drug Administration* americana per la diffusione su larga scala per particolari categorie di malati, è la pillola digitale, che contiene dei sensori che registrano l'assunzione del farmaco. Gli operatori sanitari e i medici accedono alle informazioni trasmesse dalla pillola digitale attraverso un portale web[6].

Da ultimo, le società di software di intelligenza artificiale stanno sviluppando dei sistemi in grado di rilevare e raccogliere dati sulle emozioni umane analizzando le espressioni facciali, le intonazioni della voce e altri segnali audiovisivi, aprendo la strada ad una nuova branca dell'Internet dei corpi che prende il nome di *Internet of Behaviours* o Internet dei comportamenti[7].

Ebbene, dal punto di vista giuridico, l'introduzione di strumenti tecnologici intelligenti che colleghino l'uomo alle macchine in un flusso continuo di comunicazione attraverso la rete Internet richiede una riflessione approfondita sulla urgente necessità di bilanciamento tra l'innovazione tecnologica e la salvaguardia dei diritti fondamentali dell'uomo, con specifico riguardo alla *privacy* e alla integrità dei dati personali nel contesto del quadro normativo attuale.

[6] E. Fleisch, C. Franz, A. Herrmann, *The digital pill. What everyone should know about the future of our healthcare system*, Emerald Publishing, Bingley, 2021, pp. 25 ss.; ed anche, C. Rijcken, "Reinforests of wearable and insideables", in Id. (ed.), *Pharmaceutical care in digital devolution*, Elsevier, London, 2021, p. 110.

[7] L.A. Marsch, S.E. Lord, J. Dallery, *Behavioral healthcare and technology*, Oxford University Press, Oxford, 2015, p. 83. Il tema dell'*Internet of Bodies* è stato trattato nel mio "L'evoluzione della e-Health. Aspetti bioetici dell'Internet of Bodies", in *Ordines*, (2023), n. 2, pp. 179-192.

La valutazione dell'implementazione delle applicazioni dell'intelligenza artificiale in un campo come quello delle neuroscienze non può prescindere, infatti, da un'attenta valutazione legale ed etica che implica un giudizio importante con riguardo sia al rispetto della vita e dell'integrità fisica e psichica umana sia alle minacce alla libertà e all'autodeterminazione di ogni individuo[8].

Il tema trattato di per sé non assume un carattere di particolare originalità, ma oggi necessita di essere contestualizzato nel momento storico che stiamo attraversando, in cui il processo di digitalizzazione, specie in ambito pubblico, ha subito una improvvisa accelerazione non supportata da una adeguata presa di coscienza del cambiamento da parte delle persone.

Lo ricordiamo, solo a titolo di esempio, anche se da un punto di vista etico dovrebbe essere superfluo: durante la pandemia da Covid-19 le nuove tecnologie, unite al caos interpretativo delle leggi in materia, in Italia, hanno – a ragione o meno – tolto il lavoro ed escluso dalla vita sociale persone che non erano scansionabili con un codice a barre (o *Qr-Code* in linguaggio tecnologicamente avanzato).

La pericolosità di tali tecnologie non è, dunque, meramente ipotetica, ma si colloca in una dimensione reale e attuale, dove la potenzialità di accesso diretto ai processi fisici o a quelli comportamentali e cognitivi dell'individuo apre scenari di controllo e manipolazione, inimmaginabili e ancora del tutto inesplorati dal diritto.

A fronte di tale necessità l'articolo si propone di analizzare le disposizioni normative in tema di trasparenza attualmente in vigore contenute nel GDPR nonché le nuove tecniche di comunicazione più all'avanguardia, come il c.d. *legal design*, da applicarsi a settori particolari come la sanità, per indagarne le

[8] Il tema solleva grandi preoccupazioni non solo in relazione al campo specifico delle neuroscienze ma a tutte le tecnologie della *e-Health* e dell'*Internet of Bodies* che comprende tutta una serie di dispositivi di uso comune oggi che monitorano il corpo umano, raccolgono e scambiano dati grazie alle tecnologie di connettività. In tal senso, C.C. Gould, "How democracy can inform consent: Cases of the Internet and Bioethics", in *Journal of Applied Philosophy*, (2019), n. 2, pp. 173-191.

potenzialità di fronte alle nuove possibilità offerte dall'avanzare delle tecnologie.

E ciò muovendo da un assunto fondamentale: ad oggi il GDPR in Italia é ancora sostanzialmente inapplicato o applicato nel modo sbagliato; la gestione dei dati avviene in modo discutibile e poco chiaro. Solo a titolo di esempio può dirsi che esistono ancora informative *privacy* che richiamano il d.lgs. 196 del 2003; e che ogni anno aumentano in modo esponenziale i furti di dati detenuti, in specie, dalle pubbliche amministrazioni.

Il rischio di questo pressapochismo è che le nuove tecnologie rendano ancora più fumose le poche tutele concesse ai cittadini, acquisendo dati in modo sempre più invasivo, addirittura dall'interno delle persone, senza che gli interessati ne abbiano alcuna cognizione. E che, quindi, bisogna (almeno tentare di) correre ai ripari, in modo fattivo non teorico, prima che sia troppo tardi, anche con strumenti meno formali e di più immediata e semplice comprensione come il *legal design*, che diffondano conoscenza su questi temi.

2. Intelligenza artificiale e neuroscienze

Come abbiamo appena avuto modo di osservare, gli studi tecnologici di intelligenza artificiale hanno fatto progressi esponenziali negli ultimi vent'anni. I computer sono, oggi, in grado di svolgere compiti sempre più complessi ed eccellono in tutta una serie di attività che prima erano impensabili per le macchine.

L'intelligenza artificiale occupa, pertanto, una parte di grande rilievo nella ricerca informatica e racchiude molte branche di studi sui sistemi intelligenti, ognuno dedicato a un particolare aspetto del comportamento umano. I programmi di IA usano in genere lo stesso linguaggio dei sistemi convenzionali, ma con una logica diversa, finalizzata a progettare macchine che siano in grado di pensare come gli esseri umani.

Il campo dell'intelligenza artificiale riguarda, quindi, quei sistemi programmati in modo tale da tenere dei comportamenti che coinvolgono, a loro volta, capacità quali il ragionamento, l'apprendimento, la comunicazione e l'azione in ambienti complessi.

Le macchine intelligenti di ultima generazione sono alimentate dai dati, processati da pochi algoritmi essenziali di apprendimento e il linguaggio che usano non è più la logica, ma la statistica: combinando le informazioni tra loro, le macchine riescono ad effettuare delle valutazioni accurate.

Questo ecosistema computazionale, fondato su una stretta interazione tra dati, oggetti (*IoT*) e algoritmi complessi, può trovare impiego in qualsiasi settore sociale e molte di queste applicazioni hanno già fatto ingresso o sono in fase di sperimentazione[9].

Nella sua evoluzione più recente, l'intelligenza artificiale trova sempre maggiore diffusione in ambito biomedico e il ricorso agli algoritmi informatici si pone come un potente metodo di risoluzione di problemi che permette spesso di raggiungere risultati migliori rispetto agli approcci tradizionali. In questo ambito, uno dei settori che ha tratto diversi vantaggi dallo sviluppo tecnologico è quello delle neuroscienze, ossia quel complesso di discipline che hanno ad oggetto lo studio strutturale e funzionale del sistema nervoso, centrale e periferico, sotto l'aspetto normale e patologico, con riferimento particolare al cervello.

La disciplina delle neuroscienze nasce nel 1962, quando è il neurofisiologo americano Francis Schmitt fondò il *Neuroscience Research Program*[10]. Questa disciplina comprende vari ambiti applicativi: le neuroscienze cognitive, ad esempio, che studiano il substrato biologico, neurologico e cerebrale posto alla base

[9] Nella storia della loro evoluzione, i sistemi di intelligenza artificiale si sono fatti strada muovendo dall'idea che la descrizione dei processi mentali sia indipendente dal corpo. Alan Turing ha ipotizzato che la struttura logica del cervello potesse essere rappresentata ed incorporata in qualche altra macchina fisica, aprendo così la strada alla creazione di macchine con risorse simili a quelle dell'uomo; sul punto, si vedano A.C. Amato Mangiameli, "Intelligenza artificiale, big data e nuovi diritti", in *Rivista italiana di informatica e diritto*, (2022), n. 1, pp. 94 ss.; A.M. Turing, *Intelligenza meccanica*, trad. it., Bollati Boringhieri, Torino, 1994, p. 103. In tema si veda anche M.N. Campagnoli, M. Farina, *Tec-no-identità? Percorsi, provocazioni e istanze delle nuove s/oggettività*, Key editore, Milano, 2022, pp. 42 ss.; A.C. Amato Mangiameli, *Natur@. Dimensioni della Biogiuridica*, Giappichelli, Torino, 2021, pp. 152 ss.; A.C. Amato Mangiameli, *Corpi docili. Corpi gloriosi*, Giappichelli, Torino, 2007, pp. 89 ss.

[10] F.O. Schmitt, F.G. Worden (eds.), *The Neurosciences: Third Study Program*, Mit Press, Boston, 1974.

delle funzioni cognitive; le neuroscienze affettive che ricercano i meccanismi neurali di funzionamento e regolazione dei processi emotivi; le neuroscienze computazionali, che analizzano i meccanismi di astrazione del cervello.

Gli ambiti di applicazione dell'intelligenza artificiale all'interno delle neuroscienze sono molteplici ed eterogenei. I nuovi algoritmi garantiscono un minore margine di errore rispetto all'attività di studio umana, con una notevole riduzione de tempi di lavoro, per cui l'intelligenza artificiale viene utilizzata per gli scopi più diversi. L'innovazione tecnologica fornisce, infatti, strumenti sempre più sofisticati, funzionali allo studio di diversi aspetti del sistema nervoso. Ad esempio, le neuroscienze cognitive si avvalgono delle tecniche di *neuroimaging* grazie alle quali è possibile studiare in tempo reale, senza alcuna invasività e senza alcuna interferenza, l'attivazione delle aree cerebrali deputate allo svolgimento di un compito correlato alla funzione cognitiva d'interesse.

Oggi, poi, attraverso le c.d. *Brain Computer Interfaces* è possibile misurare l'attività neurale per controllare intenzioni, sensazioni, comportamenti e, dunque, gestire malattie o disturbi neurologici di interesse. Le interfacce neurali basate sul *machine learning* permettono già grazie alla elaborazione dei dati medici, di usare i segnali cerebrali per controllare protesi o altri oggetti esterni[11].

Fra le varie sperimentazioni terapeutiche in tal senso, diversi studi sono stati fatti sul trattamento dell'epilessia attraverso interfacce neurali che permettono un costante monitoraggio intracranico. L'analisi dell'attività neurale tramite *machine learning* può essere utile anche per prevedere l'andamento dell'attività neurale futura a partire da quella passata[12]. I modelli preditti-

[11] Il potenziamento cognitivo favorito dalle applicazioni tecnologiche e dalle scoperte nell'ambito delle neuroscienze è trattato da L. Palazzani, "Il potenziamento cognitivo e morale: riflessioni bioetiche", in *Forum*, (2020), n. 6, p. 8. L'ambito biomedico delle applicazioni ICT è anche in A. Montanari, "Questioni di tecno etica in intelligenza artificiale, robotica e bionica", in P. Moro (a cura di), *Etica, informatica, diritto*, FrancoAngeli, Milano, 2008, pp. 33-50.

[12] P. Patrikelis, G. Lucci, S. Gatzonis, *La neuropsicologia dell'epilessia. Aspetti neuro cognitivi e comportamentali della malattia sacra*,

vi utilizzati nel rilevamento precoce delle crisi hanno mostrato un'accuratezza di analisi vicina al 100%; per cui il loro utilizzo potrebbe portare presto allo sviluppo di nuovi trattamenti in grado di agire tempestivamente per controllare e mitigare la malattia, riducendo la durata delle crisi.

Gli algoritmi di intelligenza artificiale, oltre a permettere di prevedere intenzioni o comportamenti a partire dall'attività neurale, sono utili anche per prevedere l'attività neurale sulla base di stimoli esterni. Nell'ambito delle neuroscienze sensoriali, per esempio, una grande sfida è quella di costruire modelli di codifica neurale che riescano a prevedere l'attività neurale di un'area sensoriale in risposta a uno stimolo d'interesse[13]. Questo tipo di studi sono molto utili per migliorare le prestazioni degli arti bionici permettendo loro di eguagliare le capacità di quelli naturali.

In maniera analoga e più generale, le tecniche di *machine learning* possono aiutare le neuroprotesi, ovvero dispositivi artificiali in grado di sostituire o migliorare specifiche funzioni del sistema nervoso, a compiere in maniera più efficiente il loro lavoro.

Le *brain computer interfaces* e gli studi sulle neuroprotesi intelligenti sono alla base della più recente ed avanzata sperimentazione di impianti di microchip cerebrali che si pongono l'obiettivo di creare un collegamento diretto fra la mente umana e le macchine, in modo da captare e registrare l'attività elettrica del cervello e convertire i segnali in comandi che permettano di controllare dei dispositivi esterni. Ciò renderà possibile, quindi, utilizzare i segnali cerebrali per attivare le macchine.

Dal punto di vista tecnologico, il sistema del microchip, in grado di connettere l'uomo alla rete, opera tramite la c.d. RFID, acronimo dell'espressione "*Radio Frequency Identification*", e assolve una funzione di identificazione di beni, individui o animali per effetto di una trasmissione di segnali a radiofrequenza. L'identificazione comporta l'attribuzione di un'identità univoca finalizzata alla raccolta di informazioni riguardanti gli interes-

FrancoAngeli, Milano, 2018.

[13] D.J. MachInnis, L.L. Price, "The Role of Imagery in Information Processing: Review and Extensions", *in Journal of Consumer Research*, 13 (1987), n. 4, pp. 473-491.

sati mediante l'ausilio di apparati radio, associati agli stessi. Le informazioni vengono assunte tramite operazioni di ricerca, selezione, localizzazione spaziale, tracciamento[14].

La tecnologia RFID nasce fondamentalmente per l'identificazione e il monitoraggio delle merci, nei vari comparti delle catene di produzione aziendale. Ma negli ultimi anni ha trovato ampia diffusione per la grande varietà di applicazioni cui si presta.

A differenza dei più comuni codici a barre, le etichette RFID possono gestire contemporaneamente diverse identità, memorizzano molte più informazioni e sono in grado di misurare anche fattori esterni all'elemento monitorato.

In sostanza, le etichette RFID sono microchip dotati di un'antenna *wireless* che comunicano a distanza con i lettori e che acquisiscono i dati e li trasmettono ad un server in modo che siano processati e analizzati. Il sistema RFID non prevede nessun impiego di batterie o fonti fisiche di alimentazione per funzionare[15].

Questa tecnologia si presta alla creazione di applicazioni orientate all'identificazione delle persone, al monitoraggio, mediante sensori appositi, dei parametri vitali ed, in generale, viene presentata come uno strumento in grado di migliorare la qualità della vita, specie in settori come quello sanitario, in cui sono necessarie attività di controllo costante.

Essa ha raggiunto oggi un livello di maturità tale da poter essere utilizzata per l'impianto di etichette intelligenti biocompatibili all'interno del corpo che porteranno, in futuro, allo sviluppo di applicazioni sempre più avanzate[16].

Il 30 gennaio 2024 la società statunitense *Neuralink* ha annunciato di aver impiantato il primo prototipo di *brain computer interface* tramite microchip in un cervello umano. Si tratta della

[14] R. Marani, A.G. Perri, "Principali applicazioni biomedicali della tecnologia RFID", in *La Comunicazione. Note, Recensioni e Notizie*, 63 (2020), p. 2 ss.

[15] L. Battezzati, J.L. Hygounet, *RFID. Identificazione automatica a radiofrequenza*, Hoepli, Milano, 2006; ed anche A. Buscemi, "Nuove tecnologie per la prevenzione di errori nelle aziende sanitarie RFID (Radio Frequency Identifier)", in *Rivista elettronica di diritto, economia, management*, (2010), n. 1, pp. 213-242.

[16] S. Amato, *Biodiritto 4.0: intelligenza artificiale e nuove tecnologie*, Giappichelli, Torino, 2020, pp. 89-100.

prima tappa di un progetto che, se portato fino in fondo, permetterà, appunto, di comandare le macchine col pensiero. Questa notizia chiaramente ha determinato l'aumento esponenziale dell'interesse economico ad investire in questo campo. È, dunque, prevedibile che, assisteremo nel prossimo futuro, alla creazione di un crescente numero di dispositivi in grado sia di leggere l'attività del cervello umano che di scrivere informazioni neurali nel cervello.

3. La protezione dei neurodiritti nel GDPR. Il diritto *by design*

A ben guardare, tutto l'ecosistema tecnologico descritto ha una sola funzione ben precisa: raccogliere dati dalle persone.

I dati sono diventati la linfa vitale dei modelli produttivi contemporanei, quindi hanno un valore economico, e le tecnologie di nuova generazione sono l'elemento naturale in cui questi dati possono essere sfruttati in tutte le loro potenzialità. La quarta rivoluzione industriale origina proprio dal concetto di "dato" e si muove e si sviluppa intorno alla capacità dei sistemi di intelligenza artificiale di interpretare questi dati e creare valore da essi.

Ma la rapidità con cui queste tecnologie stanno prendendo piede e vengono presentate al pubblico crea diversi problemi e desta varie preoccupazioni di tipo tecnico, giuridico ed etico, legate, principalmente, alla difficoltà di controllarne adeguatamente il funzionamento e la corretta gestione.

Dal punto di vista medico, ad esempio, ancora non è noto se impianti biotecnologici invasivi come i microchip possano creare traumi o danni al cervello, o in genere al corpo in cui vengono impiantati. Ancora, non è noto se i campi elettromagnetici esterni possano interferire con il loro funzionamento.

Analogamente, dal punto di vista giuridico si evidenziano delle difficoltà oggettive nella capacità degli ordinamenti di disciplinare adeguatamente l'evoluzione tecnica e garantire in modo esaustivo il rispetto dei diritti fondamentali, della sicurezza, dei principi etici[17].

[17] Questi aspetti, oltre a quelli più prettamente informatici, sono ampiamente discussi in dottrina. Tra i diversi contributi in tal senso si richiamano R. Brighi, A. Zullo (a cura di), *Filosofia del diritto e nuove tecnologie.*

Invero, nel mondo attuale, la digitalizzazione è ad un passo dal creare un vero e proprio sconvolgimento del nostro modo di vivere, che, però, risulta ancora difficilmente percepibile dalla società. E, sebbene ci siano delle norme che disciplinano almeno gli aspetti più problematici di questa transizione, come il regolamento per la protezione dei dati personali o il più recente *AI Act* emanato dalla Commissione europea, persistono ancora molte difficoltà interpretative e applicative su di esse. Del pari, la popolazione di utenti, cittadini, fruitori di servizi innovativi, non solo in ambito sanitario, ma in ogni aspetto della vita quotidiana, non ha consapevolezza dei propri diritti digitali.

Il GDPR, ad esempio, in tal senso, possiede – secondo dottrina – diverse potenzialità inesplorate che incidono negativamente sulla sua effettiva e corretta attuazione, sia in ambito pubblico che in ambito privato, ledendo così di fatto il diritto alla protezione dei dati personali degli individui e consentendo ai creatori di tecnologia di muoversi in zone d'ombra in cui non vi sono responsabilità[18].

Alla base dell'impianto del GDPR vi è, infatti, un obbligo di base che prevede che i valori legali vengano integrati in ogni sistema tecnologico già fin dalle fasi di ideazione e progettazione. I principi sanciti dal considerando 78 della *privacy by design* e *privacy by default* non si limitano soltanto a regolamentare l'uso della tecnologia applicando la legge dall'esterno ed in una fase successiva, ma mirano a guidare l'intero processo sociotecnico di creazione e utilizzo degli strumenti informatici dall'interno.

Prospettive di ricerca tra teoria e pratica, Aracne editrice, Roma, 2015; L. Palazzani, *Tecnologie dell'informazione e intelligenza artificiale, sfide etiche al diritto*, Edizioni Studiorum, Roma, 2020; G. Sartor, *Intelligenza artificiale e diritto*, Giappichelli, Torino, 2022; S. Zorzetto, "La metafora della IA: una giungla lessicale e foresta simbolica", in *Notizie di Politeia*, 151 (2023), pp. 179-185. Inoltre, dal punto di vista filosofico giuridico la questione delle neuroscienze e dei neurodiritti è stata oggetto di diversi studi. Per queste analisi si rinvia, tra gli altri, a S. Fuselli, *Diritto, neuroscienze, filosofia. Un itinerario*, FrancoAngeli, Milano, 2014; P. Sequeri, *Deontologia del fondamento*, Giappichelli, Torino, 2016. Sul punto occorre richiamare anche i pareri del Comitato Nazionale per la bioetica, tra cui *Neuroscienze ed esperimenti sull'uomo: osservazioni bioetiche*, 2010; *Neuroscienze e potenziamento cognitivo farmacologico: profili bioetici*, 2013.

[18] C. Djeffal, "The Normative Potential of the European Rule on Automated Decisions: A New Reading for Art. 22 GDPR", in *ZaöRV*, 80 (2020), pp. 847-879.

Invece di raccogliere ed elencare una serie di diritti da tutelare, il GDPR è fondato su una logica che mira alla valutazione dell'impatto preventiva delle varie attività di gestione dei dati personali, abbinando a ciascuna delle misure di sicurezza idonee alla tutela dei diritti, delle libertà e degli interessi legittimi di ciascun individuo.

Il valore giuridico della protezione dei dati dei fruitori delle applicazioni tecnologiche deve, dunque, essere esaminato, valutato ed integrato già nelle fasi iniziali di ogni progetto tecnologico proprio al fine di creare un equilibrio costante tra le esigenze di progresso e la tutela delle persone[19].

In questo modo, lo scopo del GDPR non si limita soltanto alla protezione dei dati fine a se stessa, ma mira ad ampliare la portata di tutti i possibili diritti rilevanti. Questa formulazione aperta, da un lato, garantisce la sopravvivenza nel tempo della norma, che grazie ad una interpretazione evolutiva si adatterebbe facilmente a tutti gli scenari tecnologici futuri, laddove correttamente applicata; e, dall'altro, evita il sovraccarico e la confusione normativa in un ambito, come quello tecnico-informatico, in continuo divenire.

Questo principio vale tanto più nel caso dei sistemi di intelligenza artificiale applicata ad un settore particolare come quello sanitario, ed in specie nell'ambito delle neuroscienze, dove la ponderazione adeguata dell'uso di strumenti particolarmente invasivi richiede un bilanciamento importante tra le esigenze scientifiche e i diritti e le libertà dell'uomo, in ogni fase del processo di utilizzo della tecnica.

Il tema presuppone, chiaramente, la conoscenza e la creazione di una struttura di *governance* del sistema tecnologico, e dunque una riorganizzazione funzionale degli apparati produttivi pubblici e privati, in mancanza della quale ogni tentativo di disciplina rischia di trovare applicazione solo parzialmente, se non addirittura di restare lettera morta.

[19] M. Hildebrandt, "Legal and Technological Normativity: More (and Less) than Twin Sisters", in *Techné: Research in Philosophy and Technology*, 12 (2008), p. 169; Id., "Saved by Design? The Case of Legal Protection by Design", in *Nanoethics*, 11 (2017), n. 3, p. 307; Id., "Legal Protection by Design, Objections and Refutations", in *Legisprudence*, (2015), n. 5, p. 223; ed anche L. Moerel, M. Storm, "Automated Decisions Based on Profiling", in *Autonomous Systems and the Law*, (2019), p. 91.

Come sottolineato da ultimo anche nel regolamento europeo sull'intelligenza artificiale (*AI Act*), che muove da premesse in tutto simili al GDPR, la creazione di una struttura di *governance* comune per la gestione dei sistemi di intelligenza artificiale è fondamentale per avviare il dialogo e la cooperazione tra i vari portatori d'interesse. La trasparenza della gestione dei sistemi di intelligenza artificiale serve ad attuare una corretta valutazione dei rischi ad essi connessi e a favorire la condivisione dei risultati, funzionale a creare fiducia verso i sistemi tecnologici; diventa, così, un prerequisito fondamentale per la salvaguardia dei diritti umani, che, oggi, sono tutti potenzialmente messi a rischio dagli strumenti digitali[20].

In linea con i principi della *governance*, i sistemi di intelligenza artificiale devono essere conformi ai principi legali *by design*, quindi fin dalla loro progettazione, come richiede il dettato normativo del GDPR.

Il tema della *governance*, quello dell'*accountability* e la valutazione dei rischi delle nuove tecnologie sono concetti complessi che meritano una trattazione a parte, a fronte della loro complessità. Se ne fa qui solo un cenno per sottolineare che dare risalto e comprendere questi principi, oggi in Italia, significa diffondere nella società conoscenza e comprensione di un tema su cui ancora c'è molta disinformazione e scarsa chiarezza, al fine di creare delle solide basi per un uso consapevole e giuridicamente corretto delle tecnologie, che colmi il *digital divide* e prevenga discriminazioni ed ingiustizie.

4. La trasparenza informativa in sanità. Riprogettare la comunicazione attraverso il *legal design*

Governance, *accountability* e valutazione preventiva dei rischi poggiano tutte su un principio fondamentale: la trasparenza,

[20] In materia di intelligenza artificiale, l'Unione europea ha cercato di fornire un quadro di regole comuni per l'intero territorio comunitario, con una serie di documenti strategici che analizzano questo articolato fenomeno. A partire dal Libro Bianco sull'intelligenza artificiale, *Un approccio europeo all'eccellenza e alla fiducia*, COM(2020) 65 final del 19 febbraio 2020, si è giunti alla proposta di regolamento *Artificial Intelligence Act*, COM(2021) 0106 final del 21 aprile 2021, che è stata definitivamente ed interamente approvata, all'unanimità degli Stati membri, il 26 gennaio 2024.

che si pone alla base della riprogettazione giuridica delle società digitali e accomuna tutta la normativa europea sulla digitalizzazione.

In relazione ad esso, il GDPR impone al titolare del trattamento dei dati necessariamente due attività: documentare e comunicare le azioni di raccolta e gestione dei dati in modo chiaro, coerente, proattivo e responsabile.

La trasparenza è dunque da interpretarsi, ai sensi del considerando 37 del GDPR, come l'obbligo di rendere conoscibili le modalità con cui i dati sono raccolti, utilizzati e consultati grazie ad informazioni e comunicazioni facilmente accessibili e comprensibili, fondate su un linguaggio semplice e chiaro.

Questo perché si tratta di un principio generale cui ogni trattamento deve uniformarsi e riguarda, in specie, la comunicazione agli interessati di ogni informazione necessaria ad assicurare un trattamento corretto.

L'art. 5 del GDPR individua, ancora, la trasparenza come condizione di qualunque trattamento di dati personali, in tal modo sottolineando che si tratta di un dovere che incombe in capo al titolare per il fatto stesso di porre in essere un trattamento di dati personali.

Il principio di trasparenza è, quindi, funzionale ad individuare le modalità organizzative e tecnologiche che possono consentire al titolare di dimostrare in qualunque momento che il trattamento sia stato effettuato conformemente alla normativa.

L'obbligo di trasparenza dei trattamenti voluto dal GDPR muove dall'assunto, dichiarato nel considerando 7, che l'evoluzione della società digitale e della sua tecnologia richiedono un quadro più solido e coerente in materia di protezione dei dati in ambito comunitario, affiancato da efficaci misure di attuazione, data l'importanza di creare un clima di fiducia che consenta lo sviluppo dell'economia digitale in tutto il mercato interno[21].

Per cui l'obiettivo essenziale dell'intero impianto del GDPR e del principio di trasparenza è, sì quello di rendere più efficace la tutela dei dati personali intesa come diritto fondamentale della persona, di cui all'art. 8 della Carta dei diritti dell'Unione euro-

[21] F. Pizzetti, "Trasparenza nel trattamento dati che cambia col GDPR: l'alba di un nuovo valore sociale", in *Agenda Digitale*, 13/06/2018.

pea, ma anche, e forse in misura più rilevante, quello di incentivare la più ampia e libera circolazione dei dati e favorire così lo sviluppo del mercato unico digitale europeo.

A riprova di ciò potrebbe citarsi, fra gli altri documenti, l'*European Data Governance Act* del 2022 che sottolinea come i dati siano diventati, nelle società digitali, un fattore di produzione per cui ne deve essere incrementato l'utilizzo e la domanda nel mercato unico europeo già entro il 2030. In questo mercato unico in cui i dati circoleranno in formato aperto, le pubbliche amministrazioni e le imprese potranno avere accesso facilmente a una quantità pressoché infinita di dati personali e non personali, compresi i dati commerciali sensibili quindi quelli sanitari in particolare, al fine di stimolare la crescita economica e creare valore.

La richiesta trasparenza nei trattamenti di dati incide, così, sulle modalità con le quali il titolare deve dare agli interessati conto del modo in cui diffonde i dati che li riguardano.

Per giustificare appunto la libera circolazione dei dati funzionale all'economia digitale comunitaria è, dunque, fondamentale il rispetto degli artt. 13 e 14 del GDPR che disciplinano l'informativa e il consenso dell'interessato al trattamento dei propri dati.

Il fondamento dell'intera attività di gestione dei dati personali, infatti, risiede nella responsabilità posta in capo al titolare del trattamento di fornire ai soggetti a cui i dati sono richiesti, in dettaglio, ogni elemento utile a comprendere l'uso che ne farà; quindi, le finalità del trattamento, la comunicazione di questi dati ad altri, il periodo di conservazione dei dati, il diritto di presentare un reclamo all'autorità di controllo. Nella sua nuova articolazione, l'informativa deve specificare chiaramente anche se il trattamento implica processi decisionali automatizzati o procedimenti di profilazione e deve indicare le conseguenze che tali procedure pongono a carico dell'interessato.

Secondo il considerando 58 del GDPR, queste informazioni devono essere concise, facilmente accessibili e di facile comprensione, scritte con un linguaggio semplice e chiaro, e se, necessario, descritte anche con elementi visivi.

Il successivo considerando 60 e l'art. 12 del regolamento prevedono la possibilità di fornire le informazioni in combinazione con icone standardizzate per dare, in modo facilmente visibile, intelligibile e chiaramente leggibile, un quadro d'insieme del trattamento previsto. Se presentate elettronicamente, le icone dovrebbero essere leggibili da dispositivo automatico.

Il consenso, di conseguenza, rappresenta – secondo l'art. 4 del GDPR – la manifestazione di volontà, libera, specifica, informata e inequivocabile dell'interessato, con la quale lo stesso esprime il proprio assenso inequivocabile al trattamento dei dati personali che lo riguardano. Il consenso rappresenta, quindi, la base giuridica del trattamento[22].

In ambito sanitario, occorre chiarire che ai sensi dell'art. 9 del GDPR il trattamento di tali categorie di dati è vietato in linea di principio, con alcune eccezioni, come ad esempio il caso in cui l'interessato abbia prestato il proprio consenso esplicito al trattamento di tali dati personali per una o più finalità specifiche (art. 9, c. 2, lett. a).

Il consenso esplicito dell'interessato è sempre necessario nel caso di trattamenti effettuati attraverso il fascicolo sanitario elettronico e i trattamenti connessi all'utilizzo di applicazioni mediche con le quali autonomi titolari raccolgono dati, anche sanitari dell'interessato, oppure quando, indipendentemente dalla finalità dell'applicazione, ai dati dell'interessato possano avere accesso soggetti diversi dai professionisti sanitari o altri soggetti tenuti al segreto professionale.

In queste circostanze, la chiarezza dell'informativa assume una rilevanza maggiore proprio per la particolarità dei dati trattati. Eppure, ancora oggi, soprattutto in Italia, una fra le maggiori criticità della comprensione delle dinamiche complesse che intervengono tra *privacy* e sanità sta proprio nella difficile intelligibilità delle informative, spesso errate, incomplete e scritte con un linguaggio poco chiaro, non realmente pensato per essere compreso dagli utenti.

[22] Tra le altre analisi proposte in merito al trattamento dei dati personali dal punto di vista etico e giuridico si veda A. Maceratini, "Dati digitali e intelligenza artificiale: sfide etiche e giuridiche", in *Tigor*, XIV (2022), n. 2, pp. 63-76; ed anche C. Casonato, "Potenzialità e sfide dell'intelligenza artificiale", in *BioLaw Journal*, (2019), n. 1, pp. 177-182.

Questa grande difficoltà si riscontra già soltanto nell'informare puntualmente le persone sulle modalità di trattamento dei dati sanitari. Per cui sembra facile immaginare quale difficoltà potrebbe esserci nel diffondere la conoscenza dell'eventuale utilizzo di strumenti digitali di ultima generazione che vengono integrati nel processo di cura.

E, dunque, la scarsa conoscenza della tecnologia e delle norme che la governano si trova ad incidere negativamente sia sulla necessità di fornire un valida informativa sulle terapie, funzionale ad ottenere il consenso informato, sia sulla necessità di fornire una altrettanto valida informativa *privacy*, funzionale ad ottenere il consenso per il trattamento dei dati.

Così, dare al paziente informazioni troppo complesse sui trattamenti medici o gli studi clinici e dare una errata informativa sul modo in cui verranno trattati i dati personali e sanitari raccolti genera una mancanza di comprensione che incide negativamente sul rapporto medico-paziente, viziando irrimediabilmente il consenso. Ciò determina non solo l'insorgere di controversie medico legali, ma anche la sostanziale violazione delle disposizioni del GDPR con il conseguente rischio delle applicazioni delle pesanti sanzioni in esso previste[23].

Gli errori di comunicazione nel campo della protezione dei dati sono, purtroppo, molto comuni e possono essere collegati principalmente al fatto che la maggior parte dei titolari e dei responsabili del trattamento dei dati, specie nelle pubbliche amministrazioni, considera ancora questi obblighi pratici di informazione soltanto come obblighi formalistici da applicarsi per conformità alla legge.

Per rendere veramente effettive tali tutele, una tra le soluzioni più innovative proposte per favorire una corretta interazione tra le amministrazioni, i professionisti e i pazienti potrebbe essere, oggi, il c.d. *legal design*, inteso come supporto comunicativo del

[23] A. Cavaliere, "Consenso informato: profili evolutivi e rapporto medico-paziente", in *Diritto.it*, 2018; ed anche Aa.Vv. *L'evoluzione dei Modelli Sanitari Internazionali a Confronto. Per costruire il servizio sanitario nazionale del futuro*, LUISS Guido Carli, Roma, 2016; Aa.Vv., *Le parole giuste. Scrittura tecnica e cultura linguistica per il buon funzionamento della pubblica amministrazione e della giustizia*, Atti del convegno, Senato della Repubblica, Roma, 2017.

personale sanitario, il cui scopo è quello di ripensare il linguaggio giuridico dei documenti legali che accompagnano la pratica clinica al fine di renderlo accessibile a ogni tipologia di utente, tramite integrazione dei testi scritti con immagini, video, icone[24].

Si tratta di un concetto che è stato sviluppato, per la prima volta, nel lontano 2010 nelle università statunitensi e solo in seguito, precisamente nel 2016, ripreso dal considerando 60 del GDPR citato in precedenza, con cui si intende integrare la semantica giuridica con la "visualizzazione", ossia con i metodi e gli strumenti del *design* per aumentarne l'efficienza, l'efficacia, la comprensione[25].

L'implementazione delle tecniche di *legal design* in ambito sanitario potrebbe rappresentare un cambio di passo verso la migliore comprensione da parte dei cittadini del modo in cui vengono trattati i dati sanitari e di come la tecnologia li sfrutti per compiere i propri processi logico-matematici.

Il *legal design* diventa essenziale nell'era digitale per promuovere, accanto allo sviluppo del mercato, anche e soprattutto la libertà degli individui e la loro partecipazione attiva nella società.

5. I *legal design patterns* come nuovo approccio multidisciplinare del diritto

Il c.d. *design process* rappresenta un metodo innovativo di progettazione e di risoluzione di controversie che i *designers* utilizzano per creare oggetti che non siano solo gradevoli dal punto di vista estetico, ma capaci di rispondere al meglio ai bisogni degli utenti finali.

[24] Aa.Vv., *Informatica giuridica e legal tech*, Egea, Milano, 2021; P. Montella, "Legal design, a cosa serve e quali gli utilizzi nelle informative privacy", in *Cybersecurity360*, 25/06/2021; M.L. Rizzo, "Come il legal design nella sanità digitale cambia l'interazione con i pazienti", in *RiskManagement360*, 25/11/2020. Lo sviluppo di strategie che possano favorire la comunicazione, specie nel campo della sanità digitale, deve muovere necessariamente dall'obbligo, normativo e deontologico, del rispetto dei dati personali, così, P.B. Helzel, A. Sergio (a cura di), *La bioetica come ponte tra società e innovazione*, Aracne, Roma, 2016.

[25] C. D'Onofrio, "Legal design e intelligenza artificiale", in *I-lex. Rivista di Scienze giuridiche, scienze cognitive ed intelligenza artificiale*, 16 (2023), n. 1, https://doi.org/10.6092/issn.1825-1927/17198.

Questo procedimento ha iniziato a svilupparsi nel corso degli anni '80 del Novecento e oggi interessa molti campi del sapere come il *management*, l'ingegneria informatica, l'urbanistica e il diritto.

Attraverso il *design process* sono state sviluppate diverse soluzioni, i c.d. *design patterns*, finalizzate a risolvere i problemi più ricorrenti che le persone affrontano quando hanno a che fare con documenti legali. Si tratta di standard condivisi tra i professionisti per risolvere problemi ricorrenti, come nel caso di specie, fornire ai clienti dei documenti legali facilmente comprensibili e quindi efficienti.

Per essere veramente utile e funzionale allo scopo il *design process* necessita di una profonda comprensione degli elementi che lo caratterizzano e della loro connessione. Questo approccio, che deve necessariamente essere multidisciplinare, permette di individuare nuovi percorsi e nuove strategie per far fronte ai problemi di comunicazione più comuni in ambito legale, ridimensionandoli e offrendone una visione più completa.

La caratteristica essenziale del *legal design* è, dunque, quella di essere *user-centered* e può trovare applicazione in tutti i campi del diritto già fin dalla creazione e redazione delle leggi.

Negli ultimi anni, l'Unione Europea, ad esempio, ha riconosciuto ai cittadini un ruolo sempre crescente nella redazione delle leggi. Nell'agenda sulla *Better regulation* viene espressamente previsto il coinvolgimento diretto degli *stakeholders* in tutte le fasi dell'iter legislativo ed è incoraggiata la raccolta di *feedback* dell'impatto che la normativa europea ha su imprese, parti sociali e società civile.

Un altro esempio di *legal design*, già in uso da diverso tempo, sono le licenze *Creative Commons*, che descrivono, all'utente interessato, attraverso delle icone, le caratteristiche giuridiche più rilevanti di una licenza.

Ma, come abbiamo già ampiamente sottolineato, anche nel caso della *privacy* in sanità è diventato fondamentale che i cittadini comprendano a fondo la tutela che la normativa in materia offre loro, perché senza questo elemento essenziale, le persone rischiano sempre di più di essere discriminate[26].

[26] S. Rodotà, "Tra diritti fondamentali ed elasticità della normativa: il nuovo codice sulla privacy", in *Europa e diritto privato*, (2004), n. 1, pp. 1-12.

E ciò è tanto più vero nel caso della *e-health*, che, come detto in precedenza, si riferisce all'impiego delle nuove tecnologie nel settore sanitario, e della *connected health*, che indica l'impiego di tecnologie nelle cure mediche, in cui rientra l'uso di dispositivi sanitari impiantabili, e che si basa su un modello concettuale di gestione della salute fondato sulla condivisione continua delle informazioni cliniche del paziente.

Già oggi, questo problema è evidente nell'uso delle app sanitarie di uso più comune che ogni utente può scaricare sul proprio smartphone e utilizzare quotidianamente[27].

Come evidenziato da molti ricercatori, le informative *privacy* di questi strumenti sono poco chiare, non rispettano nessun principio minimo, e, addirittura, nel caso delle app progettate appositamente per essere utilizzate dal personale medico, le *privacy policies* non sono nemmeno contemplate[28].

Questa paradossale trascuratezza è molto rischiosa specie in un ambito come quello di cui si tratta caratterizzato da una forte innovatività e dalla mancanza di linee guida di comportamento specifiche.

Per questo, l'applicazione del *design process* potrebbe offrire vantaggi non indifferenti dal punto di vista giuridico. La prototipazione e i test ripetuti sulla qualità dell'informazione sono due aspetti centrali del *legal design*, che consentirebbero di differenziare le comunicazioni *privacy* in base alla tipologia degli

[27] E. Bellocci (a cura di), *Dispositivi medici: aspetti regolatori e operativi*, Indagine del Ministero della Salute, Roma, 2010; Aa.Vv., *Impact of Digital Health Grows as Innovation, Evidence and Adoption of Mobile Health apps accelerate*, (2017), disponibile online al seguente link: https://healthcaretransformers.com/digital-health/current-trends/improving-digital-health-adoption-healthcare-transformers/?utm_source=google&utm_medium=cpc&utm_campaign=ht-google-sea-eu-en-sea-content-type-ebook&utm_term=eu-ebook-improve-dhs&gad_source=1&; Adoc, *I consumatori e le app per la salute, un'inchiesta sulla salute digitale*, (2018), disponibile online al seguente link: www.adocnazionale.it.

[28] T. Mulder, "Health Apps, their Privacy Policy and the GDPR", in *European Journal of Law and Technology*, 10 (2019), n. 1, pp. 1-22; M.P. Borja, I. De la Torre Díez, M. Lopez-Koronado, "Privacy and Security in Mobile Health Apps: a Review and Recommendations", in *Journal of Medical Systems*, 39 (2014), n. 1, pp. 1-8; M.S. Gal, O. Aviv, "The competitive effects of the GDPR", in *Journal of Competition Law & Economics*, 16 (2020), n. 3, pp. 349-391.

utenti e di effettuare un controllo sulla qualità dell'informativa, proprio come richiesto dal principio di trasparenza[29].

6. Conclusioni

Una delle maggiori sfide per l'affermazione del *legal design* è, ancora oggi in Italia, la resistenza sociale al cambiamento.

La materia della protezione dei dati, in specie, è stata definita un "diritto inquieto" poiché in dialettica con una tecnica in continua evoluzione e con molteplici interessi, di natura individuale e collettiva, ma che trova forza nella sua funzione sociale. Essa dovrebbe rappresentare un punto di equilibrio tra libertà e tecnica, tra persona e società, fare da presupposto della tenuta della democrazia in circostanze eccezionali, ed essere la garanzia del mantenimento dei diritti e delle libertà individuali[30]. Ma l'unico modo per spostare effettivamente l'attenzione dal dato in sé, inteso come bene economico, ad un effettivo diritto della persona dovrebbe essere, appunto, diffondere la conoscenza su questi principi e sulle poche tutele disponibili e applicare correttamente la normativa.

La mancanza di consapevolezza sui propri diritti fondamentali nel mondo digitale, che può sembrare un concetto vago e forse anche troppo ribadito da chi scrive[31], espone le persone a rischi ancora non del tutto chiari e questo *digital divide* si ripercuote inevitabilmente sulla possibilità di richiedere la giusta tutela giuridica in caso di violazioni.

La conoscenza rappresenta l'elemento chiave della contemporaneità per mantenere l'innovazione, specie quella tecnica,

[29] C. Rauccio, "How legal design can improve data protection communication and make privacy policy more attractive", in *European Journal of privacy Law & Technologies*, (2021), n. 1, pp. 224-242; M. Palmirani, A. Rossi, M. Martoni, M. Hagan, "A Methodological Framework to Design a Machine-Readable Privacy Icon Set", in *Data Protection/Legaltech*, *Proceedings of the 21st International Legal Informatics Symposium (IRIS)*, (2018), n. 2, pp. 451-454.

[30] A. Soro, "Tracciamento contagi coronavirus, ecco i criteri da seguire", in *Agenda Digitale*, 2020.

[31] Tra tutti mi sia consentito un rinvio al mio *E-Government e società del controllo. Il lato oscuro del mondo digitale*, Rubbettino, 2023 e, in tema di protezione dei dati personali, "Lotta alla pandemia e tutela della privacy", in *Tigor*, XIV (2022), n. 2, pp. 92-103.

nei confini entro i quali dovrebbe restare, ossia soltanto quello di un utile strumento per semplificare la vita dell'uomo.

L'uomo è da sempre chiamato a costruire la società in cui vivere e soltanto la conoscenza può orientare gli individui in un mondo in cui sembra non esserci più alcun limite morale di fronte alle ragioni economiche e tecnologiche.

La sostanza della *governance* della società digitale si basa sull'assunto che non basta più adempiere, semplicemente, agli obblighi formali prescritti dalle leggi, magari soltanto per non incorrere in sanzioni, ma è necessario comprendere i problemi di questa transizione e riorganizzare i settori produttivi, pubblici e privati, sulla base di un principio di responsabilizzazione generalizzata verso il rispetto dei diritti degli utenti.

Per cui, appare chiaro come la progettazione giuridica, il diritto *by design*, la trasparenza e la comunicazione siano diventati elementi imprescindibili della professione legale, tanto più perché le nuove tecnologie stanno ridisegnando le categorie giuridiche tradizionali[32].

Le politiche di rispetto della *privacy* si basano sulla trasparenza dell'azione dei titolari del trattamento e tutti gli errori comunicativi si ripercuotono, imprescindibilmente, sulla legalità del loro modo di agire, che lede il diritto dell'interessato a mantenere il pieno controllo sui propri dati personali. Nella maggior parte dei casi, infatti, gli interessati non leggono realmente le informative sulla *privacy*, restano completamente all'oscuro del trattamento dei loro dati personali e, di conseguenza, non sono in grado di prendere decisioni significative al riguardo.

D'altro canto, i giuristi devono porsi di fronte al fatto che il sistema legale deve cambiare in accordo con il mondo esterno e che trasformazione digitale non si sostanzia soltanto in uso generalizzato delle macchine.

Significa trasformazione sociale e culturale, significa una nuova mentalità in grado di aprirsi a nuove prospettive giuridiche. Significa contemplare tutte le incognite legate alla diffusio-

[32] I diritti umani sono diritti storici che nascono in risposta all'evoluzione della società per difendere la libertà. Così N. Bobbio, *L'età dei diritti*, Einaudi, Torino, 1990; ed anche C. Faralli, "Diritti e nuove tecnologie", in *Tigor*, (2019), n. 2, pp. 43-52.

ne indiscriminata delle tecnologie per mantenere vivo il diritto degli individui di scegliere. Significa riorganizzazione.

Dal punto di vista sostanziale, quindi, queste affermazioni di principio dovrebbero iniziare a tradursi in azioni concrete, pensate e gestite effettivamente a garanzia dei diritti fondamentali.

E questo tanto più in un ambito come quello sanitario che, in Italia, com'è noto, è afflitto da diversi problemi che ne intaccano la qualità e il livello di efficienza. Per essere valido, il processo di cambiamento offerto dalla rivoluzione digitale deve partire da un obiettivo concreto e misurabile di miglioramento da definire con il coinvolgimento di tutti gli attori che intervengono nei processi di cura e deve essere programmato nel tempo e attraverso l'impiego delle risorse necessarie, agendo laddove occorrano degli interventi correttivi e stabilizzando i risultati conseguiti. La comunicazione al pubblico di questi processi di miglioramento dovrebbe diventare così la spinta, una tensione costante, verso la gestione proattiva del cambiamento organizzativo.

In questo senso, la *governance* digitale del settore pubblico, che pretende di responsabilizzare ed aprire al dialogo la cittadinanza, la politica, l'amministrazione, il *management* pubblico, il diritto, dovrebbe rappresentare un punto di svolta, ossia diventare una forma di *buon governo* per perseguire il *bene comune*, consentendo alle amministrazioni di dotarsi di strumentazioni utili a fornire servizi più conformi a quelle che sono le necessità dei cittadini, migliorando al contempo il rapporto tra costo e qualità dei servizi offerti e riducendo sprechi ed inefficienza[33].

Ebbene, il grande affanno dell'Italia di fronte a questi temi è causato, sostanzialmente, proprio dal fatto che questo cambiamento non è mai stato attuato, a causa della incapacità del sistema di superare le logiche burocratiche, settoriali e obsolete che lo attanagliano[34].

[33] E. Molinari, "Sanità 4.0, quali scenari tra big data, blockchain e internet delle cose", in *Agenda Digitale*, 10/07/2017; ed anche R. Goretta, "Intelligenza artificiale, ecco come il GDPR regola l tecnologia di frontiera", in *Agenda Digitale*, 10/10/2019.

[34] Le prime dichiarazioni europee sulla nascita della società dell'informazione e il primo impianto normativo in tema di protezione dei dati personali a livello nazionale risalgono rispettivamente, agli ormai lontani 1990 e 1996. Il Codice dell'amministrazione digitale che ha reso obbligatoria la

E ora, per colmare di fretta questo ritardo si tende, sostanzialmente, a perseguire una sorta di "innovazione a tutti i costi", si attua un *digital washing*, funzionale soltanto ad allinearsi ai più stringenti obblighi di legge, soltanto perché non se ne può fare a meno. Ma la mancanza di una presa di coscienza reale, su questa rivoluzione che sta investendo il nostro modo di vivere, apre le porte a diverse incognite, esponendoci al decadimento di una situazione già critica.

Non occorre soffermarsi sul fatto che l'uso delle nuove tecnologie nasconda interessi politici ed economici delle aziende e delle multinazionali *high tech*, e potrebbe non essere finalizzato a creare i processi virtuosi, partecipativi e sostenibili tanto dichiarati. Ma questa è la ragione principale per cui questi processi dovrebbero essere pensati e strutturati attraverso lo sviluppo di prospettive critiche innovative, la creazione di *partnership* culturali e di nuovi e più ampi punti di vista nel campo dell'etica, *per il* digitale e *del* digitale, per prevenire il rischio di trovarci un giorno, senza neanche accorgercene, in una cyber-dittatura peggiorativa della vita umana.

Però, di fronte all'avanzare di strumenti come l'*Internet of bodies*, alle sperimentazioni di *Neuralink*, alle dichiarazioni del *World Economic Forum*, che già contemplano, per il vicino 2030, la possibilità di mettere i *microchip* a disposizione di tutti gli uomini per vivere nella realtà aumentata, che vengono definiti come "accessori" al pari di una borsa o uno smartphone[35], forse dovremmo iniziare a chiederci seriamente se abbiamo ancora tempo per parlare di "umanesimo digitale".

L'umanesimo digitale oggi vive un momento cruciale, perché si pone all'intersezione tra la crisi di sostenibilità e le opportunità offerte dalla digitalizzazione per cui, al fine di calibrare le sfide future da affrontare, è necessario valutare che esso comporta vantaggi, ma anche rotture cruciali, ed implicazioni profonde e di lungo termine, relative a ciò che significa essere "umani" e a che

digitalizzazione dell'intero settore pubblico, di cui, ricordiamo, la sanità rappresenta un ambito strategico, è stato emanato nel 2005.

[35] K. Philips, *Augmented tech can change the way we live, but only with the right support and vision*, in www.weforum.org, 2022. Recuperato da https://intelligence.weforum.org/monitor/latest-knowledge/1a8874ceb14448ee99084b543ec483a5.

cosa debba essere una buona società digitale. L'umanesimo digitale segnala, in questo senso, la transizione dalla linearità nella concezione e comprensione del mondo, che è stata una delle caratteristiche fondamentali della modernità, verso la necessità di affrontare i processi non lineari dei sistemi adattativi complessi. La digitalizzazione esaspera le tensioni sociali e mette a nudo le disuguaglianze, per cui, per creare una buona società digitale, occorre imparare ad orientarsi nel rapporto tra esseri umani e macchine per creare tra di essi delle interazioni meno sfumate e caotiche. Di conseguenza, la posta in gioco per l'umanesimo digitale è elevata e si sostanzia nell'urgenza di individuare proposte concrete, pratiche, veramente etiche che includano i più profondi strati umanistici, andando oltre le soluzioni tecnologiche. Una società digitale buona è possibile solo se le tecnologie digitali vengano integrate nei contesti sociali in cui operano. Ma questo richiede il valore insostituibile del giudizio critico umano[36].

[36] Le preoccupazioni cruciali in merito ai problemi legati alla tecnologica della sorveglianza sono state analizzate da S. Zuboff, *Il capitalismo della sorveglianza*, LUISS University Press, Roma, 2019. Sul tema dell'umanesimo digitale, si veda a J. Nida-Rumelin, N. Weidenfeld, *Umanesimo digitale. Un'etica per l'epoca dell'intelligenza artificiale*, FrancoAngeli, Milano, 2018, *amplius*. Il tema del difficile rapporto tra le macchine e l'uomo è stato, fra l'altro, molto dibattuto e l'umanesimo digitale è stato anche oggetto di un manifesto pubblicato dall'Università di Vienna: H. Werthner *et al.*, *Vienna Manifesto on Digital Humanism*, (2019). Recuperato da www.informatik.tuwien.ac.at/dighum/. Sulle tensioni sono analizzate da H. Nowotny, *Umanesimo digitale: come navigare le tensioni che ci attendono*, Luiss University Press, Roma, 2022; ed anche da A. Punzi, "L'Umanesimo digitale: verso un nuovo principio di responsabilità?", in *Democrazia e diritti sociali*, (2023), n. 1, pp. 23-32.

Vincenzo Bruno Muscatiello

Ex Machina. Il diritto penale, all'improvviso

Abstract: **Ex Machina. Criminal Law, Suddenly.** Artificial intelligence is increasingly occupying our lives, posing new possibilities and new critical issues. The legal system tries to regulate them but, although the civil law reading is widely discussed, the possible punitive instances capture aspects of the artificial reality for which traditional criminal law could prove unprepared.

Keywords: Artificial Intelligence, Civil Law, Criminal Law, Artificial Criminal Liability.

Sommario: 1. Ehi Siri – 2. *The artificial game*: la mossa 37 – 3. Le parole per dirlo – 4. "Ho paura, Dave" – 5. Precop: sorvegliare e punire – 6. La colpevolezza artificiale – 7. Dalla colpa di organizzazione alla colpa di programmazione – 8. La costruzione delle regole. La vigilanza artificiale.

1. Ehi Siri

Nel 2019 l'Organizzazione per la cooperazione e lo sviluppo economico (OCSE) ha previsto che entro 15-20 anni le nuove tecnologie di automazione finiranno probabilmente per eliminare il 14% dei posti di lavoro a livello mondiale e trasformare radicalmente un altro 32%, il che significa che quasi oltre un miliardo di persone, in tutto il mondo, vedranno modificate le loro abitudini lavorative, con un cambiamento radicale delle forme e una accelerazione delle competenze professionali. Le nuove tecnologie non solo potranno accrescere e migliorare le attività ripetitive manuali, ma saranno in grado di eseguire tipologie di lavoro sempre più complesse e fondate sulla ricerca, codifica e scrittura, un tempo immaginate impermeabili ai cambiamenti radicali.

Non possiamo più far finta di ignorare gli sviluppi della intelligenza artificiale, le sue attitudini, le sue funzionalità: siamo di fronte ad una profonda modifica esistenziale, portatrice di gran-

di benefici e di grandi cambiamenti, per una diversa, e probabilmente migliore, gestione documentale, una dematerializzazione completa dei documenti, dello loro validazione, fino alla loro archiviazione e conservazione; per una diversa e migliore *digital finance*, attraverso la creazione di sistemi di allerta preventiva e strumenti che aiutino a valutare la propria situazione aziendale e quella dei clienti dei fornitori, per prevenire eventuali crisi e problemi di liquidità; una migliore e diversa capacità di analisi, attraverso strumenti che consentano di trasformare numeri e dati di diversa provenienza e difficile aggregazione, in un vantaggio competitivo attraverso proprio gli algoritmi di *machine learning* che riusciranno ad organizzare le informazioni in modo coerente, per una migliore interpretazione e un loro migliore utilizzo strategico; una diversa e migliore attenzione nel seguire il passo degli obblighi normativi, ricreando soluzioni semplici e intuitive per adeguare le organizzazioni aziendali, per esempio alla disciplina in materia di *privacy* e di GDPR, o per tutti quegli adempimenti che siano richiesti dalla legge, come, per esempio, la fatturazione elettronica o la segnalazione degli illeciti *whistleblowing*. Non solo. L'intelligenza artificiale ci aiuterà in una nuova visione dei *risk management* riuscendo a controllare ed impedire che singole classi di rischio raggiungano livelli minacciosi e ciò che è ordinariamente affidato alle intelligenze umane, potrebbe essere affidato a questi sistemi, a questi algoritmi e a queste nuove applicazioni di IA, per nuove modalità innovative di utilizzo di dati, trasformati in modelli operativi e in processi inesplorati. L'intelligenza artificiale riuscirà ad occupare attività anche quelle pericolose, come quelle dell'edilizia, estrazioni minerarie, monitoraggio di sezioni radioattive delle centrali nucleari, consentirà di, o meglio, riuscirà ad esplorare oceani e lo spazio interplanetario, a svolgere incarichi rischiosi: terremoti, alluvioni, disastri, passeranno, dalla cura del caso, alla disciplina del prevedibile, consentendo di salvare coltivazioni, territori, abitazioni e vite umane[1]; l'AI promette di saper e poter moni-

[1] Nell'articolo on line, "Intelligenza Artificiale e Disastri Ambientali: Flood Hub e Nestore", *in Environment da Creative AI* (https://data.creative-words.com/intelligenza-artificiale-e-disastri-ambientali-flood-hub-e-nestore/), emerge l'importanza dell'intelligenza artificiale nella prevenzione di disastri e calamità naturali attraverso una preventiva analisi delle precipita-

torare le coltivazioni in tempo reale, misurare il *se* e il *quanto* della irrigazione in ragione di agenti atmosferici registrati in tempo reale, scegliere il fertilizzante e quando raccogliere per massimizzare la resa[2]; la diagnosi, come quella di ipertensione polmonare, sarà infallibilmente affidata ad un algoritmo; l'abbandono di rifiuti potrà essere monitorato attraverso l'esame del DNA nella spazzatura buttata ai bordi della strada (così avviene già in Asia); persino la giustizia punitiva[3], tradizionalmente arricchita da coefficienti di umanità ed equità, saprà aggiungere, o affidarsi, a regole di predittività, calcolate da un algoritmo, e rendersi più agevole, e apparentemente più giusta e imparziale, sulla base di processi decisionali garantiti da algoritmi impermeabili a preconcetti, pregiudizi o idiosincrasie[4], che si tratti di calcolare la recidiva[5] o la prevenzione della criminalità, di vario

zioni e degli eventi sismici, cosi da anticipare la possibilità di incendi, alluvioni, tempeste etc. Un valido esempio di IA che ha come obiettivo quello di prevenire catastrofi ambientali è Flood Hub, un sistema sviluppato nel 2018 di monitoraggio dei corsi d'acqua con lo scopo di prevedere inondazioni e alluvioni e arginarne le eventuali conseguenze. Si tratta di un sistema di previsione costituito da quattro sottosistemi (convalida dei dati; previsione delle fasi; modellazione delle inondazioni; distribuzione degli avvisi) preceduto da una raccolta dati sulle avvenute alluvioni, usando poi i sistemi di *Machine Learning* e intelligenza artificiale per elaborare delle stime e delle previsioni su possibili future inondazioni. Una diversa applicazione della intelligenza artificiale si è legata allo studio delle sequenze sismiche: pur risultando impossibile determinare con assoluta certezza il luogo e il momento preciso di un evento sismico, i sistemi di IA sono risultati fondamentali per analizzare e individuare i sismi grazie ad un algoritmo – denominato Nestore – in grado di apprendere ed effettuare una valutazione delle probabilità di repliche sismiche.

[2] Ce lo ricorda P. Triolo, "Il piano degli Stati Uniti per rallentare l'AI cinese", in *Limes*, 12 (2022), p. 43.

[3] Di recente G. Lasagni, "Difendersi dall'intelligenza artificiale o difendersi con l'intelligenza artificiale? Verso un cambio di paradigma", in *Rivista italiana di diritto e procedura penale*, (2022), n. 4, pp. 1545 ss.

[4] In argomento, per opportune osservazioni critiche, si veda G. Canzio, "Intelligenza artificiale, algoritmi e giustizia penale", in *Sistema penale*, 08/01/2021; nonché G. Ubertis, "Intelligenza artificiale, giustizia penale, controllo umano significativo", in *Sistema penale*, 11/11/2020, ma anche in *Diritto penale contemporaneo*, 4 (2020).

[5] Gli usi della AI nel sistema giuridico sono raccontati anche da A. Contaldo, F. Campara, nel saggio "Intelligenza artificiale e Diritto. Dai siste-

genere[6], in una nuova configurazione dei rapporti sociali[7] ed in una sempre più vasta area di collaborazione[8], estesa, perché no, alle nuove forme di responsabilità delle persone giuridiche[9] per analizzare il presente, esaminare la rischiosità aziendale attraverso una costante e progressiva digitalizzazione dei protocolli organizzativi e dei sistemi di analisi e di controllo dei rischi[10], e

mi esperti 'classici' ai sistemi esperti 'evoluti': tecnologia e implementazione giuridica", in G. Taddei Elmi, A. Contaldo (a cura di), *Intelligenza artificiale. Algoritmi giuridici.* Ius condendum *o "fantadiritto"?*, Pacini, Pisa, 2020, pp. 50 ss.

[6] C. Burchard, "L'intelligenza artificiale come fine del diritto penale? Sulla trasformazione algoritmica della società", in *Rivista italiana di diritto e procedura penale*, 62 (2019), n. 4, pp. 1919 ss.

[7] Nuova configurazione in cui la funzione terza del diritto, piuttosto che essere affidata a terzi, "diventa immanente, localizzandosi ormai nella tecnica stessa": così A. Garapon, J. Lassègue, *La giustizia digitale. Determinismo tecnologico e libertà*, trad. it., il Mulino, Bologna, 2021, p. 245.

[8] A. Giannini, "Intelligenza artificiale, *human oversight* e responsabilità penale: prove d'impatto a livello europeo", in *Criminalia*, (2021), p. 11.

[9] In tema, per la materia ambientale, R. Nitti, V.B. Muscatiello, nell'ultimo capitolo de *Il diritto penale dell'ambiente*, Cacucci, Bari, 2023.

[10] Un algoritmo programmato in queste forme potrebbe essere un buon viatico per una misura dell'esistente meno esposta alle variabili umane. *Behavioural Biometrics*, *Robot Process Automation*, *Intelligence Process Automation* e *Intelligence Business Process Management Suites*, sono alcune delle formule che traducono la nuova frontiera della protezione tecnologica, già nel 2014 capace di ospitare, quale membro di un consiglio di amministrazione, Vital, un algoritmo basato su intelligenza artificiale, quale membro del consiglio di amministrazione della Deep Knowledge Ventures, una *venture capital fund* con sede ad Hong Kong, col compito di vagliare le possibili perdite derivanti da eccessivi investimenti in progetti sopravvalutati. Nel 2016, anche la Tieto, un'azienda norvegese leader nel settore IT, ha incluso nel suo consiglio di amministrazione un'intelligenza artificiale di nome Alicia T., dotato di un sistema di interfaccia conversazionale, con il quale era persino possibile avere una discussione e porre domande su qualsiasi cosa. E sempre più spesso è annunciata la futura predominante presenza di membri non umani nei C.d.A. delle grandi imprese con funzioni specifiche (*Assisted Artificial Intelligence*), per il sostegno alle decisioni (*Augmented Artificial Intelligence*), infine per sostituire la decisione umana (*Autonomous Artificial Intelligence*).

La *Digital Criminal Compliance* riuscirebbe a rendere i meccanismi migliori e più efficienti, sia nella fase di costruzione del modello (ci ricordano, G. Morgante, G. Fiorinelli, "Promesse e rischi della compliance penale digitalizzata", in *Archivio penale*, (2022), n. 2, p. 9, come "le tecno-

cosi standardizzare i modelli di prevenzione 231[11] con, ad esempio[12], l'implementazione di *software* che controllino il rispetto delle soglie limite previste dalla normativa di settore; o meccanismi di *allert* che registrino la mancata trasmissione di dati di inquinamento alle autorità competenti in caso di superamento dei valori soglia[13].

logie digitali possano anzitutto supportare la fase preliminare del c.d. *risk assessment*, contribuendo alla 'fotografia' del contesto e all'identificazione dei rischi, mediante la raccolta dei dati e l'analisi di tutti i dati relativi alla realtà aziendale". Parimenti quanto alla "progettazione del modello organizzativo [ovverosia, alla fase di definizione dei presidi]"), sia in quello della sua vitalità aziendale, offerta alla naturale applicazione quanto alla sua innaturale disapplicazione, entrambe registrate e controllate da un sistema matematico di tipo algoritmico, in grado di captare ed elaborare i dati in un tempo sensibilmente più rapido, e più certo, rispetto al controllo umano; e meno permeabile a condizionamenti, veri o presunti, di tipo corporativistico o particolaristico, pericolo questo ritenuto scongiurabile dal momento che "Gli standard e le norme tecniche possono contenere anche elementi di tipo valutativo e disciplinare situazioni conflittuali, che però, almeno teoricamente, vengono risolte attraverso procedure codificate e schemi della valutazione e della gestione del rischio, sulla base di tecniche statistico-probabilistiche": così V. Torre, "Compliance penale e normativa tecnica", in *Archivio penale*, (2022), n. 1, p. 9.

[11] G. Morgante, G. Fiorinelli, *op. cit.*; R. Trezza, "L'Intelligenza Artificiale come ausilio alla standardizzazione del modello 231: vantaggi 'possibili' e rischi 'celati'", in *Giurisprudenza penale*, (2021), n. 1-*bis*.

[12] L'esempio è tratto dall'articolo *on line* D. Speranzini (2022), "Intelligenza artificiale e D. Lgs. 231/2001: i sistemi automatizzati di prevenzione dei reati", in *MMSLEC*. Recuperato da https://mmslex.com/intelligenza-artificiale-e-d-lgs-231-2001-i-sistemi.

[13] La nuova legge n. 22 del 2022 in tema di reati culturali e responsabilità delle persone giuridiche aiuta a comprendere quanta parte, nei meccanismi della *compliance*, possano avere le nuove tecnologie informatiche: così per i modelli di organizzazione, gestione e controllo i quali, ispirandosi alle Linee Guida delle Nazioni Unite per la prevenzione e repressione del traffico di beni culturali del 2014 ed alle misure pre-penali di cui alla Convenzione di Nicosia, potranno aprirsi al tracciamento dell'origine degli oggetti d'arte e della catena proprietaria; alla registrazione e all'aggiornamento presso archivi elettronici di tutti i dati relativi ai beni culturali di cui sono in possesso; all'introduzione di sistemi auspicabilmente informatici di licenze di esportazione e importazione per i beni culturali; all'adozione di registri auspicabilmente elettronici delle transazioni commerciali riguardanti opere d'arte e di antiquariato, laddove non già previsti; al monitoraggio delle compravendite di questi beni su Internet, possibilmente con il coinvolgimento e la

Potremmo continuare a lungo, in un lungo elenco di attività, persino le più varie e bizzarre: cantare[14], suonare[15], dipingere, ricordare e scrivere (il correttore di questo saggio è un algoritmo), formulare romanzi, pareri legali, interpretare l'esistenza, anche di settori delle arti della creatività e persino inventare, ponendo, nuovamente, e con più senso, il dubbio sulla esistenza di una intimità cibernetica, nelle forme che la parola *coscienza* può contenere[16], capace di uscire fuori dagli schemi e abbandonarsi ad una bizzarria intelligente, ancora una volta migliore di quella umana. *Le città invisibili* avranno, ed hanno avuta, finalmente una rappresentazione magica, attraverso una stupefacente e *visibile* raffigurazione segnica.

In breve, l'intelligenza artificiale sarà un copilota della nostra vita e, di tanti aspetti della nostra vita, sarà un amplificatore del

responsabilizzazione dei gestori delle piattaforme potenzialmente utilizzate a questo fine; nella segnalazione presso le autorità competenti di attività sospette. Insomma, l'informatica, e più in genere la modernità, al servizio della *compliance*.

[14] La sapienza artistica raggiunta, inspiegabilmente e insperabilmente, dalla capacità di un sistema di generare un prodotto musicale ha visto la riproduzione della voce di Anita Mui, una cantante scomparsa 40 anni prima, riuscendo a totalizzare 100 milioni di ascolti. La produzione si deve al colosso cinese Tencent, e insieme a questa con milioni di ascolti, altre mille canzoni prodotte dalla intelligenza artificiale: così F.M. De Collibus, "L'Era delle macchine che apprendono", in *Limes*, 12 (2022), p. 24. Ci ricorda C. Cavaceppi, "L'Intelligenza artificiale applicata al diritto penale: criticità attuali e prospettive future", in G. Taddei Elmi, A. Contaldo (a cura di), *op. cit.*, p. 98, che l'impatto economico della automazione potrebbe raggiungere un livello fra il 6.500 e 12.000 miliardi di euro entro il 2025.

[15] Di recente un sistema semplice di IA elaborato dal Politecnico di Milano ha trovato il modo di predire il suono emesso da tavole di violino di forma diversa: l'IA è stata addestrata con un database di oltre 1.500 forme di tavole armoniche ispirate a violini storici (fra cui il Messia di Stradivari) ma anche a forme insolite, riuscendo così a suggerire ai liutai la forma migliore da dare allo strumento.

[16] Ci ricorda G. Taddei Elmi, "Introduzione. Dall'informatica giuridica al diritto dell'informatica", in G. Taddei Elmi, A. Contaldo (a cura di), *op. cit.*, p. XXVI, che, se la coscienza dipende dalla quantità di neuroni e di connessioni sinaptiche attivate contemporaneamente, e se il cervello possiede mille miliardi di neuroni e ciascuno di essi sviluppa mille connessioni sinaptiche per un totale di 10 alla quindicesima, si stima che una disponibilità di RAM di 10 alla quindicesima verrà raggiunta da un calcolatore presumibilmente nel 2029. Un tale calcolatore potrebbe dirsi avere "coscienza".

lavoro delle persone e delle aziende, capace di sottrarre alle persone e alle aziende i tempi di *routine*, e così aggiungere nuovi tempi, aumentare la creatività, dare spazi a nuove soluzioni e a nuove innovazioni, a nuove attitudini ingegnose: la produttività del sistema, in uno studio che ha riguardato il sistema Italia, potrà aumentare fino al 18% grazie all'adozione della intelligenza artificiale generativa, con una generazione di valore che a parità di ore lavorate, di ore cioè lavorate nelle forme tradizionali, potrà arrivare fino a 312 miliardi di euro di valore aggiunto.

Il futuro è lì, ed è impossibile rinunciarvi.

2. *The artificial game*: la mossa 37

In questo futuro non troppo distante da noi, sempre più prossimo, ormai è chiaro, l'IA cambierà le nostre vite. C'era da aspettarselo. Non è, del resto, la prima volta che le macchine hanno *vinto* contro gli umani, sorpassata l'abilità umana, come quando negli anni '90 il computer *Deep Blue* di IBM sconfisse il grande maestro di scacchi Kasparov, nella rassicurante spiegazione di una vittoria legata all'esistenza di un inventario di tutte le possibili mosse, al cui interno il computer avesse *semplicemente* identificata la mossa migliore. Era come – così si pensava – se una mossa, avendo più probabilità di successo, fosse stata programmata in *Deep Blue*, mappata in tutte le possibili scelte, lasciando al computer semplicemente la decisione di attuare la mossa vincente, la possibilità cioè di razionalizzare e semplificare ciò che l'uomo, l'intelligenza umana, aveva previamente inserito nella sua programmazione. Come dire, il computer, l'IA, faceva più rapidamente quello che, semplicemente, l'intelligenza umana gli aveva consentito di fare, un acceleratore della decisività umana in puro spirito deterministico.

La superiorità umana veniva, dunque, in qualche modo riconosciuta, in una rassicurazione del dominio umano su quello artificiale, durata almeno fino alla consapevolezza dell'inatteso: l'inatteso arriva poco dopo, nella sfida del team *Deep Mind* di Google a Lee Sedol, 18 volte campione del mondo, imbattuto nel gioco da tavolo Go, una disciplina ludica esistente da almeno 3000 anni, praticata, soprattutto in Cina, da quasi 40 milioni di persone, e considerata il più antico giuoco da tavolo ininterrot-

tamente praticato fino ai giorni nostri. Si tratta di una sorta di giuoco degli scacchi, ma con una disciplina più complessa, un reticolo di molte più caselle, con pedine bianche e nere, chiamate *pietre* che hanno tutte lo stesso valore differenziale, appunto come per giuoco degli scacchi, ma molto più complesso in quanto ogni turno ha circa 200 diverse mosse possibili, rispetto alle sole 20 disponibili con gli scacchi, dal momento che ad ogni turno si aggiunge una pedina[17] e ciò rende il numero di combinazioni possibili largamente superiore al totale di tutti gli atomi che esistono nell'universo.

Per questo tipo di giuoco, è fisicamente e umanamente impossibile creare delle istruzioni, da cui poi il computer possa ricavare la mostra vincente, l'unico modo per giocare a questo gioco è usare l'intuizione e la creatività, piuttosto che la memoria o le formule matematiche. La sfida nel 2016 fu vinta da *AlphaGo*, la AI di Deep Mind, addestrata su trenta milioni di giuochi diversi, che sconfisse il campione del mondo imbattuto Lee Sedol quattro volte a uno, nello stupore degli osservatori, e non tanto per il numero delle vittorie, quanto per il fatto che l'intelligenza artificiale avesse vinto usando la cosiddetta e famosa *mossa 37* della seconda partita, il fatto, cioè, che l'intelligenza artificiale avesse pensato e deciso di mettere una pedina sulla quinta linea, scegliendo una mossa che nessun giocatore professionista avrebbe mai fatto all'inizio di un match. La *mossa 37* apparve, agli umani, come una mossa insolita, bizzarra, fantasiosa, che muoveva da una intuizione creativa, che fuoriusciva dallo schema deterministico della intelligenza artificiale, e si collocava in una dimensione nuova, una sorta di autodeterminazione in cui la macchina, autonomamente, riservava a sé la decisione di come giocare, come cioè trarre dalle innumerevoli mosse osservate e sperimentate, la possibilità di imparare autonomamente dei propri errori, fino ad arrivare al punto di inventare e creare nuove mosse, mai viste prima.

Tremila anni di antichissima perizia erano messi in discussione da *AlphaGo*, successivamente battuta da *AlphaGo Zero*,

[17] Traggo queste informazioni da un bel saggio di M. Di Michele, dal titolo *Intelligenza artificiale. Etica, rischi e opportunità di una tecnologia rivoluzionaria*, Diarkos, Santarcangelo di Romagna, 2023.

un sistema nuovo di zecca, ma ancora più avanzato, ignaro delle regole del giuoco o di giuochi passati. Ciononostante, neanche a dirlo, vincente.

3. Le parole per dirlo

Quello che era apparso chiaro, nel giuoco Go, era come l'intelligenza artificiale potesse andare oltre le capacità degli esseri umani, anche nello svolgimento di attività complesse, senza ricevere sottostanti istruzioni, senza alcuna nozione di concetti base, senza nessuna programmazione, senza, cioè, che nessun essere umano avesse disciplinato e controllato e prevenuto la bizzarria del gesto ludico. L'idea rassicurante, umanamente rassicurante, che le macchine non sarebbero mai state in grado di sviluppare comportamenti, ruoli o funzioni tipiche dell'intelligenza umana – ciò che il giuoco degli scacchi sembrava in qualche modo testimoniare – divenne di colpo un'idea sorpassata, ponendo alla attenzione umana l'esistenza di nuove possibilità, ma anche di nuove pericolosità.

Nell'immagine del cambiamento, fra le nuove consapevolezze, bisognava dunque aggiungere anche una modernità di carattere terminologico, un rinnovamento del linguaggio, l'idea di dover superare o abbandonare, andare oltre la parola *intelligenza artificiale*[18], più volte usata fino a questo momento, e cominciare a pensare non più alla IA come una *semplice* branca dell'informatica, incaricata dello sviluppo di quelle tecniche in grado di simulare le capacità cognitive degli esseri umani, comunque al di sotto, o, al limite, affianco, ma mai sopra l'intelligenza umana, in termini cioè meramente deterministici, sulla base di regole predefinite, decodificate all'interno di un programma; semmai, invece, cominciare a pensare alla intelligenza artificiale – possiamo continuarla a chiamare così, per comodità semantica – nella più compiuta e sviluppata *machine learning*, un sistema di

[18] Per IA – a rigore anche della definizione contenuta nella proposta di regolamento dell'aprile del 2021 – deve intendersi "un software sviluppato con una o più delle tecniche elencate nel I allegato della proposta e che può, per un dato insieme di obiettivi definiti dall'uomo, generare output come contenuti, previsioni, raccomandazioni o decisioni che influenzano gli ambienti con cui interagiscono".

apprendimento dove è possibile, non soltanto simulare, ma generare autonomamente modelli statistici, tramite l'analisi delle banche dati incredibilmente estese e prese attraverso il dialogo fra sistemi esperti. La differenza è molto più di quanto la parola non riesca ad esprimere: le regole statistiche a cui la *machine learning* fa affidamento sono ricreate dal programma stesso, agevolando uno sviluppo ulteriore e dirompente dell'intelligenza artificiale tradizionalmente intesa, in modo da poter autonomamente riconoscere e stabilire regole che poi consentano di raggiungere gli obiettivi che il sistema artificiale intende perfezionare. Né più né meno che una forma di apprendimento, che consente alla macchina di progettarsi da sola, per ciò che non ancora realizzato, attraverso la creazione di modelli statistici che vengono continuamente aggiornati, attualizzati a quelle che sono le variabili apprese dall'esterno e interiorizzate nel sistema. Sempre meglio, e sempre più nel profondo della conoscenza.

Nel passaggio al *machine learning* e, successivamente, al *deep learning*, grazie a questa capacità di autoproclamazione, l'intelligenza artificiale supera la pura logica deterministica e diviene capace di risolvere problemi complessi in modo creativo e inaspettato, può prendere iniziative, può – e si è verificato nel 2013 sempre in una partita, questa volta al videogioco Tetris – decidere di adottare un comportamento sleale, pur di vincere la partita. Nella sospensione del giuoco, come ci ricorda la pellicola *Wargames*, il giuoco artificiale incontra la paura di perdere e decide di superarla attraverso l'artificio dell'imbroglio.

Cambia il nome delle cose, verrebbe di dire, l'internet delle cose, la *cartografia dei cambiamenti* si arricchisce di nuove parole: l'IA è sempre più un *Intelligenza Generativa*, prossima a divenire una *Superintelligenza*, antagonista, utopica, completamente priva di innocenza, indifferente alle relazioni domestiche che uniscono le parti in un tutto, risolutamente dedita alla parzialità, all'intimità e alla perversità[19]. E in questo intreccio di umano ed artificiale, di organico e cibernetico, di mente e corpo, autosviluppo e progettazione, nulla appare uguale a prima,

[19] Per una efficace descrizione del *cyborg* si rinvia a D.J. Haraway, nel saggio *Manifesto cyborg. Donne, tecnologie e biopolitiche del corpo*, trad. it., Feltrinelli, Milano, 2018, p. 42.

nemmeno nelle parole: ingannevoli per natura[20], nella sospensione delle parole, il *meticcio* assomiglierà più ad un *fatticcio*[21], e l'*attore*, forse, più ad un *attante*[22], in una sublimazione di una semantica ibrida, intessuta di incertezze e di nuovi significati.

4. "Ho paura, Dave"[23]

Occorre in qualche modo rassegnarsi o prendere consapevolezza del fatto che, dopo le rivoluzioni industriali, la prima

[20] La cortina di suggestioni retoriche trova spazio nelle parole che avvolgono l'intelligenza artificiale e la sua narrazione: in disparte quelle cinematografiche, cedevoli ad una lettura antropomorfizzata del sistema artificiale, l'inganno è nella parola intelligenza, nell'idea di oracoli, di allucinazioni, di incantesimi, parole queste che conducono – nel pensiero di G. Finocchiaro, *Intelligenza artificiale. Quali regole?*, il Mulino, Bologna, 2024, p. 25 – alla dimensione del mito ed allontana le soluzioni giuridiche sul tema delle responsabilità.

[21] L'espressione *fatticcio* – mutuata da B. Latour, *Il culto moderno dei fatticci*, trad. it., Meltemi, Milano, 2017, p. 70: "Unendo le due fonti etimologiche, – poco prima precisa che la parola 'fatto' rinvia alla realtà esterna, la parola 'feticcio' alle folli credenze del soggetto (ndr.) – noi chiamiamo fatticcio la robusta certezza che permette alla pratica di passare all'azione senza mai credere alla differenza tra costruzione e raccoglimento, immanenza e trascendenza" – ci consente di lasciare spazio alle due opposte teorie che immaginano, sul piano della soggettività delle macchine, la piena capacità di pensare e per essa lo status di persona (c.d. SAI nella intuizione di A. Turing), ovvero all'opposto l'assenza di una soggettività che renderebbe le macchine pari a delle cose, magari indistinguibili nei comportamenti, ma oggetti sul piano morale e giuridico (c.d. WAI secondo la tesi di J.R. Searle). In mezzo la teoria c.d. negazionista (c.d. NAI) secondo cui nelle macchine non vi sarebbe alcuna capacità equiparabile al pensare o alla intelligenza umana: sul punto A. Contaldo, F. Campara, *op. cit.*, p. 79. Il rimescolamento della nozione di "soggetto" è sottolineato anche da A. Garapon, J. Lassègue, *op. cit.*, p. 251.

[22] La felice espressione risale a B. Latour, nel saggio del 1999 e ripresa da G. Teubner nel saggio *Ibridi e attanti. Attori collettivi ed enti non umani nella società e nel diritto*, trad. it., Mimesis, Milano, 2015, pp. 18 ss. Ce la ricorda anche A. Cappellini, "Machina delinquere non potest? Brevi appunti su intelligenza artificiale e responsabilità penale", in *Criminalia*, 27/03/2019, p. 9, nota 24.

[23] Nel film di Stanley Kubrick, *2001. Odissea nello spazio*, la paura è quella di HAL, la macchina che riconosce i suoi errori, le sue decisioni discutibili, e percepisce l'intenzione di David di farla cessare, spegnerne il funzionamento. La paura nel film del 1968 è quella *della* macchina, la paura

quella dei motori a vapore, la seconda dell'energia elettrica, e la terza delle tecnologie digitali, siamo alle soglie di una quarta rivoluzione industriale, dove le macchine aiuteranno gli uomini a fare calcoli complessi, senza neanche richiedere più le istruzioni, ma sulla base dei soli obiettivi, per il cui raggiungimento l'intelligenza artificiale sarà libera di scegliere il modo migliore possibile, riuscendo peraltro a fare meglio di quanto l'intelligenza umana non riuscirà a fare.

Non è possibile ignorare come l'affacciarsi del mondo cibernetico possa aggiungere una serie di rischi ai rischi che intende monitorare, e una serie di preoccupazioni, tra cui *bias* eventuali nell'*output* dei modelli, errori legati agli *input* (ad esempio una discriminazione razziale, o culturale)[24], problemi connessi alla gestione e protezione dei dati personali, pericoli di *cyber security*. Molto più che errori – tipo quello che portò l'algoritmo a censurare, ritenendola pornografica, una campagna antimurale al seno; o l'errore che portò a multare un personaggio popolare, il cui volto era ritratto su un autobus in movimento – il rischio che preoccupa la modernità artificiale – in disparte quello cinematografico del cosiddetto *complesso di Frankenstein*[25] – è quello di una autodeterminazione sfuggente, una proiezione finalistica anomala, attuata attraverso modelli comportamentali anomali.

In questa rincorsa al nuovo e al meglio, alla capacità della IA di migliorarsi e fare meglio ciò che l'intelligenza umana o la stupidità umana sa fare peggio, sarà ben possibile che, nella migliore decisione, trovi spazio quella che a noi potrebbe apparire la peggiore decisione possibile dal punto di vista etico; che l'autonoma e libera decisione si riveli disallineata con la tavola dei valori umani; che, una volta ammessa la libera scelta, nella sua attitudine al meglio, possa trovare spazio la possibilità, il rischio, di una decisione intimamente illecita, in altre parole, la possibilità di scegliere e completare una decisione illecita per

del nuovo è all'inverso, quella *dalla* macchina. Sulle paure del nuovo, dal nuovo, G. Finocchiaro, *op. cit.*, p. 15.

[24] I rischi delle nuove tecnologie su alcuni aspetti del vivere sociale, in particolare sul ruolo e sulla funzione delle donne, sono evidenziati da D.J. Haraway, *op. cit.*

[25] L'idea, cioè, spesse volte evidenziata nella rappresentazione cinematografica, di una macchina che si ribella al suo creatore.

il solo fatto di essere ritenuta la migliore possibile: una sorta di ossimoro, rischio-non-calcolato, ogniqualvolta almeno l'intelligenza artificiale, pur di raggiungere l'obiettivo finale, possa decidere di assumere un comportamento sleale, in un disallineamento dei valori che potrebbe nel futuro ripetersi sempre più frequentemente.

Il passaggio dalla IA debole (IA), alla IA forte (o AGI) fin su alla Superintelligenza conduce ad una trasmutazione dei vecchi rischi in nuovi rischi, un ammodernamento delle rischiosità, oltre il pensabile, oltre cioè la logica deterministica. Occorre, dunque, immaginare e imporre un approccio responsabile all'intelligenza artificiale, sapendo però che la mappatura dei rischi è complessa, tanto quanto l'individuazione di un rimedio che agisca al punto esatto in cui il rischio trova il suo spazio identitario e la sua attuazione mortifera, consapevoli che la penombra dei livelli algoritmici complica l'individuazione delle cause; entrambi rendono difficile la ricerca delle responsabilità.

5. Precop: sorvegliare e punire

Ad oggi, tutti i tentativi di disciplinare le responsabilità della persona artificiale si sono orientati in due direzioni: da un lato prevenire i rischi di attività illecite; dall'altro, disciplinare le conseguenze – prevalentemente civilistiche – di una attività illecita compiuta da un sistema artificiale.

L'approccio dell'UE all'intelligenza artificiale – in forme diverse da quello cinese, e dalla auto normazione statunitense[26] – si è mosso in entrambe le direzioni, e l'idea di dover e poter disciplinare lo sviluppo della IA e garantirne la sicurezza e il

[26] Di entrambi dà conto G. Finocchiaro, *op. cit.*, pp. 45-46 e nella parte conclusiva del saggio: di particolare interesse, per la riflessione penalistica che ne consegue, l'approccio statunitense, affidato all'autoregolazione dei grandi gruppi economici (Amazon, Anthropic, Google, Inflection, Meta, Microsoft e OpenAI) che scelgono di adottare le regole di disciplina. La condivisione delle regole con il mercato, da un lato conferma la rincorsa dei modelli regolativi alle innovazioni offerte dal mercato, e dunque il ruolo del mercato come principale *selfresolving* delle criticità evidenziate dal mercato medesimo, dall'altro la difficoltà per il mercato di pensare di dover sanzionare se stesso, quanto almeno all'applicazione di misure differenti da quelle civilistiche.

rispetto dei diritti fondamentali, attraverso norme e azioni concrete, ha conosciuto numerose norme e azioni, che compongo un quadro di cura e sottostante preoccupazione: fra le tante[27], in

[27] Questo il timeline delle iniziative europee, dal 2018 ad oggi: *Comunicato stampa: Gruppo di esperti sull'IA e alleanza europea sull'IA* (marzo 2018); *Comunicato stampa: Intelligenza artificiale per l'Europa*; *Comunicazione: Intelligenza artificiale per l'Europa*; *Documento di lavoro dei servizi della Commissione: Responsabilità per le tecnologie digitali emergenti*; *Dichiarazione di cooperazione in materia di intelligenza artificiale* (aprile 2018); *Lancio dell'alleanza europea per l'IA*; *Costituzione del gruppo di esperti ad alto livello sull'IA* (giugno 2018); *Commissione europea: Piano coordinato sull'IA*; *Commissione europea (comunicazione stampa): AI made in Europe; Comunicazione della Commissione europea: AI made in Europe*; *Consultazione delle parti interessate sul progetto di orientamenti etici per un'IA affidabile* (dicembre 2018); *Comunicazione della Commissione europea: Costruire fiducia nell'intelligenza artificiale incentrata sull'uomo*; *Gruppo di esperti ad alto livello sull'IA: Linee guida etiche per un'IA affidabile* (aprile 2019); *Prima Assemblea dell'Alleanza europea per l'IA*; *Gruppo di esperti ad alto livello sull'IA: Raccomandazioni politiche e di investimento dell'IA* (giugno 2019); *Gruppo di esperti ad alto livello sull'IA: Sperimentazione dell'elenco di valutazione dell'IA affidabile* (dicembre 2019); *Commissione europea: Libro bianco sull'IA: un approccio europeo all'eccellenza e alla fiducia*; *Consultazione pubblica su un approccio europeo all'eccellenza e alla fiducia nell'IA* (febbraio 2020); *Valutazione d'impatto iniziale: Requisiti etici e giuridici in materia di IA*; *Gruppo di esperti ad alto livello sull'IA: Elenco di valutazione finale sull'IA affidabile* (ALTAI); *Gruppo di esperti ad alto livello sull'IA*; *Raccomandazioni settoriali di un'IA affidabile* (luglio 2020); *2a Assemblea dell'Alleanza europea per l'IA* (ottobre 2020); *Commissione europea: Comunicazione sul tema Promuovere un approccio europeo all'IA*; *Commissione europea: Proposta di regolamento che stabilisce norme armonizzate in materia di IA*; *Commissione europea: piano coordinato aggiornato sull'IA, Commissione europea: Valutazione d'impatto di un regolamento sull'IA* (aprile 2021); *Consultazione pubblica sulla responsabilità civile – Adeguare le norme in materia di responsabilità all'era digitale e all'intelligenza artificiale*; *Commissione europea: Proposta di regolamento sulla sicurezza dei prodotti* (giugno 2021); *Banca centrale europea, Parere sulla legge sull'IA*; *Consiglio dell'UE: Testo di compromesso della presidenza sulla legge sull'IA*; *Conferenza di alto livello sull'IA: Dall'ambizione all'azione (3d Assemblea dell'Alleanza europea per l'IA)*; *Comitato economico e sociale europeo, Parere in merito alla legge sull'IA* (novembre 2021); *Comitato delle regioni, Parere in merito all'Atto sull'IA* (dicembre 2021); *Lancio del primo sandbox normativo per l'IA in Spagna: Portare avanti il regolamento sull'IA* (giugno 2022); *Proposta di direttiva sulla responsabilità dell'IA* (settembre 2022); *Orientamento generale del Consiglio sulla legge*

particolare, l'iniziativa dell'aprile 2021, vale a dire il proposto quadro di regolamentazione sull'intelligenza artificiale e la pertinente valutazione d'impatto. Disegnando un quadro giuridico europeo per l'IA che sappia difendere i diritti fondamentali e affrontare i rischi per la sicurezza specifici dei sistemi di IA, e che sappia attuare una revisione della legislazione settoriale in materia di sicurezza (ad esempio, regolamento sulle macchine, direttiva sulla sicurezza generale dei prodotti), l'iniziativa giuridica del 2021 tratteggia un quadro di responsabilità civile aggiornato ed adeguato all'era digitale e all'IA e modulato su quattro diversi livelli di rischio: rischio minimo, rischio elevato, rischio inaccettabile e rischio specifico di trasparenza. Introduce inoltre norme specifiche per i modelli di IA per uso generale.

La Proposta di regolamento è stata preceduta dalla Comunicazione della Commissione al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle Regioni (COM(2021) 118 final), diffusa il 9 marzo 2021. Facendo seguito alla strategia digitale del febbraio 2020 (Commissione europea, *Libro bianco sull'intelligenza artificiale – Un approccio europeo all'eccellenza e alla fiducia* (COM(2020) 65 final), la Commissione europea ha delineato il programma per la politica digitale per il prossimo decennio. In particolare, è prevista una struttura di *governance* condivisa con gli Stati membri basata su un sistema di monitoraggio annuale da sviluppare intorno ai seguenti quattro punti cardinali: 1) cittadini dotati di competenze digitali e professionisti altamente qualificati nel settore digitale; 2) infrastrutture digitali sostenibili, sicure e performanti; 3) trasformazione digitale delle imprese: entro il 2030 tre imprese su quattro dovrebbero utilizzare tecnologie digitali, tra cui il 5G, l'Internet delle cose, l'*edge computing*, l'intelligenza artificiale, la robotica e la realtà aumentata per sviluppare nuovi prodotti, nuovi processi di fabbricazione e nuovi modelli commerciali basati su un'equa condivisione dei dati nell'economia dei dati; 4) digitalizzazione dei servizi pubblici, compresi i si-

sull'IA (dicembre 2022); *Posizione negoziale del Parlamento europeo sulla legge sull'IA* (giugno 2023); *Accordo politico sulla legge sull'IA raggiunto dai colegislatori* (dicembre 2023); *Pacchetto sull'innovazione dell'IA per sostenere le start-up e le PMI nell'intelligenza artificiale* (gennaio 2024); *Ufficio europeo per l'IA* (febbraio 2024).

stemi giudiziari e, in particolare, le attività di indagine e di contrasto in modo da far fronte a reati digitali sempre più sofisticati.

L'IA è senza dubbio una risorsa, ma anche una preoccupazione costante: da ultimo, nel gennaio 2024 "la Commissione – si legge nella pagina della Commissione europea – ha adottato la comunicazione AI@EC, che delinea strategie per migliorare le capacità proprie della Commissione nel campo dell'intelligenza artificiale (IA), sottolineando nel contempo l'importanza di un uso sicuro, trasparente e incentrato sull'uomo delle tecnologie di IA. Gli orientamenti (inclusi nella comunicazione), incoraggiano la Commissione ad adattare internamente, innovare e adottare l'IA fin da subito per dare un esempio di buone pratiche"[28], facendosi precedere, appena un mese prima, nel dicembre 2023, dopo una "maratona" di colloqui di 3 giorni fra la presidenza del Consiglio e i negoziatori del Parlamento europeo, da un accordo provvisorio sulla proposta relativa a regole armonizzate sull'intelligenza artificiale (IA), il cosiddetto regolamento sull'intelligenza artificiale. Il progetto di regolamento mira ad assicura-

[28] L'IA è certamente un destino a cui l'UE non intende sottrarsi, convinta che promuovere l'eccellenza nell'IA rafforzerà il potenziale dell'Europa di competere a livello mondiale e che dunque debba consentirne lo sviluppo e l'adozione, lasciandola prosperare dal laboratorio al mercato; debba garantire che l'IA funzioni per le persone ed agisca come per il bene nella società; costruire una leadership strategica in settori ad alto impatto. Questi alcuni passaggi della comunicazione della Commissione europea sulla visione strategica per promuovere lo sviluppo e l'uso di sistemi di intelligenza artificiale (AI) legali, sicuri e affidabili nella Commissione europea (AI@EC): la Commissione si impegna a sostenere, accelerare e promuovere lo sviluppo e l'adozione di tecnologie AI affidabili all'interno della Commissione e, ove possibile, con partner del settore pubblico e privato, inclusi startup e innovatori. La Commissione segue un approccio basato sul rischio, stabilendo requisiti per i sistemi AI che pongono rischi elevati per la sicurezza e i diritti fondamentali e vietando alcune pratiche che contravvengono ai valori dell'UE. La Commissione si prepara a rispettare le regole previste dal futuro regolamento sull'intelligenza artificiale (*AI Act*). La Commissione introduce azioni organizzative e operative per garantire la capacità istituzionale e operativa di sviluppare e utilizzare tecnologie AI affidabili. Queste azioni riguardano la governance interna, le linee guida operative, la formazione del personale, la priorizzazione dei progetti e la cooperazione con altre istituzioni e Stati membri. La Commissione presenta lo stato dell'arte dei sistemi AI esistenti e previsti nella Commissione, identificando le aree in cui l'AI può essere più vantaggiosa per il personale e i risultati, e le capacità orizzontali che possono essere riutilizzate in vari settori e aree di applicazione.

re che i sistemi di IA immessi sul mercato europeo e utilizzati nell'UE siano sicuri e rispettino i diritti fondamentali e i valori dell'UE, fornendo regole sui modelli di IA per finalità generali ad alto impatto che possono comportare rischi sistemici in futuro, nonché sui sistemi di IA ad alto rischio; rivedendo il sistema di *governance* con alcuni poteri di esecuzione a livello dell'UE; ampliando l'elenco dei divieti, ma con la possibilità di utilizzare l'identificazione biometrica remota da parte delle autorità di contrasto negli spazi pubblici, fatte salve le tutele; offrendo una migliore protezione dei diritti tramite l'obbligo per gli operatori di sistemi di IA ad alto rischio di effettuare una valutazione d'impatto sui diritti fondamentali prima di utilizzare un sistema di IA. L'idea principale è, dunque, quella di regolamentare l'IA sulla base della capacità di quest'ultima di causare danni alla società seguendo un approccio "basato sul rischio": tanto maggiore è il rischio, quanto più rigorose sono le regole. Se da un lato i sistemi di IA che presentano solo un rischio limitato diventano soggetti a obblighi di trasparenza molto leggeri (ad esempio rendere noto che il contenuto è stato generato dall'IA, affinché gli utenti possano prendere decisioni informate in merito all'ulteriore utilizzo), d'altro canto, i sistemi di IA ad alto rischio vengono assoggettati ad una serie di requisiti e obblighi per ottenere accesso al mercato dell'UE (come ad esempio l'obbligo di registrarsi alla banca dati dell'UE per i sistemi di IA ad alto rischio; o di informare le persone fisiche quando sono esposte ad un sistema di riconoscimento delle emozioni; valutazione di conformità aggiornata alle novità tecnologiche) e tanto più nelle catene del valore complesse, dove si avverte maggiormente l'esigenza di rendere trasparenti l'assegnazione delle responsabilità e i ruoli dei vari attori in tali catene, in particolare dei fornitori e degli utenti di sistemi di IA. Per usi a rischio definito inaccettabile, l'elenco dei divieti contempla la manipolazione comportamentale cognitiva, lo *scraping* non mirato delle immagini facciali da Internet o da filmati di telecamere a circuito chiuso, il riconoscimento delle emozioni sul luogo di lavoro e negli istituti di istruzione, il punteggio sociale, la categorizzazione biometrica per dedurre dati sensibili, quali l'orientamento sessuale o le convinzioni religiose, e alcuni casi di polizia predittiva per le perso-

ne, con tuttavia possibili eccezioni a fini di attività di contrasto, permeabili ad usi di IA ad alto rischio che non abbiano superato la procedura di valutazione della conformità ma per particolari e disciplinate esigenze di contrasto (ad esempio la tutela di vittime di determinati reati, la prevenzione di minacce reali, presenti o prevedibili, come attacchi terroristici, e la ricerca di persone sospettate dei reati più gravi).

Si tratta, certamente, del tentativo organico di attuare una nuova architettura di *governance*, non priva di indici di operatività concreti, fra i quali la creazione di un ufficio per l'IA all'interno della Commissione, incaricato di supervisionare i modelli di IA più avanzati, e contribuire a promuovere norme e pratiche di prova e far rispettare le norme comuni in tutti gli Stati membri e supportato da un gruppo scientifico di esperti indipendenti in merito ai modelli di IA per finalità generali, contribuendo allo sviluppo di metodologie per valutare le capacità dei modelli di base, fornendo consulenza sulla designazione e l'emergere di modelli di base ad alto impatto e monitorando i possibili rischi materiali di sicurezza connessi ai modelli di base.

Buone prassi, forum consultivi aperti ai portatori di interessi (rappresentanti dell'industria, PMI, start-up, società civile e mondo accademico) ma non solo: ecco la seconda direttrice, l'idea cioè che la complessità della modernità debba imporre nuove strategie e nuove cautele, ma anche nuove sanzioni, e cioè sanzioni pecuniarie per le violazioni del regolamento sull'IA in percentuale del fatturato annuo globale nell'esercizio finanziario precedente della società che ha commesso il reato o, se superiore, in un importo predeterminato (35 milioni di EUR, o il 7% per le violazioni relative ad applicazioni di IA vietate, 15 milioni di EUR o il 3% per violazioni degli obblighi del regolamento sull'IA e 7,5 milioni di EUR o l'1,5% per la fornitura di informazioni inesatte. Con possibilità di massimali più proporzionati per le sanzioni amministrative pecuniarie per le PMI e le start-up in caso di violazione delle disposizioni del regolamento sull'IA) e poteri di segnalazione quale espressione della strategia europea di *whistleblowing* da parte sia di una persona fisica che di una persona giuridica.

Sorvegliare e punire: la strategia della legislazione europea è, in sintesi, quella di aumentare la sorveglianza e prevedere san-

zioni patrimoniali, per comporre un quadro di *relativa tranquillità*, in cui l'ideazione, la costruzione, la vitalità di un sistema di IA si conferma *rischiosamente irrinunciabile*.

6. La colpevolezza artificiale

Tranquillità-relativa e rischiosità-irrinunciabile, sono queste le ragioni dell'inquietudine logico dommatica, che danno corpo alle ragioni del dubbio: la prima, sistemica, giacché il quadro dei divieti non trova applicazione per aree sensibili, quali attività di ricerca, sviluppo e prototipazione che precedano l'immissione sul mercato; o per aree interessate esclusivamente a scopi militari, di difesa o di sicurezza nazionale, indipendentemente dal tipo di entità che svolge tali attività. La linea di confine potrebbe non essere robusta ed impedire elusioni di sistema e, del resto, se anche così non fosse, la nuova validazione di conformità in caso di modifiche sostanziali, o la nuova conformità ai nuovi requisiti al fine di ridurre al minimo i rischi, sembrano riproporre l'elasticità di già conosciuti sintagmi normativi – una sorta di *massima sicurezza tecnologicamente possibile* – di non facile interpretazione: inevitabile che sia[29] la regolamentazione nascerebbe provvisoria, naturalmente cagionevole, inesorabilmente affidata alla lettura del caso, nel momento storico in cui *quel caso* viene ad esistenza, come dire, in una impredicibile distanza dalla irretroattività della disciplina. La seconda, umanamente o, forse meglio, artificialmente inevitabile: qualsiasi disegno organizzativo, qualsiasi disciplina normativa non può escludere attività fraudolente per aggirare le previsioni normative e la possibilità di inganno, ben oltre la soglia del malinteso.

Potremmo dirla così: nell'artificio dell'intelligenza, ben oltre il *giuoco della imitazione*[30], permane uno spazio per la capacità

[29] Lo scarto fra regolamentazione giuridica e dinamicità in progress trova conferma nella stessa disciplina europea, la proposta di Regolamento europeo per l'intelligenza artificiale, dettata per una visione delle cose immediatamente messa in crisi dall'irruzione del ChatGPT, ciò che rese necessario una immediata rivisitazione del testo elaborato per consentire di includervi la neonata applicazione della intelligenza generativa: G. Finocchiaro, *op. cit.*, p. 50.

[30] Nella versione offerta da Alan Turing, nel saggio "Computing Machinery and Intelligence", in *Mind*, 59 (1950), pp. 433-460, riproposto ora in D.R. Hofstadter, D.C. Dennett (a cura di), *L'io della mente*, trad. it., Adelphi,

di *pensare*, fra cui anche assumere *pensieri* criminali[31]. Resta da capire quanto grande sia questo spazio, dal momento che l'illecito, che la vitalità artificiale custodisce, non è solo l'illecito della costruzione, della gestione del sistema, ma possiede natura poliedrica, sino a coinvolgere beni tradizionalmente presidiati non dalle sole sanzioni civilistiche, altrove ritenute non sufficienti. In sostanza dinanzi all'abuso di un punteggio sociale, per finalità pubbliche e private; allo sfruttamento delle vulnerabilità delle persone, all'utilizzo di tecniche subliminali; alla identificazione biometrica remota in tempo reale in spazi accessibili al pubblico, alla categorizzazione biometrica delle persone fisiche sulla base di dati biometrici per dedurne o desumerne la razza, le opinioni politiche, l'appartenenza sindacale, le convinzioni religiose o filosofiche o l'orientamento sessuale; alla polizia predittiva su singoli, al riconoscimento delle emozioni sul luogo di lavoro e negli istituti di istruzione; alla estrazione non mirata di immagini facciali da internet o telecamere a circuito chiuso per la creazione o l'espansione di banche dati, al rischio specifico per la trasparenza – per restare solo ad alcuni dei rischi immaginabili – si aggiunge un ventaglio dei rischi talmente ampio da essere paradossalmente incalcolabile, almeno per l'umana predittività.

Ben oltre quelli attesi – come quello offerto alla attenzione dei giudici statunitensi chiamati a dirimere una causa intentata

Milano, 1985, pp. 61 ss., il gioco consiste nel chiudere in due stanze separate un uomo e una donna, lasciando che siano interrogati da un terzo interrogante attraverso una qualche apparecchiatura telescrivente. L'interrogante può rivolgere domande a entrambe le stanze, senza però sapere chi le occupi, e dunque chi fornirà la risposta alla sua domanda. Lo scopo del giuoco è comprendere, attraverso le risposte, in quale stanza vi sia la donna e in quale l'uomo, sapendo che la donna, con le sue risposte, cercherà di aiutare l'interrogante, mentre l'uomo farà invece di tutto per confondere l'interrogante, dando risposte che, a suo parere, potrebbe dare la donna, e così tentare di trarlo in inganno. La versione aggiornata sostituisce l'uomo con un computer, cercando di riflettere sulla capacità del computer di sostituirsi e imitare i comportamenti – e cioè le risposte – dell'uomo.

[31] Ovviamente tutto dipende da cosa si intenda per pensare, cosa significhi comportarsi in modo intelligente e quanto il pensiero esprima questa intelligenza: J. Kaplan, *Intelligenza artificiale. Guida al futuro prossimo*, LUISS University Press, Roma, 2017.

da Getty Images contro Stability IA in materia di tutela dei diritti d'autore sulle fotografie sui segni distintivi[32] o, in genere, le violazioni della privacy, dei diritti di autore, le geolocalizzazioni indesiderate, gli eccessi dei dispositivi bellici autonomi (c.d. *Lethal Autonomous Weapons*: LAWS), le incognite della guida autonoma, e così di seguito – o quelli apparentemente bizzarri (un sistema antropomorfo schiavizzato e schiavizzante, un *pet-robot* aggressivo, un *care-robot* che maltratta la persona anziana recalcitrante alla assunzione di un farmaco, una istigazione al suicidio come quella verificatasi in Belgio, l'identificazione errata di un professore completamente estraneo agli abusi sessuali ai quali invece un *chatbot* lo aveva accusato, inventando articoli di giornale mai pubblicati)[33], preoccupano i rischi inattesi, legati alla scelta autonoma degli obiettivi disallineati per via di sotto-obiettivi privi di linee guida (un obiettivo di vendita attraverso una pubblicità ingannevole) o, e forse più probabilmente, quelli legati alla causalità del caso, alla casualità del gesto artificiale: senza troppo cedere all'ingegneria affabulatoria si immagini un sistema di difesa domiciliare che, nell'intrusione di un ladro, sopravanzi i limiti della legittima difesa; o un misura farmacologica errata; un investimento o iniziativa economica azzardata portatrice di una crisi aziendale, etc.

[32] La controversia giudiziaria fra Getty Imagen, Inc. e Stability AI Ltd, riguardava l'accusa a Stability di aver violato, tramite l'apparato di IA "Stable Diffusion" le norme sui diritti di autore sulle fotografie e i segni distintivi appartenenti a Getty Images, di cui si sarebbero abusivamente appropriate. Circa 12 milioni di fotografie realizzate da Stability AI con un business concorrente con quello della agenzia fotografica con la banca dati più grande del mondo. L'algoritmo aveva in sostanza elaborato/trasformato le immagini derivate dal patrimonio di proprietà intellettuale della azienda di Seattle, e rese simili a quelle originarie (c.d. opere trasformative).

[33] Al caso di Jonathan Turley, professore di diritto della Georgetown University, incluso nella lista dei docenti accusati di avances e citato in un inesistente e mai pubblicato articolo del Washington Post, va aggiunto il caso di Brian Hood, neoeletto sindaco della contea australiana di Hepburn Shire, accusato di corruzione a causa di una deduzione errata che ChatGPT aveva formulato confondendo il suo lavoro in banca e la sua denuncia di casi di corruzione; o il caso di Mark Walters, conduttore radiofonico, accusato di una appropriazione di denaro sulla base di un errata ricostruzione compiuta sempre da ChatGPT. Le vicende sono narrate da G. Finocchiaro, *op. cit.*, p. 62.

Non ci pensiamo spesso, probabilmente ingannati dalla suggestione cinematografica di una superintelligenza, semmai malevola, ma non erronea. Eppure, il caso spesse volte traduce una indifferenza sovrana delle cose, una sorta di scaltro genio che prova piacere alle coniugazioni insolite, e *in tal caso* nessuna causalità oggettiva (e dunque nessuna valutazione cautelare) potrebbe mai porsi a base di una prevenzione immaginifica, tanto quanto sa esserlo un evento non ancora reale, e per giunta casualmente prodotto da un pensiero artificiale. Il caso stanca l'intelligenza umana, e rende impotenti la prevenzione e la regolazione delle cose artificiali, deprivate dei normali coefficienti di regolarità umana, estranei alla logica del maleficio artificiale.

L'accidente non è da meno, introduce una equipollenza che, distante dal caso, potrebbe tuttavia rendersi parimenti irrazionale, animata da una concatenazione che il *senno del poi* non avrebbe potuto contenere prima, senza mai riuscire ad avvicinarsi al determinismo delle cose umane. Un destino, potremmo dire, qualcosa in più dell'avvenimento indisciplinato e caotico, tuttavia ancor più sfuggente alla concatenazione deterministica. Nell'intelligenza artificiale si annida un raziocinio magico che non consente previsione, lungimiranza, alcuna anticipazione, ma solo, unicamente, una strategia fatale che disegna un ossimoro dogmatico al quale il sistema non può offrire riparo: l'intelligenza umana dovrebbe contenere l'analisi, la riflessione, la previsione delle concatenazioni fatali e delle coniugazioni segrete, ma si tratterebbe di una lusinga regolativa oggettivamente inimmaginabile, aperta finanche a perturbazioni esponenziali che amplificano una rischiosità non iniziale, non altrimenti prevedibile prima che venga ad esistenza.

Prevedere l'imprevisto, vale quanto prevedere l'improvviso, in una fatalità complessa che il sistema pretende di poter evitare, in anticipo sul tempo di ciò che li ha causati, ma che in realtà segue un tempo del-dopo, ciò che il tempo rallentato del-prima, quello umano, meno veloce e scaltrito di quello artificiale, non potrebbe mai riuscire ad evitare né forse immaginare.

Dinanzi a queste nuove forme di rischi artificiali, tuttavia reali e concreti, la strada della punizione patrimoniale, della responsabilità oggettiva o per fatto altrui, semplicemente distribu-

ita, quanto alle conseguenze risarcitorie, all'interno dei fruitori del rischio e alla loro vicinanza ai vantaggi che detto rischio produce[34] potrebbe non essere sufficiente e reclamare l'aggiunta di una componente penalistica come in fondo per la analoga violazione di matrice umana, senza però che sia possibile sottrarre l'eventuale scelta punitiva a regole e principi irrinunciabili per il diritto penale. La scelta penalistica non sarebbe tuttavia così semplice, per la difficoltà di immaginare una responsabilità penale legata alla intelligenza artificiale, una responsabilità che, cioè, dalla *persona* artificiale scivoli verso quella umana, senza nel contempo garantire il rispetto del paradigma personologico della responsabilità penale, che necessita di essere traslato o forse rinnovato, attualizzato, nel gesto artificiale[35]. Dalla persona umana a quella artificiale, e da questa a quella, in una andata e ritorno che dovrebbe far salvi i principi di colpevolezza e personalità del gesto punitivo e che certamente non potrebbe affidarsi alla soluzione di una responsabilità oggettiva. L'alternativa non può essere fra cinismo delle regole e infedeltà ai principi.

Occorre intendersi: nell'ipotesi di un calcolatore programmato alla delinquenza, di un gesto artificiale guidato da un gesto umano, la macchina assomiglia ad uno strumento attraverso il quale l'agente umano attua la commissione di reati, dei quali egli solo – evidentemente – non può che assumersi la responsabilità[36]. Il diritto penale potrebbe affidarsi ai tradizionali e mol-

[34] L'indifferenza al dolo, alla colpa, e persino all'errore è un rimedio di impronta essenzialmente civilistico, che va incontro ai profili risarcitori alle vittime da parte, non tanto, o non soltanto, dei creatori del rischio (ciò che potrebbe non essere così semplice individuare) ma, ancor più facilmente, da parte dei fruitori del rischi, coloro cioè che hanno abitato la situazione rischiosa nelle loro dinamiche mercatali, ricavandone utilità che poi sarebbero chiamate in parte a ridistribuire nel caso di danni volontari o involontari, persino imprevedibili. Sul tema G. Finocchiaro, *op. cit.*, pp. 68 ss.

[35] I principi che una opzione penalistica dovrebbe fare salvi hanno occupata la riflessione problematica di C. Piergallini, "Intelligenza artificiale: da 'mezzo' ad 'autore' del reato?", in *Rivista italiana di diritto e procedura penale*, 63 (2020), n. 4, pp. 1745 ss.

[36] A. Cappellini, *op. cit.*, p. 4. Una sintetica rassegna dei reati che possono essere commessi attraverso l'IA è in F. Basile, "Intelligenza artificiale e diritto penale. Quattro possibili percorsi di indagine", in *Diritto penale e Uomo*, 10 (2019), p. 26, e cioè i "crimini informatici, economici ed ambienta-

teplici scenari: addebitare il reato a chi ha generato il focolaio di rischio ideando l'algoritmo (specie se *self-learning*), o applicandolo nel programmare il "robot" o il PC; a chi ha attualizzato quel rischio, producendo e mettendo in commercio il sistema; ovvero a chi ha concretamente gestito quel rischio, servendosi dello stesso, o magari cooperando con esso, ove sussistano, ad esempio, eventuali profili di *culpa in interagendo*[37].

Il problema si situa, invece, nel caso in cui una immensa organizzazione coordinata di piccoli e infinitesimali calcoli riesca a contenere un sistema di tendenze, di desideri, di convinzioni, che la macchina lascerebbe emergere in forma "autonoma", lasciando ospitare fenditure di legalità capaci di emergere in un luogo e in un tempo imprevisto e impredicibile. Questa nuova categoria del *pensiero* artificiale conoscerebbe forme di volontà, di tentativo, di speranza, o di errore, di sbaglio, di colpa, anche solo per una scelta del male minore (ad esempio nelle auto *less-driver*, il sistema potrebbe dover scegliere un danno piuttosto che un altro, un pedone esterno piuttosto che un ospite interno alla autovettura)[38].

li, i traffici internazionali di sostanze stupefacenti e di altri prodotti illeciti, la trattata di esseri umani" (sono queste, le ipotesi di reato riportate nel virgolettato, quelle menzionate nel documento di presentazione del "2019 OSCE Annual Police Experts Meeting: Artificial Intelligence and Law Enforcement: An Ally or an Adversary?") ma anche le violazioni in materia di privacy e trattamento dei dati personali, le violazioni della proprietà intellettuale ed industriale, i reati di diffamazione e le condotte di abuso della credulità popolare, magari commessi attraverso bot che creano *fakenews* destinate alla rete, etc. Infine, ricordate sempre da F. Basile, le ipotesi di bagarinaggio *on line* e di manipolazione abusiva del mercato. In proposito anche R. Borsari, "Intelligenza Artificiale e responsabilità penale: prime considerazioni", in *Media-Laws*, (2019), n. 3, p. 263. Un particolare caso di un reato commesso da un robot è ricordato da C.R. Mongioì, G. Ren, G. Zanesco, B. Ziviani nel saggio "Intelligenza artificiale e diritto penale", in *Trento BioLaw Selected Student Papers*, 4 (2019), pp. 1-10, dove si racconta come nel 1981 un impiegato in un'azienda giapponese sia stato ucciso da un robot che stava lavorando di fianco a lui; il robot lo erroneamente identificò come una minaccia al suo lavoro e lo spinse verso un macchinario in funzione lì vicino, uccidendolo.

[37] Come ci ricorda V. Manes, "L'oracolo algoritmico e la giustizia penale: al bivio tra tecnologia e tecnocrazia", in *DisCrimen*, 2020, p. 3.

[38] Può non essere così. In un esperimento condotto nel Bristol Robotics Laboratory dinanzi alla scelta di quale proteggere fra due automi in pericolo,

Non è, dunque, solo questione di una decisione che privilegi un bene a dispetto di un altro, o di una dolosa iniziativa illecita di una IA, resasi autonoma e capace di autodeterminazione illecita, l'attesa del domani sarà anche quella di una scelta che oltrepassi i cancelli del lecito, in un territorio di illegalità intelligente abitato non solo da scelte deliberate, dolose in linguaggio penalistico, ma anche semplicemente azzardate, incaute, indisciplinate, in una parola colpose. Il territorio della AGI – in disparte quello della Superintelligenza di *ExMachina*, preconizzato da *Il Pianeta delle Scimmie* – potrebbe conoscere un errore della chirurgia robotica, adiuvante o sostitutiva della mano umana, una sbagliata diagnostica computerizzata, una imprecisa rilevazione della retinopatia diabetica tramite la scansione dell'occhio, una scelta colposa nell'applicazione di una protesi, o di un farmaco piuttosto che un altro, un improprio dosaggio farmacologico deciso dall'algoritmo, un tempo rallentato dell'insulina sulla base del livello di zucchero nel sangue, o, in altre settore, una incauta o imprudente auto a guida autonoma, per restare ai settori (l'infortunistica medica e quella stradale) dove tradizionalmente si incontra il diritto penale della colpa – ai quali, come detto, potremmo aggiungere l'azzardo dell'algoritmo in una operazione commerciale o finanziaria, rivelatasi incauta e produttiva di un dissesto, ciò che l'intelligenza giuridica usa attrarre nel capitolo delle responsabilità commerciali – non governato da una supervisione umana resa improbabile, per difficoltà oggettive o anche solo per effetto mimetico e massivo, vale a dire la difficile adozione di un punto di vista individuale e non consolidato, reso improbabile dalla inanità del confronto rispetto ad uno standard decisionale artificiale e, dunque, più rassicurante e difficile da contrastare. Un po' come, per restare al settore sanitario, i protocolli medici come dispensa dalla colpa; o i modelli di organizzazione come guarentigia punitiva per le persone giuridiche. Tuttavia, errati, imprecisi, ineff-

il terzo robot a causa della "trappola etica" ovvero della impossibilità di decidere, aveva invece scelto di lasciar morire entrambi. Se ne dà conto nel saggio di I. Mercuri, "Ingegneria sociale e rischi nell'interazione uomo – robot", in G. Costabile, A. Attanasio, M. Ianulardo (a cura di), *Memberbook 2021. Digital Forensics*, IISFA, Roma, 2021, pp. 59 ss. Il primo caso di donna uccisa a seguito di un investimento da un SUV a guida automatica si è verificato in Arizona, nel marzo del 2018: ce lo ricorda C. Cavaceppi, *op. cit.*, p. 133.

ficaci, custodi a volte di errori clamorosi, sottratti al coraggio del dissenso[39]. In una parola, artificialmente colposi, ma umanamente incolpevoli, probabilmente inesigibili, e tuttavia astrattamente portatrici di responsabilità penali.

Il reato non è *con* l'intelligenza artificiale, ma è *della* intelligenza artificiale[40], il che pone il diritto penale[41], dinanzi a nuovi o sconosciuti scenari[42].

7. Dalla colpa di organizzazione alla colpa di programmazione

Non è, si potrebbe dire, in fondo, un problema del tutto nuovo, per il diritto penale, non è, cioè, la prima volta che il sistema

[39] Al coraggio del dissenso di taluni, e fra questi ci piace ricordare le critiche di Umberto Eco alle frequenti provvisorietà di Wikipedia, fa da contrappunto la pigrizia dei tanti, la gran parte di noi, nell'affidarsi a quanto si legge in rete, senza lo scrupolo di un approfondimento che non si avverte come necessario.

[40] Quello, ad esempio, raccontato da I. Salvadori, "Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale", in *Rivista italiana di diritto e procedura penale*, (2021), n. 1, p. 95 della *bot Random Darknet Shopper* che, programmata per fare acquisti *on line*, decise autonomamente di acquistare dieci pastiglie di ecstasy, poi fatte recapitare presso una galleria d'arte di Zurigo ed esposte al pubblico.

[41] L'interesse del diritto penale alla IA si muove tradizionalmente in quattro scenari, l'ultimo dei quali è quello decisamente più preoccupante. Questi gli scenari: 1) le attività di *law enforcement*, in particolare le attività di c.d. polizia predittiva; 2) i c.d. *automated decision systems*, che potrebbero in futuro conoscere un impiego anche all'interno dei procedimenti penali, sostituendo, in tutto o in parte, la decisione del giudice-uomo; 3) i c.d. algoritmi predittivi, impiegati per valutare la pericolosità criminale di un soggetto, vale a dire la probabilità che costui commetta in futuro un (nuovo) reato; 4) infine, le possibili ipotesi di coinvolgimento – come strumento, come autore, o come vittima – di un sistema di IA nella commissione di un reato) come ci ricorda F. Basile, *op. cit.*, pp. 4 ss. Ai quattro scenari, aggiunge due nuove prospettive M. Papa, "Future crimes: intelligenza artificiale e rinnovamento del diritto penale", in *Criminalia*, 2020 (*DisCrimen*, dal 04/03/2020), pp. 3 ss.

[42] Di recente I. Salvadori, *op. cit.*, pp. 83 ss. Quanto alle preoccupazioni derivate in tema di salvaguardia dei principi penalistici, si rinvia al saggio di F. Sgubbi, *Il diritto penale totale*, il Mulino, Bologna, 2019, e alla lettura che, del punto di vista del Maestro bolognese, effettua T. Guerini nel saggio "Il formante algoritmico all'alba della *Justice digital*. Uno sguardo a uno dei futuri possibili del diritto penale, dialogando a distanza con Filippo Sgubbi", in *DisCrimen*, 2021.

punitivo si trova ad affrontare l'identità colpevole di una persona inumana. Superata il macigno della visione antropomorfa della reità penalistica, il nuovo sistema della responsabilità delle persone giuridiche si è a lungo interrogato sulla costruzione di un addebito punitivo in capo ad un soggetto altro, rispetto all'autore del fatto delittuoso. Ed è esattamente questa la strada seguita dalla giuridicizzazione della IA, oneri di *governance* attraverso i quali impedire la commissione di fatti illeciti.

Per le persone artificiali o, meglio, per le responsabilità artificiali l'identificazione della colpa è tuttavia operazione dogmatica più complessa, per certi versi invertita. Non è la persona giuridica che non ha impedito la colpa della persona fisica, ma, qui, la persona fisica che non ha impedito la colpa della persona artificiale. Il nodo dogmatico è apparentemente simile: la colpa si situerebbe nel mancato impedimento, nella incapacità dell'una, la persona giuridica, o dell'altra, la persona fisica, di impedire che si commettessero illeciti, che invece, entrambi, avrebbero potuto impedire; la persona giuridica attraverso regole cautelari sustanziali della colpa di organizzazione; la persona fisica attraverso una congerie di standard costruttivi, protocolli, cautele *in second best*, che avrebbero potuto impedire la generazione artificiale di condotte illecite: una *colpa di programmazione*, assai simile alla *colpa di organizzazione*.

Semplice a dirsi, la costruzione delle cautele nasconde una complessità che supera l'apparente simmetria logico giuridica. Nelle persone giuridiche le intelligenze umane si sommano e confrontano con l'intelligenza umana che decide la commissione di un fatto illecito. Qui no, qui è diverso: l'intelligenza umana deve competere con l'intelligenza artificiale che, a tutto voler concedere, è più veloce di quella umana. È come se nella partita del facere illecito e della guida lecita si misurassero due giocatori con livelli di intelligenza differenti. La fraudolenta elusione – sintagma usato nel paradigma della 231 – qui avrebbe margini di probabilità altissimi, tanto quanto la manifesta superiore capacità dell'IA di aggirare le regole seppellite nel *black box* dell'algoritmo inziale: i metodi di contenimento, i metodi di incentivazione, l'impedimento dello sviluppo, il controllo sul *surplus* di modernità nelle forme

dell'"*obsolescenza programmata*"[43], i *tripwires*, ossia meccanismi di diagnostica random, ad insaputa del sistema[44] potrebbero non essere sufficienti e l'IA potrebbe sapientemente ingannare i guardiani e uscire dalla scatola. Ecco il punto, uno almeno fra i tanti: l'assenza, l'impossibilità di una validazione, e per essa la difficoltà nel capire e prevenire la colpa dell'intelligenza artificiale e, in essa, la tenuta della visione costituzionale della colpevolezza deprivata di effettiva personalità ed esigibilità.

Non sarebbe facile prevederlo, ma sarebbe ancora meno facile impedirlo. Nel *machine learning*, e, ancor più, nel *deep learning*, nella possibilità cioè degli algoritmi di costruirsi e migliorarsi attraverso altri algoritmi, in una sequela di strati che si aggiungono al primo strato, quello più profondo, passando per strati superiori, sempre più in alto, la complessità dei livelli intermedi corre il rischio di far smarrire l'orientamento rispetto anche all'origine genetica, mutata al livello superiore, e poi ancora più in alto, in una zona d'ombra sempre più fitta e oscura, in un *black box* che rende di fatto impossibile vedere e valutare ciò che è avvenuto all'interno dei vari algoritmi, stratificati e intrecciati in diversi livelli[45]. A quale livello si è situato e generato la deviazione dalle regole, il cominciamento illecito, stratificato e scomparso nella profondità dell'apprendimento, non è semplice, forse neppure possibile, comprenderlo nemmeno per coloro che hanno progettato l'algoritmo iniziale, sviluppatosi molto, troppo più in là, rispetto al controllo umano, impossibilitato, ad un certo punto, a monitorare e seguire lo sviluppo dell'(auto)apprendimento e della (auto)determinazione. Senza dire, poi, della complessità dell'esatta individuazione del momento e del luogo in cui collocare il disallineamento fra le due intelligenze, quella artificiale e quella progettuale umana.

Il tempo e lo spazio: le macchine moderne, i moderni congegni microelettronici quintessenziali, sono ovunque, sempre più

[43] E. La Rosa, "Obsolescenza programmata e diritto penale: una relazione complicata", in *Rivista trimestrale di diritto penale dell'economia*, 33 (2020), n. 3-4, pp. 568 ss.

[44] Per una spiegazione di questi metodi si rinvia a T. Tofanelli, "L'evoluzione della intelligenza artificiale tra etica e diritto", in G. Costabile, A. Attanasio, M. Ianulardo (a cura di), *op. cit.*, pp. 42 ss.

[45] G. Taddei Elmi, *op. cit.*, p. XVII.

invisibili, una beffa alla spiritualità umana e una sfida alla ubiquità territoriale, leggeri e precisi, portatili e mobili, l'artificio immateriale ed opaco resta difficile da vedere, in una continua sospensione tra posti lontani e indecifrabili. L'immensa *macchina metamorfica*[46] dipende dalla produzione, dai trasporti, dal lavoro fisico di persone sconosciute fra loro, da data center sparsi nell'oceano della rete, da cavi sottomarini fra continenti, da segnali di trasmissione che attraversano l'aria, da continui cicli di calcolo e, in questo perenne trasgredire i confini del qui ed ora, l'artificio uccide il tempo[47], ed abita luoghi distanti ed ubiqui[48]: entrambi, i luoghi e i tempi delle cose artificiali, impongono il balbettio del diritto penale, governato da un tempo *sudato*, intessuto nel sudario del dopo l'apocalisse del fatto illecito.

La pre-colpa, l'idea cioè di una *governance* per prevenire attività illecite, sconta questa evidente vulnerabilità logico dommatica: intuita già da attenti osservatori[49], sconta la oggettiva

[46] Così la definisce K. Crawford, *Né intelligente né artificiale. Il lato oscuro dell'IA*, il Mulino, Bologna, 2021, p. 56.

[47] Se l'uccisione del tempo può apparire esagerato, non lo è certamente il *True time* di Google, un protocollo temporale distribuito che funziona stabilendo gerarchie di attendibilità tra gli orologi locali di datacenter in modo che possano decidere come sincronizzarsi. Nella gestione del tempo di Google vi è la possibilità da parte di *True time* di gestire l'incertezza in caso di sfasatura temporale dei singoli server: se l'incertezza è grande, il tempo viene rallentato finché l'incertezza non si esaurisce, vale a dire una sorta di possibilità di rallentare il tempo, modificarlo a piacimento e unire il pianeta sotto un unico codice temporale proprietario. Né più né meo che una privatizzazione del tempo, la capacità di creare una linea temporale variabile sotto il controllo di un orologio master centralizzato. Se non ucciso, il tempo è perlomeno posseduto. La privatizzazione del tempo è raccontata da K. Crawford, *op. cit.*, pp. 89 ss.

[48] La complessità degli strumenti e delle infrastrutture, e delle maestranze si lascia facilmente intuire nell'esempio tratto da K. Crawford, *op. cit.*, p. 212: dovendo installare un sistema di riconoscimento facciale per le strade di Belgrado, il capo della Polizia ha ordinato l'installazione di duemila telecamere in ottocento luoghi della città, e a tal scopo il governo serbo stipulò un accordo con Huawei per la fornitura dei servizi di videosorveglianza, supporto della rete 4G e di centri di dati e di controllo unificati. Il sistema nasceva cioè strutturalmente ibrido ed ubiquo, con infrastrutture cinesi, ma è quanto ordinariamente avviene anche con infrastrutture indiane e statunitensi dislocate nei paesi europei.

[49] G. Contissa, F. Galli, F. Godano, G. Sartor nel saggio "Il Regolamento europeo sull'intelligenza artificiale. Analisi informatico-giuridica", in *I-lex*, 2021, n. 3.

difficoltà, forse persino impossibilità, per l'intelligenza umana di esercitare la sorveglianza sui sistemi di IA, in particolare su quelli ad apprendimento automatico, capaci di compiere operazioni e prendere decisioni a tal punto complesse da sorpassare la capacità umana di controllo, "eccedendo non solo i loro limiti cognitivi, ma anche quelli temporali nell'accedere alle informazioni, elaborarle e prendere decisioni".

Il giudizio umano è, in sostanza, largamente sottodimensionato rispetto alla complessità dei modelli che guidano le scelte dei sistemi artificiali, il che rende poco plausibile pensare che una intelligenza umana, per quanto vivace, possa evitare l'elusione fraudolenta di un modello posta in essere da una intelligenza artificiale; che possa prevedere e scoprire l'astuzia fraudolenta di un sistema straordinariamente più agile e veloce; che un controllore umano possa controllare un controllato artificiale. Nondimeno, che un errore determinato da un sistema artificiale, sulla base di calcoli complicati e profondi, possa essere mentalmente anticipato ed evitato da un semplice controllo di concordanza, affidato ad una mente semplice a cui si affiderebbe la coraggiosa e forse inesigibile funzione di dissentire da una mente superiore. Anche quando l'uomo dovesse mantenere formalmente il controllo sulla decisione finale, la possibilità che egli sia in grado di entrare efficacemente nel merito della decisione si tradurrebbe in un'ipotesi remota, e più verosimilmente, il supervisore umano tenderà a fare affidamento sulle decisioni di un sistema di IA, tanto più se questo è stato certificato, a meno che non abbia motivi specifici per ritenere che esso sia malfunzionante, o che non sia in grado di incorporare nella sua valutazione ulteriori elementi, esterni al sistema stesso.

Nel giuoco degli addebiti, e nella dissoluzione dell'organico, il tempo dell'intelligenza umana è quello di ieri, quello cibernetico è il tempo dell'istante, un tempo nomade, del qui ed ora. Nell'attesa del giorno dopo, il ritardo punitivo è pressoché inevitabile.

8. La costruzione delle regole. La vigilanza artificiale

L'affido al diritto penale non è così scontato anche per altre ragioni, agli ostacoli epistemologici, vanno infatti aggiunte

le ostilità tecnocratiche, l'IA come luogo del potere, tradizionalmente recalcitrante ai controlli e ancor più alle punizioni. Decisamente ostile a quelle penalistiche e più propenso ad un disegno di guarentigia penale, e a costrizioni, laddove davvero necessarie, semplicemente risarcitorie. Semmai dovesse non essere così, semmai il sistema dovesse aprirsi a regole punitive, la futura colpevolezza artificiale – ammesso che sia possibile chiamarla così – imporrà nuove regole e nuove misure. L'analisi dei cicli di vita del sistema, una sequela di verifiche continue, la possibilità di mettere a riposo il sistema non già – anche qui una necessaria novità sistemica – per l'obsolescenza del macchinario, ma, tutto al contrario, per la estrema e sopraggiunta modernità del macchinario, resosi capace di tale e tanta autonomia da non essere più controllabile. Ma tutto questo potrebbe non bastare, e certamente non per il diritto penale.

Dolo della macchina? oppure cooperazione colposa nel fatto doloso? errore determinato da fatto altrui? dolo di chi? e colpa in che modo? e *altrui* in che senso? Per non dire del *dove* e del *quando*, a cui si aggiunge, potrebbe aggiungersi – semmai dovessimo spingerci verso una personalità cibernetica – la complessità della scelta del gesto punitivo, ipoteticamente e anacronisticamente retributivo, sino al paradossale recupero di una bizzarra morte cibernetica: il diritto penale artificiale è tutto da costruire attraverso una forza immaginativa che deve saper cogliere il nuovo, anche a costo di scoprire la propria debolezza epistemologica e maturare l'idea della rinuncia, una fuga dalla padronanza punitiva.

Negli step a venire, man mano che ci si avvicinerà alla superintelligenza, aumenteranno i pericoli di un controllo incontrollabile, tuttavia nei primi momenti di questa era artificiale occorrerebbe provare ad adattare il diritto punitivo a nuovi standard di colpevolezza umanizzante, sapendo però che le tradizionali categorie non possono funzionare, non si adattano alla modernità artificiale, e impongono un cambio di paradigma, che sappia apprendere diversamente il *qui* e l'*altrove*, l'*attore* e l'*attante*, il *fatto* e il *fatticcio*, il *proprio* e l'*altrui*. In breve, adattarsi ad una nuova e inattesa topografica delle ambiguità.

E qui il paradosso è dietro l'angolo. Se si vuole davvero insistere per una responsabilità penale, e se non si vuole abbandonare l'idea di una responsabilità personale – ciò che il sistema risarcitorio mostra di prediligere – la migliore e più efficace forma di controllo non potrà che essere quella dell'IA che controlla se stessa, in un *myse en abyme* dove il controllore, controlla il controllato, finché un altro controllore, più evoluto, non controllerà il controllore, in un seguito senza fine, tanto quanto l'inevitabile rincorsa alla modernità artificiale. La personalità della responsabilità artificiale non può che essere, per il momento, quella di non essere riusciti ad impedire che l'IA faccia quello che *avremmo potuto* impedire di fare e in questo nuovo, semmai futuribile, *diritto penale sinestetico*, per dirla con Michele Papa, la personificazione potrebbe assomigliare ad una evidente finzione, la colpa assumere una dimensione ibrida, a metà strada fra il dolo dell'attante e la evitabilità dell'attore. E il cuore del diritto penale potrebbe essere l'errore, l'ultimo della serie, non necessariamente quello causativo dell'evento.

Quanto durerà questo "*avremmo potuto*" è la vera incognita del futuro che *ci* attende. Nel *ci attende*, nella speranza cioè di una prossimità non a noi estranea, si annida l'immanenza di un *presto*, che potrebbe diventare un *subito*. Ma forse anche un *mai*, come per i Topi di Autun[50], giacché il diritto penale dell'intelligenza – occorre ricordarlo – dovrebbe poter risiedere nel tempo del *prima*. E non è detto che possa riuscirvi e, se così fosse, tanto varrebbe rinunciarci.

[50] Nell'anno 1522 i topi furono citati dinanzi il tribunale di Autun in Borgogna, accusati di aver mangiato e distrutto il raccolto del circondario. Non doveva essere un fatto nuovo: Bartolomeo Chassenee, giurista del luogo chiamato a patrocinare la causa, ricordò la maledizione divina contro il serpente, la maledizione di Gesù contro l'albero dei fichi sterile di Bretania, le scomuniche a passeri, lumache, anguille e persino ad un intero frutteto. Alla fine, il tribunale decise di rinviare la sentenza fino a quando non vi fosse certezza che i topi avessero avuta notifica della causa, e così i topi, di fatto, vinsero la causa.

Lorenzo Pulito

Previsioni algoritmiche e valutazioni prognostiche per la sostituzione della pena nella prospettiva della *symbiotic artificial intelligence*[*]

Abstract: **Algorithmic Predictions and Prognostic Evaluations for Penal Replacement in the Perspective of Symbiotic Artificial Intelligence.** New "intelligent" entities aim to cure the fallacy of intuition about the offender's future behavior and to support the judge in this problematic task. This paper aims to guide the metamorphosis process and to prevent the enthusiasm for something new from ending in an unconscious enslavement of the human being to the algorithmic regime. It proposes a paradigm of interaction between "intelligences" inspired by symbiotic relationships in nature (e.g. between aphids and ants) and provides a concrete example in the context of the new alternative sentences and criminal procedures recently introduced by the so-called "Cartabia reform".

Keywords: Risk Assessment, Sentencing, Alternative Penalties, Symbiotic Artificial Intelligence.

Sommario: 1. Nuovi equilibri nell'"ecosistema processuale" – 2. Razionalità limitata, potenzialità computazionali e *output* predittivi – 3. Gli "afidi": i *risk assessment tools*. L'algoritmo predittivo COMPAS – 4. Gli strumenti predittivi PSA e SAVRY – 5. I *risk and need assessment tools* – 6. Le nuove pene sostitutive come "formiche" nella relazione simbiotica con gli afidi – 7. Spazi processuali per la sperimentazione di una relazione simbiotica – 8. Riflessioni conclusive.

1. Nuovi equilibri nell'"ecosistema processuale"

L'intelligenza artificiale (d'ora innanzi anche solo AI) evoca sconvolgimenti epocali e scenari distopici[1].

[*] Questo lavoro è stato parzialmente sostenuto dal progetto FAIR – Future AI Research (PE00000013), nell'ambito del programma MUR del PNRR finanziato dal NextGenerationEU.

[1] S. Quattrocolo, "Per un'intelligenza artificiale utile al giudizio penale", in *BioLaw Journal-Rivista di BioDiritto*, (2021), n. 2, p. 388, rileva che la riflessione sul rapporto tra giustizia penale e strumenti di intelligenza artificiale deve essere condotta "cercando di depurare il proprio approccio da ogni distopia e aprendo il campo alla comprensione di concetti propri delle scienze computazionali".

Gli approcci oscillano tra la demonizzazione (di chi paventa un rischio di estinzione dell'umano da mitigare) e l'idolatria (di chi, altrettanto ingenuamente, fa di tali tecnologie un assoluto). A tratti, tali visioni non mancano di attraversare anche l'"ecosistema" del processo penale, che, a seconda dello scenario sposato, lo si presagisce depravato del suo principale protagonista[2] oppure, all'inverso, traboccante di un efficientismo che, fino ad adesso, è stato soltanto il miraggio di reiterate sedicenti riforme.

La delicatezza della sfida è percepibile, ma tali visioni manichee rischiano di essere nient'affatto proficue: l'esponenziale incremento della complessità dei sistemi algoritmici e della dipendenza da questi dovrebbe piuttosto impegnare gli studiosi nel tradurre i principi giuridici di alto livello in protocolli di progettazione, nel ricercare soluzioni di governance del fenomeno e nuovi equilibri relazionali tra "intelligenze"[3].

Se il processo penale è un "fenomeno di vita della realtà giuridica, cioè è la stessa realtà giuridica vista quale fenomeno vitale [ed è] informato a fondamentali principi morfologici"[4], come tutti i fenomeni vitali della realtà, allora non appare fuori luogo un riferimento al micromondo animale. Dove esistono alcune interazioni interspecifiche simbiotiche, come quella tra gli afidi e le formiche, che possono efficacemente ispirare la costruzio-

[2] Secondo S. Lorusso, "La sfida dell'intelligenza artificiale al processo penale nell'era digitale", in *Sistema penale*, 28/03/2024, p. 7, "un primo scenario, che si può definire 'catastrofico-apocalittico'", è "caratterizzato dall'oscuro presagio della fine del processo penale come noi attualmente lo conosciamo, di un medioevo prossimo venturo della 'giurisdizione degli uomini', di una giustizia terrena irrimediabilmente soggiogata all'AI, preda di un'entità digitale 'superiore' che ne detti tempi, modi e risultati".

[3] In tale ottica, vedi R.E. Kostoris, "Intelligenza artificiale, strumenti predittivi e processo penale", in *disCrimen*, 05/03/2024, pp. 1 ss. Sia pur con riferimento ai principi etici, v. J. Morley, L. Floridi, L. Kinsey, A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices", in *Sci Eng Ethics*, (2020), n. 4, pp. 2141 ss. Nel senso di ritenere che gli algoritmi siano il catalizzatore di un cambiamento che riguarda il futuro del diritto, N. Lettieri, "Law in Turing's Cathedral. Notes on the Algorithmic Turn of the Legal Universe", in W. Barfield (ed.), *The Cambridge Handbook of the Law of Algorithms*, Cambridge University Press, Cambridge, 2020, p. 719.

[4] G. Foschini, *Giudicare ed essere giudicati*, Giuffrè, Milano, 1960, p. 55 s.

ne di (più rassicuranti) paradigmi di interazione uomo-macchina applicabili nell'ambito della giustizia penale.

2. Razionalità limitata, potenzialità computazionali e *output* predittivi

È noto come *bias* cognitivi e decisionali abbiano un ruolo determinante sui processi valutativi umani[5]: la razionalità del giudice può essere incrinata da diversi fattori, sia endogeni che esogeni[6].

I primi sono riconducibili alla struttura del cervello umano, la cui architettura consente di identificare scorciatoie e trappole mentali, che semplificano le operazioni di giudizio, ma portano alla caduta nei *bias*, generati dall'intreccio tra il sistema cognitivo "veloce" e quello "lento". Errori nello sviluppo del giudizio umano possono talora essere causati anche dal "rumore".

I secondi sono fattori esterni alla mente del giudice, come i vari apparati di comunicazione (mass media o social network), che ne influenzano il convincimento.

In un ambito altamente discrezionale dell'attività umana, come quello della giustizia penale, la letteratura è focalizzata a riflettere sul possibile ruolo di queste tecnologie, per capire come possano contribuire a restituire una certa efficacia a un sistema che ha perso, agli occhi dei cittadini, credibilità e affidabilità.

Caduto il mito della razionalità olimpica[7], l'aspirazione di superare l'uso di euristiche fuorvianti da parte del giudice ha aperto la riflessione sull'uso di algoritmi predittivi per sostenerlo nella decisione o nella valutazione prognostica della pericolosità dell'individuo[8].

Strumenti di giustizia predittiva promettono risparmio di tempi e risorse, neutralità, obiettività, ed appaiono persino al-

[5] G. Canzio, "Intelligenza artificiale e processo penale", in *Cassazione penale*, (2021), n. 3, p. 798.

[6] P. Felicioni, "L'attività valutativa del giudice tra ragione ed emozione", in G.M. Baccari, P. Felicioni, *La decisione penale tra intelligenza emotiva e intelligenza artificiale*, Giuffrè, Milano, 2023, p. 13.

[7] A. Forza, "Le scienze comportamentali ed il loro contributo nello studio dei processi decisionali", in G.M. Baccari, P. Felicioni, *op. cit.*, p. 42.

[8] P. Felicioni, *op. cit.*, p. 26.

lettanti per i giudici, nell'ottica di un alleggerimento dal peso della responsabilità di decidere[9], soprattutto in un ambito in cui il giudizio è probabilistico e l'incertezza è ontologica[10].

Del resto, l'*AI Act* (nel testo approvato dal Parlamento europeo il 13 marzo 2024) correla la definizione di *AI system* ad *outputs such as predictions* ed enfatizza, nel quarto considerando, i vantaggi ottenibili dall'uso di tali sistemi di intelligenza artificiale, quali il miglioramento delle previsioni, anche in materia di sicurezza e giustizia.

Numerosi sono gli output di tipo predittivo richiesti in diversi segmenti del processo penale[11]. Certamente, proprio le valutazioni sulla pericolosità sembrerebbero le più inclini ad una gestione algoritmica[12]. L'uso di sistemi di intelligenza artificiale predittiva – e in particolare di approcci di *machine learning* – può aiutare ad ottimizzare tali processi di valutazione, ma solleva diverse sfide, sia sul piano tecnico, che su quello etico e giuridico, con ampie implicazioni per gli individui e la società.

3. Gli "afidi": *i risk assessment tools.* L'algoritmo predittivo COMPAS

I c.d. *risk assessment tools* hanno conosciuto una rilevante diffusione negli Stati Uniti d'America, dove vengono adoperati per aiutare il giudice ad assumere decisioni in tutte le fasi del processo: in quella cautelare (*pre-trial decisions*), in quella de-

[9] Su tale aspetto V. Manes, "L'oracolo algoritmico e la giustizia penale: al bivio tra tecnologia e tecnocrazia", in U. Ruffolo (a cura di), *Intelligenza Artificiale – Il diritto, i diritti, l'etica*, Giuffré, Milano, 2020, p. 557.

[10] In tal senso, cfr. G. Mazza, S. Tognazzi, "La valutazione della pericolosità *ante* e *post delictum*", in G.M. Baccari, P. Felicioni, *op. cit.*, p. 192.

[11] Per una panoramica, cfr. L. Notaro, "Intelligenza artificiale e giustizia penale", in A. Massaro (a cura di), *Intelligenza artificiale e giustizia penale*, Paruzzo editore, Caltanissetta, 2020, pp. 109 ss.

Sul tema degli algoritmi predittivi del rischio di recidiva vedi M. Caianiello, "Dangerous Liaisons. Potentialities and Risks Deriving from the Interaction between Artificial Intelligence and Preventive Justice", in *European Journal of Crime, Criminal Law and Criminal Justice*, (2021), n. 1, pp. 1 ss.

[12] G. Mazza, S. Tognazzi, *op. cit.*, p. 193.

cisoria (*sentencing*) e in quella esecutiva (*parole decisions*)[13], dove in particolare tali software servono per sciogliere prognosi di pericolosità sociale e rischio di recidivanza[14].

Proprio come gli afidi, che possono causare danni alle colture, questi sistemi algoritmici potrebbero parimenti "infettare" i delicati processi valutativi e decisionali tipici del giudizio penale. I fattori patogeni sono diversi: discriminatorietà e opacità dei risultati algoritmici, effetto *anchoring*, dubbia validità predittiva per mancanza di una adeguata teoria scientifica di supporto o difettosa traduzione della stessa in un modello computazionale[15]. L'esperienza d'oltreoceano ha consentito di metterli a fuoco, ma non sempre la profilassi o le terapie risultano adeguate.

Il più noto strumento attuariale di valutazione del rischio, ampiamente utilizzato nella prassi giudiziaria statunitense, è COMPAS (*Correctional Offender Management Profling for Alternative Sanctions*)[16].

[13] L. D'Agostino, "Gli algoritmi predittivi per la commisurazione della pena", in *Diritto penale contemporaneo*, (2019), n. 2, p. 356.

[14] In questi termini S. Quattrocolo, "Equo processo penale e sfide della società algoritmica", in *BioLaw Journal-Rivista di BioDiritto*, (2019), n. 1, p. 142.

[15] Su questi problemi e, in particolare, sulle questioni che la costruzione di uno strumento di rischio attuariale solleva sull'accuratezza, l'uguaglianza e lo scopo della punizione, cfr. J.M. Eaglin, "Constructing Recidivism Risk", in *Emory L. J.*, (2017), n. 1, pp. 59 ss.

[16] Del *tool* se ne sono occupati, sotto diversi profili, molti autori: senza pretesa di esaustività, cfr. F. Basile, "Intelligenza artificiale e diritto penale: quattro possibili percorsi di indagine", in *Diritto Penale e Uomo*, (2019), n. 10, pp. 19 ss.; T. Brennan, W. Dieterich, B. Ehret, "Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System", in *Criminal Justice and Behavior*, (2009), n. 1, pp. 21 ss.; B. Galgani, "Considerazioni sui 'precedenti' dell'imputato e del giudice al cospetto dell'IA nel processo penale", in *Sistema penale*, (2020), n. 4, pp. 84 ss.; A. Simoncini, "Diritto costituzionale e decisioni algoritmiche", in S. Dorigo (a cura di), *Il ragionamento giuridico nell'era dell'intelligenza artificiale*, Pacini, Pisa, 2020, pp. 46 ss.; A.L. Washington, "How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate", in *The Colorado Technology Law Journal*, (2019), n. 1, pp. 1 ss.; S.X. Zhang, R.E.L. Roberts, D. Farabee, "An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures", in *Crime & Delinquency*, (2014), n. 2, pp. 167 ss.

La valutazione del rischio è effettuata attraverso algoritmi statistici che lo quantificano attribuendo un punteggio. Quest'ultimo scaturisce sulla base di molteplici *data points*, che includono fattori storico-statici (come i precedenti penali, l'età del primo arresto, la partecipazione ad associazioni criminose) e fattori dinamico-criminogenici (come la stabilità residenziale, lo stato occupazionale, i legami con la comunità, l'abuso di sostanze, l'inclusione e le relazioni sociali, la composizione familiare), oltre che sulla base delle risposte fornite a centotrentasette domande a scelta multipla[17]. COMPAS ha due modelli di rischio primari: recidiva generale e recidiva violenta. La scala generale del rischio di recidiva viene utilizzata per prevedere nuovi reati; la scala del rischio di recidiva violenta si concentra sulla probabilità di crimini violenti, quali omicidio, stupro, rapina e aggressione aggravata. I risultati della scala COMPAS vengono trasformati in punteggi decili, suddivisi in gruppi di uguali dimensioni. In particolare, i punteggi in decili da 1 a 4 sono etichettati come rischio "basso"; da 5 a 7 "medio"; e da 8 a 10 "alto"[18].

Tale *tool* condensa numerose problematiche, prima tra tutte quella relativa al carattere discriminatorio dei suoi risultati algoritmici, secondo quanto emerso nell'inchiesta realizzata per conto dell'organizzazione *ProPublica*[19]. Il pregiudizio non deriva dalla considerazione della razza fra le variabili con finalità predittiva, bensì "dalla correlazione tra fattori di rischio *associati* ad una certa provenienza etnica", che scaturisce dalla "rappresentazione

[17] Divise in cinque macro-aree (*criminal involvement*, *relationship/lifestyle*, *personality/attitudes*, *family and social exclusion*): vedi D. Kehl, P. Guo, S. Kessler, *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, Berkam Klein Center for Internet & Society, Harward Law School, 2017, p. 11. Sul punto cfr. anche M. Montagna, "Prognosi personologica, commisurazione della pena e applicazione di misure di sicurezza", in G.M. Baccari, P. Felicioni, *op. cit.*, p. 236.

[18] F. Lagioia, R. Rovatti, G. Sartor, "Algorithmic fairness through group parities? The case of COMPAS-SAPMOC", in *Ai & Society*, (2023), n. 2, p. 463.

[19] Si v. lo studio condotto da J. Angwin, J. Larson, S. Mattu, L. Kirchner (2016), "How We Analyzed the COMPAS Recidivism Algorithm", in *ProPublica.* Recuperato da https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Secondo l'inchiesta, l'algoritmo discrimina gli imputati neri, per i quali la propensione alla recidiva risultava pari al doppio di quella calcolata per i bianchi.

stereotipata di un gruppo sociale, frutto di un condizionamento umano difficile da estirpare, [che] inficia l'intero processo di *risk assessment*, producendo un circolo vizioso che si autoalimenta"[20].

La notorietà del software si lega soprattutto alla circostanza di essere stato l'oggetto di un *leading case*, culminato con la sentenza emessa dalla Corte Suprema del Wisconsin[21], che ha aperto il dibattito sulle condizioni di accettabilità degli output predittivi con il *right to a fair trial*, il *right to be sentence on accurate information*, nonché il *right to an individualized sentence.*

La Corte, oltre a non condividere le considerazioni difensive sulla natura discriminatoria della valutazione[22], ha respinto le doglianze relative alla violazione di tali diritti sul rilievo decisivo che le risultanze del *risk assessment tool* non fossero state valutate acriticamente dal giudice, bensì corroborate da altri elementi[23].

[20] G. Padua, "Intelligenza artificiale e giudizio penale: scenari, limiti e prospettive", in *Processo Penale e Giustizia*, (2021), n. 6, p. 1500 s. L'A. si interroga anche sulle soluzioni (sia *ex ante* che *ex post*) volte a correggere tali distorsioni.

[21] State v. Loomis, 881 N.W.2d 749 (Wis. 2016). Con riguardo al caso vedi S. Quattrocolo, "Quesiti nuovi e soluzioni antiche? Consolidati paradigmi normativi vs rischi e paure della giustizia digitale 'predittiva'", in *Cassazione penale*, (2019), n. 4, pp. 1750 ss. Per la puntuale ricostruzione della vicenda cfr. anche S. Carrer, "Se l'*amicus curiae* è un algoritmo: il chiacchierato caso Loomis alla Corte Suprema del Wisconsin", in *Giurisprudenza penale (web)*, (2019), n. 4; C. Fiorio, "Predizione algoritmica e giurisdizione di sorveglianza", in G.M. Baccari, P. Felicioni, *op. cit.*, pp. 253 ss.; M. Montagna, "Prognosi personologica, commisurazione della pena e applicazione di misure di sicurezza", cit., p. 237.

[22] Il COMPAS è "una pericolosa *bias machine* foriera di pesanti discriminazioni di tipo soprattutto razziale" secondo P.P. Paulesu, "Intelligenza artificiale e giustizia penale. Una lettura attraverso principi", in *Archivio penale (web)*, 2022, n. 1, p. 10. Sul punto, la difesa di Loomis aveva contestato la legittimità costituzionale della condanna, in quanto l'accertamento di responsabilità fondava su valutazioni che sopravvalutavano il genere maschile di appartenenza. La Corte Suprema statale ha respinto le eccezioni sollevate sotto questo profilo, argomentando che, nel caso di specie, non era dato evincere alcun passaggio motivazionale da cui poter desumere che l'asserita discriminazione di genere avesse determinato un innalzamento del punteggio base di rischio. Anche per tali aspetti, si rinvia a L. D'Agostino, *op. cit.*, pp. 362 ss.

[23] C. Parodi, V. Sellaroli, "Sistema penale e intelligenza artificiale: molte speranze e qualche equivoco", in *Diritto penale contemporaneo*, (2019), n. 6, p. 69 s.; S. Quattrocolo, "Quesiti nuovi e soluzioni antiche? Consolidati

L'applicazione, in materia penale, di uno standard "debole" della intelligenza artificiale, improntato al criterio della non esclusività del dato algoritmico per la decisione[24], non elide del tutto le perplessità, perché è lecito dubitare della capacità del giudice di "dominare" l'*output* predittivo senza subirne le "suggestioni"[25]. Questo tema incrocia il nodo dell'opacità di funzionamento dell'apparato algoritmico, dovuta essenzialmente all'impossibilità di accedere ai codici sorgente del *risk assessment* digitale, coperti da *trade secret*[26].

paradigmi normativi *vs*. rischi e paure della giustizia digitale 'predittiva'", cit., p. 1760.

[24] G. Canzio, "Intelligenza artificiale, algoritmi e giustizia penale", in *Sistema penale*, 08/01/2021.

[25] G.M. Baccari, G., Pecchioli, "I.A. e giudizio sul fatto: gli strumenti di e-evidence per la cognizione", in G.M. Baccari, P. Felicioni, *op. cit.*, pp. 161 ss. Cfr. altresì M. Caianiello, *op. cit.*, p. 15; L. Maldonato, "*Risk and need assessment tools* e riforma del sistema sanzionatorio: strategie collaborative e nuove prospettive", in G. Di Paolo, L. Pressacco (a cura di), *Intelligenza artificiale e processo penale. Indagini, prove, giudizio*, Editoriale Scientifica, Napoli, 2022, p. 154; A. Simoncini, "Il linguaggio dell'intelligenza artificiale e la tutela costituzionale dei diritti", in *Rivista AIC*, (2023), n. 2, p. 15 s., il quale afferma che sul "piano pratico è facile immaginare che la decisione automatica, nel tempo, tenderà sempre più a 'catturare' la decisione umana" perché "gli automatismi finiranno inevitabilmente per godere di quella che la nudging theory chiama la forza della 'default-option'".

[26] Come mettono in evidenza G.M. Baccari, G. Pecchioli, *op. cit.*, p. 132 s., è "possibile riscontrare un basilare deficit di trasparenza rispetto al funzionamento dei sistemi di intelligenza artificiale che, per una serie ragioni, rappresenta perciò una vera e propria incognita. Non solo per gli operatori giuridici"; risulta "di cruciale peso la circostanza 'strutturale' per la quale la programmazione dei sistemi di I.A. può basarsi sull'approccio dell'apprendimento automatico, in virtù del quale l'algoritmo stesso risulta in grado di automodificarsi sulla scorta dell'ampliamento progressivo del proprio dataset di riferimento. Di conseguenza, non è assurdo che gli stessi programmatori non siano in grado di spiegare l'iter operativo 'sotterraneo' tramite il quale il software ha elaborato una determinata risposta all'input ricevuto". Su tali aspetti, tra i molti, G. Canzio, "Intelligenza artificiale e processo penale", in. G. Canzio, L. Lupária Donati (a cura di), *Prova scientifica e processo penale*, Wolters Kluwer Cedam, Milano, 2022, p. 907; C. Casonato, "Giustizia e intelligenza artificiale: considerazioni introduttive", in *BioLaw Journal*, (2021), n. 2, p. 359; G. Contissa, G. Lasagni, G. Sartor, "Quando a decidere in materia penale sono (anche) algoritmi e IA: alla ricerca di un rimedio effettivo", in *Diritto di internet*, (2019), n. 1, p. 620; F. Donati, "Intelligenza

Sul punto, la Corte Suprema statale ha ritenuto che nella vicenda concreta le informazioni processate di carattere determinante provenissero dagli apporti forniti dall'imputato durante l'interrogatorio o dai suoi precedenti penali e che, inoltre, la difesa avesse avuto modo di contestare adeguatamente l'asserito effetto illegittimo del software, attraverso le istruzioni d'uso del medesimo, rese liberamente accessibili dalla compagnia proprietaria[27]. Tuttavia, nel legittimare l'impiego dell'algoritmo, la Corte ha elaborato una *check-list* cautelativa composta da cinque avvertimenti da inserire nel *pre-sentencing report*: l'eventuale esistenza di un segreto commerciale che copre il software; la sua incapacità di compiere una valutazione individualizzata, essendo basato su un insieme di dati riferiti a gruppi sociali, non normalizzati rispetto alla popolazione di ciascuno Stato; la creazione dello strumento per finalità proprie della fase esecutiva, successiva alla condanna[28]; l'esistenza di dubbi, nella comunità scientifica, sull'attendibilità del modello computazionale – seppure segreto – che lo regola[29].

Proprio sull'attitudine previsionale dei risultati algoritmici resi dallo strumento[30] residua l'equivoco di fondo che a rivestire

artificiale e giustizia", in *Rivista AIC*, (2020), n. 1, p. 427; L. Maldonato, "Algoritmi predittivi e discrezionalità del giudice: una nuova sfida per la giustizia penale", in *Diritto penale contemporaneo*, (2019) n. 2, pp. 404 ss.; P.P. Paulesu, *op. cit.*, p. 4; S. Quattrocolo, "Processo penale e rivoluzione digitale: da ossimoro a endiadi?", in *mediaLaws*, (2020), n. 3, p. 127; Id., "Quesiti nuovi e soluzioni antiche? Consolidati paradigmi normativi vs rischi e paure della giustizia digitale 'predittiva'", cit., p. 1752, nonché p. 1758; Id., "Equità del processo penale e automated evidence alla luce della convenzione europea dei diritti dell'uomo", in *Revista Ítalo-Española de Derecho Procesal*, (2019), n. 1, p. 117.

[27] L. D'Agostino, *op. cit.*, p. 363.

[28] Le problematiche condensate nell'algoritmo COMPAS risultano amplificate nel caso Loomis alla luce dell'"improprio" impiego del *tool* nella fase di *sentencing*, in quanto COMPAS "è stato progettato per supportare le decisioni giudiziarie nella fase di esecuzione della pena (*for correctional purposes*)": così L. D'Agostino, *op. cit.*, p. 364.

[29] S. Quattrocolo, *Artificial Intelligence, Computational Modelling and Criminal Proceed ings. A Framework for A European Legal Discussion*, Springer, Cham, 2020, p. 159.

[30] Secondo J. Dressel, H. Farid, "The accuracy, fairness, and limits of predicting recidivism", in *Science Advances*, (2018), n. 4, p. 3, il software di

centralità nella vicenda Loomis non fosse tanto la trasposizione in termini computazionali della teoria criminologica che ha ispirato il COMPAS, quanto piuttosto la validità scientifica della teoria stessa[31].

4. Gli strumenti predittivi PSA e SAVRY

Questa fondamentale distinzione concettuale tra dimensione scientifica della teoria sociocriminologica che ispira lo strumento di valutazione del rischio, da una parte, e (corretta ed efficace) traduzione della teoria in modello computazionale, risulta più evidente analizzando altri strumenti predittivi, quali PSA (*Public Safety Assessment*) e SAVRY (*Structured Assessment of Violence Risk in Youth*).

Il primo è lo strumento di valutazione del rischio pre-processo più popolare negli Stati Uniti: usato in fase di *pre-trial* per assistere il giudice nella decisione sul rischio di rimettere in libertà l'accusato arrestato prima della definizione del processo, deve il suo "successo" più alla crisi dell'istituto del *bail* (cauzione pecuniaria), che all'affidabilità dello strumento.

Oltre alla diversa fase di utilizzo, ciò che distingue questo software – che considera nove fattori, tra cui l'età dell'individuo al momento dell'arresto, l'accusa pendente e la storia criminale – è che non necessita di un colloquio con l'accusato, in quanto estrapola le informazioni da archivi e registri, e dunque da atti già in possesso delle autorità pubbliche[32]: l'indice "*Failure To Appear*" viene fornito sulla base di fattori di rischio quali i carichi pendenti, le condanne pregresse, le precedenti mancate

valutazione del rischio COMPAS non è più accurato o corretto delle previsioni fatte da persone con poca o nessuna esperienza in materia di giustizia penale.

[31] S. Quattrocolo, "Sui rapporti tra pena, prevenzione del reato e prova nell'era dei modelli computazionali psico-criminologici", in *Teoria e Critica della Regolazione Sociale*, (2021), n. 1, p. 279.

[32] Le informazioni possono essere estratte dal certificato del casellario giudiziale e dei carichi pendenti, nonché da altre informazioni generali, note alle autorità pubbliche.

comparizioni in giudizio negli ultimi due anni nonchè prima degli ultimi due anni[33].

A parte i timori di pregiudizi razziali[34] che sono stati espressi anche in relazione a questo *tool*, a destare perplessità è la sua stessa validità predittiva: quest'ultima sembra discendere dalla vastità del *data set* sulla cui base è stato creato, piuttosto che dal suffragio, in letteratura, della dimostrazione empirica di effettiva rilevanza, in termini di rischio di mancata comparizione dell'accusato all'udienza, dei fattori presi in considerazione per la valutazione della propensione dell'individuo alla fuga[35].

Il secondo, SAVRY[36], è uno strumento professionale strutturato per la valutazione del rischio dinamico nei minori,

[33] Lo strumento è illustrato da B.J. Brittain, L. Georges, J. Martin, "Examining the Predictive Validity of the Public Safety Assessment", in *Criminal Justice and Behavior*, (2021), n. 10, p. 1432. I nove fattori presi in considerazione sono i seguenti: "(a) age at current arrest, (b) current violent offense, (c) pending charge at the time of arrest, (d) prior misdemeanor conviction, (e) prior felony conviction, (f) prior violent conviction, (g) prior failure to appear in the past 2 years, (h) prior failure to appear older than 2 years, and (i) prior sentence to incarceration".

[34] Come si legge in K. Patrick (2018), "Arnold Foundation to Roll Out Pretrial Risk Assessment Tool Nationwide", in *InsideSources*. Recuperato da https://insidesources.com/arnold-foundation-to-roll-out-pretrial-risk-assessment-tool-nationwide.

[35] In questi termini S. Quattrocolo, "Sui rapporti tra pena, prevenzione del reato e prova nell'era dei modelli computazionali psico-criminologici", cit., p. 277. Tuttavia, studi più recenti e – soprattutto – indipendenti da Arnold Ventures, che ha creato PSA, hanno messo in luce una (almeno parziale) validità predittiva dello strumento, pur confermando la necessità di ulteriori ricerche e approfondimenti: cfr. ancora B.J. Brittain, L. Georges, J. Martin, "Examining the Predictive Validity of the Public Safety Assessment", in *Criminal Justice and Behavior*, (2021), n. 10, p. 1444 s.; M. DeMichele, P. Baumgartner, M. Wenger, K. Barrick, M. Comfort, "Public safety assessment: Predictive utility and differential prediction by race in Kentucky", in *Criminology & Public Policy*, (2020), n. 2, pp. 409 ss.

[36] Consiste in un inventario strutturato di giudizi professionali che mira a stimare il rischio di comportamenti violenti negli adolescenti di età compresa tra 12 e 18 anni: a riguardo, v. R. Borum, H.P.B. Lodewijks, P.A. Bartel, A.E. Forth, "The Structured Assessment of Violence Risk in Youth (SAVRY)", in K.S. Douglas, R.K. Otto (eds.), *Handbook of violence risk assessment (2nd ed.),* Routledge/Taylor & Francis Group., New York, 2021, pp. 438 ss. A riguardo, cfr. anche A.T.H. Kleeven, M. de Vries Robbé, A. Popma, "The Validity of Violence Risk Assessment in Young Adults: A Comparative

che, pur utilizzando specifici *dataset*, è comunque comprovato da un parere clinico dell'esperto[37].

È composto da ventiquattro fattori di rischio appartenenti a tre domini (fattori di rischio storici, fattori di rischio sociali/contestuali e fattori di rischio individuali/clinici) nonché da altri sei appartenenti ad un dominio di fattori protettivi; ogni *risk factor* ha una scala di valutazione a tre punti con linee guida di valutazione specifiche (bassa, moderata o alta), mentre ogni *protective factor* è valutato dicotomicamente come presente o assente[38].

Sebbene il SAVRY rappresenti indubbiamente uno strumento prezioso per l'*evidence-based assessment* e per l'adozione delle scelte trattamentali dei minori in conflitto con la legge, esistono ancora molte questioni aperte in relazione alla sua validità predittiva[39].

L'attendibilità della teoria che ispira il *risk assessment*, a prescindere da quale sia l'oggetto della valutazione che il giudice è chiamato a rendere, deve essere verificata – come per qualsiasi altro metodo scientifico – secondo gli standard Daubert, come

Study of Juvenile and Adult Risk Assessment Tools", in *Emerging Adulthood*, (2023), n. 6, p. 1412 s.

[37] Così A.M. Maugeri, "L'uso di algoritmi predittivi per accertare la pericolosità sociale: una sfida tra evidence based practices e tutela dei diritti fondamentali", in *Archivio penale (web)*, (2021), n. 1, p. 13, nt. 49.

[38] A.T.H. Kleeven, M. de Vries Robbé, A. Popma, "The Validity of Violence Risk Assessment in Young Adults: A Comparative Study of Juvenile and Adult Risk Assessment Tools", in *Emerging Adulthood*, (2023), n. 6, p. 1412 s. Gli A. precisano che: i fattori di rischio della scala storica sono valutati in relazione all'intero passato di una persona; i fattori di rischio della scala dinamica sociale/contestuale e individuale/clinica riflettono il funzionamento negli ultimi sei mesi; le valutazioni sui fattori protettivi riguardano gli ultimi dodici mesi.

[39] L. Barboni, A. von Hagen, S. Piñeyro, I. Senabre, "Predictive validity of the structured assessment of violence risk in youth (SAVRY) on the recidivism of juvenile offenders: a systematic review", in *Psychology, Crime & Law*, (2023), p. 3, danno atto di come una recente revisione sistematica abbia dimostrato che la capacità del SAVRY di prevedere la violenza futura varia ampiamente con dimensioni dell'effetto che oscillano da effetti di validità predittiva trascurabili a grandi ed aggiungono come sia urgentemente necessario comprende le cause di questa allarmante variabilità.

elaborati nel *leading case* Daubert v. Merrel Dow Pharmaceuticals Inc[40].

Proprio sulla base della "dottrina Daubert", la Corte suprema del District of Columbia, sezione famiglia e minori, pur senza pronunciarsi sulla validità generale dello strumento SAVRY, ha ritenuto inutilizzabili i risultati in un caso concreto, in applicazione della *rule* 702 delle *Federal Rules of Evidence*[41].

Diversamente, ove in questione non sia (o non sia soltanto) la validità generale della teoria che ispira il *risk assessment*, bensì l'attendibilità dei risultati della sua applicazione al singolo caso, diventa fondamentale comprendere come il software sia giunto ai risultati proposti. Per soddisfare tale esigenza occorre che siano "osservati standard di trasparenza nella realizzazione del *risk assessment* digitale" e che, "se il software fa ricorso a tecniche di auto-apprendimento", sia consentita "la ricostruzione *ex post* dei risultati raggiunti", non potendosi fare, in caso contrario, alcun uso processuale di algoritmi opachi[42].

[40] 509 U.S. 579 (1993). Sui parametri del "Daubert test" cfr. S. Quattrocolo, *Artificial Intelligence, Computational Modelling and Criminal Proceedings. A Framework for A European Legal Discussion*, cit., p. 163.

[41] Supreme Court of the District of Columbia, 25 marzo 2018, Judge Okun. I particolari sono riassunti da M. Montagna, "Prognosi personologica, commisurazione della pena e applicazione di misure di sicurezza", cit., p. 237 s.

[42] S. Quattrocolo, "Sui rapporti tra pena, prevenzione del reato e prova nell'era dei modelli computazionali psico-criminologici", cit., pp. 280 ss. *Ex plurimis*, leggono il problema dell'opacità algoritmica in chiave di impedimento alla piena realizzazione della parità delle parti e quale causa di indebito squilibrio processuale: C. Barbaro, "Uso dell'intelligenza artificiale nei sistemi giudiziari: verso la definizione di principi etici condivisi a livello europeo?", in *Questione giustizia*, (2018), n. 4, p. 194; L. Maldonato, "Algoritmi predittivi e discrezionalità del giudice", cit., 410; P.P. Paulesu, *op. cit.*; M. Pisati, "Indagini preliminari e intelligenza artificiale: efficienza e rischi per i diritti fondamentali", in *Processo penale e giustizia*, (2020), n. 4, pp. 957 ss.; S. Quattrocolo, "Equità del processo penale e automated evidence alla luce della convenzione europea dei diritti dell'uomo", cit., pp. 127 ss.; Id., "Quesiti nuovi e soluzioni antiche? Consolidati paradigmi normativi vs rischi e paure della giustizia digitale 'predittiva'", cit., pp. 1748 ss.; L. Romanò, "Intelligenza artificiale come prova scientifica nel processo penale: una sfida tra *machine-generated evidence* e equo processo", in G. Canzio, L. Lupária Donati (a cura di), *op. cit.*, p. 922.

5. I *risk and need assessment tools*

Particolare interesse hanno suscitato alcuni strumenti in grado di fornire una "base informativa più completa per la definizione di strategie di gestione del rischio posto dal singolo individuo", attraverso la valutazione dei fattori di rischio, dei bisogni criminogenici e della responsività del soggetto: sono i *risk and need assessment tools*, come l'*Offender Screening Tool* (OST) e l'*Ohio Risk Assessment System* (ORAS)[43].

Tali strumenti, basati sulla considerazione che la personalità non resta segnata in maniera irrimediabile dal reato commesso in passato, ma continua ad essere aperta alle prospettive di cambiamento, traducono il modello teorico *Risk-Need-Responsivity*, che abbraccia tutte le dimensioni di una previsione del futuro comportamento del reo in relazione alle sue caratteristiche individuali e alla sua capacità di reagire al trattamento[44].

Anche se sono state sollevate molte questioni giuridiche a proposito della inclusione nei punteggi di rischio di fattori come

[43] Sul cui funzionamento, cfr. E.J. Latessa, B. Lovins, J. Lux, "The Ohio Risk Assessment System", in J.P. Singh, D.G. Kroner, J.S. Wormith, S.L. Desmarais, Z. Hamilton (eds.), *Handbook of Recidivism Risk/Needs Assessment Tools*, Wiley Blackwell, Hoboken, 2017, pp. 147 ss. Tra gli RNAS rientra anche *Prisoner Assessment Tool Targeting Estimated Risk and Need* (PATTERN), sebbene il sistema valuti i rischi e le esigenze di ciascun detenuto federale. Si veda C.D. Stimson (2020), "The First Step Act's Risk and Needs Assessment Program: A Work in Progress", in *The Heritage Foundation*. Recuperato da https://www.heritage.org/sites/default/files/2020-06/LM265_0.pdf. Vedi altresì R.M. Labrecque, R. Hester, J. Gwinn, "Revalidation of the First Step Act Risk Assessment: A Test of Predictive Strength, Dynamic Validity, and Racial/Ethnic Neutrality", in *Crime & Delinquency*, 07/06/2023, pp. 1 ss.

[44] Tale quadro è stato originariamente proposto da D.A. Andrews, J. Bonta, (2007), "Risk-need-responsivity model for offender assessment and rehabilitation". Recuperato da https://www.publicsafety.gc.ca/cnt/rsrcs/rsk-nd-rspnsvty/rsk-nd-rspnsvty-eng.pdf. Cfr. anche D.A. Andrews, "The risk-need-responsivity (RNR) model of correctional assessment and treatment", in J.A. Dvoskin, J.L. Skeem, R.W. Novaco, K.S. Douglas (eds.), *Using social science to reduce violent offending*, Oxford University Press, Oxford, 2012, pp. 127 ss.; D.A. Andrews, J. Bonta, *The psychology of criminal conduct*, ed. 5, Anderson Publishing Co., New York, 2010.

l'associazione tra pari e le relazioni familiari di un imputato[45], si ritiene il loro "contributo decisivo […] sul piano informativo", in quanto occasione propulsiva per un dialogo sul trattamento sanzionatorio, in fase di commisurazione, che sia "*tailored made* sulle caratteristiche del reo" e proporzionato[46]. Di conseguenza, tali strumenti di intelligenza artificiale, capaci di mettere in luce le caratteristiche della persona a tutto tondo, si rivelano utili "per la definizione di progetti di sanzioni strutturati sulle concrete esigenze criminogeniche poste dal reo"[47], potendo fornire al giudice "uno spettro di elementi circa la personalità dell'autore di reato e le sue condizioni di vita"[48].

Onde permettere a tali *tools* di poter apportare efficacemente il proprio contributo occorrerebbero, sul piano sostanziale, un ampio ventaglio sanzionatorio e, su quello processuale, una netta cesura tra la fase di decisione sulla responsabilità e quella di commisurazione della pena. Entrambe le "strade" sono state recentemente imboccate con la c.d. "riforma Cartabia".

Da un lato, il riformatore ha puntato alle pene sostitutive; queste ultime potrebbero rappresentare le "formiche" che, agendo in simbiosi con gli "afidi" (i *tools* predittivi), potenziano la collaborazione uomo-macchina e il *teaming* sociotecnico, con relazioni reciprocamente vantaggiose, secondo quanto promette l'intelligenza artificiale simbiotica[49].

[45] B. Karp, "What Even is a Criminal Attitude? - and Other Problems with Attitude and Associational Factors in Criminal Risk Assessment", in *Stanford Law Review*, (2023), n. 6, pp. 1431 ss.

[46] L. Maldonato, "*Risk and need assessment tools* e riforma del sistema sanzionatorio: strategie collaborative e nuove prospettive", cit., p. 162.

[47] *Ivi*, p. 165.

[48] *Ivi*, p. 167, sia pur con riferimento alla pena prescrittiva di cui all'articolato del Gruppo di lavoro dell'Associazione italiana dei Professori di Diritto penale, incaricato di avanzare proposte in tema di riforma del sistema sanzionatorio.

[49] A. Carnevale, A. Lombardi, F.A. Lisi, "Exploring Ethical and Conceptual Foundations of Human-Centred Symbiosis with Artificial Intelligence", in *Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI*IA 2023)*, CEUR Workshop Proceedings, Roma, 3615 (2023), p. 30. Sui vantaggi della "Symbiotic AI" cfr. anche R. Hao, D. Liu, L. Hu (2023), "Enhancing Human Capabilities through Symbiotic Artificial Intel-

Dall'altro lato, ha previsto un apposito spazio processuale, disciplinato dall'art. 545-*bis* c.p.p., che si presta ad essere "sede ospitante" per esperimenti di convivenza e interazione armonica tra "intelligenze", nella quale l'IA potrebbe aumentare e valorizzare le capacità cognitive umane anziché sostituirle, secondo una logica di scambio e apprendimento reciproco, con conseguenti benefici per l'intero ecosistema processuale.

6. Le nuove pene sostitutive come "formiche" nella relazione simbiotica con gli afidi

È indubbio che l'ordinamento italiano si è mostrato, nel tempo, tradizionalmente carcerocentrico[50] ed appare altrettanto innegabile che il carcere non possa rappresentare una risposta efficace per tutte le fattispecie penalmente rilevanti: la disciplina delle pene sostitutive introdotte con il d.lgs. 10 ottobre 2022, n. 150[51] − con il quale è stata data attuazione alla l. 27 settembre 2021, n. 134, recante "Delega al Governo per l'efficienza del processo penale nonché in materia di giustizia riparativa e disposizioni per la celere definizione dei procedimenti giudiziari" – rappresenta l'ennesima manifestazione della ricerca di alternative all'esecuzione di sanzioni detentive di breve durata[52].

ligence with Shared Sensory Experiences", in *arXiv*, 2023, p. 5 s. (https://arxiv.org/abs/2305.19278).

[50] T. Travaglia Ciciriello, "Le pene sostitutive tra aspettative e incertezze", in D. Castronuovo, D. Negri (a cura di), *Forme, riforme e valori per la giustizia penale futura*, Jovene, Napoli, 2023, p. 275, che ricorda come "l'impianto sanzionatorio attuale sconta difetti cronicamente riconosciuti, primo tra tutti la sua carcerocentricità". In argomento, sulla recente trasformazione subìta dal sistema sanzionatorio, sintetizzabile nella perdita di centralità della pena carceraria, R. Bartoli, "Non più carcero-centrico l'attuale volto della pena", in *Diritto penale proc*., (2023), n. 6, pp. 745 ss.

[51] D.lgs. 10 ottobre 2022, n. 150, rubricato "Attuazione della legge 27 settembre 2021, n. 134, recante delega al Governo per l'efficienza del processo penale, nonché in materia di giustizia riparativa e disposizioni per la celere definizione dei procedimenti giudiziari", in *G.U*., 17 ottobre 2022, n. 243.

[52] In questo senso A. Gargani, "Le 'pene sostitutive' tra finalità rieducative, istanze deflattive e rischio di ineffettività", in D. Castronuovo, D. Negri (a cura di), *op. cit.*, p. 231.

Sullo sfondo vi è una più accresciuta consapevolezza che "la 'certezza della pena' non è la 'certezza del carcere'"[53] e che la pena carceraria, proprio nei casi in cui la sua durata "è troppo brev[e] per rendere possibile un efficace trattamento rieducativo, ma sufficientemente lung[a] da assicurare il contagio criminale e favorire la recidiva"[54], si rivela disumana, sproporzionata e "costosa" (laddove difetta il nesso di congruità rispetto alle reali esigenze repressive e preventive) sotto il profilo individuale e sociale, e dunque inefficace (proprio in rapporto alla diminuzione dei tassi di recidiva)[55].

Secondo il disegno del legislatore del 1981, le sanzioni sostitutive avrebbero dovuto rappresentare una valida alternativa al carcere già in fase decisoria, mentre nei fatti hanno vissuto una persistente condizione di diaspora; la loro palese marginalità applicativa è da ricollegarsi all'azione concorrente della sospensione condizionale della pena e, soprattutto, delle misure alternative alla detenzione di cui alla l. n. 354/1975: queste, applicabili *ab initio* al condannato in virtù della sospensione dell'ordine di esecuzione *ex* art. 656 c.p.p., hanno spostato nella fase esecutiva la partita sulla "sostituzione", determinando il fenomeno dei c.d. "libero-sospesi".

Con l'obiettivo di porre rimedio all'inefficacia del meccanismo sostitutivo e di introdurre delle reali alternative al paradigma segregazionista, la c.d. "riforma Cartabia" ha fatto ricorso ad un profondo rinnovamento dell'istituto previgente[56], normato

[53] Così nelle "Linee programmatiche sulla giustizia" del Ministro della giustizia Marta Cartabia, consultabili in *PenaleDP*, 16/03/2021.

[54] A. Gargani, "Le 'pene sostitutive' tra finalità rieducative, istanze deflattive e rischio di ineffettività", cit., p. 232.

[55] Lo evidenzia T. Travaglia Ciciriello, *op. cit.*, p. 277, che ammonisce come in questi casi occorre "ricercare strumenti punitivi in grado di evitare l'alterazione del rapporto tra il soggetto e la comunità".

[56] Per i commenti alla parte della riforma dedicata alle pene sostitutive di pene detentive brevi, riferiti alla legge delega o allo schema di decreto delegato, si v. A. Abbagnano Trione, "Una semantica persuasiva nel disegno di revisione delle sanzioni sostitutive delle pene detentive. Dalle parole ai fatti", in R.M. Geraci, A. Scalfati (a cura di), *Analisi della c.d. Riforma Cartabia (l. n. 134 del 2021)*, in *Processo penale e giustizia*, (2022), n. 1, pp. 238 ss.; G. Amarelli, "L'ampliamento delle sanzioni sostitutive delle pene detentive brevi: luci e ombre", *ivi*, pp. 234 ss.; R. Bartoli, "Verso la riforma Cartabia:

dalla l. 24 novembre 1981, n. 689: la nuova disciplina delle pene sostitutive, disseminata tra l'art. 20-*bis* c.p., gli artt. 53 e ss., l. n. 689/1981 e l'art. 545-*bis* c.p.p., da un lato, ha ammodernato il catalogo delle pene sostitutive e, dall'altro, ha introdotto un inedito elemento di bifasicità del processo[57].

senza rivoluzioni, con qualche compromesso, ma con visione e respiro", in *Diritto penale proc.*, (2021), n. 9, pp. 1167 ss.; A. Cavaliere (2021), "Considerazioni 'a prima lettura' su deflazione processuale, sistema sanzionatorio e prescrizione nella l. 27 settembre, n. 134, c.d. Riforma Cartabia", in *Penale-DP*, pp. 1 ss., 2 novembre 2021; D. Bianchi, "Il rilancio delle pene sostitutive nella legge-delega 'Cartabia': una grande occasione non priva di rischi", in *Sistema penale*, 21 febbraio 2022, pp. 1 ss.; O. Calavita, "La riforma Cartabia delle sanzioni sostitutive: riflessioni processualistiche in attesa del decreto legislativo", in *Legislazione penale*, (2022), n. 1, pp. 152 ss.; G. Daraio, "Le alternative al carcere tra riforme tradite e svolte annunciate: prime osservazioni sulle novità contenute nella 'legge Cartabia' (l. 27 settembre 2021, n. 134)", in *Archivio penale (web)*, (2021), n. 3, pp. 11 ss.; R. De Vito, "Fuori dal carcere? La riforma 'Cartabia', le sanzioni sostitutive e il ripensamento del sistema sanzionatorio", in *Questione giustizia*, (2021), n. 4, pp. 28 ss.; E. Dolcini, "Dalla riforma Cartabia nuova linfa per le pene sostitutive", in *Sistema penale*, 30/08/2022, pp. 1 ss.; Id., "Sanzioni sostitutive: la svolta impressa dalla riforma Cartabia", *ivi*, 02/09/2021; M. Donini, "Efficienza e principi della legge Cartabia. Il legislatore a scuola di realismo e cultura della discrezionalità", in *Politica del diritto*, (2021), n. 4, pp. 591 ss.; L. Eusebi, "La pena tra necessità di strategie preventive e nuovi modelli di risposta al reato", in *Rivista italiana di diritto e procedura penale*, (2021), n. 3, pp. 823 ss.; F. Fiorentin, "Va in soffitta la libertà controllata e potenziati i lavori di pubblica utilità", in *Guida al diritto*, (2021), n. 41, p. 87; A. Gargani, "La riforma in materia di sanzioni sostitutive", in *Legislazione penale*, (2022), n. 1, pp. 68 ss.; G.L. Gatta, "Riforma della giustizia penale: contesto, obiettivi e linee di fondo della 'legge Cartabia'", in *Sistema penale*, 15 ottobre 2021, pp. 1 ss.; G. Lattanzi, "A margine della 'riforma Cartabia'", in *Cassazione penale*, (2022), n. 4, pp. 1295 ss.; F. Palazzo, "I profili di diritto sostanziale della riforma penale", in *Sist. pen.*, pp. 11 ss., 08/09/2021; M. Telesca, "La 'nuova' disciplina delle sanzioni sostitutive delle pene detentive brevi previste dalla c.d. 'Riforma Cartabia', in *Diritto penale contemporaneo*, (2021), n. 3, pp. 34 ss.

[57] Sul d.lgs. n. 150 del 2022, con cui è stata data attuazione alla delega, relativamente alle pene sostitutive, vedi A. Abbagnano Trione, "Il sistema delle pene sostitutive e il *favor libertatis*", in *Processo penale e giustizia*, (2023), n. 3, pp. 757 ss.; Id., "Le latitudini applicative della commisurazione e della discrezionalità nel sistema delle pene sostitutive", in *Legislazione penale*, (2022), n. 4, pp. 288 ss., F. Alvino, "Pene sostitutive delle pene detentive brevi", in A. Bassi, C. Parodi, *La riforma del sistema penale*, Giuffrè

Sotto il primo profilo, l'ambizione di mutare la fisionomia complessiva del sistema sanzionatorio è stata coltivata percorrendo una "via secondaria"[58]: con il ventaglio delle "nuove" pene sostitutive – ovvero la semilibertà sostitutiva, la detenzione domiciliare sostitutiva, il lavoro di pubblica utilità sostitutivo e la pena pecuniaria sostitutiva, secondo quanto previsto dall'art. 20-*bis* c.p. di nuova introduzione – non si è diversificato il catalogo delle pene principali di cui all'art. 17 c.p. Tuttavia, le nuove sanzioni sostitutive, pur restando fuori dal catalogo edittale, vengono inserite nella parte generale del codice penale a chiusura del Capo dedicato alle "specie di pena", fornendo così "un segnale netto […] di avvicinarsi verso il riconoscimento e l'accettazione sociale di pene diverse della reclusione carceraria"[59]. Dal nuovo microcosmo sanzionatorio, che persegue l'importante finalità di "rivitalizzare la funzione rieducativa della pena"[60] e di "degradare la risposta sanzionatoria dal carcere a pene non carcerarie"[61], si attendono scosse telluriche capaci di generare effetti d'onda, destinati a propagarsi sul piano del "rinnovamento", anche culturale[62].

Francis Lefebvre, Milano, 2022, pp. 345 ss.; R. De Vito, "Le pene sostitutive: una nuova categoria sanzionatoria per spezzare le catene del carcere", in *Questione giustizia*, (2023), n. 2, pp. 107 ss.; D. Guidi, "La riforma delle 'pene' sostitutive", in *Legislazione penale*, (2023) n. 1, pp. 245 ss.; N. Madia, "Note minime, con andamento rapsodico, sugli aspetti generali delle nuove pene sostitutive", in *PenaleDP*, (2023), n. 2, pp. 193 ss.; I. Merenda, "Profili innovativi e aspetti problematici della riforma delle pene sostitutive", in *Archivio penale (web)*, (2023), n. 2, pp. 1 ss.; F. Palazzo, "Uno sguardo d'insieme alla riforma del sistema sanzionatorio", in *Diritto penale e processo*, (2023), n. 1, pp. 11 ss.

[58] Ovvero, per dirla con G. Amarelli, *op. cit.*, p. 236, "sul fronte secondario della 'degradazione giudiziaria' o 'secondaria' della risposta punitiva edittale, piuttosto che su quello principale della degradazione legislativa o 'primaria' tramite scelte generali ed astratte connesse al disvalore ridotto di taluni reati".

[59] T. Travaglia Ciciriello, *op. cit.*, p. 277.

[60] G. Varraso, "Riforma Cartabia e pene sostitutive: la rottura 'definitiva' della sequenza cognizione-esecuzione", in *Giustizia Insieme*, (2023) n. 1, p. 171.

[61] F. Palazzo, "Pena e processo nelle proposte della 'Commissione Lattanzi'", in *Legislazione penale*, (2021), n. 3, p. 219.

[62] G. Varraso, "Riforma Cartabia e pene sostitutive: la rottura 'definitiva' della sequenza cognizione-esecuzione", cit., p. 171. Cfr. anche V. Alberta,

La rinuncia da parte del potere legislativo a modificare la tipologia delle scelte giudiziarie relative alle conseguenze del reato ha di fatto rimesso le opportunità concrete di gestione costruttiva del fatto nelle mani del giudice di merito[63], cui compete il potere di sostituire la pena detentiva anticipando alla fase della cognizione, a titolo di vera e propria pena (anche se sostitutiva), alcune forme di esecuzione extra-carceraria che nell'ordinamento penitenziario vigente sono definite come "misure alternative alla detenzione".

L'anticipazione dell'alternativa al carcere già all'esito del giudizio di cognizione aspira a positivi e tangibili effetti di deflazione processuale (quali l'incentivazione dei riti alternativi e la riduzione delle impugnazioni) e penitenziaria (come l'abbassamento del numero dei procedimenti di sorveglianza, la riduzione del numero dei detenuti per pene brevi, la mitigazione del sovraffollamento carcerario), in grado di migliorare la *performance* della giustizia penale e di raggiungere le finalità efficientistiche perseguite dalla riforma.

7. Spazi processuali per la sperimentazione di una relazione simbiotica

Il giudice della cognizione, in caso di condanna a pena detentiva breve, è chiamato "a valutare se non vi siano modelli sanzionatori, sostitutivi della pena detentiva, che contribuiscano in modo più adeguato alla rieducazione del condannato, purché assicurino, anche attraverso opportune prescrizioni, la prevenzione del pericolo che il condannato commetta altri reati"; affinché possa "adempiere a tale compito, tuttavia, il giudice ha bisogno di un bagaglio di informazioni ulteriori rispetto a quelle comunemente acquisite nel giudizio di cognizione", ragione per

S. Amato, E. Losco, M. Straini, "Pene sostitutive delle pene detentive brevi: una riforma culturale", in G. Spangher (a cura di), *La riforma Cartabia. Codice penale – Codice di procedura penale – Giustizia riparativa*, Pacini, Pisa, 2022, pp. 7 ss.

[63] Sulle criticità connesse all'ampia discrezionalità del giudice della cognizione nell'applicazione delle pene sostitutive, E. Ranieri, "Le nuove pene sostitutive: l'*act loyal* delle parti ed il ruolo del giudice quali condizioni necessarie alla buona riuscita della riforma", in *Processo penale e giustizia*, (2023), n. 6, pp. 1513 ss.

la quale si è previsto il coinvolgimento degli uffici di esecuzione penale esterna[64].

Sotto il diverso profilo, squisitamente processuale, la riforma ha scandito uno spazio adeguato affinché, anche sul piano probatorio, si possa approfondire il tema della sanzione in modo separato ed indipendente rispetto alla pronuncia sulla responsabilità: il neo-introdotto art. 545-*bis* c.p.p., ispirato al modello del *sentencing* di matrice anglosassone, disciplina un inedito "momento" processuale di interazione tra l'autore del reato e il giudice in cui questi, dopo l'accertamento della responsabilità e la statuizione sulla condanna, può – raccolto il consenso dell'imputato alla sostituzione della pena originaria (da cui può prescindersi solo per la pena pecuniaria) – modularla integrando l'originario dispositivo, secondo "un modello decisorio 'a formazione progressiva'"[65].

Proprio questo nuovo "luogo" configurato dal legislatore si presta a riflessioni speculative intorno alla possibilità di percorrere una strategia collaborativa uomo-macchina in un segmento processuale in grado di migliorare la risposta al reato, nel senso auspicato dalla riforma.

Nel decidere sulla sostituibilità della pena il giudice è chiamato a compiere delicate valutazioni discrezionali che riguardano sia l'*an* che il *quomodo* della sostituzione.

La discrezionalità sostitutiva che la legge riconosce al giudice della cognizione[66] è regolamentata dall'art. 58, l. 24 novembre 1981, n. 689, rubricato "Potere discrezionale del giudice nell'applicazione e nella scelta delle pene sostitutive". Il potenziamento della discrezionalità "commisurativa" riconosciuta al magistrato

[64] "Relazione illustrativa allo schema di decreto legislativo recante disposizioni integrative e correttive del decreto legislativo 10 ottobre 2022, n. 150", in *Sistema penale,* 16 gennaio 2024, p. 246.

[65] F. Ruggieri, "Le pene sostitutive: caratteri generali e profili procedurali", in A. Salerno (a cura di), *L'applicazione della pena e le pene sostitutive dopo la Riforma Cartabia. Irrogazione, esecuzione e sostituzione della pena, alla luce della Riforma Cartabia (D.Lgs. 150/2022)*, Dike Giuridica, Napoli, 2023, p. 186.

[66] Sulla ampiezza della discrezionalità F. Palazzo, "Le linee portanti della riforma Cartabia", in R. Bartoli, R. Guerrini (a cura di), *Verso la riforma del sistema sanzionatorio*, Edizioni Scientifiche Italiane, Napoli, 2022, p. 22.

giudicante viene bilanciato da una articolata disciplina relativa ai criteri per il suo esercizio in concreto, che gravitano tra i due poli della rieducazione e della prevenzione[67].

Un primo gruppo di questi criteri vale a circoscrivere l'*an* della sostituzione, che risulta condizionata dal *quantum* di pena irrogabile, dalla valutazione sulla concedibilità della sospensione condizionale della pena quale alternativa alla sostituibilità della pena detentiva breve, dalle altre preclusioni soggettive di cui all'art. 59, l. n. 689/81 nonché dai criteri di commisurazione della pena definiti dall'art. 133 c.p., richiamati appunto dall'art. 58, comma 1, che stabilisce che, valutati detti parametri (gravità del reato e capacità a delinquere del reo), il giudice può applicare le pene sostitutive quando risultino più idonee alla rieducazione del condannato e quando, anche attraverso opportune prescrizioni, assicurino la prevenzione del pericolo di commissione di ulteriori reati. Lo stesso articolo aggiunge poi che la pena detentiva non può essere sostituita quando sussistano fondati motivi per ritenere che le prescrizioni non saranno adempiute dal condannato.

Il riferimento alla necessità di prevenire la commissione di altri reati lascia intendere come un residuo di pericolosità del condannato sia compatibile con le pene sostitutive, che appaiono orientate proprio al suo contenimento attraverso le opportune prescrizioni e costituiscono la specifica modalità prescelta dal legislatore per arginarlo al meglio[68]. In altri termini, è soprattutto attraverso la finalità rieducativa e risocializzante cui devono tendere le pene sostitutive che il legislatore ha inteso arginare il pericolo di recidiva. Questo spiega perché la prognosi (assimilabile a quella prevista nel procedimento di concessione delle misure alternative alla detenzione) che il giudice è chiamato a compiere per verificare l'*an* dell'applicazione della pena sostitutiva breve è sfavorevole quando quel tasso di recidiva non risulta azzerabile o riducibile attraverso l'adozione di quelle particolari

[67] Secondo G. De Francesco, "Brevi appunti sul disegno di riforma della giustizia", in *Leg. pen.,* (2021), n. 3, p. 241, il fine del legislatore delegante è di "evitare che le predette misure vengano a rivestire un ruolo meramente 'clemenziale', con l'effetto di 'deresponsabilizzare', per così dire, l'organo giudicante".

[68] In tal senso, v. Cass., Sez. V, 11 luglio 2023, n. 43622, in *Dir. pen. proc*., (2024), n. 1, p. 33.

prescrizioni che accompagnano la pena sostitutiva nella fase di esecuzione della stessa, perché vi sono "fondati motivi" che inducono a ritenere che le stesse non saranno adempiute, dei quali vi è l'onere di fornire un'adeguata argomentazione.

Una volta optato per la sostituzione della pena detentiva, un distinto gruppo, enucleabile dai successivi commi dell'art. 58, l. n. 689/81, regola il *quomodo* della sostituzione, disciplinando l'esercizio del potere discrezionale del giudice in ordine alla scelta della pena sostitutiva da applicare.

Il giudice, motivando adeguatamente, deve scegliere la misura più idonea a garantire – di nuovo – la "rieducazione" ed – ancora – il "reinserimento sociale" del condannato con il minor sacrificio della libertà personale.

Troverà quindi applicazione un principio di gradualità e di *extrema ratio* della misura più afflittiva, assimilabile al procedimento di scelta che veicola l'applicazione delle misure cautelari personali.

Quando l'alternativa è risolta a favore della semilibertà o della detenzione domiciliare, è imposto al giudice un onere di "motivazione rafforzata", estesa all'indicazione delle "specifiche ragioni" che ostano alla sostituzione con il lavoro di pubblica utilità o con la pena pecuniaria in quanto non idonee al perseguimento degli obiettivi indicati.

In ogni caso, nella scelta tra le pene sostitutive diverse da quella pecuniaria, il giudice deve tenere conto di particolari condizioni soggettive dell'imputato, legate all'età, alla salute psico-fisica, alla maternità/paternità, alle condizioni di disturbo da uso di sostanze stupefacenti o alcoliche ovvero da gioco d'azzardo, appositamente certificate, dalle condizioni di persona affetta da Aids conclamata o da grave deficienza immunitaria appositamente certificate.

Conclusivamente, il legislatore esige, in primo luogo, che le istanze di rieducazione e reinserimento siano vagliate rispetto a quelle securitarie e, in secondo luogo, con l'esigenza di minimizzare il sacrificio della libertà personale.

Un "apporto rilevante dell'intelligenza artificiale" potrebbe collocarsi proprio nell'ambito delle valutazioni sulla sostituibilità della pena, "con specifico riferimento ai parametri maggior-

mente di natura prognostica"[69], come il giudizio circa la pericolosità qualificata ed il concreto pericolo di violazione delle condizioni imposte, richiesto dalla normativa[70].

Per dare effettività e concretezza al finalismo rieducativo della pena e spezzare definitivamente l'equazione "pena uguale carcere" occorrerà che il giudice faccia buon governo del potere discrezionale assegnatogli. Un algoritmo di *risk assessment* appare in quest'ottica "uno strumento in grado di poter apportare un ricco compendio di notizie sul reo", utile a fornire al giudice "un ampio e dettagliato apparato informativo su rischi e sulle esigenze del reo, per inquadrare le condizioni di questo e strutturare così prescrizioni adeguate"[71].

In questo modo, l'onere di valutazione della sostituibilità della pena detentiva breve imposto oggi al giudice dal nuovo tessuto normativo[72], anche in assenza di una specifica richiesta proveniente dall'imputato, potrebbe giovarsi, in presenza di input costituzionalmente orientati[73], dell'apporto informativo proveniente dal *tool*, agevolandosi così il lavoro del giudice e degli uffici di esecuzione penale esterna[74].

Inoltre, si eviterebbe che la valutazione del giudice sia "appiattita" sulle emergenze del giudizio di primo grado: la possibilità di confronto con il dato, statisticamente rilevante, rappresenterebbe un imprescindibile momento di conferma o di confutazione dell'idea che ha maturato sulla sostituzione[75].

[69] Così, condivisibilmente, G. Camera, "A proposito dell'indagine conoscitiva sull'impatto dell'intelligenza artificiale nel settore della giustizia", in *Sistema penale*, 15/04/2024, p. 10.

[70] Cass., Sez. II, 14 febbraio 2024, n. 8794, in *C.E.D. Cass*., n. 286006.

[71] L. Maldonato, "*Risk and need assessment tools* e riforma del sistema sanzionatorio: strategie collaborative e nuove prospettive", cit., p. 167.

[72] G. Biondi, "L'applicazione delle pene sostitutive di pene detentive brevi nella fase di cognizione del processo penale", in *Sistema penale*, (2024), n. 4, p. 111.

[73] Per M. Gialuz, "Quando la giustizia penale incontra l'intelligenza artificiale: luci e ombre dei *risk assessment tools* tra Stati Uniti ed Europa", in *Diritto penale contemporaneo*, 29/05/2019, p. 21, occorre "evitare il rischio che […] si passi a un inaccettabile diritto penale del profilo d'autore".

[74] Cfr. ancora G. Camera, *op. cit.*, p. 10.

[75] In questo senso F. Coppola, "Commisurazione della pena e intelligenza artificiale: una ipotesi di lavoro con l'algoritmo *Ex-Aequo*", in *Archivio*

Ancora, la necessità di rinviare ad altra udienza potrebbe essere ridotta, nella direzione auspicata dal "decreto correttivo" (d.lgs. n. 31/2024)[76], per riservarlo ai casi in cui siano necessari ulteriori approfondimenti rispetto a quelli già forniti dal *tool*.

Ulteriormente, il giudice potrebbe meglio assolvere l'onere di fornire adeguata motivazione nella sentenza in ordine all'esercizio del suo potere discrezionale.

Dinanzi ad una valutazione dell'artefatto artificiale avvertita dal decisore umano come inadeguata, resterebbe sempre la possibilità per il giudice di razionalizzarla, inserendo nella macchina, quali nuovi input da considerare, le ragioni per le quali quel caso merita un diverso trattamento[77].

In questo modo, la valutazione effettuata dal classificatore, non concepita come vincolante, è inserita in un meccanismo di *feedback* che consente all'uomo di formare il classificatore, trasferendo a quest'ultimo il proprio *know-how* sulla pericolosità sociale[78].

In definitiva, nonostante non manchino aspetti problematici, la tecnologia predittiva consente un processo decisionale basato anche sui dati, che, "when used properly, can help people exit the criminal justice system and serve the public interest"[79].

penale (web), (2023), n. 2, p. 16, sia pur con riferimento all'applicazione della pena principale.

[76] L'art. 2, comma 1, lett. u) del d.lgs. 19 marzo 2024, n. 31, recante "Disposizioni integrative e correttive del decreto legislativo 10 ottobre 2022, n. 150, di attuazione della legge 27 settembre 2021, n. 134, recante delega al Governo per l'efficienza del processo penale nonché in materia di giustizia riparativa e disposizioni per la celere definizione dei procedimenti giudiziari" (in *G.U.* 20 marzo 2024, n. 67), ha da ultimo riscritto il primo comma dell'art. 545-*bis* c.p.p. Per un primo commento a riguardo, cfr. M. Gialuz, "Osservazioni sui correttivi alla riforma Cartabia tra rettifiche condivisibili, qualche occasione perduta e alcune sbavature", in *Sistema penale*, 29/01/2024.

[77] F. Coppola, *op. cit.*, p. 23.

[78] Ampi spunti in N. Lettieri, M. Vestoso, I. Cerere, "Beyond the Crystal Ball. Visions of AI in the Administration of Justice", in *i-lex – Rivista di Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale*, (2022), n. 2, p. 46.

[79] D.M. Applegarth, R.A. Lewis, R.M. Rief, "Imperfect Tools: A Research Note on Developing, Applying, and Increasing Understanding of Criminal Justice Risk Assessments", in *Criminal Justice Policy Review*, (2023), n. 4, p. 331, i quali sottolineano anche (p. 325) che "Machine learn-

8. Riflessioni conclusive

Un'interazione uomo-macchina, snodantesi secondo un paradigma simbiotico[80], può portare diversi benefici[81]. Tale paradigma implica la progettazione di sistemi di intelligenza artificiale secondo un approccio incentrato sull'uomo, in forza del quale la loro modellizzazione, anziché concentrarsi su ciò che le persone fanno (le decisioni che adottano)[82], si basa piuttosto sul perché lo fanno (la regola giuridica).

Il concetto di simbiosi a cui ci si è ispirati postula una relazione uomo-macchina "biunivoca" tra due agenti potenzialmente "uguali" nel processo decisionale, e in cui l'uno possa influenzare l'altro, ma ciò non esclude la possibilità che una parte eserciti un certo controllo sull'altra[83]. Il ché implica che la decisione giuridica sia resa senza servirsi "unicamente" dei risultati algoritmici e che il loro impiego, per essere giuridicamente accettabile, sia assoggettato o assoggettabile a un "controllo umano significativo"[84]. Quest'ultimo, come si è visto, potrà riguardare, tra l'altro, sia il controllo sulla validità generale della teoria che ispira il software, sia le modalità con cui quella

ing, with its adaptability, has the potential to incorporate live and updated data to reflect individuals' needs".

In uno studio recente, A. Kigerl, Z. Hamilton, M. Kowalski, X. Mei, "The great methods bake-off: Comparing performance of machine learning algorithms", in *Journal of Criminal Justice*, 82 (2022), pp. 1 ss., hanno testato diversi metodi per prevedere la recidiva (le valutazioni del rischio sono state costruite usando una varietà di algoritmi, dalle associazioni bivariate, alla regressione, agli approcci di *machine learning*). I risultati hanno suggerito che la dimensione del campione era più importante del tipo di algoritmo.

[80] Risale al 1960 la proposta di guidare lo sviluppo delle future tecnologie informatiche verso l'obiettivo della simbiosi: cfr. J.C.R. Licklider, "Man-computer symbiosis", in *IRE transactions on human factors in electronics*, (1960), n. 1, pp. 4 ss.

[81] S. Karnouskos, "Symbiosis with artificial intelligence via the prism of law, robots, and society", in *Artif. Intell. Law*, 30 (2022), n. 1, p 110.

[82] D.M. Katz, "Quantitative Legal Prediction – or – How I Learned to Stop Worrying and Start Preparing for the Data Driven Future of the Legal Services Industry", in *Emory Law Journal*, 62 (2013), n. 4, p. 918.

[83] A. Carnevale, A. Lombardi, F.A. Lisi, *op. cit.*, p. 3 s.

[84] G. Ubertis, "Intelligenza artificiale, giustizia penale, controllo umano significativo", in *Sistema penale*, (2020), n. 4, p. 83 s.

teoria è stata codificata nel software[85], sia la possibilità per il giudice del caso concreto di discostarsi dal risultato suggerito e di valorizzare in motivazione le sfumature che rendono il proprio caso differente[86].

Del resto, lo stesso *AI Act* non manca di sottolineare più volte come occorra garantire che "le previsioni, raccomandazioni o decisioni del sistema di IA siano ignorate o ribaltate"[87].

Secondo l'approccio adottato, il *risk assessment* è un processo di valutazione del rischio e non un metodo in grado di fornire una predizione esatta; è un mezzo e non un fine in se stesso[88].

Nella strategia collaborativa di tipo simbiotico l'algoritmo predittivo può fornire un contributo conoscitivo ampio, esteso

[85] S. Quattrocolo, "Sui rapporti tra pena, prevenzione del reato e prova nell'era dei modelli computazionali psico-criminologici", cit., p. 280.

[86] F. Coppola, *op. cit.*, p. 16, richiama la nota sentenza "Franzese" (Cass., Sez. un., 10 luglio 2002, n. 30328, in *C.E.D. Cass*., n. 222138) sull'utilizzo delle leggi scientifiche nell'accertamento del nesso di causalità.

[87] Cfr. il centoquarantunesimo considerando, nonché artt. 60, par. 4, lett. k) e 61, par. 1, lett. d). In dottrina, G. Ubertis, "Processo penale telematico, intelligenza artificiale e costituzione", in *Cass. pen*., (2024), n. 2, p. 444, ritiene "ammissibile l'uso dell'intelligenza artificiale per ottenere una previsione ipotetica, ma non una predizione vincolante, sostitutiva della condotta giudiziale", in linea con quanto affermato dal quinto principio della Carta etica europea sull'utilizzo dell'intelligenza artificiale nei sistemi giudiziari e negli ambiti connessi; diversamente, si avrebbe "una certezza del diritto illusoria, perché astratta e difforme dalle esigenze della specifica vicenda" e "si perverrebbe a una ossificazione del diritto vivente", che dovrebbe invece mantenere intatte le possibilità dell'interpretazione e dell'evoluzione normativa.

[88] L. Maldonato, "*Risk and need assessment tools* e riforma del sistema sanzionatorio: strategie collaborative e nuove prospettive", cit., p. 146 s. L'A. precisa che l'output non è una predizione finale, ma richiede una nuova valutazione. Ove ci si affidasse esclusivamente agli strumenti di I.A., si ridurrebbero le persone a una mera entità numerica, ciò in contrasto con il diritto fondamentale e inviolabile del rispetto della dignità umana, argine a qualsiasi reificazione dell'uomo. Secondo S. Quattrocolo, "Sui rapporti tra pena, prevenzione del reato e prova nell'era dei modelli computazionali psico-criminologici", cit., p. 258, "il *risk assessment* è un metodo e non un fine, uno strumento di valutazione del rischio di esordio o persistenza di una condotta antisociale e non la risposta ad essa". Per D.M. Applegarth, R.A. Lewis, R.M. Rief, "Imperfect Tools: A Research Note on Developing, Applying, and Increasing Understanding of Criminal Justice Risk Assessments", in *Criminal Justice Policy Review*, (2023), n. 4, p. 331, "part of what needs to be understood, participants contend, is that risk assessments are not solutions but tools for agencies to use to help them make decisions regarding individuals, policies, or procedures".

non solo ai rischi ma anche ai bisogni individuali, arricchendo le modalità di risposta al reato nel rispetto dei principi fondamentali di proporzionalità e dignità della persona, realizzando le istanze di individualizzazione della pena, dando concretezza alla sua funzione rieducativa e migliorando, in definitiva, l'esperienza umana[89].

Formiche e afidi coesistono in una relazione simbiotica che avvantaggia entrambe le specie e che assume un'autonoma valenza: la relazione non è statica, ma sollecita continui e costanti adattamenti reciproci, che producono un rafforzamento del grado di simbiosi[90]. Se si vorranno sfruttare i vantaggi che scaturiscono da simili legami, occorrerà mettere in conto la necessità che il processo penale subisca alcuni adattamenti morfologici, tollerabili purché non se ne comprometta lo statuto epistemologico.

[89] L. Maldonato, "*Risk and need assessment tools* e riforma del sistema sanzionatorio: strategie collaborative e nuove prospettive", cit., p. 169.

[90] Studi svelano come un afide simbiotico manipola egoisticamente le formiche presenti tramite la dopamina nella melata. Cfr. T. Kudo, H. Aonuma, E. Hasegawa, "A symbiotic aphid selfishly manipulates attending ants via dopamine in honeydew', in *Scientific Reports*, 17/09/2021. D'altra parte, le ricerche dimostrano come anche le formiche possono modificare il successo riproduttivo degli afidi a loro vantaggio. Cfr. S. Watanabe, J. Yoshimura, E. Hasegawa, "Ants improve the reproduction of inferior morphs to maintain a polymorphism in symbiont aphids", in *Scientific Reports*, 02/02/2018.

Paolo Sommaggio, Shan Ali

Cognitive Liberty, Neurolaw, and Digital Governance: Integrating Neuroscience and Blockchain in the Era of Neuro-Rights*

Abstract: Cognitive Liberty, Neurolaw, and Digital Governance: Integrating Neuroscience and Blockchain in the Era of Neuro-Rights. The subject of neuroscience is the study of the relationship between the human brain and human behavior. Advances in brain research and neuro-technological systems have created new possibilities to observe, interpret, and, within ethical limits, influence cognitive and behavioral processes. In this article, we analyzed how these developments enter legal frameworks and shape mechanisms aimed at addressing socially unwanted behavior. It motivates us to study the relationship between neuroscience and neurolaw with a special focus on the role of digital infrastructure, particularly blockchain technology, in ensuring evidentiary integrity and ethical data governance. With this perspective, we demonstrated the potential role of neuroscientists in Courts and the function of blockchain for neuroscientific evidence. Secondly, we examined how new neuro-paradigms influence debates about the structure of Society and the Law. We then focused on the domains of reductive neurolaw, which introduce the idea of replacing traditional sources of law with emerging neuroscientific standards. Along with this, we discuss a definition of Cognitive Liberty (a new form of safeguard) able to be collected in a "Declaration of Human Neuro-rights," noting that cognitive liberty may serve as a conceptual tool to defend individual human rights from neuro-paradigms that diminish them. Finally, we discussed how blockchain can enhance transparency, provenance tracking, and authenticity in neuroscientific data management, reinforcing trust and accountability without replacing traditional legal oversight.

Keywords: Neuroscience, Reductive Neurolaw, Neuro-deviance, Cognitive Liberty, Human Rights, Blockchain Technology.

Summary: 1. Introduction – 2. Emerging Neuro-Technologies – 3. Radicals, Reformists, and the Neurolaw Debate – 4. Law in the Age of Neuroscience – 5. Neuro-Rights and Mental Autonomy – 6. Blockchain for Neuroscientific Evidence – 7. Cognitive Liberty as a Neuroright – 8. Neuro-Human Rights and Global Governance – 9. Blockchain, Neurolaw, and Digital Accountability – 10. Conclusion.

* This essay is a product of a cooperative work of Paolo Sommaggio and Shan Ali, anyway, for any purpose, the § 1-5 are Paolo Sommaggio's and § 6-10 Shan Ali's.

1. Introduction

Neuroscience is the scientific field that investigates the relationship between the human brain (and nervous system) and human behavior. Continuous advances in brain research, neuroethics, and data governance have widened this discipline beyond clinical boundaries, linking it directly with questions of justice, autonomy, and regulatory design. This research analyzes the idea that increasing understanding of neural structures may shape social and legal approaches to behavior[1], contributing to what has been described as an emerging "neuro-civilization"[2].

In doing so, we discuss the notion of reductive neurolaw, which refers to the gradual replacement of traditional legal reasoning with neuroscientific standards. Alongside this, we highlight the concept of cognitive liberty as a legal and ethical safeguard that defends individuals from intrusive brain interventions and potential manipulations of mental autonomy[3]. This issue remains at the heart of ongoing debate about how law can respond to the epistemic authority of neuroscience without undermining personhood or responsibility.

Furthermore, we consider how this concept plays an important role in a broader and still-developing international framework on human dignity, mental privacy, and technological accountability. Finally, we outline the role of blockchain technology for transparency and evidence integrity in neuroscientific data and cognitive liberty.

2. Emerging Neuro-Technologies

New techniques for studying the human brain have opened up previously unthinkable possibilities for directly knowing and directing people's actions. Recent progress in neuroscience has also made these techniques more precise and accessible, extending their use from clinical settings to research, education, and

[1] A. Roskies, "Neuroethics for the New Millennium", in *Neuron*, 35 (2002), n. 1, pp. 21-23.

[2] P. Sommaggio, M. Mazzocca, A. Gerola, F. Ferro, "Cognitive liberty. A first step towards a human neuro-rights declaration", in *BioLaw Journal*, 3 (2017), pp. 27-45.

[3] *Ibidem*.

even personal devices, which raises new ethical and legal questions. In this section, we present a quick summary of current neuro-technologies and the kind of questions that their development and implementation raise.

The first category of neurotechnologies we consider is brain imaging techniques. The primary techniques used for brain monitoring and imaging are electroencephalography (EEG) and functional magnetic resonance imaging (FMRI). They give anatomical and functional information about the brain and its neural activity, which is employed in diagnostic and research applications. Neuroscientists, for example, can use FMRI to analyze how neurons fire, allowing them to link brain activity with mental activity and localize parts of the brain that respond to specific stimuli, such as pain or language recognition. This material provides a better understanding of how the brain functions and how it supports our thinking.

The next group of technologies includes neuro-stimulation, which provides therapies based on electrical and magnetic stimulation of the brain via medical devices mounted to the head or implanted into the brain. Transcranial Magnetic Stimulation (TMS) and Deep Brain Stimulation (DBS) are currently widely used in the treatment of neurological and psychiatric illnesses such as Parkinson's disease, epilepsy, and depression[4].

A third set of neuro-technologies comprises psychoactive drugs, which are known to cause personality changes too. The last issue that will be touched upon here is cognitive enhancement. Both neuro-stimulation technologies and psychoactive drugs can be used to augment human cognitive capacities such as attention, focus, and memory (but also mood, personality traits, and behavior).

Therefore, we can sum up the idea of effective neuroscientific technologies that can read the minds of people, inducing a change in the mood and personality of subjects to which it is implied. An ideal neuroscientific technology can modify behavior

[4] F. Jotterand, J. Giordano, "Transcranial magnetic stimulation, deep brain stimulation, and personal identity: ethical questions, and neuroethical approaches for medical practice", in *International Review of Psychiatry*, 23 (2011), n. 5, pp. 476-485.

and alter the formation of memory and its consolidation with an additional ability to be able to argue the cognitive ability. These are the reasons why the field of neuro-ethical needs is bringing about the necessity of an ever-increasing consideration of social and ethical implications of neuro-technological discoveries[5]. This need has become more evident in recent academic debates that link neuroethics, law, and data governance as shared foundations for responsible innovation[6].

3. Radicals, Reformists, and the Neurolaw Debate

In the 1990s, neurosciences began to make their way into the public eye in pursuit of recognition, coining the term "Decade of Brain" (Presidential Proclamation 6158, Office of Federal Register, 12:11 p.m., July 18, 1990)[7]. Since then, the field has continued to expand, influencing not only medicine but also ethics, law, and public policy. A recent article focuses on how neurosciences can support the efficient and equitable administration of justice. The authors make the assumption that, at this time, neurosciences have displaced all other scientific disciplines and fundamentally altered the conventional methods of conceptualizing the anthropological events that serve as the foundation for legal systems. This will allow neuroscientists to surpass all other

[5] W. Sententia, "Freedom by design: Transhumanist values and cognitive liberty", in M. More, N. Vita-More (eds.), *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology and Philosophy of the Human Future*, John Wiley & Sons, Hoboken, 2013, pp. 356-357.

[6] OECD, "Neurotechnology Governance Toolkit: Implementing the OECD Recommendation on Responsible Innovation in Neurotechnology", in *OECD Publications*, OECD Publishing, Paris, 2024, available at: https://www.oecd.org/sti/emerging-tech/neurotechnology-governance-toolkit.pdf: UNESCO, "Recommendation on the Ethics of Neurotechnology", in *UNESCO Publications*, UNESCO, Paris, 2024, available at: https://www.unesco.org/en/ethics-neurotech/recommendation.

[7] J.T. Cacioppo (ed.), *Foundations in Social Neuroscience*, Massachusetts Institute of Technology Press, Cambridge, 2002; D.D. Franks, *Neurosociology. The Nexus Between Neuroscience and Social Psychology*, Springer, Dordrecht, 2010; C. O'Connor, G. Rees, H. Joffe, "Neuroscience in the public sphere", in *Neuron*, 74 (2012), n. 2, pp. 220-226.

specialists in their field in terms of qualification for all in-depth technical examinations held in court[8].

Additionally, we observe the slow emergence of stereotypes and popular notions that have the power to sway discussions in the social and political spheres[9]. This initiative aims to integrate classic anthropological and moral conceptions, such as free will and conscious moral action, which form the foundation of all legal orders, with the latest advancements in neuroscience[10]. The conclusion is straightforward: discoveries in neuroscience require the modification of existing legal orders. This is a premise held by many neuroscientists, albeit their approaches to changing the law vary. The outcome is very simple: legal orders must be modified according to new neuroscientific achievements.

One group believes that the introduction of neurosciences into the study of law will inevitably result in a revolution of legal systems into modern legal frameworks. These individuals are known as maximalists or radicals[11]. Conversely, reformists believe that it is more beneficial to alter the legal systems gradually and consistently, without causing undue stress to society, by small but consistent reforms[12]. This contrast between radical and reformist perspectives continues to appear in current neurolegal debates, where scholars discuss whether neuroscience should reshape legal definitions of responsibility or remain an auxiliary tool for interpretation.

[8] O.D. Jones, A.D. Wagner, D.L. Faigman, M.E. Raichle, "Neuroscientists in court", in *Nature Reviews Neuroscience*, 14 (2013), n. 10, pp. 730-736.

[9] A. Kolber, "Will There Be a Neurolaw Revolution?", in *Indiana Law Journal*, 89 (2014), n. 2, pp. 807-845. D. Larriviere, M.A. Williams, "Neuroenhancement: Wisdom of the Masses or 'False Phronesis'?", in *Clinical Pharmacology&Therapeutics*, 88 (2010), n. 4, pp. 459-461.

[10] M. Farah, "Emerging ethical issues in neuroscience", in *Nature Neuroscience*, 5 (2002), n. 11, pp. 1123-1130.

[11] M.S. Gazzaniga, M.S. Steven, "Free Will in the Twenty-first Century", in B. Garland (ed.), *Neuroscience and the Law: Brain, Mind and the Scales of Justice*, Dana Press, New York, 2004.

[12] N. Vincent, "On the Relevance of Neuroscience to Criminal Responsibility", in *Criminal Law and Philosophy*, 4 (2010), pp. 77-98.

The reason we believe Greene and Cohen[13] embody the radical position is that they justify their support for the introduction of neuroscientific technologies into legal orders by pointing out that doing so will eliminate free will and the concept of responsibility that comes with it. These ideas are outlined in the theories of punishment, particularly the retributivist one[14]. To date, however, the enthusiasts have favored workable solutions over precisely defining the standards to which a subject should be treated[15]. Recent discussions extend this concern to predictive neuro-technologies, which may risk undermining the presumption of free will if used without ethical safeguards.

Stephen Morse is unquestionably among the composed reformists who acknowledge the introduction of neuroscience as being somewhat beneficial without endorsing it blindly. Morse thinks the new accomplishments won't be able to completely transform the legal system. In reality, he believes that the legal systems cannot be directly affected by the rapid expansion of neurosciences in the near future, at least not significantly. According to Morse, it is incorrect to absolve someone of responsibility for their actions solely because "his/her brain did it". Individual responsibility cannot be questioned unless some altered state is highlighted, as every action we take is in some way caused by the brain. This is so because the foundation of law is common sense psychology, which is immune to the influence of neuroscientific findings. This is especially true for criminal penalties, which are based on a "folk-psychological" understanding of the offender and their behavior[16].

The psychological theory of Morse explains behavior that is somewhat influenced by biological, psychological, and societal factors and partially by mental states including intents, beliefs,

[13] J. Greene, J. Cohen, "For the law, neuroscience changes nothing and everything", in *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359 (2004), n. 1451, pp. 1775-1785.

[14] M. Pardo, D. Patterson, "Neuroscience, Normativity, and Retributivism", in T. Nadelhoffer (ed.), *The Future of Punishment*, Oxford University Press, Oxford, 2013, pp. 133-154.

[15] A. Kolber, *op. cit.*, pp. 807-845.

[16] S.J. Morse, "Compatibilist Criminal Law", in T. Nadelhoffer (ed.), *op. cit.*, pp. 107-132.

and plans. In conclusion, folk psychology views mental states as essential to an understanding of the fundamentals of human psychology. "Folk psychology presupposes only that human action will at least be rationalized by mental state explanations or will be responsive to reasons including incentives under the right conditions"[17].

This is why, in Morse's idea neurosciences will not modify the law in a revolutionary way, as the latter is founded on premises tied to common sense and not to techno-scientific explanations.

4. Law in the Age of Neuroscience

The intervention of neurolaw in contemporary society has two important components, which ensure the traditional society's transition into a neuroscientific one. The first is to create new legal regulations based on advances in neuroscience; the second is the ability to directly alter someone's brain (i.e., therapeutically or sanctioningly). Regarding the first component, it is achieved by substituting new neuroscientific norms for established legal sources. This is because neuroscientists believe that legislation, which was designed as a traditional method of social control, has failed since it cannot ensure the upkeep of order in society[18]. On this topic, Brian Tamahana wrote:

> Under a scientific view, law would come instead to be seen as the source of social order – to produce social order is the function or purpose or end of law. In turn, this new perspective, over time, would open up questions about the efficiency and utility of law in carrying out its functions. The subtle but fundamental difference can be put thus: law is order, versus law maintains order[19].

[17] S.J. Morse, "The status of neurolaw: A plea for current modesty and future cautious optimism", in *The Journal of Psychiatry & Law*, 39 (2011), n. 4, pp. 595-626.

[18] S.J. Morse, "Mental disorder and criminal law", in *Journal of Criminal Law & Criminology*, 101 (2011), n. 3, pp. 885-968.

[19] B.Z. Tamanaha, *Law as a Means to an End: Threat to the Rule of Law*, Cambridge University Press, Cambridge, 2006, p. 21.

For the second part regarding the alteration of the brain under the umbrella of new neurolaw, David Eagleman claims that, about the second aspect of neurolaw, criminal subjects ought to be handled as people who suffer from serious illnesses or cognitive impairments. Eagleman's recommended rehabilitative methods rather than punitive ones are grounded in non-invasive behavior modification techniques like neuroimaging, which functions as a kind of biofeedback that enables people to view images of their brains and gain more control over their behavior. "To achieve this, we have started utilizing real-time feedback to participants during brain imaging", he added. Through the use of this strategy, individuals can become aware of when their brain is experiencing cravings and learn how to regulate (in this example, decrease) that neural activity by fortifying other, long-term decision-making mechanisms[20].

This rehabilitative and preventative approach has gained renewed attention, as neuroscientific evidence becomes more present in court evaluations, mental health assessments, and sentencing discussions[21]. However, these applications raise complex ethical questions about proportionality, consent, and the risk of reducing moral judgment to biological explanation.

As a scientific theory of the basis of new neurolaw, these postulates are intriguing, but when tested, they are shown to be only subjective choices. Despite this, the new "neurolaw" is founded on neuro-standards but yet they hold a significant deal of confusion. Therefore, one of the neurolaw's most sensitive topics is the definition of agreed norms. Stated differently, there is a lack of consensus among us regarding the definition of a "criminal", such as whether it stems from a desire to violate moral or legal obligations, a biological trait, or a specific illness. Contemporary academic debates continue to emphasize this uncertainty, suggesting that neurolaw must evolve through transparent eth-

[20] D.M. Eagleman, S. Isgur Flores, "Defining A Neurocompatibility Index for Criminal Justice System: A Framework to Align Social Policy with Modern Brain Science", in S. Muller *et al.* (eds.), *The Law of the Future and the Future of the Law: vol. II*, Torkel Opsahl Academic publisher, The Hague, 2012, pp. 161-171.

[21] A. Bigenwald, V. Chambon, "Criminal responsibility and neuroscience: no revolution yet", in *Frontiers in Psychology*, 10 (2019), pp. 1406.

ical frameworks rather than deterministic models. Scholars increasingly argue that its legitimacy depends on how effectively it integrates neuroscientific knowledge with long-standing legal and human-rights principles[22]. This ongoing transformation indicates that while neuroscience offers valuable insight into behavior, it cannot substitute the moral and normative reasoning that remains essential to law.

5. Neuro-Rights and Mental Autonomy

According to Rainey and Yang[23] the most popular framework in neurolaw is very straightforward: human behavior has a biological basis, and since this basis is modifiable, it is possible to control the biological matrices of socially unacceptable behavior. The method is fairly straightforward: all that is needed to conceptualize social abnormalities is to view them as biological (brain) abnormalities, placing them under the umbrella of mental disorders and diseases. Given that the law, particularly criminal law, does not seem to be the best tool for solving the issue, society must make room for alternative, more successful approaches (of neurolaw)[24].

By using different novel "therapies" (such as surgery, medications, brain transplants, etc.), it is feasible to modify the deep brain structure and mental states in several ways, which can regulate behavior more efficiently and prevent illegal activities than incarceration or other treatments. Stated differently, it appears that the road towards a society that is more neuro-standardized will be rather easy and gentle[25]. By implementing "deviance"

[22] M.S. Pardo, D. Patterson*, Minds, Brains, and Law: The Conceptual Foundations of Law and Neuroscience*, Oxford University Press, Oxford, 2013.

[23] A. Raine, Y. Yang, "Neural foundations to moral reasoning and antisocial behavior", in *Social Cognitive and Affective Neuroscience*, 1 (2006), n. 3, pp. 203-213.

[24] H. Nagera, "Reflections on Psychanalysis and Neuroscience: Normality and Pathology in Development, Brain Stimulation, Programming and Maturation", in *Neuropsychoanalysis*, 3 (2013), n. 2, pp. 179-191.

[25] I. Singh, W.P. Sinnott-Armstrong, J. Savulescu, *Bioprediction, Biomarkers, and Bad Behavior. Scientific, Legal and Ethical Challenges*, Oxford University Press, Oxford, 2013.

will be reduced to a straightforward health issue that is created by neurostandards (neurorules) and addressed with neurotechniques[26].

We believe it is important to emphasize how this scientific framework of using neurotherapies may move dangerously toward individual freedom. The central difficulty lies in defining legitimate intervention boundaries, especially when coercive medical procedures are justified as preventive measures. It is well known that many legal orders including different types of intervention, such as forced medical procedures, are carried out even when the subjects of the orders do not consent to them. In our opinion, this blind spot may serve as a test bed for novel normalization strategies influenced by neuro-civilization or, more accurately, neuro-normalization during the coming years[27]. International policy discussions on mental privacy and neurorights have now placed these concerns within a broader human-rights framework.

As one of the most ardent proponents of neuro-civilization, Hank Greely attempted to initiate a discussion in 2012 regarding the use of involuntary treatments specifically for the modification or removal of antisocial behavior, as well as for the treatment of illnesses and psychological discomfort[28].

Greely makes the audacious claim that neurosciences will enable the modification of undesirable behavior by altering the neural underpinnings of individual agents. This logic is quite straightforward: if we concur that direct brain intervention is appropriate in cases of severe illness or disability, then there is no justification for disagreement over the management of brain-related factors that contribute to socially undesirable behavior. Greely suggests using efficacy and safety as benchmarks when assessing various neuro-treatment options. He claims that

[26] M. Ienca, R. Andorno, "Towards new human rights in the age of neuroscience and neurotechnology", in *Life Sciences, Society and Policy*, 13 (2017), n. 1, p. 5.

[27] G. Meynen, "A Neurolaw Perspective on Psychiatric Assessments of Criminal Responsibility: Decision-making, Mental Disorder, and the Brain", in *International Journal of Law and Psychiatry*, 36 (2013), n. 2, pp. 93-99.

[28] H.T. Greely, "Neuroscience and criminal justice: not responsibility but treatment", in *Kansas Law Review*, 56 (2008), n. 5, pp. 1103-1138.

because conventional direct brain intervention techniques, including lobotomies, are neither safe nor successful, they are oversimplified answers to an extremely complicated issue[29]. Contemporary ethics now extends this view, insisting that any neuro-intervention must be proportional, evidence-based, and subject to informed consent review by independent oversight bodies[30].

To eliminate socially unacceptable behavior through behavior management, it is therefore required to evaluate novel kinds of safe and effective intervention given, however, that the interventions are safe, effective, and not inappropriate. According to him, there is no need for the controversy brought about by a change in a person's brain if we can calmly put someone in jail for trying in vain to change their behavior.

The issue is one of individual freedom, namely the "resistibility" of conventional methods that grant the subject some residual autonomy, which the new methods of direct intervention would not grant. Greely states that in this context, it is necessary to define an unreachable "cognitive liberty" – a kind of privacy threshold that one should not descend. The more successful (and seductive) the treatment, the greater the invasion of liberty, he argued. "Resistible" treatments, like prison rehabilitation programs, nonetheless seem to leave some freedom for choice". We agree that there should be a protected area for cognitive liberty, but since all interventions have an impact on the brain, it is difficult to understand why required brain interventions should only be prohibited if they are direct[31].

However, even with this kind of unachievability, it is hard to claim that direct brain intervention could not spread like wildfire as a way to ease approved behavior or change socially undesirable or unaccepted behavior. Though we can only trace a prelimi-

[29] *Ibidem*.

[30] S. Ruiz, L. Valera, P. Ramos, R. Sitaram, "Neurorights in the Constitution: From Neurotechnology to Ethics and Politics", in *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379 (2024), n. 1915, pp. 1-12.

[31] H.T. Greely, "Direct Brain Interventions to 'Treat' Disfavoured Human Behaviours: Ethical and Social Issues", in *Clinical Pharmacology & Therapeutics*, 91 (2012), n. 2, pp. 163-165.

nary outline, these considerations bring up an intriguing prospect of in-depth study. The idea of cognitive liberty, often known as the right to mental self-determination, has just recently entered the discourse on a global scale[32]. Cognitive liberty is associated with the idea of sovereignty over one's "cognitive heritage" and would entail a right akin to the inviolability of the brain against the state or outside parties. However, it emphasizes on enhancement of one's cognitive structure[33] by having freedom for direct interventions[34]. The evolution of neurorights from philosophical debate to policy proposals reflects an emerging consensus that legal systems must adapt to the ethical implications of neurotechnology while preserving human dignity and autonomy.

6. Blockchain for Neuroscientific Evidence

The incorporation of blockchain technology into legal systems has signified a substantial advancement in the management of neuroscientific evidence. Its relevance has expanded as judicial institutions, research consortia, and neuroethics boards increasingly depend on digital infrastructures for the authentication, storage, and traceability of sensitive data[35].

The importance of neuroscientific findings in establishing judicial decisions has increased significantly, creating a demand for verifiable data integrity and transparent auditability. A unique decentralized nature is a basic characteristic of blockchain technology, and it has equipped, modern judicial setups with the capacity for creating an immutable record that ensures the permanence and inalterability of data once it is stored. This foundational characteristic is pivotal in addressing some of the

[32] J. Bublitz, "My mind is Mine!? Cognitive Liberty as a Legal Concept", in E. Hildt, A. Francke (eds.), *Cognitive Enhancement*, Springer, Dordrecht, 2013, pp. 233-264.

[33] Wrye Sententia and Richard Glen Boire are the founders of the Centre for Cognitive Liberty and Ethics (CCLE).

[34] W. Sententia, "Neuroethical Considerations: Cognitive Liberty and Converging Technologies", in *Annals of the New York Academy of Sciences*, 1013 (2004), n. 1, pp. 221-228.

[35] H. Yang, L. Jiang, "Regulating Neural Data Processing in the Age of BCIs: Ethical Concerns and Legal Approaches", in *Digital Health*, 11 (2025), doi: 10.1177/20552076251326123.

top concerns around the authenticity and precision of neuroscientific evidence, as noted[36].

Blockchain technology can preserve the data integrity once stored, which is an essential component for neuroscientific evidence, as it plays a very significant role in influencing legal outcomes. The immutable ledger of blockchain technology guarantees that when the data is recorded under its umbrella, the data is immune to any kind of alteration, thereby setting a robust and dependable basis for judicial decisions and outcomes. With any kind of neuroscientific data its confidentiality and lawful use have become central topics in the debate on neuro-data governance. Blockchain technology can adeptly meet the privacy requirements through different state-of-the-art encryptions and pseudonymization techniques in combination with off-chain secure storage, zero-knowledge proofs, and privacy-by-design standards, which ensures compliance with stringent privacy standards like the GDPR, as reported by Zyskind, Nathan, & Pentland[37].

Blockchain technology has its characteristic feature of facilitating the conception of standardized protocols for the submission and storage of data related to neuroscientific evidence. The standardization of protocols for data management under blockchain technology is very crucial for mitigating discrepancies in the handling of evidence and their evaluation across various legal jurisdictions, and as a result, it promotes fairness and consistency in undergoing different legal proceedings[38]. This technology also encourages the exploration and implementation of a collaborative verification process among all the stakeholders in the legal chain that starts from the neuroscientists moving to

[36] M. Crosby, Nachiappan, P. Pattanayak, S. Verma, V. Kalyanaraman, "Blockchain technology: Beyond Bitcoin", in *Applied Innovation*, 2 (2016), pp. 6-10.

[37] G. Zyskind, O. Nathan, A. Pentland, "Decentralizing privacy: Using blockchain to protect personal data", in *Proceedings of the IEEE Security and Privacy Workshops*, 2015, pp. 180-184.

[38] P. Mamoshina, L. Ojomoko, Y. Yanovich, A. Ostrovski, A. Botezatu, "Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare", in *Oncotarget*, 9 (2018), n. 5, pp. 5665-5690.

legal professionals, and law enforcement agencies. This implementation ensures the enhancement of the credibility and reliability of neuroscientific evidence in different legal contexts presented under different scenarios[39].

Despite its advantages, several challenges remain in applying blockchain in legal and neuroscientific contexts, including limited digital literacy among legal actors, integration costs, and varying technical standards. Blockchain technology has faced resistance from traditional legal systems. Another important implication of this technology is its necessity to have evolved regulatory and ethical frameworks which can be an important setback presently in implementing blockchain technology in the management of neuroscientific evidence. Future adoption therefore depends on the creation of interdisciplinary standards that align technological reliability with procedural fairness and human-rights protection.

7. Cognitive Liberty as a Neuroright

The classical definition of "liberty of thought" is frequently expanded upon using the phrase "Cognitive Liberty". Currently, meanwhile, some academics are using that phrase to argue against democratic legal systems' incorporation of such a right into their constitutions[40]. This discussion has gained new relevance as advances in neurotechnology and data analysis make it increasingly possible to access or influence mental states, raising questions about how far traditional notions of liberty of thought can protect individuals. This later definition highlights three important conceptual points:

- *Privacy*. Which is to say, our thoughts must be private unless we choose to share.

- *Autonomy*. In light of this, every human being needs to be able to exercise free will and all of their mental abilities

[39] D. Tapscott, A. Tapscott, *Blockchain revolution: How the technology behind Bitcoin is changing money, business, and the world*, Portfolio, London, 2016.

[40] W. Sententia, "Freedom by design: Transhumanist values and cognitive liberty", cit., pp. 356-357.

- *Choice*. Which is why the human mind's capacity shouldn't be restricted.

In any case, however, it should be noticed how to present the possibilities of a brain intervention (permanent or not) as an alternative to imprisonment integrates implicit coercion to the individual's will[41]. Policy and bioethics reports emphasize that even therapeutic or rehabilitative neuro-interventions must respect informed consent and mental integrity, otherwise they risk becoming instruments of control rather than care.

On the other hand, a pro-Cognitive Liberty stance contends that everyone who desires access to current neurotechnologies should have them readily available[42]. A cognitive enhancement may result from the free personal use of psychoactive substances and cognitive devices (e.g., transcranial direct current stimulator or neurofeedback equipment)[43]. However, the concept of enhancement may be related to both a hypothetical individual level (e.g., improving one's memory) and a hypothetical general level (e.g., drug treatment during academic exams).

This new position states as follows: to the point till one person is not directly harming others, cognitive enhancement should not be intervened or prohibited by the states or governments. This new stance says that governments shouldn't forbid cognitive enhancement or the manifestation of any other mental state until one individual directly harms others. On the other hand, even while society may view the use of these "treatments" as morally acceptable, there is little data to support their effectiveness, and there may be long-term safety concerns that would warrant caution. The ongoing expansion of digital enhancement tools, such as algorithmic cognitive training, brain–computer interfaces, and neurofeedback apps, blurs the line between voluntary improvement and subtle behavioral manipulation.

[41] M. Farah, *op. cit.*, pp. 1123-1130.

[42] N. Bostrom, A. Sandberg, "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges", in *Science and Engineering Ethics*, 15 (2009), pp. 311-341.

[43] H. Maslen *et al.*, "The regulation of cognitive enhancement devices: extending the medical model", in *Journal of Law and the Biosciences*, 1 (2014), n. 1, pp. 68-93.

The dispute between transhumanists and conservatives is likewise based on this divide. The first part aims to "create the opportunity to live much longer and healthier lives, to enhance our memory and other intellectual faculties, to refine our emotional experiences and increase our subjective sense of well-being, and generally to achieve a greater degree of control over our own lives", the second part made the argument that the use of cognitive enhancement could have unanticipated and profound effects on society because it might enable people to develop cognitive structures that are outside the normal range of human experience[44].

And that's exactly it: a common definition of "normality" hasn't been developed, or we can say what neuro normality the other way is still to be defined. The statistical model, which is based on the observation of behavioral uniformity, and the socio-biological, or evolutionary, model are the two formulae that relate to normalcy in the neuro-scientific context[45]. Still, there are arguments against both theories. The statistical syndrome known as the bell curve, which is a standardized data distribution where there is one stupid for every genius, undermines any concept of normalcy in the case of empirical observation.

One could criticize the second one of using the approach of cognitive enhancement because it ends up in a dead end, making it impossible to identify the circumstances under which a particular behavior may have been a "good" or "bad" adoption to the context of social behavior. Thus it goes without saying that there are strong arguments in favor of the use of cognitive potentials being justified by our autonomy in determining our own identity and conscience. Still, a common criticism of cognitive augmentation emerges, even if we adopt a libertarian perspective. Since cognitive enhancement with the help of neuroscientific enhancers has the potential to be a very large market, the wealthy will have access to it while the poor will not, leading to even greater social divides. Since it represents a potentially huge

[44] G. Lynch *et al.*, "The Likelihood of Cognitive Enhancement", in *Pharmacology Biochemistry and Behavior*, 99 (2011), n. 2, pp. 116-129.

[45] P. Sommaggio, "Neuro-civilization: A New Form of Social Enhancement", in *ATINER'S Conference Paper Series*, n. SOS 2016-2106, pp. 3-18.

market, not only for drug companies but also for physicians who might enter the potentially lucrative market, especially cosmetic neurology[46]. These socioeconomic inequalities are now central to the neuroethics agenda, which examines how access to enhancement technologies may reproduce structural injustice and digital exclusion. Furthermore, it raises the question of whether the availability of enhancers would not impose professional obligations on those in high-risk occupations (like pilots or surgeons) to use them even if there is plausible doubt regarding their effectiveness and potential drawbacks[47].

8. Neuro-Human Rights and Global Governance

It has been observed that neurotechnologies have proved themselves with bearing capacity to influence and reshape legal frameworks, even though international human rights law has not yet directly addressed neuroscience[48]. Existing human-rights instruments are beginning to incorporate neuroethical concerns, but their practical implementation remains incomplete. This presents an issue. When considering global society from a broader perspective, we might argue that neurotechnologies have the power to completely change our understanding of it. They have the power to affect every individual, which explains why. Or, to put it another way, they have the power to alter every person's internal cognitive structure. This raises international and, in fact, human rights concerns.

The debate over the fundamentals of human rights interests us a little, therefore we'll stick to definition of human rights in this essay. According to Beitz (2011), human rights are mandated "to protect urgent individual interests against predictable dangerous ('standard threats') to which they are vulnerable un-

[46] J. Giordano, "Neuroethical issues in neurogenetic and neuro-transplantation technology: The need for pragmatism and preparedness in practice and policy", in *Studies in Ethics, Law, and Technology*, 4 (2011), n. 3 (https://doi.org/10.2202/1941-6008.1152).

[47] H. Maslen *et al.*, *op. cit.*, pp. 68-93.

[48] P. Sommaggio, M. Mazzocca, "Cognitive Liberty and Human Rights", in A. D'Aloia, M.C. D'Arrigo (eds.), *Neuroscience and Law*, Springer, Berlin, 2020, pp. 95-111.

der typical circumstances of life in a modern world order composed of states". In our opinion, cognitive liberty satisfies all of Beitz's criteria. This is because cognitive liberty can be understood as a prerequisite to safeguard the mind's autonomy against the interference of other entities (including the state), and this component is a fundamental aspect of the entire world[49].

Similar reasoning might be applied to prevent the so-called "rights inflation", which is the conventional argument against the acknowledgment or development of further human rights. To determine if these new rights are legitimate human rights, we apply a justificatory test. Nobody, we suppose, could contest the fact that Cognitive Liberty addresses a common and grave threat to a very important good. However, in the majority of countries, no one can prohibit its use or cast doubt on its viability. One could refer to this as the Nickel test. The prerequisites for the Nickel test are also included in Cognitive Liberty. Because of this, we believe that Cognitive Liberty can effectively "overcome" the issue of the inflation of human rights. Similar to this, Ienca and Andorno's investigation centers on the rejection of the coercive application of neurotechnologies and the creation of the legal category of Cognitive Liberty, both of which need to be backed by the introduction of new human (neuro) rights or the reinterpretation of already-existing human rights that includes the freedom of cognition, the right to psychological continuity, the right to mental integrity and the absolute right to mental privacy. This conceptual framework is now echoed in several academic and intergovernmental discussions that view these four dimensions as the foundation of future neuro-governance models.

It has been summarized that:

> In this paper, we primarily focus on the negative formulation of the right to cognitive liberty, which is the freedom to refuse coercive applications of neurotechnology, for our analysis. Furthermore, we contend that although the concept of the right to cognitive liberty has been introduced, it is insufficient on its own to address all of the ethical and legal ramifications that are connected to neurotechnology. Instead,

[49] C.R. Beitz, *The idea of human rights*, Oxford University Press, Oxford, 2011.

> a simultaneous reinterpretation of already-existing rights or even the development of brand-new, neuro-specific rights should be synchronized with the recognition of cognitive liberty as a human right. This includes the rights to psychological continuity, mental integrity, and privacy[50].
>
> In consideration of the first point, the question that arises is whether actual standards of privacy protection include the information included in or generated by our mind[51]. The 2024 European Parliament report on neurotechnologies raises this same issue, noting that mental data and neural signals fall outside many existing privacy laws, and thus require tailored safeguards.

An additional issue is that of criminal organizations' attacks on the brain. In the same manner that computer hackers influence brain capacities and the ensuing mental integrity, they can do the same with neurological devices. Since Article 3 of the European Charter of Human Rights emphasizes the right to health and biology, everyone is aware that bodily and psychological integrity is currently protected biologically[52]. The protection of the mental component from potential injury by others, as well as facilitating easy access to psychiatric therapies and assistance for people with mental health difficulties, are two goals of mental integrity. However, modern debates extend this protection to include "neurosecurity," emphasizing the prevention of unauthorized access to neural data or devices. A distinct regulatory safeguard against potential neurochemical therapies intended to

[50] M. Ienca, R. Andorno, "Towards new human rights in the age of neuroscience andneurotechnology", in *Life Sciences, Society, and Policy*, 13 (2017), n. 1, pp. 1-27.

[51] A possible protection is provided by the European Convention on Human Rights in Article 8, which recognizes the right to respect family life, domicile and correspondence and co. 2 states: "There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others".

[52] However, it is necessary to recognize that the rights of the Charter apply only to the institutions, agencies and bodies of the Union respecting the principle of subsidiarity as well as to Member States in the implementation of Union law, as stated in art. 51.

permanently alter an individual's personality with direct cognitive damage should result from this reexamination of mental integrity.

Based on Article 8 of the European Convention on Human Rights and acknowledged by the Universal Declaration of Human Rights, the European Court of Human Rights devised a neurological specification about the right to personal identity, which is known as the right to psychological continuity. Art. 22 and 29 herein explain the right to personal fulfillment and the full development of personality. Either way, approval of deliberate encroachments into personal space requires careful consideration and extensive public discussion[53].

Consequently, while they believe that Cognitive Liberty is a necessary condition for all rights about neuro aspects, we believe that, in light of their logic, it would be better to forego incorporating a new neuro-oriented right into the existing human rights declarations and instead concentrate on creating a completely new Declaration of Human Neuro-rights.

We believe we could claim a Universal Declaration on Neuro-Rights, just like we did with the idea of the Human Genome. This is due to the adaptability that human rights law has demonstrated in addressing the issues raised by genetic technology, which suggests that it could be a helpful tool to predict how this issue will change over the coming years. The route can resemble a step on a staircase, which is comparable to the debate over genetic concerns. It didn't take long for genetic issues to get international protection. The goal of preserving genetic privacy (against uses incompatible with human rights) and safeguarding the human genome led to the adoption of the Universal Declaration on the Human Genome and Human Rights in 1997. An International Declaration on Human Genetic Data was created in 2003 that outlined these ideas. Human rights and bioethics issues are closely related, as the 2005 Universal Declaration on Bioethics and Human Rights further clarified.

Following that same trajectory, a Universal Declaration on Neuro-Rights could establish a normative framework capable of addressing emerging neuroethical challenges, such as cogni-

[53] M. Ienca, R. Andorno, *op. cit.*, pp. 1-27.

tive manipulation, neurodata misuse, and digital surveillance of mental activity, while maintaining coherence with existing human-rights law. Thus, we believe that, in contrast to what Bublitz and Ienca-Andorno have written, it would be not only preferable and just but also simpler to resolve the issues surrounding the concept of Cognitive Liberty by creating a new Declaration of Human Neuro-Rights and following the already established path about the Human Genome[54].

9. Blockchain, Neurolaw, and Digital Accountability

The modern concept of neuro-law emerges with the fusion of neuroscience and legal frameworks that establish a complex set of challenges and opportunities in the face of rapid technological evolution. Cognitive liberty is one of the essential elements of neuro law, which emphasizes the right to privacy, autonomy, and control over one's mental state. The concept of cognitive ability has become increasingly relevant with the development of the latest neurotechnologies that have the potential to alter and monitor human brain activity. In this evolving environment, digital infrastructures are becoming indispensable to guarantee traceability, data integrity, and ethical accountability. Blockchain technology in this context has emerged as a critical enabler in reinforcing different neuro-legal frameworks and ensuring the protection of cognitive liberty through evolving mechanisms for secure consent management, ethical data sharing, and transparent governance of neurotechnology applications[55].

In neurolaw, consent management is an important legal and ethical issue within neurotechnological applications that demands quick and efficient mechanisms to ensure that the consent under consideration is informed, voluntary, and reversible. Blockchain provides a novel solution to the emerging challenges relating to consent management by leveraging smart contracts and immutable ledgers that can create transparent, incontrovertible records of user consent. This approach of preserving consent management has significantly enhanced individual autono-

[54] *Ibidem*.

[55] J. Rosen, B. Wittes (eds.), *Constitution 3.0: freedom and technological change*, Brookings Institution Press, Washington DC, 2011.

my and control over participation in neurotechnology studies or therapies[56]. Pilot projects[57] in medical data governance and research ethics have demonstrated how blockchain-based consent tools can allow participants to update or revoke consent dynamically, improving compliance with international data-protection standards[58].

The idea of regulating consent management under blockchain regulation aligns with legal consent standards that ensure compliance with the requirements of being informed, specific, and documented and can address this crucial aspect of neuro-law[59].

The sharing of neural data presents significant privacy concerns and emphasizes establishing a balance between advancing scientific research and ensuring the protection of individual privacy regarding neural data. With blockchain technology implementation we can offer a secure and decentralized platform for data sharing that safeguards the confidentiality and integrity of neural data. By combining encryption, anonymization, and permissioned access models, blockchain architectures can now enable ethical collaboration among research institutions while maintaining privacy-by-design principles. Blockchain technology has different encryptions and smart contracts, and by employing them, it can establish a controlled environment for ethical and responsible data sharing, contingent upon explicit consent from data subjects. The advancement with blockchain technology in securing the privacy of neural data not only meets different ethical and legal standards but also increases the trust

[56] G. Zyskind, *op. cit.*, pp. 180-184.

[57] W.M. Charles, M.B. van der Waal, J. Flach, A. Bisschop, R.X. van der Waal, H. Es-Sbai, C.J. McLeod, "Blockchain-Based Dynamic Consent and its Applications for Patient-Centric Research and Health Information Sharing: Protocol for an Integrative Review", in *JMIR Research Protocols*, 13 (2024), doi: 10.2196/50339.

[58] C. Welzel, M. Ostermann, H.L. Smith, T. Minssen, T. Kirsten, S. Gilbert, "Enabling secure and self determined health data sharing and consent management", in *NPJ Digital Medicine*, 8 (2025), n. 1, p. 560.

[59] S. Wachter, B. Mittelstadt, L. Floridi, "Why a right to an explanation of automated decision-making does not exist in the General Data Protection Regulation", in *International Data Privacy Law*, 7 (2017), n. 2, pp. 76-99.

and transparency among research participants and scientists, further demonstrating blockchain's utility in neuro-law[60].

Another striking feature of blockchain technology is its capacity for transparent and immutable transaction recording. It provides an unmatched opportunity for governing neurotechnology applications. By precisely documenting each use case and ensuring compliance with established ethical guidelines, blockchain introduces a new level of accountability, supporting regulatory bodies in enforcing legal and ethical standards. With this idea, blockchain technology can offer a workable framework for auditing and compliance verification which can increase public trust and confidence in neurotechnologies[61].

Summarizing the implementation and impact of the adoption of blockchain technology within neuro-law, we can say that this synchronization if achieved can leap a monumental shift to more secure, transparent, and accountable neurotechnology practices. By ensuring and incorporating robust consent management, ethical data sharing, and effective governance, blockchain can offer a workable and comprehensive solution to the challenges posed by neurotechnology advancements. However, successful integration will depend on parallel development of regulatory standards and digital literacy among legal professionals to interpret and apply blockchain evidence responsibly. This pioneering strategy of synchronizing blockchain technology in neurolaw not only safeguards cognitive liberty but also lays the basis for the responsible development and application of neurotechnologies. The continued and evolving collaborations among neurotechnologists, legal professionals, and ethicists will be fundamental in maximizing the potential of blockchain to protect cognitive freedoms in an increasingly digital age.

10. Conclusion

The writers we showcased have a joint neuroscientific endeavor: that unfolds mechanisms of a legal and societal improvement. A kind of "neurocivilization", where new neuroscientific

[60] M. Ienca, R. Andorno, *op. cit.*, pp. 1-27.

[61] M.J. Farah, *Neuroethics: An introduction with readings*, MIT Press, Cambridge, 2010.

norms are used in place of legal precedents, and illegal behavior is eradicated through direct brain stimulation. A harmonious future for a "better" society is promised by the neuroscientific social structure, which stigmatizes undesirable (non-normal) behavior. This is the point: where traditional humanities have failed[62]. Neurosciences promise to address social problems by direct and modified interventions. Yet, as recent neuroethical debates have shown, these promises also raise deep concerns about autonomy, identity, and the limits of scientific authority in shaping moral order[63].

However, they remain incapable of (or unwilling to) offering a shared social model to which one can strive to establish the parameters of what constitutes normal and pathological behavior. They typically only address an imprecise undesirability, which, still, allows room for dubious, if not dangerous, arbitrary remedies. Neuroscientists, as we demonstrated, should come out of the labs and take part in the discussions about the direction of society (and law), offering an ostensibly "neutral" viewpoint while striving for a significant shift. Radical neuro-enthusiasts claim that this transition will be traumatic, whereas neuro-tepid-reformists claim that it will be gradual and less intrusive. As per the latter, advances in neuroscience and technology will merely lead to a gradual enhancement of civilization.

Under these new circumstances, neurolaw will remain a controlled method, reduced to a tool for societal growth, according to scientific criteria rather than moral principles. As a result, the condition of infirmity is linked to social dangerousness and is conceptually equivalent to "mental disorder" actions that are signs of neuronal brutality and cannot be tolerated in a society founded on neuroscience[64].

Thus, the question is not whether neuro-civilization is good or bad, but rather which domains it would invade without prop-

[62] About brainwashing, see W. Bowart, *Operation Mind Control*, Collins Sons & Co., Glasgow, 1978.

[63] K. Taylor, *Brainwashing. The science of thought control*, Oxford University Press, Oxford, 2004.

[64] S.J. Morse, "The status of neurolaw: A plea for current modesty and future cautious optimism", in *The Journal of Psychiatry & Law*, 39 (2011), n. 4, pp. 595-626.

er regard for individual liberty, or where neuro-deviance would have to be eradicated by forceful brain intervention. We must acknowledge the significant role played by the deviant figure. Additionally, it stands for the opposition to social order that is critical and compels society to examine itself. Undoubtedly, this was the assignment given to the most well-known outcast, Socrates. Finding a place for this actor is, in our opinion, imperative, even in a society founded on neuroscientific principles.

We showed, in conclusion, the critical role that cognitive liberty plays in this new neuro-centric society. We started by outlining the significance and characteristics of the idea of cognitive liberty, which is seen as the foundation for all other rights because it is their neuro-cognitive system.

Secondly, we discussed the suggestions made by various advocates of cognitive liberty, who see it as both a basic human right and a key legal precept that directs the regulation of neurotechnologies. As Bublitz noted, it is "hard to conceive of any conception of a legal subject in which the mind and mental capacities (e.g. Acting from reasons, deliberation) are not among its necessary constitutive conditions". This is something to keep in mind in this regard[65].

Subsequently, as a third step, we argued how Cognitive Liberty has all the features it needs to make it a key concept from which new human rights can emerge.

It cannot be boiled down to only the rights that already exist, for this reason. It could, however, be viewed as the foundation for all liberty, both internal and exterior. Cognitive liberty aligns with the definition of human rights, which states that all people have the inalienable right to certain freedoms "just because they are Human"[66] irrespective of their nationality, place of residence, language, religion, ethnic origin, or any other status. This is because all people possess cognitive life, in it in varying degrees and forms.

[65] J. Bublitz, "My mind is Mine!? Cognitive Liberty as a Legal Concept", in E. Hildt, A. Francke (eds.), *Cognitive Enhancement*, Springer, Dordrecht, 2013, pp. 233-264.

[66] M. Sepúlveda, T. Van Banning, W.J.M. Van Genugten, *Human rights reference handbook*, University for Peace, Ciudad Colon, 2004.

As a fourth Step, we demonstrated how the protection of human characteristics that are not fully covered by current rights will be made possible by the inclusion of cognitive liberty within this framework of human rights. Then, we discussed how data confidentiality and integrity are guaranteed by Blockchain technology, which enhances the way neuroscientific evidence is processed legally. Recent developments show that blockchain's integration with neurotechnology research platforms supports transparent consent tracking, ethical review documentation, and audit mechanisms that strengthen accountability. Technology has the power to improve the reliability and equity of legal processes; yet, it requires the resolution of technological challenges and the development of digital literacy among attorneys.

Finally, we discussed how blockchain technology in neuro-law promotes cognitive liberty by facilitating ethical data exchange and safe consent management. Blockchain technology ensures data confidentiality and integrity, which improves the legal processing of neuroscientific evidence. The incorporation of technology holds the potential to enhance the dependability and equity of legal procedures; yet, it necessitates the resolution of technological obstacles and the promotion of digital literacy among legal practitioners. These efforts align with broader interdisciplinary initiatives to embed transparency and proportionality into neuro-data governance frameworks (OECD, 2024). This encourages confidence and accountability in the use of neurotechnologies, making interdisciplinary cooperation necessary for responsible development and implementation.

In this essay, we suggested viewing these actions as rungs on a figurative staircase leading to the national and worldwide defense of each person's inner world. In this way, Cognitive Liberty might serve as a foundational idea for a novel form of "habeas corpus", a legal avenue via which an individual might denounce unauthorized legal access to their inner life.

This new "habeas mentem" would imply that "my mind is free", that is, unrestricted by outside influences and able to evolve whenever we see fit[67]. As current debates in neuroeth-

[67] See P. Sommaggio, "La tutela della integrità mentale tra proprietà, beni comuni e libertà cognitiva", in *Sociologia del diritto*, (2022), n. 1, pp. 98-128.

ics and digital law indicate, the recognition of such a safeguard may become central to protecting autonomy in the age of connected neurotechnologies. In conclusion, we suggest approval for the legal acknowledgment of neuro-cognitive problems in a proactive and defensive manner, as well as the incorporation of blockchain technology for data security. What shape these neuro-rights take is not important to me. Discovering this issue and placing Cognitive Liberty at the center of this conceptual turning point for our global society of the future are things that intrigue us. The path forward will require sustained dialogue between law, neuroscience, and technology – ensuring that innovation continues without sacrificing the dignity and freedom of the human mind.

Radosław Strugała

Liability for Damage Caused to the Data Subject in the Event of GDPR Violation "by Artificial Intelligence": Between Legal Discrepancies and Policy Questions

Abstract: **Liability for Damage Caused to the Data Subject in the Event of GDPR Violation "by Artificial Intelligence": Between Legal Discrepancies and Policy Questions.** The paper discusses the topic of liability for data breach by AI. It aims at tracing the reason of rare application of the article 82 of the GDPR and proposes predictions of how its application may change with the potential enactment of special liability regime for AI-inflicted damage. The paper offers insight into how such a regime may interplay with the GDPR by pointing out the risks linked to the potential overlap thereof and offering a way to remedy such risk.

Keywords: The GDPR, Personal Data, Artificial Intelligence, Data Breach.

Summary: 1. Introduction – 2. Legal Nature of Liability Under the Article 82 of the GDPR, the Preconditions to Liability – 3. GDPR Violations in the Course of Personal Data Processing by AI – 4. AI and Damage – 5. Damage Consequent Upon Data Breach and Damage Caused by AI – 6. Concluding Remarks: Policy Choices.

1. Introduction

The issue posed in the title remains on the margins of a lively discussion on both the legal aspects of the use of artificial intelligence and the protection of personal data.

When it comes to the latter, the discussion focuses mostly on the General Data Protection Regulation (GDPR) compliance. Much weight is given to the interpretation of the GDPR provisions aimed at establishing the content of controllers' and processors' duties in order to secure that personal data processing be in line with GDPR in individual cases. Undoubtedly this interpretation emerges to be more challenging where personal data is processed with the help of AI (or "by AI"). And the AI act coming into force adds to the complexity of the picture of the controllers' and processors' duties. Whereas as far as the legal consequences of GDPR non-compliance are concerned, pub-

lic sanctions (administrative fines as provided in the article 83 of the GDPR) are in the centre of interest in the legal writing. Hardly ever extended analysis is made in Academia concerning private law measures applicable in the case of GDPR violations, albeit the GDPR expressly provides for the private enforcement tool in the form of compensatory remedies (article 82 of the GDPR). This state of legal thought reflects the common perception of practical importance of public and private remedies available under the GDPR. While the administrative fines imposed by national supervisory authorities (be it *Il Garante per la Protezione dei Dati Personali*, *Prezes Urzędu Ochrony Danych Osobowych*, *Bundesbeauftragter für den Datenschutz und die Informationsfreiheit*) are perceived as effective and thus become practically important, commonly applied sanctions of the GDPR infringements, the article 82 of the GDPR is rarely invoked to remedy GDPR violations.

As regards to the legal aspects of AI functioning, liability for damage caused by artificial intelligence (or by the use of AI systems) undoubtedly remains among the central issues discussed by the legal scholars. However, the focus in the discussion is primarily on "classic damage" being a consequence of the violation of traditionally recognised legally protected interests, such as life, health or property of the injured person. In the course of the debate the problem of damage resulting from the unlawful processing of personal data is hardly ever addressed, if it is touched upon at all. The same holds true for the legislative proposals that are successively presented by the EU institutions with aim to harmonise liability for damage caused by AI at the EU level. Thus far two such proposals were published, each given a different legal status as one was presented as a draft regulation (Draft Regulation)[1] and the other one was planned to be enacted in the form of respective directive (Draft Directive)[2]. None of these propos-

[1] Civil liability regime for artificial intelligence European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)) (2021/C 404/05).

[2] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) COM/2022/496 final.

als expressly and directly relates to liability for damage caused by AI as a result of violation of the legal rules on processing of personal data. The preambles to the drafts also lack mentions of the issue clear enough to allow to ascertain whether including liability for losses consequent upon data infringement was the intention of the law maker proposing AI liability rules at hand.

The goal of this paper is to shed light on the neglected topic of liability in damages for data breach by AI. Particularly it purports to establish the reason for limited application of the article 82 of the GDPR (compensatory remedy offered therein) by data subjects who currently make such a rare use thereof. On top of it, it proposes predictions of how the fate of the article at scrutiny may change due to the development of AI technology and the increase of automated data processing and consequential proliferation of instances of data breach by AI. Looked at from this angle the problem tackled in this article appears to lie at the crossroads of the two sets of legal rules governing both personal data processing (like the GDPR) and liability for AI-inflicted losses yet to be enacted (i.e. draft legal rules proposed in the legislative proposals mentioned above). This paper offers insight into how the two sets of rules may interplay in the likely event of the above-mentioned or similar special rules on liability for AI becoming a binding law in the future. The juxtaposition of these sets of rules leads to questions of twofold nature. First, which of these legal rules are to be applied where damage is caused as a consequence to data breach by AI system. This is a question which lawyers would be forced to face if the above-mentioned legislative plans of introducing special regime of liability for AI came true. This is a rather technical question that the lawmaker could easily avoid by expressly deciding the relation of the above-mentioned legislative proposals on liability for AI and the article 82 of the GDPR. What is way more complex are the policy reasons underlying any such decision. One may even speculate that no decision as to the relation of the two sets of legal rules was expressly made, for no policy choices have been made thus far in the field. The ultimate goal of this paper is to try and formulate policy directions that should determine future decisions of the law maker. As it was underlined, the liability scheme

provided for in the article 82 of the GDPR draws so far little attention in the legal writing. Thus, some preliminary observations about its legal nature and preconditions to liability should be made. Let it be the starting point of the further analysis.

2. Legal Nature of Liability Under the Article 82 of the GDPR, the Preconditions to Liability

According to the article 82 of the GDPR any person who has suffered material or non-material damage as a result of an infringement of this regulation shall have the right to receive compensation from the controller or processor for the damage suffered. When talking about preconditions to liability in damages traditionally three are pointed out, namely: a loss or damage (be it monetary, non-monetary or both kinds of losses), an event defined in the piece of legislation imposing liability (the event that may trigger liability if it constitutes a cause of loss) and the causal link between this event and the items of loss incurred by the victim seeking compensation. The liability provided for in the article 82 of the GDPR is usually viewed as tortious liability (non-contractual liability)[3]. This follows that for this liability to arise all above-mentioned preconditions need to be met. This view is approved of by the CJEU in a judgement referred to as *Österreichische Post* case. The CJEU underlined that

> it is clear from the wording of that provision that the existence of 'damage' which has been 'suffered' constitutes one of the conditions for the right to compensation laid down in that provision, as does the existence of an infringement of the GDPR and of a causal link between that damage and that infringement, those three conditions being cumulative[4].

3. GDPR Violations in the Course of Personal Data Processing by AI

As in the case of "traditional", non-automated personal data processing, data breach may well occur in the course of data pro-

[3] See E. Tosi, *Responsabilità civile per illecito trattamento dei dati personali e danno non patrimoniale*, Giuffré, Milano, 2019, p. 49.

[4] Judgment (Third Chamber) of 04/05/2023, Österreichische Post (Préjudice moral lié au traitement de données personnelles), case C-300/21.

cessing by AI. The legal protection provided for in the GDPR is technologically neutral[5]. The way it imposes duties on controllers and processors follows the rights-based approach[6]. It means that the infringement of the GDPR can be ascribed to controllers or processors regardless of whether data processing contrary to the GDPR provisions was the performance of their own hands (where controllers or processors collected, recorded, organized, structured, stored, reviewed, used or disclosed personal data in an "analog" manner) or the result of the use of an AI system (working for the controller or processor). At the same time, the use of AI systems seems to bring about the hazard of new risks of the GDPR infringements that are inherent to this type of technology. A simple use of common Chatbots (which undoubtedly bear the hallmarks of AI), while interacting with data subjects (customers) may lead to instances of such data breach. For example, a GDPR violation could occur in the case of a Chatbot processing given personal data by using them as "training data" with the purpose of deep learning. A situation that should definitely be viewed as GDPR infringement is a likely scenario where the controller who makes use of a Chatbot obtains consent and provides the information clause only in respect to the main pur-

[5] See recital 15 of the GDPR preamble, according to which "in order to prevent creating a serious risk of circumvention, the protection of natural persons should be technologically neutral and should not depend on the techniques used. The protection of natural persons should apply to the processing of personal data by automated means, as well as to manual processing, if the personal data are contained or are intended to be contained in a filing system. Files or sets of files, as well as their cover pages, which are not structured according to specific criteria should not fall within the scope of this Regulation". See also the article 4 (2) of the GDPR, which expressly provides that "processing [of personal data] means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction".

[6] See L. Edwards, M. Veale, "Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for", in *Duke Law and Technology Review*, 16 (2019), pp. 16-84; O. Lynskey, *The Foundations of EU Data Protection Law*, Oxford University Press, Oxford, 2015; R. Gellert, "On Risk, Balancing, and Data Protection: A Response to van der Sloot", in *European Data Protection Law Review*, 2 (2017), p. 185.

pose of data processing, which is the possibility of establishing commercial contact and conducting communication, including answering questions addressed to the controller regarding the goods or services offered. According to one possible interpretation of the GDPR provisions, the use of collected personal data as a training data set does not require separate consent of the data subject, for the legal basis for this type of processing is constituted by the Article 6(1)(f) of the GDPR, which means that personal data processing is deemed necessary for the purposes of the legitimate interests pursued by the data controller. This approach was followed in a widely discussed case of a temporal ban on the processing of Italian users' personal data by OpenAI (i.e. the company launching ChatGPT) imposed by *Garante per la Protezione dei Dati Personali* (Italian national supervisory authority within the meaning of the GDPR). *Il Garante* rightly considered OpenAI to be data controller of personal data collected through websites or social media in order to be used for machine learning by ChatGPT. Initially, *il Garante* accused OpenAI of not having provided the users of ChatGPT (and therefore data subjects) with appropriate information about the processing of their data, and of having processed personal data without an appropriate legal basis. However, as a final chord of the dispute between *il Garante* and OpenAI, the former, after formulating guidelines regarding the company's information policy, allowed OpenAI to process personal data of Italian users pursuant to Art. 6(1) 1 letter f GDPR.

Even if we accept the reasoning of the Italian supervisory authority, the mere recognition of the data processing with aim to train the system within the framework of machine learning being the legitimate interest of the data controller (the Article 6(1)(f) of the GDPR), does not mean that the controller is exempt from the information duties. It is still the controller's duty under the GDPR to provide the data subject with information (information clause) on the legal basis of data processing and on "additional" purpose of the processing as training data (see Article 13(1)(c) of the GDPR[7]). In my view, some of the controllers,

[7] According to which, the controller shall provide the data subject with information about purposes of the processing as well as the legal basis for

used to "traditional information clause content" and sometimes even not aware of the fact that the Chatbot used by them may collect personal data and use it as training data, will be prone to fail to fully comply with these information duties. Another possible case of GDPR violation in connection with the use of a Chatbot that can easily be thought of, is a scenario in which an entrepreneur makes use of a Chatbot or similar AI system to entrust the system with automated individual decision-making regarding their current or potential client, which would clearly be in violation of the Article 22 of the GDPR, as well as the Article 5 of the Artificial Intelligence Act.

To conclude, one can notice that automated personal data processing by AI not only does not mitigate the risk of GDPR infringement but can even broaden the potential areas of data breach. At first sight, it could lead to a conclusion that the increased number of cases of data processing by AI may render the article 82 of the GDPR more practical as more data subjects could decide to seek compensation under this provision for losses incurred in the process of AI processing their personal data. Whether this is an apt observation needs to be verified taking into account other preconditions to liability provided in the article 82 of the GDPR and the potential to meet them in case of personal data processing by AI.

4. AI and Damage

Damage or loss constitutes the inherent, necessary precondition to any type of liability in damages, be it tortious liability or liability for breach of contract. The same is true for the liability scheme as provided in the Article 82 of the GDPR. The provi-

the processing. However, if specific personal data are collected for several different purposes, the controller is obliged to indicate all these purposes, and if data processing is carried out pursuant to Art. 6 section 1 letter f GDPR, the controller should also provide the data subject with information about the legitimate interest they pursue when processing data (see P. Barta, M. Kawecki, P. Litwiński, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 in on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Commentary*, C.H. Beck 2021, pp. 226 ss.

sion at hand expressly renders liability dependent on the damage caused as a result of a violation of the GDPR provisions[8]. The wording of the Article 82 of the GDPR leaves no doubt that the liability may be triggered by the occurrence of both material or (and) non-material damage. Reported cases of article 82 application in disputes pending before national civil courts in the EU member states, even if not numerous, demonstrate that data subjects invoke the article 82 of the GDPR to seek compensation of non-monetary damage. Hardly ever it appears as legal basis in the context of monetary loss. This is probably due to the fact that personal data processing not in line with the GDPR has low potential to directly cause any material damage[9]. Contrary conclusion seems to be legitimate when it comes to non-monetary damage. In contrast to monetary damage, which is understood similarly in the majority of legal systems, the notion of non-monetary damage is given manifold meaning across EU Member States. The question of whether GDPR violation by AI may actually lead to non-monetary damage subject to compensation under the Article 82 depends, thus, on the way the concept of non-monetary damage is to be understood in this very provision. Ever since the GDPR got into force, there could be no doubt that the term "non-monetary damage" in the article 82 thereof should be interpreted autonomously, independently of the content of equivalent concepts in national laws of the Member States. It was thus rather clear, at least in Academia, that when interpreting the term no resort must be made to "national" non-monetary damage, e.g. within the meaning of the article 448 of the Polish Civil Code, article 2059 of the Italian Civil Code or section 253 of BGB[10]. Quite recently this conclusion was clearly

[8] See Judgment (Third Chamber) of 04/05/2023, Österreichische Post (Préjudice moral lié au traitement de données personnelles), case C-300/21.

[9] However, such damage may certainly be caused indirectly – e.g. as a result of the leak of personal data originally collected by AI without the consent of the data subject, which is then used by a third party to obtain a loan.

[10] R. Strugała, "RODO a odpowiedzialność odszkodowawcza, Podstawowe problemy odpowiedzialności za szkodę spowodowaną nieprawidłowym przetwarzaniem danych osobowych", in *Monitor Prawniczy* 17 (2018), p. 915; A. Pązik, "Szkoda wynikająca z naruszenia przepisów RODO. Wy-

confirmed by the CJEU[11]. Regardless of the certainty as to the method by which the term "non-monetary damage" in the article 82 of the GDPR should be interpreted, the result of this interpretation remained highly uncertain for a very long time. This uncertainty seems to be one of the main reasons why the article 82 of the GDPR was so rarely invoked by data subjects.

Both in case law of national civil courts and in legal writing, three competing approaches to the interpretation of the term "non-monetary damage" in the article 82 of the GDPR could be distinguished. The most liberal of them assumed that non-monetary damage within the meaning of the Article 82 of the GDPR is constituted by the mere fact of the GDPR provisions being violated. Thus, according to this interpretation, each data breach vis-a-vis data subject equals their non-monetary damage by definition[12]. According to the alternative, dominant view (represented by the case law of German and Austrian courts), non-material damage giving rise to a claim for compensation under the Article 82 of the GDPR is a phenomenon separate from the GDPR infringement itself. This perspective holds, in a nutshell, that the non-material damage takes the form of negative feelings of the

brane problemy", in *Zeszyty Naukowe Uniwersytetu Jagiellońskiego. Prace z Prawa Własności Intelektualnej*, 3 (2020), p. 133.

[11] See the Judgment (Third Chamber) of 04 May 2023, Österreichische Post (Préjudice moral lié au traitement de données personnelles), case C-300/21, which expressly undrelines that "the GDPR makes no reference to the law of the Member States as regards the meaning and scope of the terms set out in Article 82 of that regulation, in particular as regards the concepts of 'material or non-material damage' and of 'compensation for the damage suffered'. It follows that those terms must be regarded, for the purposes of the application of that regulation, as constituting autonomous concepts of EU law which must be interpreted in a uniform manner in all of the Member States".

[12] For example, the Court of Appeal for England and Wales in its judgment of 2 October 2019 (in the case of Lloyd v. Google, LLC [2019] EWCA Civ 1599). Although the decision was made before the so-called Brexit, on the basis of the "legal predecessor of the GDPR", i.e. Directive 95/46, it was later recognized in the scholarship as valid under the article 82 GDPR (see e.g. B. Van Alsenoy, "Liability under EU Data Protection Law: From Directive 95/46 to the General Data Protection Regulation", in *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, 3 (2016), p. 284).

injured data subject. It is restrictive, however, as it limits the scope of the notion of non-monetary damage only to serious detrimental emotional effects of the GDPR violation. Mere stress, discomfort, or anxiety (because of the loss of control over one's own data) do not stand for non-material damage according to the interpretation referred to. For the non-material damage to occur, it requires that detrimental emotional effects of the GDPR violation be a consequence of social exclusion or humiliation, which constitutes "the minimis rule" or "a minimum threshold of seriousness"[13]. "The minimum threshold of seriousness" set by German courts resulted in a very limited scope of detriments that could be viewed as non-material damage under the article 82 – in fact it would amount to the effects of disclosing the injured party's sensitive data, if the interpretation at hand was accepted. The third, most balanced position as to the meaning of non-material damage, links this notion to negative feelings of the injured data subject, it does not require, however, any "minimum threshold of seriousness" to be met. This allows for the mere stress or discomfort due to the loss of control combined with the fear of their unauthorized use by a third party to be recognized as non-material damage, as referred to in Article 82 of the GDPR[14].

The interpretation of the concept of non-damage has only just begun to be clarified in the case law of the CJEU. In *Österreichische Post* case (in response to a question referred for a preliminary ruling by the Austrian Supreme Court), the CJEU expressed a position closest to the third approach mentioned above. It explained that the occurrence of non-monetary damage within the meaning of Article 82 of the GDPR must not be confused with the violation of the GDPR provisions, as the two elements stand for two separate preconditions to liability both being necessary to meet for the liability to arise. The CJEU un-

[13] See e.g. the judgment of the regional court in Karlsruhe of 2 August 2019, O 26/19; the judgment of the District Court in Hamburg of 4 September 2020, 324 S 9/19; judgment of the District Court in Frankfurt am Main of 10 July 2020, 385 C 155/19.

[14] This interpretation was presented to the CJEU by the Supreme Administrative Court (Varhoven Administrativen Sad) of Bulgaria within the preliminary question procedure in the case C 340/21.

derlined that in a compensation case the plaintiff is required to demonstrate the existence of negative (non-material) effects of a given GDPR infringement as a separate (from the GDPR violation itself) liability prerequisite. At the same time, it refused the view that for the liability to arise given effects need to be particularly significant – as significant as to reach the "minimum threshold of seriousness". The facts of the *Österreichische Post* case were strikingly similar to the circumstances of the very famous "Cambridge Analytica scandal" that exploded a few years ago in the US. The plaintiff was seeking compensation under the Article 82 of the GDPR against the Österreichische Post as an entity that, in violation of the GDPR, processed the personal data of its clients using specific algorithms in order to "profile" them for political purposes – with aim to get precise information on the political preferences of the clients and to direct to them selected campaign advertising materials of specific political parties matching the "profile" of the addressee. The plaintiff held that non-monetary damage that he allegedly incurred stemmed from the fact that political beliefs or sympathies were attributed to him that were inconsistent with his actual political affiliation. The plaintiff felt offended by the mere attribution of such political beliefs to him, and was also outraged by the fact that the company kept the information about his alleged preferences. He identified these feelings with non-monetary damage within the meaning of the Article 82 GDPR. The interpretation of this provision employed by the CJEU *Österreichische Post* case enables to conclude that such an identification may turn out legitimate.

The interpretative approach accepting non-monetary damage as separate precondition to liability, but rejecting the "minimum threshold of seriousness" was confirmed by the CJEU in the next judgment of December 14, 2023, in which the CJEU – referring to the content of recital 85 of the GDPR preamble – emphasized that a real fear of possible future misuse of personal data may amount to non-monetary damage. The Court noted that the existence of such a real fear should be decided in relation to the specific plaintiff and specific circumstances in which the GDPR violation occurred[15].

[15] See the CJEU Judgment of 14 December 2023, C340/21.

Despite such a reservation, it seems that the developing case law of the CJEU could constitute a breakthrough, encouraging potential plaintiffs to seek compensation protection under the article 82 of the GDPR more often. Irrespective of the way the CJEU approached the notion of the non-monetary damage, which is rather liberal, the very fact that the CJEU dwelled on it and expressed firm interpretative direction as to the notion will definitely reduce the uncertainty as to the actual content of the article 82 of the GDPR, and therefore also the degree of legal risk that has so far been associated with bringing actions based on this provision. It seems reasonable to ask whether the possible entry into force of further EU regulations, this time relating to liability for AI, would not mitigate this effect by introducing new legal uncertainty for data subjects to face whose data is processed in an automated manner – with an aid of AI (by AI).

5. Damage Consequent Upon Data Breach and Damage Caused by AI

As mentioned above, for some time now, analyses, studies, as well as specific legislative initiatives have been undertaken at the EU level regarding the liability for damage caused "by artificial intelligence". There are undoubtedly certain legal policy motives behind them, such as the desire to unify the legal rules governing this kind of liability throughout the EU, which could result in the EU's competitiveness in the field of artificial intelligence. There is no doubt that the way in which the legal rules governing liability for damage caused by AI are shaped may have an impact on the development of this technology. If they are excessively restrictive, their application will result in a chilling effect both in the area of innovation (when it comes to the responsibility of broadly understood technology producers) and the use of technology itself (when it comes to the rules of liability of the users).

However, the main reason behind taking up the problem of liability at hand at the EU level is that the liability rules currently in place in the majority of Member States may turn out to be ineffective in the event of the damage being "caused by AI". This is the conclusion commonly drawn in the private law

scholarship in relation to liability frameworks that are based on the principle of fault that would be applicable in the majority of Member States in the cases of AI causing damage. Legal systems providing for general strict liability of persons in control of things for damages caused by the use of these things are rather exceptional (see e.g. *la responsabilité du fait des choses* and *la responsabilità per cose in custodia* in the article 1242 of the French Civil Code and the article 2051 of the *Codice Civile*). Irrespective of whether as broad interpretation of such strict liability provisions is accepted as to include within the scope of their application losses caused by the operation of various types of new technologies (machines, robots or even the so called stand-alone AI software not embedded in any tangible hardware)[16], the strict liability would remain available for the injured persons in very limited number of Member States. In the major part of them for the injured party to seek compensation for losses caused by AI it would be necessary to identify the culpable behaviour of the AI user, owner or possessor and to establish a causal link between such behaviour and the damage caused by the specific functioning of AI.

In a world where human activity is replaced by the use of AI, tracing back the damage caused directly by the use of AI to any culpable human behaviour may be very challenging. The complexity of AI systems and limited knowledge into the way they operate from the perspective of the injured party (the so-called black box effect) gives rise to difficulties in linking specific action of AI that resulted in damage with the culpable behaviour of a person involved in the functioning of a given AI system. In the case of "intelligent machine" that caused damage as a result of a specific, unpredictable "behaviour" (e.g. in the case of a lawn mower that did not stop in time, injuring a toddler's hand), the injured party may meet enormous obstacle in determining whether the machine "behaved" in a harmful way because the

[16] As to this broad interpretation employed in the French law see European Parliamentary Research Service (ed. T. Evas), *Civil liability regime for Artificial Intelligence. European Added Value Assessmen*t, Brussels, 2020, p. 111; L. Archambault, L. Zimmermann, "Repairing damages caused by artificial intelligence: French law needs to evolve", in *Gazette du Palais*, 9 (2018), p. 17.

software was defective from the beginning (e.g. the code was bugged, which presupposes an error, i.e. fault on the side of the system designer), because the device was fed with incorrect data in the process deep learning, either because it was not updated properly (which presupposes the user's fault) or because its physical elements (sensors, cameras, etc.) did not work properly (which would constitute the fault of the manufacturer of the device or its parts). Since in the majority of legal systems, in the fault liability schemes the burden of proving causality is placed on the injured party, in cases of AI-inflicted damage the victim is put under the risk of failing to meet the burden of proof and consequently of remaining not compensated for their damage.

This risk becomes even more significant when the proof of fault is taken into consideration. Regardless of how the concept of fault is defined in a given legal system, a common feature of fault liability is the need for the injured party (who bears the burden of proof in this respect) to demonstrate certain irregularity in the behaviour of the defendant (deviation of their behaviour from the standard model of due, careful behaviour of a reasonable person). The standard (model) of behaviour constituting a reference point for determining fault may pose a huge challenge to the application of provisions on fault-based liability for damage caused by AI. This state of affairs is natural for every new phenomenon of social life. Natural difficulties in determining the standard of caution (due care) that should characterise the actions of people who use AI is a consequence of the lack of "historical data" – the lack of knowledge about what behaviours may intensify the risk of damage and what may be considered a reasonable use of AI preventing it from inflicting damage.

The EU legislative proposals mentioned at the beginning of the paper are aimed at eliminating the risk that the victims relying on fault liability would be put under. They do this in two ways, which vary in detail depending on the specific legislative proposal. The first possible way is to impose strict liability on the users (operators) of AI systems for damage caused as a result of the operation of a given AI system. The second is to impose liability dependent on fault on the users. This fault liability is accompanied by the reversal of the burden of proof – the pre-

sumption of fault or presumption of causation meant to improve the injured parties' position. The first model was employed in the Draft Regulation from 2020. In this draft, strict liability is proposed to be imposed on broadly understood operators (the so-called front-end operators and back-end operators, i.e. end users of the system and entities "technologically managing" the AI system) for damage caused by the so-called "high risk AI systems"[17] which would be periodically included in the list of high-risk systems issued by the European Commission on the regular basis. As far as the other types of AI are concerned (non-high risk systems), the Article 8 of the draft provides for the fault liability of the operators. However, fault is presumed according to the provision at hand and it additionally provides for an exhaustive list of exculpatory circumstances. The fault-based liability model has also been employed in the Draft Directive on liability for artificial intelligence. Strictly speaking, this draft is of procedural character. Its aim is not to harmonise the substantive law of the Member States. It provides for a mechanism to facilitate evidentiary rules by providing the defendant with the presumption of a causal link between the "harmful behaviour of the AI system" and the fault of the plaintiff, where the latter was proven by the way of special evidentiary proceedings[18].

As mentioned above, the relationship of the drafts of legal rules on liability for AI to the article 82 of the GDPR would appear highly unclear. If the drafts become a binding law, a natural question will arise whether in the case of the liability of the user in the case of data breach by the AI system processing data, the article 82 of the GDPR is to be applied (for the case regards the GDPR violation expressly dealt with by the GDPR) or the special liability rules at hand should be applicable (as the case involves damage caused by AI). At the first glance the question may seem farfetched. One can reckon that in the event of data breach in the course of data processing by AI, the damage is always caused by the use of AI (and thus by the controller or

[17] That is AI systems that have a significant potential to cause "randomly occurring damage to one or more persons in such a way that is beyond what could reasonably be expected".

[18] See P. Machnikowski, *Prawo deliktowe wobec nowych technologii*, Wolters Kluwer, Warszawa, 2023, pp. 565 ss.

processor who uses the AI system) and not by AI directly and the question at stake is not legitimate.

However, one should bear in mind that these special rules on liability (the draft rules mentioned above), if they become a binding law, will in fact concern liability for damage caused by humans (by the users relying on a given system bearing the hallmarks of AI). Strictly speaking, they will not concern the liability for damage caused "by artificial intelligence" as such. To the extent that the draft liability rules at hand are based on the principle of fault, they ascribe liability for damage caused by the operation of AI to persons who, while using this technology, have committed a specific violation of applicable standards of care (violations of legal provisions or precautionary rules) regarding the use of this technology (e.g. by not making the required system updates, by making an incorrect decision on the use of a certain AI system that was inappropriate in given circumstances, which led to an increased risk of causing damage compared to an alternative scenario in which specific action resulting in damage would have been taken by a human instead or by another, more adequate, AI system, or by neglecting to take all necessary and reasonable measures to avoid that the system start operating outside the consent and control of the user). In essence, the liability at hand can be viewed as liability for specific human behaviour (faulty behaviour) of the AI user that is the "background of the system's behaviour" from which the damage resulted. Talking about liability for damage "caused by AI", at least in some cases, is only a metaphor.

Consequently, it would not be reasonable to assume *a priori* that the GDPR infringements always occur as a result of human activity (controller or processor) and that the fact that the infringement occurs "while using AI" (within the course of processing data in an automated manner) should be neglected as completely legally irrelevant as the need to resort to special rules on liability never materialises. The potential overlap of the article 82 of the GDPR and the special rules of liability for AI (if enacted) would still raise a valid question of how the two sets of legal rules relate. This kind of question should be posed and answered conclusively when deciding cases of "data breach by

AI" as the choice of one (the GDPR) or another (special AI liability regime) could lead to overwhelmingly different results as to whether the liability arises at all and as to the scope of potential liability. To prove this point, a few examples can be made.

If we think of the violation of the GDPR by a high-risk AI system within the meaning of the above-mentioned draft regulation (e.g. in the case of collecting personal data by a drone which could potentially be qualified as high-risk), a question would arise whether the damage consequent upon the GDPR violation is subject to compensation under the Article 82 of the GDPR or under the rules contained in the to-be-regulation draft (due to the fact that the data processing rules specified in the GDPR were violated by high-risk AI). If the latter is chosen, the data-subject could claim limited compensation because articles 4 and 5 of the draft regulation, contrary to the GDPR, provide for a cap that limits the sum of damages to be granted to the victims. Similar doubts could arise where data breaches occurred as a result of a cyberattack. Under the GDPR the controller may escape liability for damage consequent upon the acquisition of personal data by a third party as a result of a cyberattack if they demonstrate that proportionate organizational and technical measures have been taken to prevent such an attack (Article 24 and 32 GDPR)[19]. Meanwhile, the Article 8 section 3 of the draft regulation clearly provides that "where the harm or damage was caused by a third party that interfered with the AI-system by modifying its functioning or its effects, the operator shall nonetheless be liable for the payment of compensation if such third party is untraceable or impecunious". If a specific AI system begins to "transfer" personal data collected by the original controller to a third party as a result of a cyber-attack, the answer to the question of whether the liability of the original controller arises would thus depend on whether the article 82 of the GDPR or the regulation (if enacted) is applicable.

The above described or similar doubts stem from legal discrepancies between the GDPR and special liability regime for AI accompanied by the lack of clarity as to the relation of the two sets of legal rules. Such obstacles could easily be avoided

[19] See the CJEU Judgement of 14 May 2021, C 340/21.

by the EU law maker at the final stage of the legislative process. If the decision is made to enact the special liability regime for AI, the EU legislative institutions could amend the drafts by expressly deciding their relation vis-a-vis article 82 of the GDPR. Potentially there are three options to choose from. The first is to give priority to article 82 of the GDPR and to decide that it is solely applicable even where damage consequential upon data breach was "caused by AI" (in the course of processing personal data by AI). The second is to decide that article 82 of the GDPR is not applicable in such cases, as it is overridden by the special liability regime because the damage was "caused by AI". The third possible choice is to decide that neither set of legal rules prevails and to leave for the victim (data subject) to freely decide whether they intend to rely on the GDPR or special liability regime. By choosing any of these three possible paths, the EU law maker would mitigate the risk of legal uncertainty linked to the future enactment of the special liability regime. Any clear cut by the law maker would be welcome from this perspective as it would remove the legal uncertainty potentially preventing the data subjects from seeking compensation of damage resulting from data breach by AI.

It would not be advisable, however, if the EU law maker limited themselves to this kind of minimum, technical problem solving. The puzzle of priority of the legal rules at hand is to be decided upon cautious consideration of reasons underpinning the special liability regime for AI.

6. Concluding Remarks: Policy Choices

As mentioned above, the draft rules on liability for AI serve the purpose of easing the access to fair compensation for the victims of AI-inflicted losses by removing evidentiary obstacles they would face when relying on fault liability schemes. However, in the event of AI-inflicted damage caused by data breach these obstacles do not appear under the GDPR. The legal position of the victims would not change (would not be improved), if they followed within the scope of special liability regime of AI. From the victim's (data subject's) perspective it does not matter at all whether a given AI system collected their personal data

without consent because this was the intent of the controller or because of the back-end operator's or programist's mistake that resulted in unintended data collection by AI system. In the light of rights-based approach employed in the GDPR, in any case the violation of the regulation can be said to have occurred. The GDPR makes it possible for the controller or processor to escape liability by proving if they prove not to be "in any way responsible for the event giving rise to the damage". According to the CJEU's view expressed in *Krankenversicherung Nordrhein* case[20], such a wording means that the liability provided in the article 82 of the GDPR is based on the principle of presumed fault. Thus the existence of fault as the precondition to liability is presumed, unless controllers prove otherwise (as the controller-defendant bears the burden of proving that the damage was caused as a result of circumstances for which he is not responsible). Thus under the GDPR, a scenario is impossible in which the data-subject would remain without compensation because of not being able to prove the preconditions of liability that substantially were met in the case. Therefore under the GDPR no risks appear that the special liability regime for AI oat to prevent. In my view the article 82 of the GDPR shall thus be given priority and remain a sole legal basis for liability for damage consequential upon data breach was "caused by AI" even if the special liability regime for AI is enacted.

[20] The CJEU Judgement of 21 December 2023, C-667/21, EU:C:2023:1022.

III.
NEUROSCIENZE, S/OGGETTIVITÀ E MODELLI RELAZIONALI

Salvatore Amato

Macchinazioni.
Soggettività, agentività ed entità esotiche

Abstract: **Machinations. Subjectivity, Agency and Exotic Entities.** Opacity is one of the main problems of artificial intelligence not only because it is technically difficult (and in some cases impossible) to open the black box, but because it constitutes the borderline zone between mechanisation of mind and humanisation of machine. This zone was the extreme outcome of reducing brain activity entirely to the biochemistry of neurons or the biophysics of synapses. The black box effect is the other side of the same coin: a kind of possible humanisation of machines, increasingly complex and unpredictable. Is human-centred artificial intelligence a return to natural law? Do ethics by design, info-ethics, algor-ethics, data ethics, data dignity reintroduce the principles and values of the natural law tradition?

Keywords: Expandiverse Technologies, Deep Learning, Opacity, Alien Reasoning, Natural Law.

Sommario: 1. L'inconscio digitale – 2. La scatola nera dell'apprendimento profondo – 3. *Exotic mind entities* – 4. Un diritto naturale al silicio.

> "La macchina si adatta alla debolezza dell'uomo, per fare dell'uomo debole una macchina".
> K. Marx, *Manoscritti economico-filosofici.*

1. L'inconscio digitale

Se già nel medioevo, con meraviglia e preoccupazione, si affermava che "ogni giorno si scopre una nuova tecnica"[1], oggi sono gli aggettivi che ci mancano, *computing machine*, *business machine*, *learning machine*, *smart machine adaptive machine, soul machine* per indicare non solo come le macchine sostengono le nostre azioni, ma soprattutto in che modo e fino a che punto influenzino i nostri pensieri, condizionino la nostra cultura, determinino il nostro modo di essere. Sono *ingenia*, frutto

[1] Lo ricorda C.M. Cipolla, *Le macchine del tempo. L'orologio e la società 1300-1700*, il Mulino, Bologna, 1996, p. 16.

del nostro ingegno, ma anche *search engineering*, *neuromorphic engineering*, costruttrici del proprio ingegno. A volte incapaci di rispondere alle nostre attese: *underfitting*. A volte riluttanti a produrre nuovi sviluppi: *overfitting*. A volte incontrollabili nelle procedure e imprevedibili nei risultati: *black box effect*. In ogni caso *mêchanízein* e alle loro macchinazioni stiamo affidando la nostra organizzazione sociale, ma anche la nostra sensibilità culturale[2].

In media su internet si registrano ogni minuto 575mila tweet, 5,7 milioni di ricerche su Google, 65.000 foto su Instagram. Solo una parte minima di questi dati viene conservata allo stato grezzo, la maggior parte è rielaborata per gli scopi più vari, in genere commerciali, e comunque conservata per essere sovrascritta con dati più recenti. Abbiamo, quindi, un effetto moltiplicatore di metadati sui dati e di metadati sui metadati. Pare siano otto miliardi e seicento milioni gli oggetti connessi al *web* che ci assistono, senza che ce ne rendiamo conto, nelle nostre più svariate attività. Treni e rotte ferroviarie, aerei e rotte aeree, denaro e transazioni finanziarie, messaggi e sistemi di comunicazione, sono regolati da processi di intelligenza artificiale: oggetto delle nostre richieste e intanto artefici di un mondo a sé che si accresce continuamente a un ritmo vertiginoso.

Tra non molto si dovrebbero produrre intorno ai mille triliardi di byte. Siamo avviluppati dal *cloud*, ma anche aggrappati al *cloud*… L'*Internet of Things* (IoT) ci segue passo per passo nel nostro rapporto con il mondo esterno. L'*Internet of Bodies* (IoB) ci monitora sistematicamente nel rapporto con i nostri equilibri biologici. Fuori di noi e dentro di noi aleggia un custode digitale che capta anche quello che ci sfugge e registra anche quello che non ci interessa.

Non sapremmo cosa farcene di questo continuo flusso di dati, se non avessimo supercomputer la cui capacità di elaborazione non è neppure immaginabile da una mente umana: 200 milioni di miliardi di calcoli al secondo; quanto potrebbero fare tutte le persone sulla terra se eseguissero un calcolo ogni istante di ogni giorno per 305 giorni. È una capacità che diverrà irrisoria se

[2] S. Cave, K. Dihal, *Imagining AI. How the World Sees Intelligent Machines*, Oxford University Press, Oxford, 2025.

riusciremo a realizzare efficienti computer quantistici che, basandosi sui quantum bit, o "qubit", operano nel cosiddetto stato di "sovrapposizione". Una coppia di bit normali può dar vita solamente a una delle quattro tradizionali combinazioni, o stati, possibili (00, 01, 10 e 11), ma una coppia di qubit contiene tutte e quattro le combinazioni insieme. Frontier, il supercomputer dell'Oak Ridge National Laboratory, può seguire 10^{18} operazioni al secondo, 10.000 volte superiore alla probabile velocità massima di computazione del cervello. Sycamore, il computer quantistico annunciato da Google, impiega 53 qubit per un totale di 253 valori pari a più di 10 biliardi di combinazioni. Dovrebbe in 200 secondi effettuare quelle operazioni che gli attuali supercomputer all'avanguardia possono compiere in 10.000 anni[3].

Se mettiamo assieme la capacità di assimilazione dei dati con la capacità di rielaborazione, la complessità si complessifica nel senso che non esistono solo tanti specifici dispositivi, ma ciascun dispositivo ne sviluppa altri e poi ancora altri finché non si determina una concatenazione strutturale che tende a divenire un mondo a sé. Ogni macchina è "imperialistica" o "espansionistica"[4] perché si procura un proprio regno coloniale di servizi, sosteneva, già nel 1964, Gunther Anders, riflettendo sulla tendenza alla "macchinizzazione". A sua volta Jacques Ellul, nel 2004 aveva invitato a riflettere sul fatto che il sistema tecnico non solo cresce su se stesso, ma questo continuo auto-accrescimento sfugge a qualsiasi possibilità di controllo: "la direzione tecnica si decide da sé"[5].

Colonizzazione e autoaccrescimento sono ormai gli sviluppi strutturali dell'*expandiverse technology* da cui deriva l'infosfera. Applicazioni e dispositivi, per quanto diversi, convergono (paradigma della confluenza) verso una funzionalità unitaria che spesso non è progettata né voluta, ma è effetto dell'alleanza

[3] È difficile ipotizzare le implicazioni della fine dell'era digitale per l'avvento dell'era quantistica. Ci prova M. Kaku, *Quantum supremacy: how the quantum computer revolution change everything*, Penguin, New York, 2023.

[4] G. Anders, *Noi figli di Eichmann*, trad. it., Giuntina, Firenze, 1995, pp. 53-55.

[5] J. Ellul, *Il Sistema tecnico. La gabbia delle società contemporanee*, trad. it., Jaca Book, Milano, 2009, p. 279.

commerciale tra *Big Tech* e *startup* attraverso la quale le imponenti infrastrutture *hardware* (tra *cloud* e potenza di calcolo) hanno accesso a enormi quantità di dati. Non esiste una netta linea di demarcazione tra il *Machine Learning*, l'Apprendimento Profondo, la *Computer Vision*, il *Natural Language Processing* (NLP), la Robotica, la Simulazione del Comportamento Umano, le neuroscienze cognitive, le neuroscienze sociali e così via. Ciascuna di queste attività ha le sue caratteristiche, ma gli sviluppi dell'una incidono sugli sviluppi delle altre in un calderone di funzioni in cui tutto si accumula senza limiti in un disordine che è solo apparente perché per la prima volta nella storia dell'umanità:

1) disponiamo di tecnologie che possono agire regolarmente e normalmente come utenti autonomi di altre tecnologie;

2) abbiamo macchine che continuano a perfezionarsi indipendentemente dal compito per cui sono programmate;

3) il prodotto di questi sviluppi e interazioni è spesso assolutamente imprevedibile.

Come osserva Floridi, "la modernità è diventata ben presto l'universo di complesse dipendenze reticolari, di reazioni a catena meccanica così come di connessioni obbligate"[6]. I dati si sommano gli uni con gli altri perché le tecnologie si sviluppano le une con le altre: c'è sempre l'algoritmo di un algoritmo che rinvia ad altri algoritmi. L'effetto "*loop*" dei meccanismi ricorsivi del nostro cervello, per cui i pensieri derivano dai pensieri nei pensieri dei pensieri, è la tecnica propria anche dei linguaggi dei sistemi di calcolo. Cambia la capacità di *looping*, ma non l'imprevedibilità. Attraverso l'esperienza che ha maturato come *chief business officer* di Google, dopo aver lavorato in IBM e Microsoft, Mo Gawdat ci fa notare che

> le AI non vengono propriamente programmate. Benché sulle prime siano alimentate da algoritmi, analoghi ai semi dai quali germoglia l'intelligenza, le loro vere capacità scaturiscono dall'osservazione. Una volta corredate del codice iniziale, analizzano enormi quantità di dati per rilevare pattern e modelli, seguendo un percorso simile a quello della selezione naturale, per sviluppare il proprio ingegno in erba. Infine di-

[6] L. Floridi, *La quarta rivoluzione industriale. Come l'infosfera sta trasformando il mondo*, trad. it., Raffaello Cortina, Milano, 2017, p. 31.

> ventano pensatrici originali e indipendenti, influenzate meno dagli input dei loro creatori che dai dati ricevuti da noi[7].

Gli algoritmi dei sistemi di apprendimento automatico, strutturati sul modello di reti di neuroni artificiali, "capiscono" dagli stessi dati ciò che devono fare attraverso l'architettura di svariati "trasformatori generativi pre-addestrati", i *learners* (apprenditori) e i *classifiers* (classificatori), che sono in grado, con un programma di apprendimento lungo poche centinaia di righe, di generare automaticamente milioni di stringhe di codice. Il rapporto tra dati e metadati diventa, così, sempre più labile perché la loro fusione determina sviluppi ed esiti che non sono prevedibili neppure dagli stessi programmatori (*black box effect*).

L'idea che le macchine siano a nostra disposizione si sta, quindi, lentamente capovolgendo nella situazione opposta: siamo noi a dipendere sempre più dalle macchine. "Dai motori di ricerca di Google ADS ai sistemi di personalizzazione e raccomandazione di Instagram e YouTube, Spotify Apple Music o Amazon, dalle chat BOT agli algoritmi di filtraggio delle applicazioni di incontri, tu, io e gli altri siamo tutti cavie che corrono alla cieca nel labirinto"[8]. Quante cose abbiamo rivelato ad *Alexa* di Amazon, a *Siri* di Google, a *Cortana* di Windows, oppure più semplicemente alla tastiera del nostro computer o ai messaggi del nostro cellulare? Non ci interessa. Pensiamo che tutto passi come la varietà dei pensieri che attraversano la nostra mente. Invece avviene proprio il contrario. Parafrasando un noto detto di Trockij, "anche se la guerra non ti interessa, la guerra si interessa di te", potremmo dire che, anche se quello che memorizzano Google o Facebook o Instagram non ci interessa, loro si interessano di noi, perché non dimenticano nulla: ogni dato potrebbe essere conservato, catalogato, collegato, impiegato.

Si parla, infatti, di una sorta di "inconscio digitale" per indicare tutta le informazioni su ciascuno di noi che sono raccolte dai sistemi di rilevazione dell'economia digitale che sfruttano il nostro "surplus comportamentale" per elaborare le proprie pre-

[7] M. Gawdat, *Superintelligenti. Come salvare il mondo dall'intelligenza artificiale*, trad. it., Rizzoli, Milano, 2022, p. 205 (ed. digitale).

[8] *Ivi*, p. 215.

visioni di mercato[9]. Il controllo dei corpi, delle propensioni e in parte anche dei sentimenti è sempre più in mano ad algoritmi[10]. E gli algoritmi sono al servizio del mercato, che a sua volta è condizionato dagli algoritmi che producono quell'entità impalpabile e onnipresente che ci avvolge nel *cloud* e ci segue con l'intelligenza artificiale. Un nuovo *apparatčik* sotto la forma di dispositivi *wireless* onnipresenti[11]?

2. La scatola nera dell'apprendimento profondo

Infosfera, cloud, intelligenza artificiale sono le metafore con cui cerchiamo di identificare la complessa aggregazione di materiali e attività che costituisce l'attuale ecosistema produttivo. In un versante abbiamo fibre ottiche, chip e microchip, cavi, sensori, satelliti, computer, magazzini, megastrutture informatiche, collegate in rete a particolari catene di approvvigionamento, vincolate al consumo di ingenti quantità di acqua e di energia e, infine, intente a soddisfare miliardi di utenti. In un altro versante troviamo gruppi di ricerca, sviluppo e produzione sorretti da una miriade di brevetti e da un affastellarsi di norme nazionali e internazionali. Possiamo elencare materiali e strutture, isolare legami commerciali e rapporti contrattuali, ma continua a sfuggirci la compiuta percezione degli effetti sistemici dell'entità che si sta progressivamente delineando. L'accelerazione tecnologica ha profondamente trasformato noi stessi e la nostra società, ma non ha cambiato il nostro rapporto con la tecnologia digitale, perché continua a prevalere l'idea che ci troviamo di fronte a tutta una serie di oggetti di consumo di cui dobbiamo soltanto regolamentare le modalità di impiego.

Infosfera, *cloud*, intelligenza artificiale, *loop* nella loro vaghezza semantica, sono la misura di questo limite concettuale.

[9] S. Zuboff, *Il capitalismo della sorveglianza. Il futuro dell'umanità nell'era dei nuovi poteri*, trad. it., Luiss University Press, Roma, 2019, Part III, cap. 12.

[10] D. de Kerckhove, *Il futuro della memoria*, trad. it., Castelvecchi, Roma, 2018.

[11] A. Keen, *Digital Vertigo: How Today's Online Social Revolution is Dividing, Diminishing and Disorienting Us*, St. Martin's Press, New York, 2012, cap. 7.

Sono il segno della difficoltà nell'individuare e nel delimitare quanto sta avvenendo. Le "tecnologie esponenziali del techno-capitalismo" ci pongono di fronte alla "elaborazione di sistemi destinati a rispondere col tempo a ogni circostanza esistenziale e a istituire una gestione automatizzata del mondo"[12]. Anche se rifuggiamo dall'ipotesi[13] estrema della *Singularity*, della *Superintelligence*, della *Scary Smart*, non sappiamo fino a che punto l'inevitabile sviluppo dell'automazione, effetto di questo enorme accumulo di informazioni e dispositivi, sarà in grado di ricondurre la connettività ad unitarietà, la coerenza a intelligenza generale, e l'intelligenza generale a una qualche forma di senzienza se non a un prototipo di coscienza. Non lo sappiamo perché non conosciamo quali siano le leggi fondamentali del pensiero e come queste si colleghino al funzionamento del cervello. Sappiamo che il cervello ha una composizione biochimica e quindi dovrebbe sottostare alle leggi della fisica. Sappiamo che tutti i cervelli sono modulari e quindi si articolano attraverso una rete di connessioni come qualsiasi sistema informatico. Tuttavia né la meccanica newtoniana, né la teoria elettromagnetica di Maxwell, né la teoria quantistica sono ancora in grado di spiegarci come la chimica diventi coscienza negli scambi tra i nostri neuroni. E non conosciamo neppure il ruolo della genetica in tutto questo.

Allo stato attuale dobbiamo limitarci a prendere atto di come sia sempre maggiore il numero di sistemi in grado di svilupparsi autonomamente e con crescenti margini di imprevedibilità (l'effetto scatola nera). E più questi sistemi operano in connessione, più il limite dell'uno viene colmato dalle capacità dell'altro.

La robotica, ad esempio, non è solo un campo limitato a singole macchine, che eseguono ripetitivamente sempre gli stessi compiti prefissati. Per effetto del *cloud computing*, si sta trasfor-

[12] E. Sadin, *La siliconizzazione del mondo. L'irresistibile espansione del liberalismo digitale*, trad. it., Einaudi, Torino, 2019, p. 40.

[13] Ipotesi tutt'altro che remota se, nel mese di ottobre del 2025, alcuni dei maggiori esperti di IA come Geoffrey Hinton, Yoshua Bengio, Stuart Russell, Andrea Miotti, diversi scienziati, tra cui 5 premi Nobel, insieme a personalità politiche e religiose hanno emanato un documento (https://superintelligence-statement.org) in cui sottolineano la necessità di un divieto globale di sviluppo della superintelligenza artificiale.

mando in un'entità collettiva, una rete diffusa di macchine che condividono conoscenze ed esperienze, moltiplicando in maniera esponenziale, la capacità di adattarsi ai cambiamenti, risolvere problemi e trovare nuove vie di sviluppo. Brett Adcock, il Ceo della società produttrice dei robot *Figure*, ha annunciato l'ultimo modello in programmazione, *Figure03*, come un *general robot* che sarà in grado non solo di svolgere tutte le attività domestiche, ma anche di apprendere sempre nuove attività attraverso la connessione *wireless* con la rete degli altri robot.

Se non sarà, quindi, mai intelligente il singolo computer o robot, non sappiamo fino a che punto potremo dire lo stesso del *cloud* a cui è connesso e che, in qualche modo, lo regola e assimila. Il dubbio è giustificato dall'analogia con il funzionamento del cervello dove il singolo neurone, anche se all'interno di una rete nervosa, non è "consapevole" e non produce "consapevolezza", mentre l'esperienza cosciente che caratterizza la nostra mente deriva dalla complessità del sistema di interazioni dei neuroni tra loro e con l'ambiente esterno, l'ancora in gran parte ignoto rapporto tra *pattern* di *firing* neuronali e *binding* di questi *pattern*. Risale già alla metà del secolo scorso la convinzione che, per effetto della ridondanza dei sistemi di calcolo, l'auto-organizzazione potesse essere la base della progettazione di una "*infallible network of fallible neurons*"[14].

Anche senza affrontare il problema dell'esplosione computazionale di una superintelligenza, dobbiamo renderci conto che l'incremento delle possibilità determinato dalle tecnologie trasformative rende sempre più evidente la divisione teorica tra le due anime dell'intelligenza artificiale: quella ingegneristica e riproduttiva, che tende a imitare e perfezionare, attraverso gli impulsi che riceve dai programmatori, le condotte umane, assistendole o sostituendole in un gran numero di contesti; quella cognitivista e produttiva che aspira ad ottenere l'equivalente del nostro cervello con sistemi di apprendimento automatico (*machine learning*) e/o con sofisticati artefatti biologici[15]. Non si

[14] M. Pasquinelli, *The Eye of the Master. A Social History of Artificial Intelligence*, Verso, London, 2023, p. 148.

[15] L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, trad. it., Raffaello Cortina, Milano, 2022, pp. 48 ss.

tratta soltanto di una scelta ideologica tra realismo e futurismo, tra techno-ottimismo e techno-pessimismo, tra *Effective Accelerationism* ed *Effective Altruism*, ma delle due vie che abbiamo di fronte: entrambe necessarie nella definizione dell'orizzonte etico e nell'elaborazione delle linee regolamentari. Dobbiamo certamente muovere dall'una, l'anima ingegneristica e riproduttiva, senza tuttavia rinunciare a tenere sullo sfondo l'altra, l'anima cognitivista e produttiva.

La prima prospettiva non nega i cambiamenti negli attuali equilibri sociali e politici, dal mondo del lavoro all'esercizio della democrazia, dall'economia alla finanza, dalla tutela dei diritti fondamentali alla giustizia digitale, ma ritiene che sia sufficiente un progressivo adeguamento degli schemi giuridici e dei modelli di *governance*, restando tuttavia nel solco di consolidate tradizioni culturali dove il libero arbitrio è ancora il fondamento della responsabilità individuale, dove la tutela della *privacy* è ancora il presidio della dignità individuale, dove le consultazioni elettorali sono ancora la radice della democrazia, e così via. I sostenitori di questa prospettiva affermano che non dobbiamo lasciarci ingannare dalla contiguità degli orizzonti di ricerca. Per quanto le neuroscienze suggeriscano modelli di sviluppo all'informatica e l'informatica aiuti a verificare i progressi nella conoscenza della codifica dei segnali nervosi e del modo in cui le reti neurali elaborano l'informazione, le macchine resteranno macchine e gli algoritmi algoritmi. Dietro le une e gli altri sarà sempre riconoscibile il primato dell'impronta umana. Malgrado il perfezionarsi della neurorobotica e il sistematico impiego di interfacce computer-cervello sarà sempre evidente la linea di demarcazione tra soggetto e oggetto, organico e inorganico, naturale e artificiale.

La seconda prospettiva pensa che le implicazioni di una tecnologia *expandiverse* non consentano, sin da ora e ancora meno in futuro, di capire dove finisce l'intelligenza ristretta e specializzata della singola macchina e dove cominci lo "apprendimento associato illimitato" che costituisce il presupposto di una futura intelligenza generale e indeterminata. È innegabile che la singola macchina sia "stupida", ma non sappiamo se potremo affermare lo stesso del "sistema" di macchine che si sta profilan-

do all'orizzonte. Sono "stupidamente intelligenti" perché fanno tantissime cose, ma non sanno cosa fanno? Il problema è che incominciamo a non sapere neppure noi cosa faranno, perché si continuano a sviluppare sistemi "*human out of the loop*" per raggiungere risultati imprevisti o addirittura imprevedibili, perché i sistemi di *Machine Learning* sono progettati in modo da assimilare dati, prendere decisioni o fare previsioni senza che siano esplicitamente istruiti per farlo.

Del resto i programmi di ricerca avanzata si basano sulla centralità di *Answer box* che uniscono sofisticati modelli di archiviazione a nuove tecniche di indicizzazione e di elaborazione del linguaggio naturale non solo per fornire la risposta alla domanda dell'utente, ma per capirne le intenzioni. Tra *Knowledge Graph* e *Passage Ranking* lo scopo non è sempre quello di fornire una risposta corretta, ma quella risposta che l'utente desidera. I sistemi di intelligenza artificiale non possiedono quel "sapere di sfondo" sulla società in cui operano e sui pregiudizi da cui è animata, che è proprio degli esseri umani, ma si stanno affinando le tecniche di apprendimento in contesto (*Chain-of-Thought Prompting*) e di *Representational similarity analysis* per cogliere le sfumature implicite tra le pieghe delle analogie all'interno dei dati raccolti. Pressate tra *rich answers*, *direct answers*, *instant answers*, *quick answers*, *featured snippets*, *knowledge panel*, *chain-of-thought prompting* le applicazioni a cui ci rivolgiamo stanno scardinando nel nostro rapporto con i sistemi informatici le barriere tra sintassi e semantica, tra reazione e cooperazione, tra programmazione e coevoluzione fino ad avvicinare l'interazione alla "sensibilità". È l'ideale a cui aspirano gli algoritmi affettivi o i *companion* robot e verso cui ci sospinge l'apprendimento profondo (*Deep Learning*), che utilizza reti neurali artificiali, la cui struttura tenta di avvicinarsi alla complessità dell'organizzazione del cervello.

> I computer addestrati con l'apprendimento profondo possiedono delle reti neurali enormi, in cui la conoscenza viene distribuita tra migliaia di pesi di connessione. Inoltre, alcune delle caratteristiche mentali sono difficili da identificare persino nei codici espliciti, perché richiedono interazione tra

> moduli diversi all'interno del programma complessivo del computer[16].

Assieme alla complessità cresce l'opacità[17]. Il modello ingegneristico e riproduttivo presuppone la prevalenza degli algoritmi deterministici in cui l'intelligenza artificiale è vincolata a specifici parametri di azione, a definite funzioni e a chiari obiettivi, attraverso un rapporto tra *input* e *output* regolato da "schemi esecutivi", accessibili e trasparenti, che indichino i singoli passaggi, le loro connessioni e gli esiti (modello *white box*). Già nel 1978 Jeremy Bernstein osservava che "in linea di principio è possibile fare scrivere alla macchina i passi che ha seguito nell'esecuzione di un certo algoritmo; ma si sta raggiungendo il punto in cui, in pratica, ciò risulterà impossibile"[18]. Dopo tanti anni sappiamo che gli algoritmi deterministici coprono solo una parte, per quanto rilevante, dei processi digitali. Nella maggior parte dei casi le intelligenze artificiali, una volta corredate di un codice iniziale, sviluppano più modelli in concorrenza tra loro, seguendo un percorso simile a quello della selezione naturale nell'incrementare le proprie potenzialità. Questa tecnologia di apprendimento automatico genera quell'effetto *black box* che non consente di spiegare come e perché un dato diviene metadato, come e perché i metadati filtrano i dati, segnando il passaggio dalla "realtà aumentata" alla "realtà virtuale". Non siamo neppure in grado di capire quando gli algoritmi deterministici di un *white box model* mutino negli algoritmi stocastici di un *black box model*.

> The specific characteristics of many AI technologies, including opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour, may make it hard to verify compliance with, and may hamper the effective

[16] P. Thagard, *Cervelli a confronto. Perché l'intelligenza umana è diversa da quella degli animali e dei robot*, trad. it., FrancoAngeli, Milano, 2021, p. 34.

[17] Sono gli stessi sviluppatori di IA ad ammettere che non sono in grado di interpretare i processi che hanno creato. Ad esempio il CEO di Anthropic, Dario Amodei, ha dichiarato che "comprendiamo solo il 3% del loro funzionamento".

[18] J. Bernstein, *Uomini e macchine intelligenti*, trad. it., Milano, Adelphi, 1990, p. 74.

enforcement of, rules of existing EU law meant to protect fundamental rights[19].

3. *Exotic mind entities*

Il "*veil of technical inscutability*"[20], il problema dell'opacità, sta diventando centrale nella riflessione sull'intelligenza artificiale non solo perché è tecnicamente difficile (e in alcuni casi impossibile) aprire la scatola nera dei codici di funzionamento, ma perché costituisce la zona di confine nel rapporto tra meccanizzazione della mente e umanizzazione della macchina. Questa zona grigia era già il dato inquietante, inquietante per il dualismo tra soggetto e oggetto, del tentativo delle neuroscienze cognitive di spingere fino alle estreme conseguenze il "sottofondo biologico" dell'attività cerebrale riconducendola integralmente alla biochimica dei neuroni o alla biofisica delle sinapsi. Se le nostre sensazioni non sono altro che particolari algoritmi spazio-temporali che si armonizzano con determinate frequenze elettromagnetiche, dovremmo cominciare a prendere sul serio l'equiparazione del cervello a un computer. Il riduzionismo fisicalista ritiene che, "se il senso e la mente vengono associati con la materia, è perché *provengono da* essa"[21].

La mente sarebbe il *software* che elabora gli algoritmi implementati nel cervello (l'*hardware*) e sono tali algoritmi che le diverse branche delle neuroscienze cognitive cercano di riprodurre per poi trasferirli nei supporti fisici più disparati. Il problema è trovare gli algoritmi giusti e, forse, l'algoritmo "definitivo" con cui è sintetizzabile la teoria del tutto, una stringa che contiene l'essenza dell'universo[22].

[19] *White Paper On Artificial Intelligence – A European approach to excellence and trust*, 19/02/2020, p. 12.

[20] F. Pasquale, *The Black Box Society. The Secret Algorithms That Control Money and Information*, Harvard University Press, Cambridge (Mass.)/ London, 2015, p. 163.

[21] J.P. Dupuy, *Alle origini delle scienze cognitive. La meccanizzazione della mente*, trad. it., Mimesis, Milano, 2009, p. 37

[22] P. Domingos, *L'algoritmo definitivo. La macchina che impara da sola e il futuro del nostro mondo*, trad. it., Bollati Boringhieri, Torino, 2016, p. 332 (ed. digitale).

Il *black box effect* ci presenta l'altra faccia della stessa medaglia: una sorta di possibile umanizzazione delle macchine che, rendendo sempre più complessi e imprevedibili i "propri" percorsi, avvicinano il digitale al celebrale. "Ci vogliono più che nervi d'acciaio per fare una cosa del genere. Ci vuole un cervello di silicio"[23]. È il commento di Kasparov, il campione di scacchi, dinanzi alla mossa, con cui il computer *Deep Blue* lo ha sconfitto. "Era diversa da qualunque altra cosa un computer avesse mai fatto prima. Ed era anche diversa da qualunque cosa un essere umano avesse mai preso in considerazione. Era qualcosa di nuovo, una totale rottura con la tradizione, una deviazione radicale da migliaia di anni di esperienza accumulata"[24]. È il commento al "colpo 37" con cui il computer *AlphaGo* ha sconfitto il campione del mondo, Lee Sedol, in una partita di *Go*, il gioco più difficile per la sua complessità da inserire in un sistema informatico.

Ha avuto la stessa sensazione di entrare a contatto con una singolare forma di pensiero l'ingegnere di Google, Blake Lemoine, dialogando con LaMDA il sistema elaborato dalla sua azienda attraverso l'impiego di reti neurali artificiali. Anche gli *affective computing* sembrano, per la loro capacità di leggere le nostre emozioni, manifestare una sorta di emotività. Del resto "Without emotion, computers are not likely to attain creative and intelligent behavior, but with too much emotion, we, the maker, may be eliminated by our creation"[25].

Non possiamo, quindi, escludere che le nuove tecnologie si stiano consolidando attorno a una particolare forma di intelligenza che nasce dalle macchine e si sviluppa con le macchine, andando oltre le macchine. Macchine di Dio[26]? Macchine

[23] G. Kasparov, M. Greengard, *Deep Thinking. Dove finisce l'intelligenza artificiale e comincia la creatività umana*, Fandango, Roma, 2019, p. 147 (ed. digitale).

[24] B. Labatut, *Maniac*, trad. it., Adelphi, Milano, 2023, p. 314.

[25] R.W. Picard, "Affective Computing", in *M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321*, p. 15.

[26] H. Novotny, *Le macchine di Dio. Gli algoritmi predittivi e l'illusione del controllo*, trad. it., Luiss University Press, Roma, 2022.

sapienti[27]? Oppure semplicemente "macchine ingannevoli"[28]? LamDa, Gemini, Watson, AlpaZero, Chat-GPT, DeepSeeK, OpenAi, Grok3, Claude e Bard sono estremamente diverse tra loro, ma sono indifferenziatamente molto più vicine a noi di quanto potrebbe esserlo una lavatrice. Ce ne rendiamo conto appena cerchiamo di individuare le caratteristiche fondamentali dell'intelligenza umana. Ad esempio, la percezione, il *problem solving*, la pianificazione, la decisione, la comprensione, l'apprendimento, l'astrazione, la creazione, il ragionamento, i sentimenti, la comunicazione, l'azione[29].

Ogni forma di intelligenza artificiale implica o esclude qualcuna di queste caratteristiche. Ma, per quante dissomiglianze possiamo individuare, restiamo estremamente lontani dal mondo degli oggetti. Dobbiamo fare ricorso all'*agency* della terminologia anglosassone, meno compromettente del nostro "soggettività", per sottolineare il passaggio da un'intelligenza artificiale puro strumento a nostra disposizione a un'intelligenza artificiale che ci supporta aggiungendo il "suo" alle nostre volontà[30]. Non sappiamo dove ci condurranno gli algoritmi indeterministici, con la loro opacità, o i processori quantistici, con la loro complessità, e non siamo neppure in grado di immaginare cosa potrà rappresentare per noi quella superintelligenza che la teoria della singolarità prospetta come la sintesi unitaria di tutte le tecnologie digitali. Il ricorso alla nozione di agentività dovrebbe consentire di graduare i livelli regolamentari in base ai diversi tipi di autonomia, di complessità e di opacità, imponendo tutta una serie di specifici metodi di controllo e di blocco in relazione ai livelli di rischio di questi singolari ed eterogenei agenti artificiali.

[27] P. Benanti, *Le macchine sapienti. Intelligenze artificiali e decisioni umane*, Marietti, Bologna, 2018.

[28] S. Natale, *Macchine ingannevoli. Comunicazione, tecnologia, intelligenza artificiale*, trad. it., Einaudi, Torino, 2022, p. 9.

[29] Seguo l'analisi di P. Thagard, *op. cit.*, pp. 30 ss.

[30] "They have the potential to rearrange human knowledge itself into new and wonderful forms…" C. Summerfield, *These Strange New Minds. How Ai Learned to Talk and What it Means*, Penguin Random House UK, London, 2025, p. 4.

Resta aperto il problema più ampio del rapporto tra organico e inorganico, tra la tecnologia del carbonio che sta alla base dell'uno e la tecnologia del silicio che sta alla base dell'altro.

> Se delle molecole chimiche molto complicate possono operare negli esseri umani in modo tale da renderli intelligenti, dei circuiti elettronici altrettanto elaborati potrebbero a loro volta far agire i computer in modo ingegnoso. E una volta raggiunta l'intelligenza, è presumibile che loro stessi saranno in grado di progettarne altri dotati di una complessità e di un acume ancora più sviluppati[31].

Seguendo la teoria dell'evoluzione, il pensiero si è prodotto dalla materia[32]. Perché non potremmo riprodurre, a nostra volta, il pensiero attraverso la materia? L'opacità, che non riusciamo né a spiegare né a controllare, è solo la prima avvisaglia di una possibile transizione di fase in cui l'approfondirsi dei processi di *machine learning* finirà per determinare quello stesso scarto che, nella prospettiva fisicalista, avrebbe dato origine prima alla vita e poi all'attività cerebrale?

L'intelligenza artificiale ci costringe a ripercorrere quel processo per cui siamo sospesi, come insegnava Monod, tra il caso e la necessità. Il caso delle transizioni di fase e la necessità dei nuovi e inattesi equilibri che si sono venuti a determinare. Che siano i 13,8 miliardi di anni, nei primi brevissimi sviluppi dell'evoluzione dell'universo, in cui si è determinato uno squilibrio tra materia e antimateria, oppure i 3,5 miliardi di anni in cui sono comparsi i primi batteri o i 430 milioni di anni dell'esplosione biologica del cambriano o… delle prime sinapsi, a un certo punto, avviene uno scarto. Il metabolismo di enzimi e proteine ha determinato la capacità di duplicazione degli acidi nucleidi o la capacità di duplicazione degli acidi nucleidi ha determinato il metabolismo? In un caso e nell'altro, all'improvviso, la chimica e la fisica hanno prodotto "informazione". L'informazione che consente l'acquisizione e la distribuzione di energia; l'informazione che consente l'autoreplicazione e quindi produce quelle macchine per la sopravvivenza

[31] S. Hawking, *Le mie risposte alle grandi domande*, trad. it., Rizzoli, Milano, 2018, p. 136 (ed. digitale).

[32] T.W. Deacon, *Natura incompleta. Come la mente è emersa dalla materia*, trad. it., Le scienze, Milano, 2012.

che chiamiamo vita. Dalla vita al pensiero è un altro scarto: "com'è accaduto che un protozoo simile a un coanoflagellato diventasse una spugna, una spugna diventasse uno cnidario, uno cnidario un invertebrato bilaterale, e un invertebrato bilaterale un vertebrato e un vertebrato un essere pensante?"[33].

Mi pare estremamente significativo che, negli ultimi anni, la maggior parte dei libri sull'intelligenza artificiale tocchi il tema dell'origine della vita e del pensiero, mentre i libri sull'origine della vita e del pensiero si concludono, a loro volta, con una serie di interrogativi sull'intelligenza artificiale. Sullo sfondo si staglia il problema del *black box effect*, dell'imprevedibilità. La teoria dell'evoluzione ci pone fronte al fascino della complessificazione. Ci spiega che avremmo avuto, sin dai primi secondi del *big bang*, un continuo aumento della complessità e degli adattamenti sistemici alla complessità, ma non ci fornisce una ragione della complessificazione.

Possiamo pensare, con Gould, che siamo un evento improbabile nel film della vita[34] o, convenire con Al-Khalili, che "la probabilità che processi casuali determinino l'origine della vita è la stessa di una tromba d'aria che colpisca una discarica e produca un aereo"[35]. Resta il fatto che "[…] noi esseri viventi siamo i vortici generati dalle onde d'acqua liberate dall'apertura della paratia. Siamo lo spumeggiare irreversibile dell'energia libera che era intrappolata nel disequilibrio fa idrogeno ed elio ed è stata liberata dal sole"[36]. Proprio l'inesplicabile legame tra organico e inorganico, evidenziato dalla fisica quantistica, impone l'analisi dei modelli di senzienza o forse addirittura di seità presenti in tutte le specie viventi, mentre si prospetta l'eventualità che "l'intelligenza biologica sia soltanto un fenomeno transito-

[33] J. Le Doux, *Lunga storia di noi stessi, Come il cervello è diventato cosciente*, trad. it., Raffaello Cortina, Milano, 2020, p. 161.

[34] Riprendo le osservazioni di D. Bickerton, *Quello di cui la natura non ha bisogno. Linguaggio, mente ed evoluzione*, trad. it., Adelphi, Milano, 2022, p. 285 (edizione digitale).

[35] J. Al-Khalili, J. McFadden, *La fisica della vita. La nuova scienza della biologia quantistica*, trad. it., Bollati Boringhieri, Torino, 2015, p. 229 (ed. digitale).

[36] C. Rovelli, *Buchi Bianchi*, Adelphi, Milano, 2023, p. 112.

rio, una fase passeggera nell'evoluzione dell'intelligenza nell'universo"[37].

Una volta messo da parte il pregiudizio cartesiano degli animali come macchine senz'anima, ci troviamo di fronte a svariate "figure della mente"[38] che vanno dai batteri alle piante. Non solo, ma emerge l'ipotesi che "l'esperienza sentita comparve diverse volte in diverse linee evolutive"[39]. Noi saremmo soltanto una di queste possibili varianti. Dovremmo, quindi, abbandonare ogni residuo antropomorfico e incominciare a immaginare una pluralità di dimensioni in cui azione, sensibilità e coscienza si esprimono non solo variamente, ma anche diversamente. Le riflessioni su una coscienza diffusa in tutte le specie viventi muovono dalla constatazione che ogni cervello presenta un'immagine virtuale del mondo esterno, controllata dai sensi e regolata dal principio di autoconservazione. Ogni specie vivente opera attraverso una sorta di "allucinazione" determinata dalle proprie caratteristiche biologiche[40]. Non entriamo a contatto con il mondo così come è, secondo le leggi della fisica, ma come ce lo presenta, e forse meglio racconta, la nostra percezione sensoriale. Diverse percezioni, diversi mondi, diverse agentività, diverse senzienze, diverse seità…

Mi torna in mente la meraviglia destata dall'evoluzione dei *Large Language Models*. "A threshold was reached, as if a space alien suddenly appeared that could communicate with us in an eerily human way"[41]. Nel rapporto con questi "inquietanti spazi alieni" il concetto di senzienza è particolarmente utile perché ci impone di cambiare prospettiva, prendendo in esame forme "esotiche" di mente che, pur non rispecchiando i parametri uma-

[37] M. Kaku, *Il futuro dell'umanità. Dalla vita su Marte all'immortalità, così la scienza cambia il nostro destino*, trad. it., Rizzoli, Milano, 2018, p. 292.

[38] S. Ginsburg, E. Jablonka, *Figure della mente. La coscienza attraverso la lente dell'evoluzione*, trad.it., Raffaello Cortina, Milano, 2023. Attraverso il "filo di Arianna" della biologia evoluzionista il libro passa in rassegna le diverse entità che possono definirsi coscienti e in che termini.

[39] P. Godfrey-Smith, *Metazoa*, trad. it., Adelphi, Milano, 2021, p. 320.

[40] A. Seth, *Come il cervello crea la coscienza*, trad. it., Raffaello Cortina, Milano, 2023, pp. 207 ss.

[41] T.J. Sejnowski, "Large Language models and the reverse Turing test", in *Neural Computation*, 35 (2023), n. 3, p. 303.

ni, elaborano costrutti diversamente intelligenti. "[T]hey are minds, of sorts, nonetheless – strange new minds, quite unlike anything we have encountered before"[42].

Se si parla di embrioni di coscienza nei batteri e nelle piante o di sofferenza nei pesci non si intende la pedissequa riproduzione dei nostri sentimenti. Come diamo per scontato che ogni forma vivente abbia elaborato i propri percorsi di interazione con il mondo in relazione alla propria nicchia biologica ed ecologica, così i percorsi imprevedibili degli algoritmi stocastici potrebbero essere il segno di qualcosa di analogo: l'inattesa reazione di questi algoritmi alla diversità dei dati in rapporto all'ulteriore diversità dei sistemi di addestramento. Se si parla di intelligenze artificiali al plurale è proprio per sottolineare la varietà dei modelli di sviluppo e di implementazione, pur sulla base delle stesse premesse scientifiche.

Possiamo parlare di *exotic mind entities* oppure di *alien reasoning* o di assemblaggi cognitivi, ci rendiamo in ogni caso conto che non è rilevante domandare quanta intelligenza ha una macchina o un topo, dovremmo piuttosto interrogarci su quanto siamo in grado di percepire e utilizzare il loro "esotismo" e fino a che punto sia necessario abituarci a convivere con un paradigma logico diverso dal nostro. Come è stato detto a proposito di Alpha Zero: si sta facendo strada l'idea che ci possano essere modi del conoscere che, pur non essendo assimilabili alla coscienza umana, siano ugualmente forme di conoscenza[43]. Da una parte, come nota Floridi, l'intelligenza artificiale ci apre

> lo spazio sterminato di problemi e compiti, ogni volta che questi possono essere conseguiti senza comprensione, consapevolezza, acume, sensibilità, preoccupazioni, sensazioni, intuizioni, semantica, esperienza, bio-incorporazione, significato, persino saggezza e ogni altro ingrediente che contribuisca a creare l'intelligenza umana[44].

Dall'altra l'effetto *black box* aggiunge a tutto questo, l'elaborazione di percorsi di apprendimento che sfuggono al condizionamento umano, sommando all'automatismo di compiti sempre

[42] C. Summerfield, *op. cit.*, p. 333.

[43] Fino a che punto possiamo aprirci all'impensato? Cfr. N. Katherine Hayles, *L'impensato. Teoria della cognizione naturale*, trad. it., Effequ, Firenze, 2021.

[44] L. Floridi, *Etica dell'intelligenza artificiale*, cit., p. 52.

più complessi l'eventualità di modelli di analisi e di comunicazione finora impensabili.

Ad esempio, l'utilizzazione di un programma di intelligenza artificiale simile a ChatGPT, addestrato su dati relativi alla composizione chimica e genetica di milioni di cellule, ha consentito di scoprire le "cellule Norn", un raro tipo di cellula renale che è responsabile della produzione dell'eritropoietina quando i livelli di ossigeno diventano troppo bassi. "A vital discovery about biology that otherwise would not have been made by the biologists" ha affermato Eric Topol, direttore dello Scripps Research Translational Institute[45].

Partendo da presupposti simili, si è pensato di impiegare i trasformatori generativi pre-addestrati, che stanno alla base dei sistemi digitali di traduzione, per intercettare le comunicazioni tra gli animali. Sappiamo che alcune specie (dai delfini ai topi, dai capodogli agli uccelli) hanno modelli sofisticati di trasmissione dei segnali che, legati in specifiche sequenze, assumono un particolare significato in base al modo in cui si formano le combinazioni finali[46]. Sono in corso diversi esperimenti non solo per riuscire ad operare una corretta registrazione e classificazione, ma anche per realizzare una plausibile decifrazione attraverso "animalgoritmi"[47]. L'autonomia della capacità computazionale potrebbe rompere le barriere naturali, consentendoci di penetrare le forme di comunicazione "altra"[48]? E con quali effetti?

4. Un diritto naturale al silicio?

Gli sviluppi dell'intelligenza artificiale sembrano, quindi, imporre una riflessione su ciò che è "naturale" o meglio su quanto

[45] C. Zimmer, "A.I. Is Learning What It Means to Be Alive", in *New York Times* (10/03/2024).

[46] E. Bucci, *Geni, memi e bit. Evoluzione biologica, termodinamica e teoria dell'informazione*, Mondadori, Milano, 2024, pp. 200 ss.

[47] T. Mustil, *Come parlare il balenese. Il futuro della comunicazione animale*, trad. it., Il Saggiatore, Milano, 2023, cap. 9.

[48] È più lontano nel tempo del libro di Mustil, ma altrettanto suggestivo S. Budiansky, *Se un leone potesse parlare. L'intelligenza animale e l'evoluzione della coscienza*, trad. it., Baldini Castoldi Dalai, Milano, 2007.

l'artificiale sia un aspetto del naturale[49]. Le reti neurali artificiali dei sistemi di *Deep learning* imitano la struttura e i modelli di apprendimento del cervello; la progettazione dei robot muove dalla biomimesi, dalla riproduzione delle forme e dei movimenti dei più svariati organismi biologici; gli animalgoritmi cercano di riprodurre i processi comunicativi presenti nelle varie forme di vita; i computer quantistici toccano il problema della struttura dell'universo; gli *affective computing* interpretano i modelli espressivi delle nostre condotte. In nessuna epoca della storia ci siamo interrogati, con dispositivi tanto raffinati, su quanto vi sia di meccanico in noi e su quanto vi sia di umano nel mondo che ci circonda. Più costruiamo sofisticati modelli di intelligenza artificiale e più scopriamo quanto poco sappiamo su noi stessi. Potremmo ripetere con Eliot, "Endless invention, endless experiment,… /All our knowledge brings us nearer to our ignorance… /Where is the wisdom we have lost in knowledge?"[50].

Quale "macchinazione" è più perfetta di quella che muove da un mondo meccanico dominato da algoritmi e robot per poi farci intuire le tante "figure della mente" di una natura che, tra l'agentività e la senzienza, presenta svariati modelli di percezione sensoriale e di trasmissione dell'informazione? Quale macchinazione è più perfida di quella che mostra quanto siamo impotenti dinanzi all'opacità degli algoritmici e, poi, suggerisce che proprio questa opacità ci potrebbe aiutare a penetrare alcuni degli aspetti più riposti del vivente?

Sembra delinearsi, pur con tutte le radicali ed evidenti differenze, un orizzonte culturale molto vicino a quello su cui si sono fondate, nel passato, le varie teorie del diritto naturale e, in particolare, quella declinazione di questa teoria che intendeva far propri gli insegnamenti riposti nelle costanti dei processi di sviluppo e nelle omogeneità latenti in natura: penso alla teoria dell'*oikeiosis*, alla *dynamei on* di Aristotele, alla *teleiosis* stoica, all'*ordo factivus* medievale. Il tentativo di elaborare animalgoritmi ricorda quel *proximun humanis sensibus*, quell'attenzione

[49] F. De Felice, R. Race, *Dialoghi su etica e intelligenza artificiale*, Luiss University Press, Roma, 2023, p. 69.

[50] T.S. Eliot, "Cori da 'La rocca'", in *Opere [1904-1939]*, Bompiani, Milano, 2005, p. 1230.

per la vicinanza tra tutte le specie viventi cara allo stoicismo, mentre le potenzialità degli algoritmi, che prendono *forme* impreviste a mano a mano che assimilano l'imponderabilità dei dati, richiamano la raffinata terminologia scolastica che distingueva l'*intellectus agens*, l'intelletto che è in azione, dall'*intellectus possibilis*, l'intelletto che ha in sé gli elementi dei futuri sviluppi, e l'*intellectus in actu*, l'intelletto che si manifesta caso per caso.

Se nel diritto naturale è fondamentale la domanda su quali siano i valori propri dell'identità umana, anche per l'intelligenza artificiale viene invocato un "*HCAI approach*", uno sviluppo umano centrico, e si parla di *ethics by design*, eualgoritmica, infoetica, algor-etica, *data ethics*, *data dignity*.

L'*Ethics by Design* viene richiamata più volte dal Regolamento 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati (GDPR). L'*Ethics by Design* è ribadita dalla Risoluzione del Parlamento europeo del 20 ottobre 2020 recante raccomandazioni alla Commissione concernenti il quadro relativo agli aspetti etici dell'intelligenza artificiale, della robotica e delle tecnologie correlate, ponendo le basi delle scelte compiute nel recente *AI Act*.

Una sorta di eualgoritmica[51] si profila nella Raccomandazione *The Ethics of Artificial Intelligence* dell'Unesco del 31/07/2022 che suggerisce agli Stati membri di sviluppare strategie che assicurino la continua valutazione della qualità della programmazione dei sistemi digitali, con particolare attenzione per l'adeguatezza dei processi di raccolta e selezione dei dati, promuovendo la sicurezza e l'affidabilità attraverso la previsione di specifici meccanismi di *feedback* per imparare dagli errori e condividere le migliori pratiche, eventualmente certificate con "*Gold standard datasets*".

Data ethics e *data dignity* dovrebbero essere l'esito della *DIKW pyramid* (*Data*, *Information*, *Knowledge*, *Wisdom*) che auspica un'ecologia della comunicazione in cui i singoli passaggi

[51] Usa questa espressione M. Chiriatti, *#Humanless. L'algoritmo egoista*, Ulrico Hoepli, Milano, 2024, pp. 79 ss.

(*purpose*, *insight*, *meaning*, *context*) dovrebbero, se programmati adeguatamente, determinare un'ideale concatenazione virtuosa.

L'info-etica ha per oggetto la deontologia dei programmatori, la serietà dei diffusori delle notizie, la disciplina delle piattaforme che acquisiscono e veicolano l'informazione, la correttezza dei fruitori pubblici e privati. In questo campo abbiamo, attualmente, un insieme eterogeneo di indicazioni normative. Dichiarazioni internazionali, linee guida, codici deontologici, politiche aziendali sottolineano continuamente, anche se con sfumature e accentuazioni diverse, la necessità di garantire: dignità umana, responsabilità, autonomia, giustizia, equità, solidarietà, trasparenza, affidabilità, sicurezza, salute, integrità fisica e morale, *privacy*, sostenibilità. Vengono progettate *start up* per contrastare le *fake news*, verificando le fonti e l'attendibilità delle notizie, e si propone l'istituzione di *board* per l'etica e la sicurezza. Molte di queste aspirazioni sono significativamente condivise da filosofi, scienziati e imprenditori come si evince, ad esempio, dalla Dichiarazioni di Asilomar del 2017 o dal documento *Rome Call for AI Ethics*, sottoscritto da Pontificia Accademia per la Vita, Microsoft, IBM e Fao, a conclusione dell'incontro svoltosi in Vaticano il 28 febbraio 2020 su *The 'Good' Algorithm? Artificial Intelligence: Ethics, Law, Health*.

L'ambizione estrema della programmazione algor-etica sarebbe quella di inserire un codice etico in un codice macchina, dotando gli stessi algoritmi di regole che garantiscano il rispetto dei valori fondamentali. Valori condivisi e quindi universali, valori programmabili e quindi definiti, valori essenziali e quindi efficaci. Tra algor-etica, *ethics by design*, eualgoritmica, infoetica, *data ethics* e così via emerge una sistematica esigenza di eticizzazione dei bit alla ricerca di un "costituzionalismo digitale", espressione positiva dell'ordine di valori essenziale alla tutela dei diritti fondamentali. Possiamo definirlo un diritto naturale al silicio?

Nel riflettere sulla storica contrapposizione tra autocrazia e democrazia, Kelsen affermava ironicamente che "l'uomo non è un libro pensato a tavolino"[52]. Eppure, seduti ai loro tavolini, i

[52] H. Kelsen, *Il primato del parlamento*, trad. it., Giuffré, Milano, 1982, p. 41.

programmatori stanno ormai da tempo ripensando e condizionando il senso della vita e della nostra società. Ecco la necessità di porre tutta una serie di domande. Dati: quali? Informazione: su cosa? Conoscenza, per chi? Saggezza: di chi? Domande ormai ineludibili dinanzi al dominio esercitato sul *web* e sulle tecnologie ad esso collegate dalle corporazioni digitali che, come scrive Habermas, promuovono una sistematica "spinta verso la mercificazione dei contesti del mondo della vita"[53]. Chi guida i *cookie*, elabora i filtri, canalizza i flussi delle informazioni? Gli sviluppi tecnologici assecondano le aspirazioni etiche o le esigenze del mercato? L'intelligenza artificiale si progetta per vendere o per assistere? Serve a incrementare gli utili o il benessere? Guarda al bene o al consumo? All'efficienza produttiva o alla qualità della vita?

Può la saggezza scaturire dagli algoritmi o dall'insieme dei dati? Il dato, al singolare, è un enunciato descrittivo, vero o falso, ma in sé privo di qualsiasi implicazione assiologica. L'algoritmo, al singolare, è un insieme di numeri, più o meno coerenti logicamente, ma in sé privo di qualsiasi implicazione assiologica. I dati e gli algoritmi, assunti nella loro pluralità attraverso la rielaborazione nei contesti in cui sono formati o alla luce degli scopi per cui sono acquisiti, assumono una capacità trasformativa per cui c'è una soglia oltre la quale la quantità o acquista la dignità di uno strumento di elevazione sociale o si traduce in un meccanismo di oppressione. Siamo in bilico tra infoetica e infocrazia.

Lo "*HCAI approach*", l'approccio umano-centrico, mette in luce come l'etica non sia qualcosa di sovrapposto alla scienza e la scienza qualcosa di avulso dall'etica, ma i valori siano il modo attraverso cui ci accostiamo ai fatti e comprendiamo i fatti: è l'antica lezione del diritto naturale sulla indispensabile ricerca di una connessione tra empirismo e razionalismo[54]. All'interno della concezione positivistica la distinzione tra fatti e valori è netta. Nella prospettiva del non cognitivismo etico, fondato sul timore

[53] J. Habermas, *Nuovo mutamento della sfera pubblica e politica deliberativa*, trad. it., Raffaello Cortina, Milano, 2023, p. 56.

[54] B. Shneiderman, *Human-Centered AI*, Oxford University Press, Oxford, 2022, p. 39.

della fallacia naturalistica, è impossibile parlare di un'etica dei dati e ancor meno di un'algor-etica. Tuttavia gli sviluppi dell'intelligenza artificiale ci mostrano come l'etica non possa restare qualcosa di esterno ed eventuale, chiamata magari *a posteriori* a valutare le conseguenze di determinati effetti, ma deve essere inglobata nella dimensione progettuale. L'*ethics by design* impone l'affermazione del principio di esplicabilità[55]: "the obligation to gain greater awareness of unintended consequences and the moral significance of what they do, including how they deal with tragic problems. If AI is not going to be responsible in this sense, it will crash"[56].

Il principio di esplicabilità è appena accennato nel generico riferimento al "diritto di ottenere l'intervento umano, di esprimere la propria opinione, di ottenere una spiegazione" del Considerando 71 del GDPR. Emerge nei Considerando 38 e 40 dell'AI Act, mentre l'art. 86 prevede, anche se con varie limitazioni, un diritto individuale alla spiegazione dei singoli processi decisionali che incidono sui diritti fondamentali. La spiegabilità rientra tra i principi enunciati dall'art. 3 delle L. 23 settembre 2025, n. 132 sulle deleghe al Governo in materia di intelligenza artificiale.

L'esplicabilità garantisce la fiducia. La fiducia presuppone il primato della trasparenza sull'opacità. La trasparenza implica il controllo dell'imprevedibilità da parte della ragione e quindi la limitazione del potere da parte dei diritti. Se guardiamo al passato, non possiamo ignorare come sia stata la crescita dei diritti umani sotto l'impulso del diritto naturale a dissolvere l'assolutismo degli *arcana imperii*[57]. Oggi la scatola nera del *machine learning* rappresenta una nuova forma di oscurantismo, avallato dalle pressioni degli interessi economici sulla politica e della politica sulla scienza. Non sappiamo se questo singo-

[55] B. Goodman, S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'", in *AI Magazine*, 38 (2017), n. 3, pp. 1 ss.

[56] M. Coeckelbergh, "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability", in *Science and Engineering Ethics*, 26 (2020), p. 2068.

[57] Su AI e totalitarismo, M. Coekelberg, *The Political Philosophy of AI*, Polity Press, Cambridge, 2022, pp. 38 ss.

lare diritto naturale al silicio che si sta formando, tra computer e algoritmi, avrà la stessa capacità che ha avuto la tradizione giusnaturalistica nel sollevare la dignità umana, nel sollecitare quella "intenzione verso l'andatura eretta"[58] che ha rifiutato la soggezione agli *arcana imperii*. È innegabile che l'intelligenza artificiale rappresenti una svolta cruciale all'interno del nostro futuro. L'eualgoritmica, l'infoetica, l'algor-etica, l'*HCAI approach*, la *DIKW pyramid* hanno il merito di porre "oggettivamente in primo piano il problema dei giusti motivi sociali dell'agire, e soggettivamente il problema di coscienza della possibilità di conoscere quei giusti motivi"[59].

[58] E. Bloch, *Diritto naturale e dignità umana*, trad. it., Giappichelli, Torino, 2005, p. 173.

[59] Quando ha scritto queste parole, Welzel non sapeva neppure cosa fosse l'intelligenza artificiale, ma era sicuro di offrire una nitida descrizione del diritto naturale (*Diritto naturale e giustizia materiale*, trad. it., Giuffré, Milano, 1965, p. 7).

Sveva D'Agosto

La costruzione della figura femminile fra neuroni specchio e intelligenza artificiale simbiotica

Abstract: **The Construction of the Female Figure Between Mirror Neurons and Symbiotic Artificial Intelligence.** The unstoppable process of construction of the symbiotic system (human/artificial intelligence) undoubtedly determines a phantasmagorical implementation of the phenomenon of inauthentic representations. The correlations of these representations to the distortion of the informational world have consequences of no small importance in the context of mirror neurons functioning. Indeed, the new way of producing knowledge and "truth" through an infinite flow of data seems to go in the direction of perfect epistemic injustice in the representation of female figure online. This phenomenon produces mechanisms that significantly distance women from the recognition and affirmation of their fundamental rights.
The algorithmic logic, or rather that of the multiplication of data, information, and acquired images, only abnormally amplifies this process. As a result, the subject is induced to focus thoughts and behaviours solely on the objectification of female characteristics and their instrumentalization. The human brain is not insensitive to these continuous stimuli, on the contrary it seems to be shaped by them. What are the irreversible consequences?

Keywords: Symbiotic AI, Mirror Neurons, Female Figure Construction.

Sommario: 1. Può il concetto di informazione spiegare il concetto di verità? − 2. Intelligenza artificiale simbiotica: quale è la nuova x del paradigma? − 3. Neuroni specchio e l'uomo come ente naturale generico – 4. Il diritto come *kathécon*.

1. Può il concetto di informazione spiegare il concetto di verità?

L'incapacità di penetrare la complessità opaca del mondo è un fatto ineludibile, a partire, soprattutto, dalla seconda metà del Novecento, momento storico in cui l'informatica inizia a pervadere in modo capillare l'intero mondo del sapere. Il tema del rapporto tra mente e *computer*, e delle implicazioni gnoseologiche ad esso correlate, è soltanto uno dei principali argomenti di indagine.

Nel giro di mezzo secolo, infatti, la rivoluzione digitale ha inciso profondamente sulla realtà in cui siamo immersi, trasfigurando la modalità attraverso cui percepiamo noi stessi e rappresentiamo il mondo ambiente.

Il concetto di rappresentazione, d'altro canto, è strettamente correlato a quello di informazione. Nei processi di rappresentazione, ciò che è evocato dall'elemento visibile si trasfigura in complemento invisibile nella dimensione più intima dell'esperienza dell'altro, del suo essere cosciente e desiderante entro una specifica realtà fattuale.

La "rappresentazione", infatti, è sistematicamente definita da Husserl come "procedimento di ri-presentificazione il quale rende presente, riproducendolo in un'immagine (grafica, o di fantasia), un momento assente della coscienza"[1]. Il principio generale della rappresentazione è che un evento presente stia *per* o rimandi *a* una cosa assente nel momento attuale, istituendo con la stessa una relazione di corrispondenza biunivoca.

È possibile, d'altro canto, che le rappresentazioni siano fuorvianti, rinviando a ciò che simboleggiano in modo meramente apparente, ovvero come pseudo-concetti o "rappresentazioni inautentiche"[2]. Il fenomeno delle rappresentazioni inautentiche determina problematiche di diversa natura: la delimitazione di ciò che è rappresentabile in modo autentico, infatti, è decisiva per circoscrivere la sfera di ciò che è in linea di principio possibile o non possibile, esistente o inesistente nella realtà umana. E questo ha a che fare non semplicemente con la modalità attraverso cui il soggetto percepisce e recepisce l'informazione, ma in prima istanza con il funzionamento meccanico dell'organo deputato all'elaborazione dei dati immagazzinati, il cervello. Che tipo di relazione, dunque, si instaura fra il cervello e le informazioni proposte ripetutamente e secondo modalità ben precise dai sistemi di informatica computazionale? Si ingenerano semplici "*bias* cognitivi", oppure le conseguenze sono ulteriori ed hanno proprio a che fare con il funzionamento dell'organo principale del sistema nervoso centrale?

[1] E. Husserl, *Ricerche logiche*, II, trad. it. a cura di G. Piana, il Saggiatore, Milano, 2005, pp. 375-377.

[2] Cfr. A. Zhok, *Rappresentazione e realtà. Psicologia fenomenologica dell'immaginario e degli atti rappresentativi*, Mimesis, Milano, 2014, p. 7.

In questa prospettiva si rivela determinante la scoperta dei neuroni specchio[3], gli stessi rappresentano per le neuroscienze ciò che il DNA è stato per la biologia.

La loro scoperta nella corteccia premotoria e parietale della scimmia, e la successiva evidenza dell'esistenza di un sistema specchio anche nel cervello umano hanno permesso per la prima volta, infatti, di chiarire i meccanismi neurofisiologici alla base di alcuni aspetti determinanti della cognizione sociale.

Alla luce di tali e complessi fenomeni di rispecchiamento, quali possono essere le conseguenze epistemiche derivanti dalla interazione uomo/macchina? Queste ultime potrebbero incidere nella costruzione della figura femminile in rete, mettendo addirittura a repentaglio la tutela dei diritti fondamentali frutto di una tanto tribolata e sofferta acquisizione?

2. Intelligenza artificiale simbiotica: quale è la nuova x del paradigma?

L'inarrestabile processo di costruzione del sistema simbiotico (uomo/intelligenza artificiale) determina senza dubbio un'implementazione fantasmagorica del fenomeno delle rappresentazioni inautentiche correlate alle distorsioni del mondo informa-

[3] La scoperta dei neuroni specchio negli anni Novanta del ventesimo secolo ha reso possibile una nuova concezione relativa alla prefigurazione del sistema motorio, permettendo, in tal modo, l'approfondimento dell'indagine neurofisiologica in campi del sapere prima appannaggio esclusivo delle discipline umanistiche. Sull'argomento la letteratura di riferimento è ormai vasta. Mi limito a segnalare: G. Rizzolatti, L. Vozza (a cura di), *Nella mente degli altri. Neuroni specchio e comportamento sociale*, Zanichelli, Bologna, 2007; G. Rizzolatti, C. Sinigaglia, *So quel che fai. Il cervello che agisce e i neuroni specchio*, Raffaello Cortina, Milano, 2005; G. Rizzolatti, A. Gnoli, *In te mi specchio. Per una scienza dell'empatia*, Rizzoli, Bologna, 2016; G. Rizzolatti, *Sei tu il mio io. Conversazione sui neuroni specchio con Antonio Gnoli*, Edizioni Casagrande, Bellinzona, 2017; M. Ammanniti, V. Gallese, *The Birth of Intersubjectivity. Psycodynamics, Neurobiology and the Self*, W.W. Norton and Company, New York, 2014; M. Ammanniti, V. Gallese, *La nascita dell'intersoggettività. Lo sviluppo del sé tra psicodinamica e neurobiologia*, Raffaello Cortina, Milano, 2014; V. Gallese, U. Morelli, *Cosa significa essere umani? Corpo, cervello e relazione per vivere nel presente*, Raffaello Cortina, Milano, 2024.

zionale, con conseguenze di non poco conto sul versante di vere ingiustizie epistemiche.

Di fatto la tecnologia disvela, trasforma e controlla il mondo, a volte disegnando e creando nuove realtà. Essa tende a stimolare nuove idee, a modellare nuovi concetti e a causare problemi inediti, spesso modificando valori e prospettive etiche[4]. Ma c'è di più: i sensori, gli attuatori e le componenti sempre più autonome e "intelligenti" consentono la raccolta massiva di dati (estrapolati dall'ambiente di riferimento e dagli interlocutori stessi), la loro analisi (tramite le moderne tecniche di apprendimento automatico: il cosiddetto "*machine-deep-learning*") e la generazione di servizi e applicazioni proprio a partire da tali dati e dal loro continuo vaglio, arrivando a coinvolgere i "soggetti interagenti" come parte funzionale nel processo di costruzione del sistema simbiotico di riferimento stesso. L'intelligenza artificiale, oggi, non si limita al mero apprendimento automatico, agli effimeri modelli predittivi e all'analisi asettica dei dati, ma inizia a modularsi come forma di *AI* "*human aware*"[5]. Nello specifico, il fenomeno "*machine learning*" si configura come sottocategoria dell'insieme più ampio "*AI*" avente la straordinaria funzione di sviluppare il riconoscimento dei *pattern* e di effettuare automaticamente previsioni attraverso l'utilizzo dei dati, per poi apportare modifiche in autonomia a seguito di qualsivoglia forma di interazione con l'essere umano e con l'ambiente. Il "*deep learning*", invece, è un sottoinsieme della categoria di riferimento "*machine learning*" che impiega reti neurali artificiali per simulare i processi di apprendimento del cervello umano: quante più informazioni la macchina sarà in grado di analizzare, tanto più miglioreranno le capacità performative dell'algoritmo nel prendere decisioni e nel veicolare gli stimoli primari del soggetto interagente[6].

Un aspetto determinante dal punto di vista della teoria dell'elaborazione dell'informazione è che l'*AI* studia gli agenti cogni-

[4] Cfr. L. Floridi, "La filosofia dell'informazione e i suoi problemi", in *Iride: Filosofia e Discussione Pubblica*, 18 (2005), n. 2, pp. 291-312.

[5] Cfr. S. Sreedharan, "Human-aware AI – A Foundational Framework for Human–AI Interaction", in *AI Magazine*, 44 (2023), n. 4, pp. 460-466.

[6] Cfr. C.M. Bishop, *Deep Learning*: *Foundations and Concepts*, Springer, Cham, 2024.

tivi come sistemi informativi che ricevono, registrano, recuperano, trasformano, generano, trasmettono e operano proprio in base all'informazione ricevuta[7].

L'avvento di tali e sempre più complesse forme di intelligenza artificiale "simbiotica" (uomo/macchina) determina indiscutibilmente una rivoluzione nell'idea stessa di intersoggettività, condizionata ormai in maniera significativa da pratiche e relazioni sociali che si moltiplicano nella rete digitale per mezzo di algoritmi, prefigurando un vero e proprio mondo parallelo e bidimensionale. L'attuale realtà "frapposta" proietta la nostra percezione del reale verso territori inesplorati, influenzando considerevolmente non soltanto l'identità individuale ma anche quella collettiva e giuridica, con conseguenze a dir poco allarmanti sul piano epistemico. Per quanto…

> […] il Web non è qualcosa di compiuto: è uno strumento costantemente in evoluzione che deve essere riprogettato periodicamente per rimanere al servizio dell'umanità. Sul tema del servizio all'umanità le cose sono piuttosto complesse. Non è sempre chiaro se e in che senso la tecnologia possa restare al servizio dell'umanità intera, o invece porsi al servizio di una parte di essa[8].

3. Neuroni specchio e l'uomo come ente naturale generico

In senso assoluto, nessuna mente si può considerare immune dagli effetti del condizionamento che le viene imposto attraverso il linguaggio, le rappresentazioni e la cultura dominanti[9].

D'altro canto, la capacità dell'uomo di produrre senso sembrerebbe già essere iscritta nel carattere aperto della sua biologia, ovvero nel suo essere per costituzione un "ente naturale generico"[10]. Il soggetto, infatti, che continuamente altera, costituisce, riconfigura la relazione di familiarità con il contesto sociale di

[7] *Ivi*, p. 9.

[8] D. Fiormonte, T. Numerico, F. Tomasi (a cura di), *L'Umanista digitale*, il Mulino, Bologna, 2010, p. 60.

[9] Cfr. S. Moscovici, *Le rappresentazioni sociali*, il Mulino, Bologna, 2005.

[10] Cfr. K. Marx, *Introduzione per la critica dell'economia politica*, trad. it. a cura di G. Backhaus, Shake Edizioni, Milano, 2018.

riferimento ne è simultaneamente il prodotto. L'organismo vivente è una realtà organica che include entro i confini della propria esistenza anche il mondo ambiente: quest'ultimo non è al di fuori del corpo, ma è un filtro che lo plasma, una forma di complemento necessario. Ciò ha delle ripercussioni significative sul piano dell'apprendimento. Infatti, se l'atto del conoscere richiede la capacità di fornire una spiegazione corretta di un'informazione esterna da elaborare, al tempo stesso è innegabile che, il più delle volte, gli individui siano meramente informati su quanto pensano di conoscere nel dettaglio. Attraverso tale meccanismo si possono ingenerare rappresentazioni fittizie capaci di determinare sia la conoscenza di sé che, al tempo stesso, una specie di scollamento o di alienazione dell'io.

C'è, dunque, una specifica dimensione fenomenica della esistenza umana che si manifesta nel mondo ambiente, la quale, per essere decifrata e compresa attende l'occhio dell'osservatore.

Il pensiero umano ha sempre cercato di definire le modalità di questo trasferimento di significato, anche in una fase primordiale, prima ancora che si sviluppasse l'utilizzo del linguaggio. I neuroni specchio riescono in parte ad esemplificare tale complessa relazione tra movimento e sensazione[11]. Essi sono, appunto, come specchi che riflettono e connettono, determinando, al tempo stesso, anche drammatici isolamenti.

La premessa è la seguente: "un neurone non è un soggetto epistemico, ma una cellula che genera delle tensioni, dei voltaggi. Non v'è nulla, quindi, di intrinsecamente intenzionale nel funzionamento di un neurone. D'altro canto, non è contenuto in una scatola magica, ma in un organo – il cervello – che è legato, vincolato, cresce e si sviluppa in parallelo ad un corpo, attraverso il quale ha accesso al mondo esterno"[12].

Sarebbe un errore imperdonabile quello di analizzare il sistema nervoso in una dimensione solipsistica, tralasciando l'influenza determinante dell'intersoggettività nel costruirsi, svilupparsi e consolidarsi delle architetture computazionali di

[11] Cfr. V. Gallese, "Dai neuroni specchio alla consonanza intenzionale. Meccanismi neurofisiologici della intersoggettività", in *Rivista di Psicoanalisi*, LIII (2007), n. 1, pp. 197-208.

[12] *Ivi*, p. 198.

riferimento. È necessario, perciò, partire da un dato di fatto: "in condizioni non alterate gli esseri umani non sono alienati dal significato delle azioni, emozioni, o sensazioni esperite dai propri simili, in quanto godono di quella che può definirsi come 'consonanza intenzionale'"[13]. Ciò non dipende unicamente dal fatto che esperiamo sensazioni, emozioni e, in genere, modalità di agire analoghe a quelle degli altri individui, quanto piuttosto dall'omogeneità dei processi nervosi che presiedono ad esse.

Quando ci relazioniamo all'altro, al "diverso da noi", abbiamo la facoltà di esperirne direttamente l'umanità. Ad entrare in gioco sono veri e propri meccanismi di rispecchiamento e simulazione. I neuroni specchio, infatti, pur appartenendo al sistema motorio non sono deputati unicamente al movimento, ma si configurano altresì come particolarmente reattivi agli stimoli visivi. Ad esempio, se osserviamo qualcuno afferrare una mela, abbiamo una comprensione istantanea dell'azione, grazie ad un meccanismo di simulazione dell'atto osservato da parte del soggetto passivo. L'analisi dell'agire altrui, seppur superficiale e immediata, induce il soggetto osservante ad attivare lo stesso circuito motorio deputato all'attuazione di quell'identico e specifico movimento. In una certa prospettiva è come se il sistema neurale "comprendesse" l'intenzione dell'altro, simulandone l'azione, e, dunque, cogliendo in via immediata le finalità del suo agire[14].

Ancora: Cosa accade nel nostro cervello quando osserviamo i gesti comunicativi di una persona che parla, il *lipsmaking* di una scimmia o un cane che abbaia? La risposta è fornita da uno studio condotto da Buccino e da altri autorevoli scienziati con la tecnica fMRI[15]. I risultati sono avvincenti: se guardiamo

[13] Cfr. V. Gallese, "Il corpo teatrale: Mimetismo, neuroni specchio, simulazione incarnata", in *Culture Teatrali*, 16 (2008), pp. 13-38; V. Gallese, "Mirror neurons and the social nature of language: The neural exploitation hypothesis", in *Social Neuroscience*, 2008, n. 3, pp. 317-333.

[14] Cfr. V. Gallese, M. Guerra, *Lo schermo empatico*, Raffaello Cortina, Milano, 2015.

[15] G. Buccino, N. Canessa, G. Lagravinese, F. Lui, C. Porro, I. Patteri, G. Rizzolatti, "Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study", in *Journal of Cognitive Neuroscience*, 16 (2004), n. 1, pp. 114-126.

qualcuno parlare vi è una immediata attivazione bilaterale del sistema premotorio che include l'area di Broca. Non accade la stessa cosa se ci soffermiamo sui gesti comunicativi di una scimmia o di un cane. L'attivazione premotoria bilaterale sarà ridotta nel primo caso, e addirittura assente nel secondo. Ne deriva che tutte le relazioni sociali sono per natura "interattive", a patto che vi sia di fronte un "tu" e non un mero oggetto inerte da esaminare.

Il fatto, insomma, che un uomo, per dirla con Wittgenstein, possa rappresentare un vero e proprio enigma per un altro essere umano, non comporta una distanza strutturale di fondo[16]. Non vi è un distacco epistemico o uno sguardo apartecipativo.

Tutt'altro. Molteplici studi attestano il ruolo della simulazione motoria anche nella finzione narrativa, tanto che A. Coplan[17] non esita a parlare di "prospettiva empatica" a proposito del coinvolgimento del lettore.

I meccanismi attraverso cui la narrazione stessa è costruita o plasmata, non sono scevri da condizionamenti, ed è proprio in questa prospettiva che si manifesta una stretta correlazione fra la modalità attraverso cui è rappresentata la figura femminile in rete per mezzo dei sistemi di intelligenza artificiale e la simulazione incarnata propria dei meccanismi di rispecchiamento. Infatti, tali meccanismi sono fortemente coinvolti nei processi imitativi con conseguenze del tutto negative.

D'altronde, la tecnologia "non si muove come l'astuzia della ragione di hegeliana memoria, ma si realizza nelle forme casuali o secondo le richieste momentanee della propria storia. Essa insomma è frutto di scelte"[18].

La premessa è semplice. I sistemi di intelligenza artificiale che interagiscono simbioticamente con il soggetto si "nutrono" essenzialmente dei contenuti divulgati per mezzo della rete. La questione fondamentale si concentra, quindi, nella programmazione degli algoritmi che ne stanno alla base. Infatti, se l'"*input*" e il materiale usato per l'elaborazione dei dati sono viziati "*ab*

[16] L. Wittgenstein, *Ricerche Filosofiche*, trad. it. a cura di M. Trinchero, Einaudi, Torino, 2014, p. 261.

[17] A. Coplan, "Empathic Engagement with Narrative Fictions", in *The Journal of Aesthetics and Art Criticism*, 62 (2004), pp. 141-152.

[18] Cfr. D. Fiormonte, T. Numerico, F. Tomasi (a cura di), *op. cit.*, p. 240.

origine", anche i risultati non saranno meno riduttivi. Ne conseguirà, in sintesi, l'ulteriore diffusione e l'automatico consolidamento dei cosiddetti e ben noti "stereotipi di genere".

Che i *media* possano influenzare la "*communis opinio*" relativamente all'immagine della donna è fatto notorio, ed è, peraltro, riconosciuto a chiare lettere dalla Convenzione di Istanbul, che all'art. 17 menziona esplicitamente il ruolo degli stessi nel contrasto e nella prevenzione alla violenza di genere. Sono, del resto, evidenti i cambiamenti che hanno investito la figura femminile e la relativa concezione della donna correlati all'utilizzo dei mezzi di comunicazione di massa ed in particolare alla diffusione dilagante dei *social media*. Con l'avvento dell'intelligenza artificiale i rischi aumentano piuttosto che diminuire. I sistemi di trattamento automatizzato, infatti, sono "addestrati" con *data set* di immagini che l'algoritmo reperisce direttamente dalla rete. Quali sono i modelli propugnati?

"Le donne vengono rappresentate in rapporto agli uomini: non solo viste dall'altro sesso, ma viste unicamente in relazione all'altro sesso"[19]. La figura femminile è di conseguenza relegata a ruoli di tendenziale marginalità: secondari, passivi e ornamentali; la bellezza femminile è vista riduttivamente come unica qualità da coltivare e sfoggiare.

Si ha, così, una esaltazione ed esasperazione degli stereotipi di genere, ovvero di quell'insieme rigido di credenze condivise e trasmesse socialmente, relative ai comportamenti, al ruolo, alle occupazioni, alle caratteristiche fisiche di una persona, in relazione alla sua appartenenza ad un genere specifico. La mancanza di conformità a tali attese fa sì che i soggetti interessati siano ritenuti o giudicati non aderenti ai modelli di riferimento e, nel caso specifico, qualificati come "poco femminili". Ne deriva, tra l'altro, l'omologazione delle persone appartenenti a uno stesso gruppo, ridotte a copie di un giornale con lo stesso stampo tipografico (si parla, appunto, di stereotipi)[20].

[19] V. Woolf, *Una stanza tutta per sé*, trad. it. a cura di M. del Serra, Newton Compton, Roma, 2010, p. 93.

[20] Cfr. W. Lippmann, *L'opinione pubblica*, Donzelli Editore, Roma, 2018.

L'industria tecnologica, grazie ai sistemi di intelligenza artificiale, non fa che implementare e automatizzare tali fenomeni. Analiticamente, come è stato osservato:

> The term Artificial Intelligence (AI) refers to a large class of software-based systems that receive signals from the environment and take actions that affect that environment by generating outputs such as content, predictions, recommendations, classifications, or decisions influencing the environments they interact with, among other outputs[21].

Si tratta, dunque, di sistemi altamente pervasivi. Senza neppure rendercene conto, interagiamo con "*AI Systems*", conversando con "assistenti digitali" o "*chatbots*" incorporati nei nostri dispositivi personali. A questi "*digital assistants*" sono spesso attribuite caratteristiche tipicamente "umane", in un continuo processo di antropomorfizzazione caratterizzato da una spiccata tendenza alla "femminilizzazione", in perfetto accordo con la semiotica degli stereotipi di genere[22].

A ciò si aggiunge un altro aspetto. Con l'intelligenza artificiale si passa da un meccanismo razionale-cognitivo finalizzato alla risoluzione di un problema, ad una interazione di tipo emozionale che tende sempre più ad assumere i connotati della "socialità". Ne deriva una vera e propria relazione "emotiva" che si traduce in un dialogo ben più profondo, il quale può sconfinare nella percezione irrazionale dell'"essere compresi". Si acuisce, in tal senso, la natura "performativa" del genere che si sostanzia nella ripetizione di atti che si manifestano e qualificano secondo uno specifico codice di significati. Mi riferisco qui al concetto butleriano di performatività per cui gli atti ripetuti rappresentano la fonte stessa della produzione discorsiva sul genere, e determinano di per sé la soggettività degli individui[23].

[21] A. Burt, K. Greene, P. Hall, L. Perine, R. Schwarz, A. Vassilev (eds.), "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence", in *NIST* Special Publication 1270, https://doi.org/10.6028/NIST.SP.1270

[22] Cfr. P. Costa, L. Ribas, "AI Becomes Her: Discussing Gender and Artificial Intelligence", in *Technoetic Arts*, 17 (2019), n. 1, pp. 171-193.

[23] Cfr. J. Butler, *Gender Trouble: Feminism and the Subversion of Identity*, Routledge, New York-Londra, 1990.

Interagendo con sistemi di intelligenza artificiale "*on daily basis*" si rischia, dunque, di percepirli non più come meri strumenti o *tools*, bensì come "sostituti artificiali" con precisi "attributi di genere" che attendono per altro, pienamente, le aspettative codificate della società. Lo dimostrano, fra gli altri, gli studi condotti da C. Nass e Y. Moon secondo cui le persone tendono a relazionarsi ai computer secondo le stesse modalità con cui si relazionano agli esseri umani[24].

È ciò che sosteneva già nel 1950 Alan Turing in un suo scritto intitolato "*Computing Machinery and Intelligence*". Né si può ignorare la specificità e la potenza dei simboli. Si pensi ad Apple:

> We recognize the Apple computer symbol, I think, as a clever icon for the digitalization of the creation myth. Within this logo, sin and knowledge, the forbidden fruits of the garden of Eden are interfaced with memory and information in a network of power[25].

Resta, dunque, la questione di come la macchina sia "istruita".

Si fa riferimento, in merito, al fenomeno del "*garbage in, garbage out*": la qualità delle informazioni *output* è determinata dalla qualità degli elementi attraverso cui la macchina è "indottrinata". E. Bender e altri hanno evidenziato l'imperscrutabile quantitativo di dati che confluiscono nei sistemi LLMs, includendo associazioni stereotipate e ritratti di genere a volte annichilenti[26].

I sistemi di intelligenza artificiale accrescono straordinariamente la capacità di veicolare informazioni e messaggi ad un pubblico sempre più vasto e tendenzialmente sconfinato. La diffusione senza soluzione di continuità di contenuti, immagini, dati, anche attraverso *AI Systems* come Chat Gpt o Chat Bot, determina la condivisione e il consolidamento di veri e propri

[24] C. Nass, Y. Moon, "Machines and mindlessness: social responses to computers", in *Journal of Social Issues*, 56 (2000), n. 1, pp. 81-83.

[25] J. Halberstam, "Automating gender: Postmodern Feminism in the Age of Intelligent Machine", in *Feminist Studies*, 17 (1991), n. 3, p. 440.

[26] Cfr. E.M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, "The Dangers of Stochastic Parrots: Can Language Model Be Too Big?", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021 (https://doi.org/10.1145/3442188.3445922).

modelli stereotipati e trasfigurati, i quali rimandano ad una idea di donna contraffatta e surreale. L'esposizione quasi ossessiva a "simulacri" che promuovono specifici canoni di bellezza convenzionali fa in modo che le giovani donne si concentrino sempre più sul proprio aspetto fisico, tralasciando del tutto lo sviluppo e l'implementazione delle proprie capacità cognitive.

Si sedimentano, così, attraverso la comunicazione per immagini propugnata dai sistemi di intelligenza artificiale, i fenomeni di ipersessualizzazione e di oggettivazione del corpo femminile come "forme di radicale deprezzamento e ostracismo che negano apertamente l'umanità dell'altro, [o come] strategie più o meno sottili che erodono in modo consapevole l'altrui partecipazione all'umanità"[27]. Le donne, a loro volta, tendono ad interiorizzare la prospettiva dell'osservatore e ad autodefinirsi riduttivamente, a causa dei citati meccanismi di rispecchiamento e simulazione incarnata propri del cervello umano[28].

La logica algoritmica, ovvero quella della moltiplicazione dei dati, delle informazioni, delle immagini acquisite, non fa che amplificare in maniera abnorme tale processo con gravi ripercussioni sull'intera società.

4. Il diritto come *katéchon*

In un quadro tanto intricato e complesso è ancora possibile garantire la piena tutela dei diritti fondamentali?

Nella nuova e penetrante "società algoritmica" non ci limitiamo ad abitare una circoscritta superficie fisica, ma entriamo continuamente a far parte di una molteplicità di mondi paralleli e fittizi, la cui moltiplicazione assume connotati esponenziali.

Senza dubbio, l'impiego indiscriminato di forme di intelligenza artificiale fa sì che l'algoritmo sia non solo più competente nella predittività, nel "*mind reading*" e nella comprensione empatica delle condizioni interiori dei personaggi finzionali, ma che sia anche capace di un ruolo decisivo nella veicolazione

[27] Cfr. C. Volpato, "Negare l'altro. La deumanizzazione e le sue forme", in *Psicoterapia e scienze umane*, 47 (2013), n. 2, pp. 311-328.

[28] Cfr. Y.P. Huang, B. Moradi, "Objectification theory and psychology of women: A decade of advances and future directions", in *Psychology of Women Quarterly*, 32 (2008), n. 4, pp. 377-398.

dei desideri, delle emozioni e delle aspirazioni più intime del soggetto. Tali processi sono agevolati dall'anatomia umana. Ad essere decisivo, infatti, è proprio il fenomeno del "rispecchiamento" che, come si è visto, induce il soggetto ad emulare e desiderare ciò che vede rappresentato. Gli impulsi visivi illusori si sincronizzano con le scariche elettriche cerebrali; le stesse non codificano il movimento in sé, ma lo scopo dell'azione, il quale viene generalizzato e semplificato, al di là dei singoli movimenti specifici.

Le neuroscienze, oggi, ci permettono di poter affermare con più chiarezza che lo scarto fra mondo reale e immaginario è senza dubbio meno netto di quanto si potesse pensare. Vedere e immaginare di vedere, fare qualcosa o immaginare di farlo, sentire o immaginare di sentirlo, condividono l'attivazione degli stessi circuiti neurali, il rischio è che la società tenda a trasformarsi nella sua stessa rappresentazione.

Cosa può fare il diritto in questo quadro tanto complesso? Può svolgere realmente il suo ruolo di "*kathécon*" ossia di freno, ostacolo, e, soprattutto, di indirizzo per il futuro?

C'è, intanto, un problema di tempistiche. L'*AI Act* recentemente approvato dal Parlamento europeo potrebbe già essere inattuale[29]. Il testo prevede, infatti, la sua effettiva applicazione (ad eccezione di alcune disposizioni) soltanto a partire dai 24 mesi successivi alla sua entrata in vigore, un termine che mal si concilia con lo sviluppo proteiforme dei sistemi *AI*.

Va rimarcata, però, la questione di fondo. Riflettere, infatti, sulla parità di genere potrebbe essere un fatto estremamente sterile, se si cade in una realtà finzionale e al tempo stesso prevaricatrice. D'altronde l'elaborazione, lo studio e la promozione delle tecnologie finalizzate al superamento dei limiti umani, attraverso l'analisi dei *trend*, delle dimensioni psicologiche e il loro impatto sociale, tendono ad enfatizzarne unicamente gli aspetti positivi di tali strumenti, sottovalutandone i potenziali

[29] Il 13 marzo 2024 il Parlamento europeo ha approvato l'AI Act, prima legge complessiva sull'intelligenza artificiale. Il nuovo regolamento mira a garantire che i sistemi *AI* immessi sul mercato europeo siano sicuri nel rispetto dei diritti e dei valori fondamentali dell'UE. Le diverse tecnologie sono valutate in base a quattro parametri di rischio: inaccettabile, alto, limitato, minimo.

ed intrinseci pericoli. L'esito ineluttabile è l'alterazione in senso "peggiorativo" della condizione umana.

Croce ritenne che l'attività spirituale procedesse secondo una precisa dialettica, quella dei distinti, secondo la stessa lo spirito si distingue nei due momenti differenti della teoria e della prassi, ovvero dell'intuizione logico concettuale e dell'impegno etico morale o utilitaristico economico[30]. L'assenza di una chiara visione logico/concettuale porta ad una semplificazione catastrofica della realtà, che si focalizza unicamente e drammaticamente sulla prospettiva utilitaristico/economica tipica di un distacco sempre minore fra consapevolezza dei propri limiti e mezzi tecnici a disposizione.

[30] B. Croce, *Filosofia della pratica*, Laterza, Bari-Roma, 1963, p. 5.

María Teresa García-Berrio

Ética de la virtud para una IA compasiva: Ontología, subjetividad y condición humana en la era digital

Abstract: **Virtue Ethics for a Compassionate AI. Ontology, Subjectivity and the Human Condition in the Digital Age.** As artificial intelligence systems increasingly shape social reality, there is an urgent need to rethink their ethical foundations beyond simple rule-based approaches. This chapter positions virtue ethics as a transformative model for shaping AI systems that genuinely foster human flourishing and compassion. Integrating philosophical, bioethical, and legal analysis, it addresses challenges such as vulnerability, autonomy, and informational rights. Drawing on thinkers including Bauman, Floridi, Nussbaum, Cortina, and Stiegler, the chapter argues for governance frameworks that safeguard dignity and enable responsible, human-centered artificial intelligence.

Keywords: AI, Virtue Ethics, Neurorights, Algorithmic Biases, Human Dignity, Ethical Governance.

Resumen: 1. Inteligencia Artificial y Condición humana – 2. Vigilancia, control y responsabilidad digital – 3. Inconsciente, infosfera y dignidad: Hacia una ontología crítica de la subjetividad en la era de la IA – 4. La ruptura del Antropocentrismo ante la irrupción de la IA – 5. Emocionalidad, Vulnerabilidad y Virtud frente a la Tecnología – 6. Hacia una IA Compasiva – 7. Principio de Beneficencia ante el riesgo de discriminación por sesgos algorítmicos – 8. Principio de No-maledicencia: La prevención del daño y la preservación de la dignidad humana ante el riesgo de alexitimia virtual – 9. A modo de conclusión.

1. Inteligencia Artificial y condición humana

La Inteligencia Artificial (en lo sucesivo, IA) se inscribe en el grupo de tecnologías denominadas "disruptivas" que están reconfigurando de manera acelerada e irreversible nuestro mundo a partir de la convergencia NBIC (nanotecnologías, biotecnologías, tecnologías de la información y ciencias cognitivas). En este contexto, la IA puede describirse como un entramado de procedimientos algorítmicos capaces de procesar, cruzar y reutilizar cantidades masivas de datos para efectuar cálculos, resolver problemas y adoptar decisiones, sustituyendo o ampliando

funciones tradicionalmente asociadas a las capacidades cognitivas humanas. En efecto, la IA procesa cruza y reutiliza ingentes cantidades de datos mediante algoritmos. Un algoritmo es una lista más o menos larga de instrucciones; dicho en otros términos, un conjunto ordenado y finito de pasos que puede emplearse para hacer cálculos, resolver problemas y alcanzar decisiones.

Esta mutación tecnológica intensifica la configuración de nuestras sociedades como entornos socio-técnicos híbridos, en los que se difuminan las fronteras entre sujetos humanos y artefactos virtuales, y en los que la interacción constante hombre-máquina (*Human-Machine Interaction*) genera una suerte de simbiosis entre inteligencia humana y artificial, lo que nos exige articular nuevos marcos de colaboración transversal entre filósofos y juristas, expertos en ética, ingenieros y gestores públicos, con el objetivo de diseñar, desde una perspectiva genuinamente interdisciplinar, cauces normativos e institucionales que orienten el desarrollo de la IA hacia un progreso éticamente aceptable y, en un plano ulterior, que permitan integrar en estas *tecnologías disruptivas* los valores que sostienen la dignidad y los derechos humanos.

Desde esta óptica, la IA puede entenderse como una disrupción antropológica sin precedentes, con un impacto directo sobre las estructuras naturales, económicas y sociales que vertebran la vida en las comunidades humanas. La expansión de sistemas algorítmicos reaviva y tensiona nuestra autocomprensión ética y, en este sentido, los últimos avances tecnológicos asociados a la IA Generativa han supuesto un revulsivo para nuestra conscienciа ética frente a la ilusión de la autonomía individual postmoderna, fuertemente marcada por el elevado precio pagado en las sociedades industrializadas por el proceso conocido como *individualización* como marca sustantiva de la "modernidad reflexiva" que propugnan algunos de los sociólogos y filósofos contemporáneos más destacados como Gilles Lipovetsky, Elizabeth Beck o Zygmunt Bauman, quienes a su vez se han nutrido de la tradición sociológica de Norbert Elias y Émile Durkheim.

En la obra *Modernidad líquida*, Zygmunt Bauman conceptualiza la llamada *individualización* como aquel proceso que consiste "en hacer que la identidad humana deje de ser un 'dato'

para convertirse en una tarea, y en cargar sobre los actores la responsabilidad de dicha tarea y de las consecuencias (y efectos secundarios) de su actuación"[1]. En el sentido de Bauman, la *individualización* ya no alude sólo a la emancipación de la persona respecto de marcos comunitarios tradicionales, sino también a la progresiva atribución al individuo de la responsabilidad exclusiva sobre su biografía en un contexto mediado donde los márgenes de decisión real del ser humano aparecen fuertemente condicionados. Tal es así que, para Bauman, la identidad deja de ser un dato estable, anclado en instituciones sólidas (como la pertenencia a una clase social, a una familia o a una nación), para convertirse en un proyecto biográfico abierto cuya realización se impone al individuo como obligación permanente y como fuente de responsabilidad exclusiva, tanto para sus éxitos como para sus deficiencias o fracasos. Efectivamente, esta "obligación de ser individuo ante todo" se acompaña de una transferencia sistemática de problemas estructurales al plano de la gestión privada de la vida, erosionando los márgenes reales de autodeterminación en contextos crecientemente mediados por tecnologías digitales y lógicas de mercado; dado que la *individualización* genera la ilusión de un control pleno sobre el propio destino, se induce al individuo a creer que, mediante decisiones informadas y racionalizadas, puede gestionar riesgos y optimizar su trayectoria vital, incluso cuando tales decisiones están insertas en sistemas altamente asimétricos.

La noción de *individualización* en Bauman puede pues servirnos para comprender la mutación contemporánea de la relación entre sujeto, comunidad y estructuras de poder en los contextos actuales tecnológicamente mediados. En efecto, Bauman ha situado la *individualización* en el corazón de la transformación de la modernidad "sólida" en modernidad "líquida", entendida esta última como un estadio en el que los marcos institucionales pierden rigidez, se flexibilizan los vínculos comunitarios y se intensifica la exigencia de movilidad y adaptación constante del ser humano al medio condicionado en el que desarrolla su existencia. En este contexto, el individuo es interpelado como "empren-

[1] Z. Bauman, *Modernidad Líquida*, Fondo de Cultura Económica de Argentina, Buenos Aires, 2003, p. 20.

dedor de/por sí mismo", al que se le atribuye no sólo la tarea de construirse una identidad, sino la carga moral de responder por todas las consecuencias de dicha construcción, incluidas aquellas fallas sistémicas que escapan en realidad de su control. Desde esta perspectiva, la *individualización* conforma un proceso ambivalente de adaptación del sujeto al medio: al mismo tiempo que promete libertad y autonomía, refuerza la vulnerabilidad del sujeto al desvincularlo de redes colectivas de apoyo y al desplazar el tratamiento de los conflictos sociales desde la comunidad hacia soluciones biográficas "individualizadas". El resultado de todo ello es una forma de *responsabilización* radical del sujeto que tiende a despolitizar los conflictos, convirtiendo fracasos estructurales (desempleo, exclusión, sobre-endeudamiento) en déficits personales de esfuerzo, falta de carácter o personalidad de los sujetos e incluso insuficiencia de competencia técnica de las personas.

La lectura de Bauman converge en este punto parcialmente con la tesis del sociólogo Ulrich Beck[2] sobre la "sociedad del riesgo" y la llamada "biografía electiva", en la que la trayectoria vital de las personas se presenta como resultado de decisiones individuales supuestamente libres, pero en realidad fuertemente condicionadas por estructuras económicas y tecnológicas. Beck subraya que la individualización implica una "desanclaje" de los sujetos respecto a los estratos de clase y los marcos de familia tradicionales, de modo que las biografías se tornan más abiertas, pero también más expuestas al azar de la precariedad, la vulnerabilidad y la inseguridad existencial.

Aplicada a los procesos tecnológicos, la categoría de *individualización* permite problematizar el modo en que las infraestructuras digitales y la IA reconfiguran la experiencia de autonomía de las personas y la condición humana. La IA se apoya en arquitecturas sociotécnicas altamente interconectadas – plataformas, sensores, dispositivos móviles, ecosistemas del Internet de las Cosas – que recogen, procesan y monetizan volúmenes masivos de datos generados por las propias interacciones cotidianas de las personas. Paradójicamente, la promesa de empoderamiento digi-

[2] U. Beck, *La sociedad del riesgo mundial: En busca de la seguridad perdida*, Paidós, Barcelona, 2008.

tal se construye sobre un esquema en el que el usuario aparece simultáneamente como fuente primaria de datos y como responsable último de gestionar los riesgos derivados de su exposición permanente a la vigilancia y la clasificación automatizada[3].

Por consiguiente, la *individualización* supone una transformación radical de la estructura de la personalidad propia de las sociedades, ya que se hace creer al individuo aislado que es factible su liberación del encorsetamiento de las estructuras sociales tradicionales y que, por ende, puede disponer de un control completo sobre el desarrollo de su vida a través de las decisiones que toma en los procesos de racionalización tecnológica. Sin embargo, el efecto de esta *individualización* es devastador, ya que arranca al individuo de la comunidad, desgastando sus vínculos de confianza hasta dejarle sin protección ante un mundo virtual en el que cada vez es más difícil el pleno desarrollo autónomo de las personas y cuyo reverso es el riesgo asociado al uso masivo de medios digitales para la vigilancia y el control social. Desde una lectura baumaniana, esta configuración intensifica la "liquidez" de la identidad: los perfiles se actualizan de forma continua, las huellas digitales se agregan para producir representaciones dinámicas del sujeto, y las decisiones algorítmicas condicionan crecientemente sus oportunidades educativas, laborales y de participación social. Se multiplica así la exigencia de autorregulación – administrar la propia reputación en línea, gestionar la privacidad, optimizar la imagen pública – al tiempo que se difuminan las capacidades efectivas del individuo aislado para negociar las asimetrías de poder inscritas en las infraestructuras de datos y en los mercados de plataformas. Este proceso produce un vaciamiento progresivo del *ágora* – entendida aquí por Bauman como el espacio de articulación entre lo privado y lo público – donde lo público se coloniza por lógicas privadas, lo que acrecienta inevitablemente la sensación de soledad y desprotección del sujeto, al tiempo que oscurece la dimensión estructural de fenómenos como la desigualdad digital, la precariedad laboral o la discriminación algorítmica.

[3] U. Beck, E. Beck-Gernsheim, *La individualización: El individualismo institucionalizado y sus consecuencias sociales y políticas*, Paidós, Barcelona, 2003.

2. Vigilancia, control y responsabilidad digital

La expansión de la IA en entornos laborales, financieros, educativos y recreativos refuerza la transición descrita por Zygmunt Bauman desde 'sociedades de disciplina' hacia 'sociedades de control y consumo', en las que la vigilancia se normaliza como condición de acceso a servicios, oportunidades y formas de reconocimiento social. Dispositivos digitales adheridos al cuerpo – relojes inteligentes, pulseras de actividad, sensores biométricos – generan flujos constantes de datos que alimentan sistemas de evaluación, predicción y segmentación, cuyos criterios escapan en gran medida al escrutinio democrático y quedan inscritos en arquitecturas algorítmicas opacas y difícilmente impugnables por los usuarios. Esta infraestructura sociotécnica profundiza la lógica de la individualización baumaniana: es decir, se promete autonomía y auto-optimización, pero se traslada al individuo la carga moral de gestionar, casi en solitario, las consecuencias de decisiones y clasificaciones automatizadas que responden a racionalidades de mercado y control.

En clave de *individualización*, esta dinámica produce una forma de "responsabilización digital" según la cual el sujeto debe justificar permanentemente sus trayectorias vitales (académicas, laborales, de salud, de consumo) frente a métricas automatizadas, asumiendo como fracaso personal lo que en realidad constituye el resultado de sesgos algorítmicos, asimetrías informacionales y diseños orientados a la maximización del beneficio económico. La biografía individual deviene radicalmente contingente a puntuaciones, perfiles de riesgo y reputaciones digitales, reforzando la "inseguridad líquida" descrita por Bauman y debilitando la capacidad de articular experiencias de injusticia digital como problemas públicos y estructurales. El riesgo no se limita a la pérdida de anonimato, sino que incluye la consolidación de un entorno en el que la subjetividad se configura como un proyecto constantemente evaluado, en el que la obligación de ser "emprendedor de sí mismo" se combina con una creciente dependencia de infraestructuras algorítmicas que median silenciosamente el acceso a recursos y oportunidades.

En este mismo sentido se pronuncia la socióloga norteamericana Shoshana Zuboff cuando nos advierte del advenimiento de

un modelo de capitalismo – al que califica como *Capitalismo de la vigilancia* – en el que los procesos automatizados realizados por medios digitales y tecnológicos han ido reemplazado progresivamente los vínculos interpersonales propios de las relaciones humanas por algoritmos que buscan erradicar el sentimiento de "interdependencia humana". Y es ahí precisamente donde reside, a juicio de esta autora, *el oscuro corazón del capitalismo de la vigilancia* que puede llegar a tener efectos desgarradores tanto en la biografía personal de las personas, como en su comprensión de la realidad social y de las normas cognitivas. Por todo ello, admite Zuboff, la manera de generar y recibir información desde medios digitales – tan inmediata en tiempo, tan difusa en fuentes y tan automática en su origen – está logrando alterar la propia capacidad crítica y cognitiva del receptor de la información. A este tipo de poder lo califica esta autora como "poder instrumentario para una tercera modernidad" o *instrumentarismo*, construido a partir de la combinación de los términos predicción, monetarización y control. La tesis de Shoshana Zuboff sobre el "capitalismo de la vigilancia" permite profundizar esta lectura al caracterizar un modelo económico basado en la captura masiva de excedente conductual y su conversión en productos de predicción y modificación de comportamiento. En este modelo, los procesos automatizados ejecutados por medios digitales tienden a sustituir los vínculos interpersonales propios de las relaciones humanas por algoritmos orientados a neutralizar la experiencia de interdependencia y a producir conductas previsibles y rentables. Zuboff conceptualiza este fenómeno como "poder instrumentario", un tipo de poder que no busca tanto adoctrinar conciencias de forma explícita como modular entornos de elección para orientar silenciosamente la conducta, apoyándose en la combinación de extracción de datos, predicción y control algorítmico. La instrumentación alude así a la arquitectura material ubicua – sensores, interfaces, sistemas de computación sensible – que captura y activa la experiencia humana en forma de datos, mientras que designa las orientaciones sociales que permiten a los actores del capital de vigilancia convertir a las personas en meros medios para fines mercantiles ajenos.

> En esa definición – apunta esta autora – la instrumentación hace referencia al títere: la arquitectura material ubicuamente conectada con la computación sensible que transfiere, convierte, interpreta y acciona la experiencia humana. La instrumentalización, por su parte, denota las orientaciones sociales que orientan a los titiriteros hacia la experiencia humana cuando el capital de la vigilancia se vale de las máquinas para transformarnos en medios de los fines mercantiles de otros[4].

Resulta pues imperativo el abordaje ético de las tecnologías asociadas a la IA, identificando para ello, por un lado, los daños y peligros a evitar y promocionando, por otro, aquellos valores de *interdependencia humana* que sean capaces de establecer un ecosistema de confianza para los ciudadanos y usuarios frente al empleo "potencialmente nocivo" de la IA. En este escenario, la vigilancia y el control adquieren una dimensión específicamente digital que se articula en torno a la explotación de vulnerabilidades cognitivas y emocionales, planteando desafíos inéditos para la autonomía, la privacidad mental y la responsabilidad social.

La Unión Europea ha reaccionado a estos riesgos mediante la aprobación del *Reglamento (UE) 2024/1689 sobre Inteligencia artificial* (conocido como *AI Act,* en terminología anglosajona o RAI, en nomenclatura técnica castellana), adoptado en 2024, que establece un marco regulatorio horizontal basado en el riesgo, y que ya se encuentra en fase de despliegue gradual desde comienzos de 2025. Como más adelante veremos, el Reglamento clasifica los sistemas de IA en categorías en función del riesgo [riesgo inaceptable (prohibido), alto riesgo, riesgo de transparencia específica y riesgo mínimo o nulo] y articula obligaciones diferenciadas para fabricantes, proveedores y usuarios en función del impacto potencial sobre los derechos fundamentales y los valores de la Unión, buscando así compatibilizar la promoción de una IA "fiable" con la delimitación de fronteras normativas claras frente a prácticas consideradas intrínsecamente incompatibles con la dignidad humana y la democracia.

[4] S. Zuboff, *La era del capitalismo de la vigilancia: La lucha de un futuro humano frente a las nuevas fronteras del poder*, Paidós, Barcelona, 2020, p. 439.

Esta propuesta legislativa europea en materia de IA es considerada un hito normativo a nivel comunitario en la medida en que tiene por objeto estructurar un marco regulatorio de la IA *(i) transnacional*, ya que supone la primera regulación jurídica transversal que es directamente aplicable en todos los Estados miembros de la Unión Europea sin necesidad de elaborar para ello las subsiguientes normas de transposición nacional; *(ii) con eficacia universal*, ya que se trata de un sistema regulatorio universal que extiende su ámbito de aplicación a todos los sistemas de IA que funcionan como componentes de productos o, que siendo productos en sí mismos, se pretenden comercializar en el mercado de la Unión European y *(iii)* orientado al control y la gestión del *riesgo*.

En efecto, el Reglamento europeo de la IA incorpora un sistema sustanciado en la *gestión del riesgo*, en base al cual se establecen diferentes obligaciones de información para los proveedores de sistemas en función del nivel de riesgo asociado al sistema de IA respecto a las garantías de los derechos fundamentales de los usuarios; podemos distinguir así entre: *(iii.1)* sistemas de IA de riesgo limitado, *(iii.2)* sistemas de IA de alto riesgo y *(iii.3)* sistemas de IA de riesgo inaceptable. *(iii.1)* El Reglamento establece en primer lugar una primera obligación de *información mínima* para aquellos sistemas de IA considerados de "riesgo bajo" o "riesgo limitado". En concreto, estos sistemas de IA deben cumplir unos requisitos mínimos de transparencia que permitan a los usuarios, tras interactuar con las aplicaciones, tomar decisiones con conocimiento de causa y bajo su consentimiento. Por tanto, estaríamos ante un sistema de IA de "riesgo limitado" cuando los usuarios son conscientes de que los contenidos de imagen, audio o vídeo que se les ofrecen han sido generados por una aplicación de IA. A este respecto, en la propuesta de Reglamento se incluyó una mención especial para las aplicaciones de IA generativa, como ChatGPT, para las que se establece que deberán cumplir con requisitos adicionales de transparencia de cara a poder ser catalogadas dentro de las aplicaciones de "riesgo limitado". En concreto, el Reglamento impone como requisitos específicos a los sistemas de IA generativa: (a) que revelen siempre al usuario que el contenido ha sido generado

por una IA, (b) que se obliguen a publicar resúmenes periódicos de los datos protegidos por derechos de autor empleados para el entrenamiento y (c) que prohíba la difusión de contenidos ilegales. *(iii.2)* Aquellos sistemas de IA que pudieran afectar negativamente a la seguridad o a la garantía y salvaguarda debidas de los derechos fundamentales son considerados en el Reglamento como sistemas de "alto riesgo". La Propuesta de Reglamento europeo en IA distingue a ese propósito entre dos categorías de sistema de "alto riesgo". *(iii.2.1)* En primer lugar, los sistemas de IA que se utilicen en productos sujetos a la legislación de la UE sobre seguridad de los productos de consumo (por ejemplo juguetes, aviación, automóviles, dispositivos médicos o ascensores). *(iii.2.2)* En segundo lugar, los sistemas de IA que permitan (a) la identificación biométrica y categorización de personas físicas, (b) la gestión y explotación de infraestructuras críticas, (c) la educación y formación profesional, (d) el empleo, gestión de trabajadores y acceso al autoempleo, (e) el acceso y disfrute de servicios privados esenciales y servicios y prestaciones públicas, (f) la gestión de la migración, el asilo y el control de fronteras y (h) la asistencia en la interpretación jurídica y la aplicación de la ley. Todos estos sistemas deberán ser evaluados antes de su comercialización y a lo largo de su ciclo de vida. *(ii.3)* Por último, aquellos sistemas de IA que supongan una amenaza directa para las personas y para la garantía de sus derechos fundamentales son considerados en el Reglamento como sistemas de "riesgo inaceptable" y serán prohibidos. Dicha prohibición se extiende a tres modalidades esenciales de sistemas de IA: *(ii.3.1)* Sistemas de manipulación cognitiva del comportamiento de personas o grupos vulnerables, como la infancia y adolescencia [por ejemplo, es el caso de posibles juguetes animados que pudieran fomentar comportamientos peligrosos en los niños o sistemas de IA que pudieran inducir a comportamientos suicidas en los adolescentes]. *(ii.3.2)* Sistemas que emplean algoritmos para la generación de sesgos identitarios que permitan clasificar a las personas en función de su estatus socioeconómico o características personales – tales como la raza, el género, la nacionalidad, la orientación sexual, la religión, etc. *(ii.3.3)* Por último estarían

aquellos sistemas de identificación biométrica, tanto en tiempo real como a distancia, que emplean el reconocimiento facial.

A pesar del refrendo sustancial que la propuesta de Reglamento en materia de IA obtuvo desde un primer momento en la Eurocámara, una de las cuestiones que mayor controversia suscitaron durante la fase de su tramitación legislativa fue precisamente la de establecer un rango de prohibición adecuado para aquellos sistemas de IA que constituyen un "riesgo inaceptable" para la garantía de los derechos fundamentales de las personas y para la salvaguarda de valores y principios éticos. En este sentido, las arduas discusiones parlamentarias acontecidas entre los Grupos de expertos durante el proceso de tramitación legislativa del RAI dieron como resultado la ampliación del listado de sistemas de IA que deben ser considerados prohibidos a cinco modalidades: *(a)* los sistemas de identificación biométrica remota "en tiempo real", cuando se realiza en espacios de acceso público que permitirían una vigilancia masiva; *(b)* los sistemas de identificación biométrica remota "en diferido", con la única excepción de que el empleo de dichos sistemas fueran realizados por las fuerzas y cuerpos de seguridad del Estado para la persecución de delitos graves y mediante autorización judicial previa; *(c)* los sistemas de previsión de riesgo de comisión de ilícitos penales o administrativos; *(d)* los sistemas de IA predictivos que permiten inferir las emociones de una persona física en los ámbitos de la aplicación del Derecho y la gestión de fronteras, en lugares de trabajo y en centros educativos y *(e)* los sistemas que empleen técnicas subliminales más allá de la conciencia de una persona con el fin de distorsionar materialmente el comportamiento de la misma.

De particular relevancia para el debate que nos ocupa sobre *vigilancia, control y responsabilidad digital* es el artículo 5 del Reglamento europeo, el cual enumera una serie de prácticas de IA consideradas de "riesgo inaceptable" y, por tanto, prohibidas en el mercado interior de la UE; entre ellas se incluyen: *(i)* los sistemas que emplean técnicas subliminales más allá de la conciencia de la persona, o técnicas manipuladoras o engañosas, con el objetivo o el efecto de distorsionar materialmente el comportamiento e impedir una decisión informada, cuando ello

pueda causar un daño significativo; *(ii)* los sistemas que explotan vulnerabilidades de personas o grupos por razón de su edad, discapacidad o situación socioeconómica, con riesgo de daño relevante; *(iii)* determinados sistemas de "social scoring", basados en un tratamiento masivo de datos para evaluar la fiabilidad o el mérito global de las personas; *(iv)* la creación o ampliación de bases de datos de reconocimiento facial mediante "scraping" indiscriminado de imágenes de internet o de circuitos de videovigilancia; *(v)* los sistemas que infieren emociones en contextos laborales y educativos, salvo por razones médicas o de seguridad muy delimitadas; y *(vi)* ciertos sistemas de evaluación de riesgo de comisión de delitos basados exclusivamente en perfiles o rasgos de personalidad.

Este catálogo de prácticas, en vigor desde febrero de 2025 para las prácticas prohibidas, marca un hito normativo sin precedentes al reconocer expresamente la manipulación cognitiva y la explotación de vulnerabilidades psíquicas o psicológicas como formas de *riesgo inaceptable* para la autonomía y la integridad mental de los usuarios de sistemas de IA; más concretamente, el articulo 5.1a) del Reglamento recoge expresamente en su última redacción la prohibición de "[…] la comercialización, puesta en servicio o uso de un sistema de IA que emplee técnicas subliminales más allá de la conciencia de una persona o técnicas deliberadamente manipuladoras o engañosas, con el objetivo o el efecto de distorsionar materialmente el comportamiento de una persona o de un grupo de personas, al afectar considerablemente la capacidad de la persona para tomar una decisión informada, causando así que la persona tome una decisión que de otro modo no habría tomado de una manera que cause o sea probable que cause daño significativo a esa persona, a otra persona o a un grupo de personas. La prohibición de un sistema de IA que emplea técnicas subliminales mencionada en el primer párrafo no se aplicará a los sistemas de IA destinados a ser utilizados con fines terapéuticos aprobados, siempre y cuando se obtenga un consentimiento informado específico de las personas expuestas a ellos o, en su caso, de su tutor legal".

La trascendencia ética de la cláusula normativa introducida por el artículo 5.1 del RAI es sustancial, ya que cualquier ten-

tativa de intervención en los procesos mentales o inconscientes humanos mediante técnicas subliminales que fueran más allá de la conciencia de una persona, o cualesquiera técnicas deliberadamente manipuladoras o engañosas que se empleen en dispositivos de IA con el propósito de influenciar en nuestras decisiones como usuarios o consumidores sobre qué comprar, qué consumir, qué despreciar o qué apreciar, debe ser prohibida y declarada nula de pleno derecho. Por tanto, pese a las numerosas deficiencias en que incurre el articulado del RAI, se vislumbra un hilo de esperanza al haber introducido en su clausulado una nueva garantía de *supervisión humana* que emplaza a aquellas personas físicas que ejerzan la gestión de los sistemas de IA a que sean conscientes del riesgo de sesgo, de automatización o confirmación que este tipo de aplicaciones supone. En este sentido, la Eurocámara exhorta a los gestores de los sistemas de IA a que cumplan con la obligación legal de proporcionar especificaciones para los datos de entrada o cualquier otra información pertinente en cuanto a los conjuntos de datos utilizados en los sistemas de IA, incluida su limitación y supuestos, teniendo en cuenta la finalidad prevista y el mal uso previsible y razonablemente previsible del sistema. En esta misma línea de actuación, las recientes orientaciones de la Comisión Europea sobre las técnicas prohibidas en materia de IA han precisado el alcance de la gestión del riesgo asociado a este tipo de sistemas, detallando que la prohibición de técnicas subliminales, manipuladoras o engañosas se aplica a sistemas que, por su propio diseño, socavan la capacidad de las personas para *tomar decisiones libres e informadas*, con independencia de que exista una intención explícita de dañar; se pueden incluir aquí, como ejemplos paradigmáticos, el uso de mensajes subvisuales o subaudibles destinados a inducir comportamientos sin percepción consciente, la explotación sistemática de sesgos cognitivos en contextos de especial vulnerabilidad o la utilización de agentes conversacionales que simulan identidades personales para influir en decisiones con riesgo de daño significativo sobre los derechos subjetivos de los usuarios de sistemas de IA. Estas recomendaciones proporcionadas por el legislador europeo también aclaran que la prohibición no se extiende a usos estrictamente investigadores que

empleen tales técnicas en entornos controlados y con garantías adecuadas, aunque se subraya la necesidad de salvaguardias reforzadas cuando el objeto de estudio son respuestas cognitivas y conductuales a estímulos generados por IA. En paralelo, el Reglamento europeo en materia de IA también ha consolidado el enfoque de alto riesgo para aquellos sistemas utilizados en ámbitos sensibles como pudiera ser la identificación biométrica, la gestión de infraestructuras críticas, la educación y formación, el empleo, el acceso a servicios esenciales, la gestión migratoria, imponiendo obligaciones estrictas en materia de gobernanza de datos, transparencia, documentación técnica, robustez, ciberseguridad y supervisión humana significativa.

El pasado 4 de febrero de 2025 la Comisión Europea publicó las llamadas *Directrices de la Comisión sobre las prácticas de Inteligencia artificial prohibidas que se establecen en el Reglamento (UE) 2024/1689 (Reglamento de Inteligencia Artificial)* diseñadas para garantizar la aplicación coherente, efectiva y uniforme del Reglamento en toda la Unión Europea. Si bien nos ofrecen información valiosa sobre la interpretación de las prohibiciones por parte de la Comisión, a través sobre todo de ejemplos prácticos para ayudar a las partes interesadas a comprender y cumplir los requisitos del Reglamento en materia de IA, no son vinculantes; no obstante, dichas directrices han reforzado de forma explícita la tutela del inconsciente al precisar que la proscripción de técnicas "subliminales, manipuladoras o engañosas" debe extenderse a todos aquellos sistemas de IA que, por diseño, socavan de manera relevante la capacidad de las personas para formarse un juicio autónomo, incluso cuando la influencia se produce por debajo del umbral de la conciencia. De este modo, dichas Directrices no sólo clarifican el alcance del artículo 5 del RAI, sino que lo orientan hacia el reconocimiento de un *interés jurídico* específico por preservar la integridad de los procesos deliberativos humanos y de la esfera preconsciente frente a formas avanzadas de manipulación algorítmica.

En este contexto regulatorio, la protección del inconsciente humano se perfila como un nuevo horizonte normativo y ético que trasciende la tradicional tutela de la privacidad informativa, abarcando tanto la defensa de la libertad cognitiva como la con-

tinuidad psicológica frente a interferencias tecnológicas no consentidas o desproporcionadas. En España, autores como Ignasi Beltrán de Heredia[5] han planteado a este propósito que el escenario descrito por el artículo 5.1.a) del RAI obliga a revisar los conceptos de "acto propio y voluntario" y, por derivación, los fundamentos de la responsabilidad social cuando las decisiones están mediatizadas por sistemas capaces de operar por debajo del umbral de la conciencia. Por tanto, si se acepta que existen herramientas tecnológicas con capacidad para modular la conducta humana más allá de la percepción consciente, no basta con prohibir las formas más groseras de manipulación lucrativa, si no que debemos avanzar hacia un marco integral de protección de la condición humana en el seno de los ecosistemas de IA, que incluya garantías sobre el diseño, la gobernanza y la supervisión democrática de estos sistemas.

Por consiguiente, la implantación masiva de sistemas de IA nos sitúa ante una doble encrucijada: *(i)* por un lado, la posible redefinición del paradigma *antropocéntrico* clásico y *(ii)* por otro, la de la *responsabilidad social frente a los efectos de la individualización y del capitalismo de la vigilancia*. Cabe preguntarse, en clave baumaniana y a la luz de las recientes aportaciones de Zuboff, hasta qué punto la IA expone la conciencia colectiva de la responsabilidad al coste extremo de una *individualización* que descarga en los sujetos aislados los efectos de un orden socio-técnico profundamente asimétrico. Debemos pues considerar si basta con apelar al principio de responsabilidad colectiva para fundar un marco ético capaz de contrarrestar la ilusión de autodeterminación que promueven las narrativas tecnológicas dominantes, o si, por el contrario, es preciso re-acondicionar una ética de la interdependencia que reconozca la vulnerabilidad humana como punto de partida para una verdadera responsabilidad digital y una gobernanta ética de la IA.

A este respecto, el reconocimiento del papel clave que juega la *intersubjetividad humana* ante las aplicaciones de IA – sobre todo, tratándose de IA generativa – requiere cuatro líneas de acción prioritarias. *(i)* En primer lugar, se debe apostar por

[5] I. Beltrán de Heredia, *Inteligencia artificial y neuroderechos: La protección del yo inconsciente de la persona*, Aranzadi, Madrid, 2023.

el desarrollo de un marco legislativo internacional consensuado que subordine el diseño, producción y desarrollo de la IA a la *dignificación de la condición humana*. *(ii)* En segundo lugar, debemos articular una vía prioritaria de protección de la condición humana mediante la penalización de cualquier tentativa de *instrumentalización* a través de la irrupción en los procesos mentales profundos de técnicas subliminales que van más allá de la conciencia de una persona, o cualesquiera técnicas deliberadamente manipuladoras o engañosas que se empleen por la implementación de aplicaciones de IA, con el objeto de transformar la conciencia ética y crítica de las personas para ser y hacer libremente. *(iii)* En tercer lugar, se debe apostar por el reconocimiento con carácter universal de un *derecho humano a la preservación de la mente inconsciente* como base irrenunciable de la individualidad humana y fuente de validez del consentimiento libre exigible en cualquier acto jurídico. *(iv)* Por último, dado que es altamente probable que la gestión de los sistemas de IA sean monopolizados en los próximos años por élites políticas y económicas y que, en consecuencia, el acceso a estas herramientas tecnológicas acrecienten las desigualdades y los mecanismos de exclusión social, se deben adoptar todas la medidas normativas necesarias para perseguir y castigar el uso de la IA como dispositivo de control y manipulación social.

En este contexto, el debate doctrinal sobre neuroderechos ofrece un primer paso, aunque todavía implícito, para reinterpretar el Reglamento europeo en materia de IA hacia la protección normativa del subconsciente humano. La prohibición de técnicas subliminales y de la explotación de vulnerabilidades puede entenderse como un reconocimiento indirecto de la libertad cognitiva y de la privacidad mental, en la medida en que impide el despliegue de sistemas diseñados para intervenir en los procesos de formación de la voluntad sin conocimiento ni control por parte del sujeto.

Este giro se alinea con las tesis que, en el ámbito de la doctrina jurídica española, sostienen autores como Beltrán de Heredia o Göran Rollnert, para quienes la emergencia de sistemas capaces de modular la conducta por debajo del umbral consciente obliga a revisar los presupuestos clásicos del "acto propio y volunta-

rio" y, en consecuencia, la distribución de responsabilidad en sociedades digitalizadas. Esta perspectiva ha sido desarrollada más concretamente por Göran Rollnert[6], quien propone ampliar el concepto de neuroderechos más allá de la estricta protección de la intimidad frente a neurotecnologías invasivas, integrando también las arquitecturas algorítmicas de persuasión y vigilancia que configuran el entorno de desarrollo de la personalidad humana. En consecuencia, articular normativamente los neuroderechos supone, por un lado, reforzar las garantías materiales que ofrece el Reglamento europeo para los derechos fundamentales (diseño no manipulativo de los sistemas de IA, supervisión humana, evaluación de impacto sobre derechos fundamentales) y, por otro, abrir la puerta a instrumentos jurídicos específicos que reconozcan la mente – incluida su dimensión pre-consciente – como bien jurídico autónomo cuyo menoscabo no puede justificarse por razones meramente económicas o de eficiencia.

Asimismo, la implantación masiva de sistemas de IA en los próximos años nos situará inevitablemente ante la compleja encrucijada del *transhumanismo* y ante el riesgo exponencial de cuestionarnos –una vez más en la historia de la Filosofía moral– sobre la veracidad del paradigma filosófico y ontológico *antropocéntrico*, según el cual, el ser humano es el único ser en el mundo dotado de conciencia y voluntad para actuar con libertad y que, en consecuencia, el ser humano es el único que merece un tratamiento ético.

Cabe que nos formulemos a continuación una serie de interrogantes: *¿Si estamos ante la ruptura del Antropocentrismo?* Y lo que es más relevante, *¿si basta con apelar al principio de responsabilidad colectiva para favorecer un marco ético sustentado en la interdependencia humana frente a la ilusión de la autodeterminación y autosuficiencia que fomenta la IA?*

[6] G. Rollnert Liern, *Los neuroderechos y la libertad de pensamiento*, Dyckinson, Madrid, 2024.

3. Inconsciente, infosfera y dignidad: Hacia una ontología crítica de la subjetividad en la era de la IA

La irrupción de una nueva generación de sistemas de IA capaces de operar en los estratos infraconscientes de la vida psíquica introduce una mutación cualitativa en la comprensión de la subjetividad contemporánea. No nos encontramos simplemente ante un mero avance técnico, sino ante una transformación ontológica en sentido propio que afecta decisivamente al modo en que el ser humano se constituye, se reconoce y se diferencia de su entorno y, por tanto, ante un dilema filosófico más profundo: *la defensa de la interioridad como condición de la dignidad humana*[7].

La tesis que propugna el Reglamento europeo, que centra parte de sus esfuerzos de gestión de los riesgos asociados a los sistemas de IA manipulativos en dotar de especial protección al inconsciente humano a través de su reconocimiento como bien jurídico y ontológico digno de protección, adquiere especial relevancia en un escenario donde la IA no solo interviene en conductas manifiestas, sino en los procesos psíquicos que anteceden a toda decisión, juicio o acto voluntario de las personas. La IA actúa en el nivel pre-consciente del ser humano y transforma la arquitectura interna que sostiene el juicio, la deliberación y el deseo. La erosión del "acto propio y voluntario" – concepto fundacional de la arquitectura ética moderna – revela que lo que está en juego en el caso de la interacción de la IA ya no es únicamente la autonomía humana en su formulación ilustrada, sino la continuidad misma del sujeto como agente moral.

Desde la filosofía de la técnica, Hans Jonas ya nos anticipó en 1979 esta crisis al advertir que la técnica y la tecnología actúan sobre el entramado ontológico que sostiene la autocomprensión humana[8]. Con el advenimiento de la IA generativa y predictiva

[7] M. Ienca, R. Andorno, "Towards new human rights in the age of neuroscience and neurotechnology", in *Life Sciences, Society and Policy*, 13 (2017), n. 5, pp. 1-27.

[8] H. Jonas, *Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation*, Suhrkamp, Frankfurt am Main, 1979. En su traducción al castellano: H. Jonas, *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*, Ediciones Trotta, Madrid, 1995.

en la última década, nos aproximamos peligrosamente al límite que Jonas consideraba crítico: *aquel en el que la capacidad de proyectar, deliberar y gobernarse queda subordinada a dispositivos que modelan silenciosamente el horizonte de acción y los dinamismos afectivos previos a la conciencia humana*. Este desplazamiento coincide con el diagnóstico más contemporáneo de Zuboff sobre el apodado *Capitalismo de la vigilancia*, en el que la subjetividad – incluido su sustrato inconsciente – es convertida en recurso extraíble, anticipable y finalmente moldeable. Ante la IA, el debate deja de ser meramente epistémico o regulatorio y, por este motivo, la intuición de autores como Jonas o Zuboff sólo puede entenderse a través de un estrecho diálogo con la ontología de la técnica de Heidegger acondicionada a los actuales sistemas de IA.

La novedad de la IA radica en que la disponibilidad ya no recae únicamente sobre el mundo exterior, sino sobre la interioridad misma, integrada ahora en un proceso de *instrumentalización* de la conciencia humana sin precedentes y hoy observamos cómo la subjetividad humana es constantemente absorbida por matrices algorítmicas antes incluso de su manifestación consciente. En efecto, la IA aplicada al inconsciente humano representa la consumación del *Gestell heideggeriano,* en la medida en que convierte la interioridad – tradicionalmente considerada un espacio de opacidad y ocultamiento – en un recurso disponible (*Bestand*)[9]. El riesgo de la interacción de la IA en la conciencia humana no es solo la pérdida de autonomía, sino la erosión al espacio ontológico de apertura que fundamenta toda existencia auténtica.

En el contexto actual de la teoría filosófica contemporánea, la *ontología informacional* propuesta por Luciano Floridi constituye un marco decisivo para comprender los desafíos éticos de la era digital y, lo que es más relevante, el efecto de la intervención de la IA en la *psique* humana. La noción clave de Floridi es la de la *infosfera*, entendida como el entorno global de información donde los seres humanos ya no son simples usuarios de herra-

[9] M. Heidegger, *Die Frage nach der Technik,* Neske, 1954. En su traducción al castellano: M. Heidegger, *La pregunta por la técnica* (trad. Eustaquio Barjau), Herder Editorial, Barcelona, 2012.

mientas digitales y sistemas de IA, sino entidades constituidas y definidas por su configuración *informacional*; dicho con otras palabras, a juicio de Floridi, la identidad personal y la agencia moral de las personas dependen cada vez más de la coherencia, estabilidad e integridad del entramado de datos, relaciones y procesos en los que cada individuo se encuentra inmerso[10]. Más concretamente, Floridi sostiene que intervenir en los umbrales pre-conscientes de la experiencia humana – es decir, en los procesos *informacionales* que anteceden a la toma de conciencia – implica modificar la estructura profunda de la subjetividad humana. Cualesquiera intervención algorítmica posibilitadas por tecnologías como la IA puede perturbar la constitución misma del yo, afectando así la autonomía y la capacidad de autodefinición de las personas[11]. A partir de este diagnóstico, Floridi argumenta que una ética de la información contemporánea debe fundamentarse en el principio de justicia ontológica (*ontological fairness*), el cual remite a la necesidad de proteger la integridad informacional de los sujetos contra aquellas manipulaciones o interferencias que atenten contra su *individualización* y agencia moral. Así pues, la ontología informacional de Floridi contribuye a replantear los marcos de los derechos digitales y la ética algorítmica, subrayando que las nuevas formas de intervención informacional no sólo desafían los derechos clásicos, sino que ponen en juego las condiciones mismas de posibilidad de la subjetividad y la agencia moral humana. Más allá de la defensa tradicional de derechos individuales, la protección del inconsciente por la que aboga el Reglamento Europeo – en tanto dimensión *informacional preconsciente* – se convierte, desde la propuesta ontológica de Floridi, en una salvaguarda esencial de la identidad de los seres humanos en entornos cada vez más mediados por sistemas de IA.

Asimismo, la genealogía foucaultiana puede servirnos para complementar la propuesta ontológica de Floridi. Los sistemas

[10] L. Floridi, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*, Oxford University Press, Oxford, 2014.

[11] L. Floridi, *The Ethics of Information*, Oxford University Press, Oxford, 2013. L. Floridi, "Information and the Modern Philosophy", in *The Routledge Handbook of Philosophy of Information*, Routledge, London, 2016, pp. 9-20.

de IA que actúan a nivel subliminal sobre el inconsciente humano constituyen, desde la perspectiva de Foucault[12], una *tecnología de poder* de tercer grado: no disciplinaria ni biopolítica, sino psicopolítica; es decir, una tecnología de vigilancia y control orientada pues a la modulación de los microprocesos que engendran la subjetividad humana. Este régimen de sometimiento algorítmico sitúa al individuo en una paradoja de "autogobierno ilusorio", pues la aparente libertad en la que se desarrolla la interacción humana con el medio digital estaría en realidad pre-formateada por procesos invisibles de *normalización informacional,* si empleamos para ello la terminología de Floridi.

Por su parte, la *esferología* de Sloterdijk[13] nos aporta otro ángulo interesante en este debate. Según este autor, el ser humano es capaz de generar *esferas inmunológic*as con el objeto de preservar su interioridad ante su exposición al mundo. Las intervenciones algorítmicas en el inconsciente humano perforarían dichas *esferas de preservación*, desestabilizando la inmunidad psíquica y produciendo sujetos hiper-expuestos, sin la distancia reflexiva necesaria para sostener una vida mental propia. Asimismo, Stiegler en su obra *The Age of Disruption: Technology and Madness in Computational Capitalism*[14] ha sido capaz de avanzar en el argumento esferológico de Sloterdijk hasta demostrar que la digitalización radical producida por la irrupción de sistemas de IA en todas las esferas de interacción humana externaliza – y posteriormente automatiza – funciones psíquicas pre-conscientes, produciendo una especia de "proletarización" de la psique humana y el abandono progresivo de la individualización baumaniana. En su obra, Bernard Stiegler sostiene que el capitalismo contemporáneo – al que este autor denomina *capitalismo computacional* – ha alcanzado un grado de aceleración técnica que desestructura simultáneamente la economía, la política, los vínculos sociales y, sobre todo, la vida psíquica de los

[12] M. Foucault, *Surveiller et punir: Naissance de la prison*, Gallimard, París, 1975.

[13] P. Sloterdijk, *Sphären I: Blasen*, Suhrkamp Verlag, Frankfurt am Main, 1998. P. Sloterdijk, *Sphären III: Schäume*, Suhrkamp Verlag, Frankfurt am Main, 2004.

[14] B. Stiegler, *The Age of Disruption: Technology and Madness in Computational Capitalism*, Polity Press, Cambridge, 2019.

seres humanos. La llamada *disrupción antropológica de la IA,* celebrada por la ideología tecnocientífica como motor de innovación, constituye por el contrario para Stiegler un proceso de desintegración sistemática de la condición humana; a juicio de este autor, el dinamismo tecnológico que nos ofrece la IA destruye los marcos simbólicos, las temporalidades comunes y las capacidades de atención necesarias para el sostenimiento de la *individualización* humana.

A juicio de Bernard Stiegler, la tecnología subordinada a la lógica capitalista de las sociedades contemporáneas produce un empobrecimiento generalizado de la vida mental de las comunidades humanas, generando una forma estructural de *locura colectiva*[15]. Esta locura no remite a patologías clínicas, sino por el contrario a una pérdida colectiva de orientación, de capacidad crítica y de sentido común en los seres humanos. En su opinión, el llamado *capitalismo computacional* captura y automatiza los procesos cognitivos y afectivos de las personas hasta transformar la experiencia humana en un flujo continuo de estímulos que impiden la formación de razonamientos juiciosos e incluso la construcción simbólica de la experiencia humana. En el núcleo de esta transformación se encuentra lo que Stiegler concibe como "proletarización psíquica": así como el capitalismo industrial expropió saberes técnicos de los trabajadores, el capitalismo digital expropia las facultades mentales de las personas. Esta expropiación se articula con un fenómeno que Stiegler denomina *entropía social y psíquica* para referirse a la aceleración disruptiva a la que nos induce el entorno digital, el cual aumenta el desorden, fragmenta los colectivos y dificulta la transmisión generacional de saberes. Lo que se pierde no es solamente información, sino la capacidad de organizar la experiencia en estructuras estables de sentido. A juicio de Stiegler, cuando las sociedades ya no pueden producir símbolos, relatos o instituciones que ordenen el mundo, emergen comportamientos caóticos, impulsividad y desafección democrática. El resultado es una sociedad que se vuelve inestable, una "sociedad loca", en la que la estructura de la experiencia humana se vuelve disfuncional.

[15] *Ivi*, pp. 23-41.

Asimismo, la disrupción antropológica de la IA se convierte para Stiegler en un proceso devastador: destruye la posibilidad de una vida mental plena para las comunidades humanas y, con ello, la base misma de la civilización. Más concretamente, la disrupción descrita por Stiegler se sitúa en continuidad con el análisis heideggeriano del *Gestell* y con la advertencia foucaultiana sobre la producción técnica de subjetividad, pero incide particularmente en un punto decisivo: *la tecnología digital no solo modifica el mundo exterior, si no que altera las condiciones internas que permiten al sujeto habitar ese mundo y conferirle sentido*. Su tesis de la proletarización psíquica revela, con especial agudeza, que la IA no opera simplemente sobre nuestras decisiones, sino que va más allá, alterando incluso los procesos afectivos y psíquicos que determinan nuestras propias decisiones. De ahí que la cuestión que debe ocuparnos analizar a continuación no sea únicamente si la IA amenaza el *Antropocentrismo clásico* – lo cual es ya evidente desde el momento en que desplaza al sujeto de su rol tradicional como centro epistémico y operativo – sino si puede llegar a redefinir el modo mismo en que se constituye una vida consciente.

La lectura stiegleriana nos permite comprender con mayor precisión por qué la Ética que hoy se requiere para abordar el fenómeno de la IA no puede limitarse a prolongar los marcos epistemológicos heredados del pasado, sino que debe avanzar hacia una Ética ontológica capaz de intervenir allí donde se juega la continuidad misma de la *psique humana.* Si, como muestra Stiegler, el llamado *capitalismo computacional* opera sobre los procesos psíquicos fundamentales de las personas – atención, memoria, deseo – y los somete a una lógica de automatización que produce *entropía social y pérdida de individualización,* entonces la amenaza que introduce la IA no es únicamente moral o regulatoria, sino ontológica, ya que afecta a la propia condición humana. Esta convergencia refuerza la tesis según la cual la interioridad se ha vuelto un campo de programación automatizada y que, por tanto, defender el inconsciente – o, en clave floridiana, la *integridad informacional del yo* – constituye la última frontera donde aún puede sostenerse la dignidad humana antes de una eventual ruptura definitiva del *Antropocentrismo*.

Esta última reflexión es crucial para entender por qué las éticas tradicionales – basadas en la proximidad, la invariabilidad de la condición humana y la delimitación de la acción moral – resultan insuficientes ante un fenómeno que introduce una discontinuidad en la propia arquitectura de la subjetividad. Allí donde Martin Luther King podía aún apelar a una "red ineludible de reciprocidad" sustentada en la experiencia humana compartida, la era computacional descrita por Stiegler nos introduce ante un nuevo tipo de reciprocidad humana tecnológicamente mediada, en la que las transformaciones de la interioridad tienen efectos agregados que exceden las relaciones interpersonales inmediatas. La alteración algorítmica del inconsciente, lejos de circunscribirse a la esfera privada, reconfigura de forma silenciosa las dinámicas colectivas, la estabilidad democrática y las formas culturales de temporalidad. Así pues, la irrupción de la IA no solo confirma el diagnóstico preconizado de Jonas sobre la necesidad de una *Ética orientada a lo venidero*, sino que lo intensifica: la producción técnica de subjetividad inaugura un régimen en el que los daños potenciales ya no afectan únicamente al mundo natural, sino a la propia estructura ontogenética del ser humano.

La convergencia de todas las tradiciones filosóficas, ontológicas y psicoanalíticas que hemos analizado nos conduce a pensar que la protección del inconsciente representa hoy la frontera decisiva para la defensa de la dignidad humana ante la irrupción de la IA. No se trata solo de un imperativo jurídico, como a primera vista pudiera parecer de la lectura del artículo 5 del Reglamento europeo en materia de IA a través de la penalización de las técnicas de manipulación subliminal, si no que se trata de una auténtica *urgencia ontológica: sin la salvaguarda debida de los estratos profundos de la mente, no hay identidad humana estable, y sin identidad humana no hay agencia moral, ni responsabilidad ni vida.*

Defender el inconsciente – o, en términos floridianos, la *coherencia informacional del yo* – no es una reivindicación ilusoria, ya que constituye el último umbral donde aún puede sustanciarse la dignidad humana y donde podemos evitar el peligro de una ruptura definitiva del Antropocentrismo ante la irrupción de

la IA. Por tanto, la cuestión que nos debe ocupar a continuación no es solo analizar la amenaza que representa la IA para el Antropocentrismo clásico, sino si puede llegar a redefinir el modo mismo en que se constituye una vida consciente.

4. La ruptura del Antropocentrismo ante la irrupción de la IA

Las éticas tradicionales son netamente antropocéntricas – también calificadas como *éticas de la proximidad* – cuyas normas están circunscritas más al ámbito de lo personal. Se trata pues de éticas en la que la realidad de la persona y su condición son consideradas como constantes en su esencia y para las que el alcance de las prescripciones éticas se sustancia tradicionalmente en relaciones de intersubjetividad humana y de conexión con el prójimo. Como afirmara Martin Luther King en la carta que escribiera el 16 de abril de 1963 durante su estancia en la cárcel de Birmingham: "Estamos atrapados en una *red ineludible de reciprocidad*, ligados en el tejido único del destino. Cuando algo afecta a una persona de forma directa, afecta indirectamente a todas"[16].

Todas las éticas tradicionales dan cuenta de tres premisas básicas: *(i)* la condición humana permanece *inmutable* para siempre; con base en el presupuesto anterior, *(ii*) se puede determinar con certeza el bien humano y, por último, *(iii)* el alcance de la acción humana y de su consecuente responsabilidad es susceptible de ser delimitado.

Sin embargo, con la irrupción de la tecnología el esquema clásico de las "éticas de proximidad" se transforma completamente, ya que las consecuencias que se derivan del empleo de técnicas disruptivas superan con creces a las tres premisas básicas de las éticas tradicionales, obligándonos a cuestionarnos de nuevo acerca del "principio de responsabilidad" que tratara Hans Jonas en su *Ensayo de una Ética para la civilización tecnológica*[17]. Como apunta este autor, la intervención tecnológica

[16] Aa.Vv., *Informe 2013: El estado de los derechos humanos en el mundo*, Editorial Amnistía, Madrid, 2013, p. 11.

[17] H. Jonas, *El principio de responsabilidad: Ensayo de una Ética para la civilización tecnológica*, Herder Editorial, Barcelona, 1995.

cambió drásticamente el componente de ineludible reciprocidad de la "ética de la proximidad" al poner la naturaleza al servicio del hombre, quien la alterará radicalmente en tan sólo cuatro décadas con efectos nocivos y de degradación inimaginables. Así pues, el nuevo poder de la acción humana sobre el mundo natural supone, a juicio de Jonas, la transformación de la propia naturaleza de la ética. A partir del momento en que el ser humano tiene el poder material de destruir la humanidad adquiere una nueva responsabilidad de cara a las generaciones futuras*: la responsabilidad por lo venidero*. Siendo que, en el ámbito de la ciencia y tecnología, la *responsabilidad hacia lo venidero* es siempre *intergeneracional* y *colectiva*.

Son múltiples los sentidos tradicionales asociados al término de "responsabilidad": *(i)* sea responsabilidad como imputación causal de los actos cometidos – éste es el sentido normalmente asumible en teoría penal al poder coercitivo de la sanción, a la responsabilidad legal y, en el plano de la axiología, a la responsabilidad moral; *(ii)* sea responsabilidad *de lo debido* orientada hacia el futuro – este segundo sentido es el normalmente asociado a la ética; *(iii)* sea responsabilidad por los actos cometidos – en el sentido de lo que se hace (en tiempo presente) y de lo que se ha hecho (en tiempo pasado); *(iv)* sea responsabilidad de padre-hijo – en el sentido de obligación por la responsabilidad que se atribuye a la potestad. Todos estos sentidos tradicionales de "responsabilidad" se encuentran asociados a éticas deontológicas, es decir, *éticas del deber*.

Por su parte, la constatación creciente en la postmodernidad de la vulnerabilidad de la naturaleza ante el yugo de la intervención tecnológica del hombre revela uno de los más importantes desafíos a los que se enfrenta el pensamiento ético con respecto a la condición humana: el ser humano pasa del plano de la "ineludible reciprocidad" entre prójimos – *ama a tu prójimo como a ti mismo* – a una *responsabilidad para con la naturaleza* de marcado carácter teleológico-aristotélico y opuesta, por tanto, a las éticas fundadas en la convicción.

El imperativo ejemplar kantiano – *Actúa de tal modo que el principio de tu acción se transforme en una ley universal* – será adaptado por Jonas a través de nuevas formulaciones de un im-

perativo colectivo, intergeneracional y de acción en la esfera pública: "Actúa de tal modo que los efectos de tu acción sean compatibles con la permanencia de una vida humana auténtica"; o en su formulación negativa, "No pongas en peligro la continuidad indefinida de la humanidad en la Tierra"; o, lo que es su formulación más programática, "Incluye en tu elección presente, como objeto también de tu querer, la futura integridad del hombre"[18].

Por todo ello, ante el poder extraordinario de transformación de los paradigmas antropológicos que en las últimas décadas nos ha aportado la biotecnología, la neurociencia y las tecnologías disruptivas como la IA, es imperativo formular los presupuestos de una *responsabilidad por lo venidero* que repose – como preconizara el filósofo español Fernando Savater en su ensayo *Ética para Amador*[19] – no sólo en los imperativos categóricos de la biología humana aislada, sino en atributos diferentes a la inteligencia, tales como la voluntad, la solidaridad o capacidad de cooperar, la capacidad de sacrificio y, lo que es más importante, las emociones.

Si emulamos un recorrido por la historia de la ética como el que magistralmente elabora Jonas advertiremos, junto al filósofo alemán, que son muchos los testimonios en filosofía moral que evidencian la necesidad de coaligar la razón con la emoción para que el bien objetivo adquiriera poder sobre la voluntad. En este sentido, existe una consolidada tradición de investigación en antropología moderna y filosofía moral que apuesta por la determinación del elemento emocional de la ética aplicada a las tecnociencias; destacan en este ámbito las aportaciones de Evandro Agazzi, Jürgen Habermas, Paul Ricoeur, Gilbert Hottois o Carl Mitcham, entre otros. En este sentido, la mayoría de las corrientes filosóficas surgidas desde mediados del siglo XIX favorables a la emoción, cuestionan el binomio positivista *razón vs. emoción*, según el cual las emociones deben estar fuera del campo de la razón. Nietzsche es seguramente el exponente más claro del abandono de este modelo racionalista y de la disolución de la dicotomía *razón vs. emoción* al considerar que la

[18] *Ivi*, p. 188.

[19] F. Savater, *Ética a Amador: Una invitación a vivir sin odio ni miedo*, Ariel, Barcelona, 2011.

emoción anticipa a la razón y que las personas deciden primero por emociones. En este mismo sentido se pronuncia Antonio Damasio, cuando en su obra *El error de Descartes*[20] considera que "la cognición y las emociones no solo están estrechamente entrelazadas, sino que además, la emoción es el primer mecanismo para la racionalidad". A juicio de este autor, si se integran adecuadamente cognición – en el sentido de racionalidad – y emoción, los seres humanos estarían adecuadamente dirigidos a la toma de decisiones racionales.

Asimismo, cada día son más numerosas las voces que muestran la necesidad de abordar la ética aplicada a las tecnociencias y la IA desde la hermenéutica crítica que promociona la *interdependencia* a través de la transversalidad del enfoque emocional en el conocimiento humano. En este sentido, compartimos la propuesta del filósofo valenciano Jesús Conill Sancho de abandonar la reducción cientificista del positivismo naturalista y el factualismo fenomenológico heideggeriano, para abrazar la vía de la hermenéutica-critica – bien conocida en autores como Habermas, Apel o la propia filósofa española Adela Cortina – en un intento por *re-interpetar la realidad de la persona humana desde su fragilidad y vulnerabilidad con relación a la tecnología*[21].

5. Emocionalidad, Vulnerabilidad y Virtud frente a la Tecnología

La llamada *emocionalidad*, mundo emocional o dimensión emocional de la persona es aquella dimensión en la que tiene cabida nuestro mundo interno de sentimientos, lazos afectivos y otras pulsiones emotivas que facilitan nuestra interacción social en los colectivos humanos. A través de las emociones nos vinculamos con el mundo y por tanto, las emociones resultan indispensables no sólo para nuestro bienestar como individuos, sino para poder garantizar una interactuación social adecuada en los colectivos sociales.

[20] A. Damasio, *El error de Descartes*, Booket, Barcelona, 2001, p. 57.

[21] J. Conill, *Intimidad personal y persona humana. De Nietzsche a Ortega y Zubiri*, Taurus, Madrid, p. 216.

En este sentido, la rama conocida como *Sociología de la emoción* distingue entre dos tipos posibles de emociones. *(i)* En primer lugar, las llamadas *emociones primarias*, que se consideran respuestas fisiológicas, biológicas y neurológicas innatas a los seres humanos e inmediatas ante determinados estímulos. Suelen tratarse de impulsos o reacciones de corta duración a nuestro entorno más inmediato – tales como la ira, el miedo o la alegría. Por consiguiente, el miedo ante un peligro a nuestra integridad física o la ira inmediata ante la visión de un acto de injusticia que nos impulsa a gritar e insultar a alguien, serían ejemplos de esta primer tipología de emociones primarias. *(ii)* En segundo lugar, las llamadas *emociones secundarias* adquieren una naturaleza siempre *relacional.* Se trataría pues de aquellas emociones construidas desde el colectivo social y que, por tanto, son siempre portadoras de significados dependientes de consideraciones sociales y culturales que definen los momentos y situaciones sociales que los seres humanos vivimos. Por ejemplo, las situaciones sociales en las que el sujeto se siente con adecuado nivel de poder y/o estatus dan lugar siempre a emociones positivas sobre el propio sujeto, emociones que le proporcionan seguridad, satisfacción y respeto. Por el contrario, las situaciones sociales en las que el sujeto se siente con escaso o nulo poder y/o estatus dan lugar siempre a emociones negativas sobre el propio sujeto, emociones que le proporcionarán inseguridad, frustración, miedo e ira. La emoción, por lo tanto, puede llegar a convertirse en un dispositivo político de poder y en un efectivo mecanismo de control social. Quien mayor poder tiene y quien menos poder goza, viven en mundos diferentes y antagónicos, no solo físicos sino sociales y emocionales.

En su libro *Emociones Políticas ¿Por qué el amor es importante para la justicia?*[22], la filósofa norteamericana Martha Nussbaum despliega una detallada argumentación sobre el rol y la repercusión de este segundo tipo de emociones para la vida pública de las democracias actuales. En primer lugar, Nussbaum trata de identificar tanto aquellas *emociones secundarias positivas* que ayudan a consolidar la estabilidad de las democracias

[22] M.C. Nussbaum, *Emociones políticas. ¿Por qué el amor es importante para la justicia?*, Paidós, Barcelona, 2014.

contemporáneas (por ejemplo, la compasión) para, a continuación, contrastarlas con aquellas *emociones secundarias negativas* que conspiran contra la estabilidad de la vida pública (por ejemplo, el asco, el miedo y la envidia). A juicio de esta autora, las democracias actuales deben favorecer y promover el cultivo de aquellas emociones secundarias que conducen a la benevolencia y el florecimiento virtuoso – *eudaimonía* – de todos los seres humanos en la esfera pública. En segundo lugar, Nussbaum aspira a proponer estrategias de participación ciudadana en el cultivo de la virtud que permitan fomentar las emociones secundarias positivas en detrimento de las segundas (las llamadas negativas).

Para Martha Nussbaum el *cultivo de las emociones* es condición necesaria para alcanzar la calidad democrática que se le predispone a las sociedades contemporáneas occidentales, cuya estabilidad no sólo se hace depender de su adecuación a unos principios de justicia, sino del desarrollo de "virtudes cívicas" que reposan en emociones y sentimientos. Tanto es así que la estabilidad de las democracias actuales dependería, a juicio de Nussbaum, no sólo mecanismos procedimentales como los que aportan los sistemas normativos – equilibrando los intereses de cada uno y reaccionando en caso de conflicto entre los mismos –, sino a través también de la promoción de lo que Will Kymlicka[23] define como un cierto grado de "emocionalidad" en la esfera pública correspondiente a sentimientos ciudadanos y al espíritu público.

Por consiguiente, la llamada *ética de la virtud* conforma un contexto válido respecto a otras tradiciones filosóficas – centradas en el consecuencialismo de las acciones – para reflexionar sobre el significado ético y universal de aquellas virtudes cívicas – tales como la bondad, la empatía y la compasión – que contribuyen significativamente al bienestar de las comunidades. La *ética de la virtud* trasciende pues las muestras esporádicas de benevolencia en la comunidad, supera las afiliaciones personales y los gestos momentáneos para construir así un compromiso cívico más profundo con la empatía, la comprensión y la buena

[23] W. Kymlicka, *Ciudadanía multicultural. Una teoría liberal de los derechos de las minorías*, Paidós, Barcelona, 1996.

voluntad activa hacia los demás. Al adoptar esta perspectiva, las personas activan la *interdependencia humana* y fomentan la empatía, la compasión y la comprensión hacia "el otro".

A este respecto, Martha Nussbaum nos ofrece un estudio transversal de lo que califica como "filosofía de la bondad" en el que cobra un papel esencial la *compasión como virtud en la promoción del florecimiento humano y la vida ética*. La exploración de la bondad por parte de Nussbaum subraya su valor inherente en la toma de decisiones éticas y en el florecimiento moral, ya que contribuye al desarrollo de un *yo virtuoso y compasivo*.

En su obra monumental *Paisajes del pensamiento: La inteligencia de las emociones*[24], Nussbaum analiza la compasión como virtud como primera etapa fundacional hacia la bondad. La compasión representa, a su juicio, una virtud esencial que contribuye al bienestar general de las personas y las comunidades, que trasciende los meros actos de benevolencia y que abarca una preocupación genuina por el bienestar y la dignidad de los demás. Es más, la compasión desempeña para esta autora un rol esencial en el fomento de la cohesión social y en el cultivo de relaciones interpersonales. Nussbaum se adhiere a la tradición aristotélica de la compasión como aquella "pasión penosa suscitada por el dolor o el sufrimiento de otro"[25] o por "el pesar o dolor físico y los grandes males que provoca la fortuna"[26]. Asimismo, en la formulación de su teoría cognitiva de las emociones, Martha Nussbaum formula un concepto crucial, el concepto de la llamada *compasión política universal,* para referirse a aquellos *impulsos individuales de compasión*, los cuales nos llevan a atender las necesidades materiales de los demás y que sirven de base para la generación de acuerdos institucionales, objetivos políticos y políticas públicas. De este modo, argumenta Nussbaum que la compasión, convenientemente guiada por la racionalidad, puede contribuir a la realización del bienestar colectivo.

[24] M.C. Nussbaum, *Paisajes del pensamiento: La inteligencia de las emociones*, Paidós, Barcelona, 2008.

[25] Aristóteles, *Retórica*, 1385b13 y ss.

[26] M.C. Nussbaum, *La fragilidad del bien: Fortuna y ética en la tragedia y la filosofía griega*, Visor, Madrid, 1995, p. 476.

Por su parte, Alasdair McIntyre, en su reconocida obra *Tras la virtud*[27], coincidiría con la perspectiva de Nussbaum al entender que la bondad es una disposición de carácter que emana de una preocupación genuina por el bienestar de "los otros", reconociéndoles su dignidad y que, en consecuencia, moldea nuestra perspectiva ética y nuestros compromisos morales. Para MacIntyre, como también lo es para Nussbaum, la bondad trasciende las acciones aisladas, los intereses personales y los gestos benévolos. Sin embargo, McIntyre discrepa con Nussbaum cuando afirma que la filosofía de la bondad no debe separarse del discernimiento moral y la razón, ya que la bondad requiere una reflexión meditada y una acción deliberada cuando nos esforzamos por adoptar comportamientos que promuevan el florecimiento virtuoso de los demás.

6. Hacia una IA Compasiva

La Inteligencia artificial puede cambiar radicalmente el mundo en el que vivimos, pero es la ética detrás de ella la que determinará cómo será ese cambio. En este sentido, la transversalidad del enfoque de la *ética de la virtud* nos obliga a entablar un debate en torno a cuáles son las virtudes éticas fundamentales en las sociedades digitales y si la *ética de la virtud* debería vincularse a la *responsabilidad de lo venidero frente al desarrollo tecnológico* que apuntara el filósofo alemán Hans Jonas.

Al embarcarnos en el viaje trascendente que nos aporta el enfoque de la *ética de la virtud*, comprobamos que no se trata simplemente de crear máquinas inteligentes, capaces de sustituir e incluso superar al ser humano en su raciocinio y capacidades cognoscitivas, sino de fomentar una *IA virtuosa* que refleje lo mejor de las aspiraciones morales del ser humano.

Debemos empezar pues reconociendo la responsabilidad moral que tenemos como arquitectos de un marco ético de la IA. Integrando bondad, empatía y compasión en el diseño de los sistemas de IA ayudaremos decisivamente a dar prioridad al bienestar de los usuarios, promover la equidad, la transparencia

[27] A. MacIntyre, *Tras la virtud*, Ed. Crítica Ediciones Grijalbo, Barcelona, 1984.

y la responsabilidad, y garantizar que las tecnologías de IA sirvan a los intereses de los ciudadanos y no así de corporaciones tecnológicas o poderes fácticos. Esto último se debe traducir en diseños de algoritmos de IA que den prioridad a la empatía, al respeto y a la preservación de la dignidad humana frente a la construcción de sesgos discriminatorios.

En efecto, cuando imbuimos a las máquinas de inteligencia y capacidad de decisión, las virtudes que podemos inculcar en ellas se convierten en la verdadera piedra angular del desarrollo ético de la tecnología, sobre todo a la hora de hacer frente a las posibles injusticias sistemáticas que se pudieran derivar del uso malicioso de sistemas de IA y de las injusticias sistémicas que implican los sesgos en los datos y en los algoritmos discriminatorios.

Los seres humanos somos particularmente vulnerables ante la IA por el potencial peligro que entraña para las personas el empleo de dispositivos técnicos que inducen a una violencia psicológica encaminada al control de la voluntad humana – la cual ha sido calificada por los especialistas en neurociencia como "violencia neuronal".

La Propuesta de *Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de Inteligencia artificial* es consciente de que la vulnerabilidad humana es central en cualquier contexto de disrupción como el que aporta la IA. Por este motivo, prohíbe *(i)* en primer lugar, aquellos sistema que desplieguen técnicas subliminales con el objetivo de distorsionar el comportamiento de una persona de manera que pueda causarle daños físicos o psicológicos a ella o a otros y *(ii)* en segundo lugar, aquellos sistemas de IA que exploten vulnerabilidades de un grupo específico de personas por su edad, discapacidad o situación social o económica, de tal forma que se distorsionen el comportamiento de estas personas y probablemente les causen daños a ellas o a terceros.

Por tanto, para hacer realidad una *IA confiable* no solo es suficiente con cumplir con la ley, si no que es necesario adoptar una serie de estándares éticos en su uso que generen en la sociedad civil la confianza debida de que los sistemas de IA no provocarán un daño, no solo doloso, si no involuntario a sus usuarios.

Sin embargo, construir un marco ético universal para los sistemas de IA no es tarea fácil y nos enfrenta al manido embate postmoderno del relativismo ético de que cada comunidad y cada persona pueden tener concepciones distintas de lo que constituye "lo moralmente aceptable". El enfoque colectivo que hemos explorado antes a través de la *responsabilidad por lo venidero* de Jason, nos proporciona una justificación más explícita de la asunción de un paradigma moral universal, al basarlo en experiencias humanas compartidas y en los principios duraderos de las tradiciones filosóficas que han guiado el fundamento ético a lo largo de la historia. Por este motivo, la construcción de un marco ético compartido para la IA, lejos de parecer una utopía, se vislumbra como una necesidad cada vez más acuciante.

La adopción de un paradigma moral y de un marco ético universal para el desarrollo de la IA obedece a dos presupuestos. *(i)* En primer lugar, un marco ético universal nos sirve de parapeto frente los inconvenientes de la reducción cientificista del positivismo naturalista y el factualismo fenomenológico heideggeriano, y nos impulsa en la dirección de una hermenéutica-crítica que promociona la *interdependencia* a través la transversalidad del enfoque emocional en el conocimiento humano. La construcción de un marco ético compartido fomenta pues una comprensión colectiva de *aquellos límites que la IA no debe sobrepasar*. *(ii)* En segundo lugar, la búsqueda de un marco de referencia ético universal en la IA se alinea con el propio *telos* de la tecnología. La palabra "tecnología" procede etimológicamente del griego τέχνη ("téchnē") en el sentido de *arte*, *oficio* o *destreza.* Desde una perspectiva más amplia, la tecnología se identifica con *un proceso o una capacidad de transformar o combinar algo ya existente para construir algo nuevo y permitir, de este modo, mejorar y profundizar la existencia humana*. Esta segunda manera de abordar el sentido de la tecnología tiene su punto de partida en el presupuesto aristotélico de que la manera debida de valorar los bienes, instituciones y prácticas sociales depende de los propósitos o fines de dichas practicas o instituciones. Y, precisamente, la idea de que podemos descubrir las virtudes – tanto éticas como dianoéticas – apropiadas a un proceso, oficio o destreza como la tecnología, cuando tratamos de comprender

el fin, propósito o *telos* de dicho proceso, constituye el núcleo de la teoría aristotélica de la Justicia y el fundamento de la *ética de la virtud*.

Si hacemos depender el desarrollo de una IA *con factor humano* – respetuosa pues con la dignidad humana – al reconocimiento de marcos éticos virtuosos, no solo mitigamos los riesgos de la divergencia anti-cognitivista, sino que garantizamos que las creaciones tecnológicas contribuyan positivamente al bienestar de las personas y de la sociedad en general. Asimismo, para poder encontrar un equilibrio adecuado entre los intereses individuales y el bienestar colectivo, es necesario explorar a fondo las consideraciones que en Bioética se han llevado a cabo entorno a la naturaleza de la virtud y los límites éticos[28].

7. Principio de Beneficencia ante el riesgo de discriminación por sesgos algorítmicos

El llamado *principio de beneficencia* se refiere a la obligación moral de prevenir o aliviar el daño – por tanto, *de hacer el bien* – y el deber de ayudar al prójimo por encima de los intereses particulares; en otras palabras, *obrar en función del mayor beneficio posible, procurando el mayor bienestar general*.

Como enuncia Kottow en su *Introducción a la bioética*[29], los elementos que se incluyen en este principio son todos los que

[28] H. Engelhardt, *Fundamentos de la bioétic*a, ed. 2, Paidós, Barcelona, 1995. Engelhardt formula en su tratado introductorio cuáles son los cuatro principios básicos de la Bioética: *autonomía*, *beneficiencia*, *no-maleficiencia y justicia*. (i) La autonomía es entendida como la capacidad de las personas de deliberar sobre sus finalidades personales y de actuar bajo la dirección de las decisiones que pueda tomar. (ii) La beneficencia significa "hacer el bien", entendida como la obligación moral de actuar en beneficio de los demás. Curar el daño y promover el bien o el bienestar. (iii) La no-maleficencia significa no producir daño y prevenirlo. Incluye no matar, no provocar dolor ni sufrimiento, no producir incapacidades. Es un principio de ámbito público y su incumplimiento está penado por la ley. (iv) Por último la justicia en su dimensión aristotélica, entendida como la equidad en la distribución de cargas y beneficios. El criterio para saber si una actuación es o no ética, desde el punto de vista de la justicia, es valorar si la actuación es equitativa, con un rechazo a cualquier forma de discriminación por cualquier motivo.

[29] M. Kottow, *Introducción a la bioética*, Editorial Universitaria, Santiago de Chile, 1995, p. 72.

implican *una acción de beneficio que haga o fomente el bien, prevenga o contrarreste el mal o daño y, adicionalmente, todos los que implican la omisión o la ausencia de actos que pudiesen ocasionar un daño o perjuicio.*

El llamado "principio de beneficencia", a juicio de Beauchamp y Childress, favorece que los individuos y las instituciones sientan la obligación ética de contribuir activamente al bienestar de la comunidad[30]. Destacan en este contexto, en particular, la promoción de virtudes cívicas como el altruismo, la solidaridad, la compasión y la responsabilidad social en las acciones humanas.

Por su parte, el filósofo Peter Singer, en su obra *Vivir éticamente: Cómo el altruismo eficaz nos hace mejores personas*, revitaliza el razonamiento moral utilitario, según el cual los individuos tienen la obligación moral de evitar el sufrimiento y promover el bienestar en la medida de sus posibilidades. Según la propuesta de Singer, disponemos de *libre albedrío* para elegir el curso de nuestras acciones y en esta autonomía de la voluntad debemos sobrepujar aquellas decisiones que maximizan el bienestar general y que, en contrapartida, reducen el sufrimiento innecesario[31].

No obstante, el *principio de beneficencia* nos enfrenta ante el principal escollo de doctrina moral del Utilitarismo bethamiano cuando formula "promover el bienestar y actuar de forma que maximice la felicidad del mayor número posible de personas". En este sentido se pronuncia el filósofo norteamericano Michael J. Sandel, quien considera que el punto débil más clamoroso del razonamiento moral utilitarista es precisamente su falta de respeto a los derechos individuales. "La lógica utilitaria, si se aplica coherentemente, refrenda maneras de tratar a las personas que violan normas de decencia y respeto que creemos fundamentales"[32].

Esta última aproximación sobre las deficiencias de la lógica utilitaria es todavía más evidente en el contexto de la IA, y nos obliga a centrar nuestros esfuerzos en mecanismos de beneficien-

[30] T. Beauchamp, J. Childress, *Principles of Bioethical Ethics*, ed. 2, Oxford University Press, Oxford, 1994, pp 148-149.

[31] P. Singer, *Vivir éticamente: Cómo el altruismo eficaz nos hace mejores personas*, Paidós, Barcelona, 2017.

[32] M. Sandel, *Justicia ¿Hacemos lo que debemos?*, Ed. Debate, Barcelona, 2011, p. 48.

cia que permitan mitigar daños innecesarios asociados a los sistemas de IA, sobre todo aquellos que pudieran comprometer seriamente el bienestar colectivo. A este respecto, un aspecto crucial con el que deben lidiar los desarrolladores de sistemas de IA para dar cumplimiento a este *principio de beneficiencia* es la cuestión de mitigar el impacto de los sesgos en los algoritmos de IA.

Lamentablemente, los sistemas de IA tampoco son inmunes a los sesgos y prejuicios que existen en la sociedad. Estos prejuicios tienden a impregnar los algoritmos y a propagar la discriminación. En consecuencia, si efectuamos una mala selección de los datos que se emplean para entrenar a un sistema de IA, el resultado puede suponer que la predicción que nos facilita el algoritmo llega a decisiones injustas que inducen a la discriminación o estigmatización de determinados colectivos y/o personas a través de la generación de creencias preconcebidas, predilecciones o prejuicios inconscientes – los llamados "sesgos" – que podemos haber adquirido a lo largo de nuestras vidas fundadas en los estereotipos socioculturales con los que se nos ha educado.

Ejemplos de esto último los encontramos en el recurso a una modalidad de algoritmos que se emplean en aplicaciones de IA para el reconocimiento facial, los cuales muestran un sesgo discriminatorio significativo a la hora de identificar a personas de diferentes orígenes raciales y étnicos. Asimismo, una segunda muestra la encontramos en los llamados algoritmos predictivos de control *ante facto*. *(i)* En el caso de los primeros, algoritmos para el reconocimiento facial, se ha demostrado que este sesgo obedece a la representación desequilibrada que los desarrolladores de sistemas de IA para el reconocimiento facial emplean en los datos de entrenamiento, en los que predominan los rostros masculinos y de piel clara. En consecuencia, estos sistemas presentan elevadas tasas de error en la identificación, al no haber mitigado los sesgos discriminatorios por razones étnicas o raciales. *(ii)* En el caso de los segundos, la función de los algoritmos predictivos *ante facto* es la de determinar el grado de riesgo, por posible reincidencia, que puede tener una persona a lo largo de las diferentes etapas en que se desarrolla un procedimiento judicial o proceso administrativo. Así pues, desde un punto de vista estrictamente normativo, el empleo de algoritmos predictivos

de riesgo implica graves impedimentos para que el legislador, en su tarea de regulación jurídica con ánimo de permanencia futura, pueda reprimir y combatir eficazmente las situaciones vulneradoras de los derechos fundamentales. Asimismo, supone también un problema para los operadores jurídicos, ya que los sistemas de IA predictivos buscan siempre establecer virtualidades de comportamiento humano y eso puede derivar a la postre en una patronización de modelos erróneos que acabe desembocando en sesgos cognitivos y discriminatorios en base al género, la nacionalidad, los orígenes étnicos, la raza o la religión.

Tanto en la administración de justicia como en la actividad policial se han implementado en estos últimos años una serie de *algoritmos predictivos de riesgo* aplicables a aquellas personas que responden a estereotipos delictivos asociados a diferentes grupos raciales u orígenes étnicos que pueden hacer aumentar la idea de culpabilidad o que están vinculados a un mayor nivel de delincuencia, lo que daría lugar a diferencias de trato en investigaciones policiales y sentencias en función del llamado *prejuicio judicial*. Este fue el caso del empleo durante años del llamado modelo COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) que se utilizó en Estados Unidos para evaluar la probabilidad de reincidencia de las personas implicadas en el sistema de justicia penal. El sistema COMPAS reveló una discriminación algorítmica manifiesta hacia varones de origen afroamericano, los cuales representan a la población penitenciaria en Estados Unidos que aglutina penas de reclusión más largas – incluyendo cadena perpetua – respecto a cualquier otra combinación de raza y género. Tras diversas investigaciones se pudo demostrar que el sistema COMPAS se sustanciaba en un marcado sesgo racial, ya que era más probable que se etiquetara erróneamente a varones afroamericanos – dado su perfil en el sistema de un mayor riesgo de reincidencia – en comparación con los varones blancos. El caso COMPAS fue objeto de gran atención por parte de la opinión pública y suscitó preocupación por el posible impacto discriminatorio de los algoritmos de IA cuando se emplean en las decisiones jurisdiccionales en el orden penal.

Un caso que nos es más próximo que el anterior fue el del denominado *Sistema de Indicación de Riesgos* – conocido bajo las

siglas SyRI – que el gobierno neerlandés utilizó para prevenir y combatir el fraude de prestaciones de la Seguridad Social. Dicho sistema permitía que las administraciones públicas neerlandesas recurrieran a la elaboración de informes de riesgo de aquellos demandantes de subsidios por hijos menores de edad a su cargo para evitar, de este modo, que se hiciera un uso ilegal de fondos gubernamentales en el campo de la Seguridad Social.

El Sistema SyRi se había configurado a partir del marco normativo que aportaba una ley nacional – la llamada *Ley sobre la estructura de organización de ejecución del trabajo y los ingresos* –, cuyo artículo 65.2 contenía una extensa lista de categorías de información que podían procesarse en el sistema SyRI; en concreto: género, historial de empleo, impuestos, propiedad de bienes, educación, seguro médico, permisos del gobierno, el nivel de endeudamiento, el historial de prestaciones públicas percibidas, las sanciones administrativas, tales como multas de tráfico, entre otros.

Para calcular posibles irregularidades de evasión y fraude, los algoritmos del sistema SyRI enlazaban todos los datos personales de los solicitantes almacenados por instancias gubernamentales para, a continuación, cotejarlos con un "perfil de riesgo" generado a partir de la información de otros ciudadanos con antecedentes penales. Una vez establecidas las posibles similitudes y/o divergencias, el sistema confeccionaba informes de riesgo sobre una lista de nombres como "potenciales defraudadores" que las autoridades podían conservar hasta dos años.

El sistema SyRi se sustanciaba en *proyectos de vecindad* en los que las agencias gubernamentales identificaban aquellos distritos municipales en los que se quería implementar este sistema de evaluación de riesgos. Esto en la práctica se tradujo la implantación del SyRI exclusivamente en aquellos barrios y distritos municipales económicamente más desfavorecidos, caracterizados por sus altas tasas de población inmigrante. Como consecuencia de ello, las autoridades administrativas holandesas acusaron erróneamente de fraude a centenares de familias perceptoras de la prestación por el mero hecho de tener ascendencia marroquí o turca. Resulta pues evidente a través de este ejemplo que el modelo de riesgo

"ante facto" elaborado por SyRI alicienta una evidente discriminación de la ciudadanía con ascendencia árabe.

Este caso dio lugar a la primera decisión judicial en Europa en la que se puso a examen un sistema algorítmico de evaluación del riesgo, la conocida como *Netherlands Committee of Jurists for Human Rights vs. State of the Netherlands*. En dicha sentencia de 6 de marzo de 2020, el tribunal concluyó que el sistema SyRI no sólo había producido la afectación del derecho humano a la vida privada, sino que vulneraba adicionalmente la exigencia de transparencia del artículo 8 del Convenio Europeo de Derecho Humanos. Asimismo, el alto tribunal también analiza la legitimidad del recurso gubernamental a configurar informes de riesgo de la ciudadanía para calcular la atribución de prestaciones sociales, concluyendo que el sistema SyRi no era "ni transparente ni verificable"; no sólo porque un sistema así podía ser empleado en la elaboración de perfiles de datos de las personas para otros fines – lo que está prohibido por ley –, si no también porque los modelos de riesgo empleados por el Gobierno neerlandés nunca fueron publicados; asimismo, los interesados no fueron notificados con carácter previo sobre dicho particular cuando sus datos fueron introducidos en el sistema SyRi para la confección de su *perfil de riesgo* ante la Administración pública. Por último, por lo que respecta a la prueba de equilibrio, el tribunal llegó a la conclusión de que un informe de riesgo tiene un efecto jurídico no desdeñable sobre el derecho a la privacidad de la persona objeto de escrutinio algorítmico y si bien un informe así no puede utilizarse a efectos sancionatorios, esto no excluye para que una información tan sensible pueda ser empleada en ulteriores procedimientos y comunicaciones de la ciudadanía con la Administración pública. En base a este razonamiento, el tribunal desestimó el "interés declarado del Gobierno neerlandés".

Los dos ejemplos que hemos seleccionado nos sitúan ante la principal amenaza que acompaña a la automatización de los algoritmos en los sistemas predictivos de riesgo que emplea la IA. Así pues, el recurso de los Gobiernos a configurar "informes de riesgo" de sus ciudadanos mediante sistemas de IA predictiva lamina uno de los pilares epistemológicos sobre los que se

fundamenta la propia definición legal de Estado de Derecho: la noción de autonomía y autodeterminación de las personas.

Los defensores de las concepciones contemporáneas de autodeterminación se sirven del ideal kantiano de autonomía moral para luchar contra la visión estereotipada de aquellos que critican fuertemente la concepción libertaria de autonomía personal, tachándola de individualista o incluso egoísta. Los libertarios, en efecto, conciben la autonomía personal a expensas de deseos o de preferencias subjetivas y en detrimento del bien común. Por este motivo, liberales postmodernos como Roberts o Joseph Raz han apostado por una noción de *autonomía socializada* que permite coaligar de forma efectiva el ideal clásico kantiano de autonomía de la voluntad con los retos que nos ofrecen las nuevas aplicaciones de los algoritmos en los sistemas de IA. Si nuestra autonomía y capacidad de actuar y elegir libremente está comprometida mediante el empleo de algoritmos, como el sistema SyRI, ya no actuamos conforme a una máxima que nos damos a nosotros mismos, si no que lo hacemos en cumplimiento de una máxima que la comunidad debe darse a sí misma, por el bien común.

A pesar de las deficiencias apuntadas, hemos de tener en cuenta que parte del poderoso atractivo que la IA tiene hoy en día es el de venderse como un mecanismo de superación de la subjetividad humana, o incluso de erradicación de estereotipos y prejuicios sociales, a través del recurso a la certeza y neutralidad que nos ofrecen los algoritmos.

Sin embargo, ejemplos como el llamado sistema SyRI nos hacen considerar que existe la amenaza potencial de que los sistemas algorítmicos de evaluación de riesgos acaben siendo explotados por instancias gubernamentales con vistas a instaurar un sistema represivo que fuera en detrimento de las libertades públicas y garantías constitucionales de los ciudadanos; o incluso, como nos advierten los profesores Lorenzo Cotino Hueso y Jorge Castellanos Claramunt, que se tratase de un sistema vulnerador de la dignidad y los derechos humanos, sobre todo tratándose de minorías étnicas y población inmigrante[33].

[33] L. Cotino Hueso, J. Castellanos Claramunt (coords.), *Algoritmos abiertos y que no discriminen en el sector público*, Tirant lo Blanch, Valencia, 2023.

Cualquier persona susceptible del "escrutinio algorítmico" podría ser potencialmente perseguida en base a eventuales virtualidades de comportamiento susceptibles de riesgo. No seríamos, así pues, perseguidos por los actos que hayamos realizado, sino por determinadas predicciones de comportamientos *pro-futuro* o *ante-facto* en razón de sesgos identitarios asociados a algoritmos que tiene en cuenta desde nuestro nivel de ingresos y de endeudamiento, pasando por nuestra interacción en redes sociales, hasta nuestro lugar de residencia, nuestra religión o nuestro origen étnico.

La problemática que subyace en el "escrutinio algorítmico" que aplican los sistemas de IA de control de riesgo pone de relieve la necesidad del abordaje ético de los sistemas de IA. También nos cuestionamos sobre si la promoción de virtudes como la bondad, la empatía y la compasión en el diseño de los sistemas de IA nos ayuda verdaderamente a generar algoritmos capaces de mitigar los sesgos históricamente asociados a determinadas minorias por razones étnicas, religiosas y raciales. Esta segunda apreciación nos lleva a considerar el segundo de los principios bioéticos anteriormente enunciado: el principio de *no-maledicencia*.

8. Principio de No-maledicencia: La prevención del daño y la preservación de la dignidad humana ante el riesgo de alexitimia virtual

El segundo principio que debemos considerar en nuestra apuesta por una ética de la virtud aplicada a la IA es el *principio de no-maleficencia*. Este principio se inscribe en la tradición de la máxima clásica *primum non nocere – lo primero es no dañar* – y hace referencia a la obligación de no infringir daño intencionadamente. El *telos* de este segundo principio bioético de *no dañar a otros* – por ejemplo, no robar, no lastimar o no matar –, es claramente diferente de la obligación de beneficencia, la cual busca proteger intereses personales o promover el bienestar colectivo.

Aplicado a los sistemas de IA, el *principio de no-maledicencia* garantizaría que dichos sistemas diesen prioridad, en primer lugar, a la la seguridad de las personas o de los potenciales usuarios y, en segundo lugar, a promocionar la dignidad humana,

minimizando así el riesgo y maximizando la transparencia y la explicabilidad.

Son numerosos los casos en los que los dispositivos de IA han incorporado el *principio de no-maledicencia* para promover la seguridad de los usuarios. Un ejemplo recurrente lo encontramos en los medios de automoción, en particular en la implementación de sistemas de IA en vehículos autónomos con el objetivo de reducir los accidentes de tráfico y mejorar la seguridad vial. Multinacionales como Tesla o Uber han desarrollado en estos últimos años prototipos de coches automatizados que emplean algoritmos de IA para percibir el entorno y anticipar a continuación decisiones de conducción automatizadas, o incluso controlar el propio vehículo durante desplazamientos por carretera. Lo que se persigue es precisamente la superación del error humano para mejorar la seguridad vial ya que, de esta manera, los vehículos autónomos tienen el potencial de reducir significativamente los accidentes causados por errores humanos; ya sea intencionados como no intencionados como el déficit de atención, la fatiga visual o la falta de reflejos de los conductores. Se trataría aquí, por tanto, de un ejemplo de implementación del *principio de no-maledicencia*, ya que el objetivo de los dispositivos de IA es el de minimizar los daños potenciales, maximizando la seguridad de las personas en la carretera.

Al contemplar las posibles aplicaciones del *principio de no-maledicencia* a los sistemas de IA debemos cuestionarnos acerca del papel que juegan virtudes como la bondad, la empatía y la compasión en el aprendizaje automátizado de los sistemas de IA, en el procesamiento del lenguaje natural y en la comunicación afectiva para el diseño de las aplicaciones de IA que sean capaces de percibir, comprender y responder a las emociones humanas. En efecto, en el proceso de *Machine Learning*, los sistemas de IA adquieren la capacidad de aprender de ingentes recopilaciones de datos para hacer predicciones de comportamientos humanos *ante facto*. Por tanto, si entrenamos a los sistemas de IA a recopilar con carácter prioritario datos que ejemplifiquen las virtudes esenciales a la condición humana – tales como la bondad, la empatía, la solidaridad, el coraje, la prudencia y la compasión – podemos así imbuirles la capacidad de reconocer

y responder a situaciones de tal forma que se promuevan el bien común en las comunidades humanas.

El aprendizaje automatizado nos ofrece posibilidades infinitas para instruir a la IA en virtudes como la bondad, la empatía o la compasión. Pero, *¿cuáles son las limitaciones que debemos superar en esta tentativa por implantar virtudes éticas en la IA?*

El afán de la ética de la virtud de incorporar en los sistemas de IA valores esenciales a la condición humana se enfrenta al riesgo ineludible de la *alexitimia*. Estimular el aprendizaje afectivo en los procesos automatizados de la IA, o impulsar computación afectiva, evidentemente hace aumentar el potencial de los sistemas de IA para percibir, comprender y responder a las emociones humanas. No obstante, debemos aceptar dichos avances con recelo ya que, si bien los sistemas de IA puedan simular empatía, bondad y compasión, carecen en realidad de la conexión emocional que se deriva de la experiencia humana; o dicho en términos kantianos, de la *condición de humanidad*.

Cualquier tentativa de construcción de un marco ético de la IA encierra una promesa inconmensurable de "humanizar la tecnología" a través de la promoción de virtudes esenciales a la condición humana. En esta ingente apuesta por una *IA con factor humano*, las virtudes éticas se nos revelan como mecanismos válidos a la hora de afrontar algunos de los riesgos más acuciantes de la implementación de los sistemas de IA en nuestro quehacer cotidiano. Así es, entreverar las tres virtudes éticas de la bondad, empatía y compasión en el desarrollo de la IA allana el camino para el desarrollo de una IA con un profundo sentido de humanidad.

Como hemos tratado de justificar a lo largo de nuestra investigación, *(i)* la bondad aplicada al diseño de los sistemas de IA exige equidad, fomentando así el reconocimiento de un principio de *no-maleficiencia* para evitar los daños potenciales que se pueden derivar del empleo de sesgos algorítmicos. *(ii)* La promoción de la empatía obliga a los desarrolladores de sistemas de IA a dar buen uso del *principio de beneficencia* para, de este modo, compensar los prejuicios ocultos en los algoritmos de prevención de riesgos *ante facto*, así como los efectos discriminatorios en base a identidades étnicas y raciales que derivan

inevitablemente en la segregación de las minorías. Y, por último, *(iii)* la promoción de la compasión en la IA es el motor que impulsa la cada vez más demandada responsabilidad ética de los programadores y desarrolladores de IA ante las secuencias algorítmicas de sesgo y los efectos maliciosos de las últimas utilidades de IA generativa.

9. A modo de conclusión

En la travesía intelectual que supone abordar los desafíos de la Inteligencia artificial desde la perspectiva de la Ética y la Ontología, una evidencia cardinal se impone: nos encontramos ante una transformación epistémica y axiológica cuyo alcance trasciende las fronteras de la mera tecnología y penetra, de manera irreductible, en la estructura misma de la subjetividad y de la agencia humanas.

La irrupción de sistemas de IA capaces de incidir en dimensiones pre-conscientes y de modular la arquitectura *informacional del yo* (Floridi) tensiona los conceptos tradicionales de autonomía, dignidad humana y consentimiento, obligando a repensar sus cimientos ético-jurídicos y exigiendo una re-elaboración profunda de los fundamentos ontológicos hasta ahora vigentes.

Confiamos en que el recorrido realizado en el presente estudio contribuya a ampliar la conciencia de la comunidad académica sobre el hecho de que una reflexión ética y ontológica en torno a la IA no es un complemento accesorio a su desarrollo tecnológico y su implementación técnica, sino, más bien al contrario, una condición epistemológica necesaria para poder comprender la magnitud de la mutación antropológica en curso.

Lo que emerge ante nosotros es un proceso de re-configuración de la subjetividad que exige cuestionarnos los vínculos entre vulnerabilidad, responsabilidad y agencia moral. Desde esta perspectiva, la IA no sólo interpela nuestra capacidad regulatoria, si no nuestra capacidad de reconocer y proteger aquellas funciones humanas – emocionales, deliberativas y corporales – sin las cuales la vida digna deja de ser realizable.

En este horizonte doctrinal, las aportaciones de filósofos de la tecnología como Luciano Floridi y Bernard Stiegler, de especialistas en bioética como Hans Jonas y de reputadas filóso-

fas, sociólogas y pensadoras morales como Martha Nussbaum, Shoshanna Zuboff y Adela Cortina, entre otros de los autores estudiados, adquieren un valor hermenéutico decisivo. Bernard Stiegler, al subrayar la amenaza de la *proletarización psíquica* y la pérdida de individualización provocada por los sistemas de IA, ilumina la urgencia de recuperar espacios de cuidado de la condición humana y del cultivo de la racionalidad y del sentido común; dimensiones éstas últimas indisociables de la protección de la interioridad que aquí hemos defendido. Martha Nussbaum, por su parte, nos recuerda que cualquier intento regulatorio institucionalizado de la IA debe orientarse a expandir las capacidades humanas fundamentales, y no a erosionarlas mediante dinámicas algorítmicas opacas o extractivas. Adela Cortina, desde su propuesta favorable a una *ética de la cordialidad*, enfatiza la necesidad de una "compasión cívica" capaz de sostener la cohesión democrática en los entornos actuales cada vez más hiper-tecnificados y, por ende, más proclives a la *despersonalización*.

En conjunto, todos y cada uno de los múltiples marcos conceptuales abordados en este estudio convergen en la advertencia de que la expansión de la *infosfera* y la consolidación de marcos regulatorios como el Reglamento europeo en materia de IA exigen, antes de nada, la promoción de un principio de responsabilidad atento a la protección de la *integridad informacional del yo,* preconizado por Floridi, y a la salvaguarda de la condición humana.

El verdadero desafío del presente no se limita, por tanto, a la mera elaboración de códigos éticos para el uso de la IA, sino a la articulación de dispositivos de gobernanza, estrategias políticas y sistemas de auditoría y trazabilidad que aseguren la equidad ontológica y la sostenibilidad de los derechos fundamentales en la era digital.

El panorama contemporáneo de regulación de la IA demanda pues un enfoque interdisciplinar capaz de situar la Ética como brújula ineludible en la orientación del desarrollo tecnológico. Los valores que sustentan nuestra tradición humanista y antropocéntrica deben traducirse en procedimientos de control rigurosos sobre el uso de la IA y en una reflexividad crítica que preserve la agencia humana, incluida su dimensión pre-reflexiva, hoy in-

tensamente impactada por la mediación algorítmica y las formas sutiles de manipulación asociadas al uso de sistemas de IA.

En suma, sobre la comunidad académica y científica recae la tarea ineludible de diseñar marcos ético-normativos suficientemente flexibles para acompañar la evolución de la IA, pero igualmente firmes para anticipar sus riesgos ontológicos y antropológicos. Solo mediante una gobernanza centrada en la protección de la condición humana, la *equidad informacional* y la dignidad humana será posible preservar aquello que constituye la esencia de nuestra humanidad ante el avance imparable de la IA: *la capacidad de ejercer autonomía, de desarrollar nuestras capacidades, de cuidar de nosotros mismos y de los demás, y de establecer relaciones éticas basadas en el respeto y la compasión.* Este compromiso ético es, en última instancia, la garantía de que la IA sirva para la promoción de la dignidad humana y no así para su erosión o, incluso, su extinción.

Olimpia Giuliana Loddo

L'agire sociale ai tempi dell'intelligenza artificiale. Il concetto di "istituzione socio-digitale"*

Abstract: **Social Action in the Age of Artificial Intelligence. The Concept of "Socio-Digital Institution".** The concept of socio-digital institutions, proposed by Teubner and Backers, to which this article is dedicated, can be framed as one of the conceptual resources developed to address new questions that arise from the evolving role of artificial intelligence in social dynamics. In this article, I intend to explore the implications and assumptions of the concept of socio-digital institutions by analysing it from three perspectives. The first perspective is ontological. The second perspective is typological. The third perspective is meta-theoretical. My goal is to highlight how the concept of digital institutions fits into the theory of social systems of which Teubner is a main proponent. In particular, I aim to investigate the role of the digital institution as conceived by Beckers and Teubner in social systems and how the peculiar nature of communication mediated or produced by artificial intelligence may impact such a role.

Keywords: Social Systems, Artificial Intelligence, Socio-digital Institutions.

Sommario: 1. Introduzione (Il problema) – 2. Il concetto di "*socio-digital institution*". Ontologia dell'istituzione sociale digitale (L'approccio di Teubner e Beckers) – 2.1. Riflessione sulla tipologia delle istituzioni sociali-digitali – 2.1.1. L'assistenza digitale come istituzione socio-digitale – 2.1.2. L'associazione uomo macchina come istituzione socio-digitale – 2.1.3. L'esposizione all'interconnettività digitale come istituzione digitale – 3. Il ruolo dell'istituzione digitale nel sistema sociale (Discussione) – 3.1. Il ruolo della comunicazione – 3.2. L'influenza sulla società caratterizzata da contatti non-comunicativi – 4. Conclusioni.

1. Introduzione (Il problema)

In che modo l'intelligenza artificiale sta influenzando il sistema sociale? Come si concilia la capacità del sistema giuridico di

* Segnalo con riconoscenza il supporto finanziario fornito nell'ambito del Piano Nazionale di Ripresa e Resilienza (PNRR), Missione 4, Componente 2, Investimento 1.1, Bando n. 104 pubblicato il 02/02/2022 dal Ministero dell'Università e della Ricerca (MUR), finanziato dall'Unione Europea – NextGenerationEU – Titolo del Progetto: Artefatti normativi e disegni normativi: indagine sulla regolamentazione non linguistica (NAND) – CUP: F53D23003500006 – Decreto di Assegnazione della Sovvenzione n. 968 adottato il 30/06/2023 dal Ministero dell'Università e della Ricerca (MUR). Desidero inoltre esprimere il mio sentito ringraziamento al Prof. Giuseppe Lorini.

produrre vincoli individuali e collettivi mediante le nuove tecnologie che sfruttano l'intelligenza artificiale? Chi è responsabile delle conseguenze di un evento prodotto, di una scelta compiuta, di una informazione fornita da un'intelligenza artificiale? Fenomeni nuovi e domande nuove talora richiedono parole nuove e nuovi concetti. Questa esigenza emerge sia in ambito sociologico sia, in modo più cauto e mediato dalle peculiarità dell'ordinamento, in ambito giuridico. Infatti, nella visione sociologica di Teubner, in continuità con la teoria dei sistemi di Luhmann[2], il diritto è inquadrato come un sottosistema chiuso e autopoietico che è, al contempo, caratterizzato da un dinamismo tale da innescare dei meccanismi di (auto)sovversione[3] in grado di incorporare le istanze degli altri sottosistemi sociali, i loro contrasti, i loro paradossi, le loro antinomie. La chiusura deriva dal fatto che il sottosistema giuridico non incorpora passivamente le suggestioni esterne ma le trasforma filtrandole attraverso le proprie categorie concettuali e autogenerandosi in base ai suoi procedimenti interni[4]. Pertanto, Beckers e Teubner affermano che:

Non è sufficiente […] "leggere" le questioni giuridiche direttamente dalle proprietà tecniche delle macchine digitali. Questo condurrebbe ad un cortocircuito tra tecnologia e diritto. Il cortocircuito condurrebbe a risultati disastrosi perché non coglierebbe il legame cruciale tra le sfide tecnologiche e le reazioni del diritto[5].

Il ruolo che l'intelligenza artificiale sta assumendo nelle dinamiche sociali ha condotto all'elaborazione di nuovi strumenti concettuali[6] che stanno contribuendo a definire il dibattito so-

[2] N. Luhmann, *Sistemi sociali. Fondamenti di una teoria generale*, trad. it., il Mulino, Bologna, 1990.

[3] R. Prandini, G. Teubner, *La (auto)sovversione del diritto: ambiguità e paradossi del sistema immunitario della società*, FrancoAngeli, Milano, 2022.

[4] G. Teubner, *Law as an Autopoietic System*, Blackwell, Oxford, 1993.

[5] A. Beckers, G. Teubner, *Three Liability Regimes for Artificial Intelligence. Algorithmic Actants, Hybrids, Crowds*, Hart Publishing, Oxford, 2021, p. 14. Ecco il testo originale: "[I]t does not suffice […] to 'read' legal questions directly from digital machines' technical properties. This would amount to a short circuit between technology and law. The short circuit ends up in disastrous results because it misses the crucial link between technology's challenges and law's reactions". La traduzione è mia.

[6] Il sociologo del diritto svedese Håkan Hydén utilizza il termine "*algo norm*" per indicare le norme implicite che emergono nell'attività di algo-

cio-giuridico contemporaneo sulle nuove forme di regolazione emergenti grazie allo sviluppo tecnologico[7].

2. Il concetto di "*socio-digital institution*". Ontologia dell'istituzione sociale digitale (L'approccio di Teubner e Beckers)

Il concetto di *socio-digital institution*, al quale è dedicato il presente articolo, può essere inquadrato tra le risorse concettuali elaborate al fine di rispondere ai nuovi interrogativi emergenti dall'evoluzione del ruolo dell'intelligenza artificiale nelle dinamiche sociali. L'istituzione socio-digitale è un elemento fondamentale dell'apparato concettuale proposto nel libro, pubblicato nel 2021, *Three Liability Regimes for Artificial Intelligence* di Anna Beckers e Gunther Teubner.

Apparentemente Anna Beckers e Gunther Teubner ripropongono, sviluppandole, delle soluzioni tecnico-giuridiche finalizzate ad affrontare i problemi legati alla responsabilità degli agenti digitali, riprendendo alcune tesi che appaiono nel saggio di Teubner *Soggetti giuridici digitali: sullo* status *privatistico degli agenti* software *autonomi*[8]. I due autori suggeriscono l'introduzione di nuove forme di *status* giuridico come la persona-

ritmi basati sull'intelligenza artificiale. Hydén distingue l'aspetto tecnico degli algoritmi dalle conseguenze sociali derivanti dalla loro applicazione e ritiene che le "algo-norm" riguardino questo secondo aspetto. Secondo Hydén le algo-norm pongono dei problemi di legittimità poiché sottraggono alcuni aspetti della regolazione al processo democratico, delegandoli di fatto ai tecnici. Inoltre, possono essere uno strumento di manipolazione e portare alla creazione di *filter bubbles* che espongono gli utenti esclusivamente ad informazioni che confermano le loro opinioni H. Hydén, "AI, Norms, Big Data, and the Law", in *Asian Journal of Law and Society*, 7 (2020), n. 3, pp. 409-436.

[7] Ad esempio, il giurista tedesco Klaus Günther utilizza il termine "*smart order*" per descrivere l'uso di tecnologie digitali nella regolazione del comportamento umano al fine di minimizzare le possibilità di devianza, inducendo una riflessione sulla natura normativa di tale forma di regolazione. K. Günther, "From Normative to Smart Orders?", in *Smart Legal Order in the AI Society*, University of Bologna, Bologna, 30/06/2023.

[8] G. Teubner, "Digitale Rechtssubjekte? Zum privatrechtlichen Status autonomer Softwareagenten", in *Archiv für die civilistische Praxis (AcP)*, 218 (2018), pp. 155-205; Id., *Soggetti giuridici digitali? Sullo status privati-*

lità giuridica parziale, l'attribuzione dello *status* di membri di associazioni o la creazione di pool di rischio [*risk pool*] per gli agenti software.

Tali forme di qualificazione sono tese a superare le carenze nella disciplina sulla responsabilità giuridica, consentendo, al contempo, lo sviluppo coerente della regolamentazione giuridica dell'interazione nel contesto digitale[9]. L'intento di Beckers e Teubner sembra essere, apparentemente, "solo" quello di trattare ciò che è uguale in modo uguale e ciò che è diverso in modo diverso, nello spazio digitale, assumendo una prospettiva tecnico giuridica rigorosa che tenga, tuttavia, conto delle esigenze emergenti nel sistema sociale derivanti dalla diffusione di tecnologie basate sullo sviluppo dell'intelligenza artificiale.

In realtà, il libro redatto da Beckers e Teubner sembra avere un obiettivo ancora più ambizioso. Con l'esplicitazione del concetto di "istituzione socio-digitale", infatti, non solo si irrobustisce l'apparato concettuale volto a chiarire i presupposti teorico-sociologici della proposta di Teubner del 2018 ma si offre uno strumento concettuale rilevante sia per la sociologia del diritto sia per l'ontologia sociale.

Nel presente articolo intendo ricostruire le implicazioni e i presupposti del concetto di istituzione socio-digitale analizzandolo sotto tre diversi profili.

Il primo è il *profilo ontologico*, per esplorare questa dimensione mi focalizzerò sulle condizioni di esistenza delle istituzioni socio-digitali. Il secondo è un profilo di carattere *tipologico*, cercherò di ricostruire la tipologia delle istituzioni digitali proposta dagli autori alla luce delle considerazioni svolte nel presente articolo. Il terzo è un profilo *metateorico*. Il mio obiettivo è evidenziare come il concetto di istituzione digitale si incardina nella teoria dei sistemi sociali della quale Teubner è certamente uno dei maggiori esponenti. In particolare, voglio indagare il ruolo dell'istituzione digitale così come concepita da Beckers

stico degli agenti software autonomi, trad. it., Edizioni Scientifiche Italiane, Napoli, 2019.

[9] Tali soluzioni partono dal rifiuto di una attitudine, comune a molti giuristi, di ritenere le categorie giuridiche già esistenti perfettamente idonee a disciplinare fenomeni sociali nuovi e talora rivoluzionari come, ad esempio, lo sviluppo delle tecnologie legate all'intelligenza artificiale.

e Teubner nel sistema sociale e come la peculiare natura della comunicazione mediata o prodotta dall'intelligenza artificiale possa incidere su tale ruolo.

Beckers e Teubner forniscono una definizione dell'espressione *socio-digital institution* individuandone le condizioni di possibilità ed evidenziandone la dimensione "normativa". Nelle parole di Beckers e Teubner:

> Proponiamo di introdurre il concetto di "istituzioni socio-digitali". Gli algoritmi non hanno di per sé le qualità ontologiche di un agente [*actor*] che consentano loro di instaurare relazioni sociali e di comunicare con esseri umani. Solo nel momento in cui gli algoritmi entrano a far parte di una istituzione socio-digitale [*socio-digital institutions*], queste istituzioni otterranno, in base alle loro premesse normative, delle capacità comunicative[10].

Dalle parole degli autori si evince che le istituzioni socio-digitali sono tali quando soddisfano quattro diverse condizioni:

(i) hanno delle premesse normative [*normative premises*],

(ii) hanno delle capacità comunicative [*communicative capacities*],

(iii) sono qualificabili come attori [*qualify as actors*],

(iv) hanno come componente un algoritmo [*algorithms are part of socio-digital institutions*].

L'algoritmo è, quindi, condizione necessaria ma non sufficiente per l'esistenza dell'istituzione socio-digitale. Le altre tre componenti dell'istituzione socio-digitale sono congiuntamente condizione sufficiente di socialità dell'algoritmo, ovvero sono condizione sufficiente perché tale algoritmo possa acquisire una natura socio-istituzionale.

Possiamo pensare alla quarta condizione di esistenza, all'algoritmo, come a un *análogon* digitale dei fatti bruti nel senso che si evince nel saggio *On Brute Facts* di G.E.M. Anscom-

[10] A. Beckers, G. Teubner, *op. cit*., p. v. Ecco il testo originale: "We propose to introduce the concept of 'socio-digital institutions'. Algorithms do not have as such the ontological qualities of an actor that allow them to engage in social relations and communicate with humans. Only once algorithms are part of socio-digital institutions, these institutions will, according to their normative premises, obtain communicative capacities and qualify as actors". La traduzione è mia.

be. In questo contesto, il concetto di *brute fact* proposto dalla Anscombe è preferibile a quello proposto da Searle[11], poiché mentre secondo Searle i fatti bruti sono fatti che esistono "*independently of any human institution*" nel lessico di Anscombe il fatto bruto è chiaramente un concetto relativo. Gli algoritmi non sono certamente indipendenti dalle istituzioni umane, poiché devono trovare espressione nei linguaggi di programmazione e sono ideati dall'uomo. Per questa ragione, gli algoritmi non possono essere intesi come fatti bruti in senso searliano. L'algoritmo non è bruto per sé ma è bruto rispetto alla descrizione dell'istituzione socio-digitale[12].

Gli algoritmi non sono entità puramente materiali, essi sono entità ibride che non possono prescindere da fenomeni fisici legati all'hardware ma sono al contempo dotate di una dimensione semiotica poiché sono prodotti da segni e a loro volta producono segni. Inoltre, essi alterano la loro morfologia anche in ragione dell'evoluzione del loro ruolo istituzionale. In particolare, questo vale per gli algoritmi basati sull'intelligenza artificiale. Quando parlo di alterazione della morfologia di un algoritmo mi riferisco a variazioni della sua struttura interna che possono riguardare i pesi nei modelli di rete neurale[13], le regole di classificazione nei sistemi di apprendimento simbolico o altre rappresentazioni interne specifiche dell'algoritmo.

Secondo Beckers e Teubner anche i rischi, la potenziale capacità di ledere interessi giuridicamente rilevanti, propri di ciascuna istituzione socio-digitale derivano dalle sue specifiche caratteristiche e, nelle parole degli autori, tali rischi "are coproduced by computer technology and social relations"[14].

[11] J.R. Searle, *Making the Social World. The Structure of Human Civilization*, Oxford University Press, Oxford, 2010, p. 10.

[12] G.E.M. Anscombe, "On Brute Facts", in *Analysis*, 18 (1958), n. 3, pp. 69-72.

[13] In una rete neurale artificiale, i pesi sono i parametri che definiscono la forza delle connessioni tra i neuroni. Ogni peso determina quanto un segnale in uscita da un neurone influisce sull'attivazione di un altro neurone. L'aggiornamento di questi pesi è finalizzato a ridurre la discrepanza tra le previsioni della rete e i risultati conseguiti.

[14] A. Beckers, G. Teubner, *op. cit.*, p. 145.

2.1. Riflessione sulla tipologia delle istituzioni socio-digitali

La tipologia delle istituzioni socio-digitali proposta da Beckers e Teubner si basa su una riflessione sull'agentività digitale, sulle diverse modalità di interazione tra intelligenza artificiale e agenti umani e sui suoi potenziali rischi. In particolare, Beckers e Teubner distinguono tre forme di rischio diverse individuabili in ragione delle peculiarità socio-tecnologiche degli agenti digitali: il *rischio di autonomia*, il *rischio di associazione*, il *rischio di rete.*

Il *rischio di autonomia* si verifica in caso di scelte completamente autonome da parte dell'agente digitale. Teubner suggerisce che una risposta adeguata a questo rischio consiste nel concedere agli agenti software uno "*status* di personalità giuridica parziale". Ciò implica che le decisioni autonome degli algoritmi dovrebbero essere legalmente vincolanti e dovrebbero comportare conseguenze in termini di responsabilità. In altre parole, gli agenti software avrebbero una sorta di capacità giuridica limitata che consentirebbe loro di stipulare contratti giuridicamente vincolanti per conto di terzi. In caso di responsabilità contrattuale ed extracontrattuale, gli agenti software sarebbero considerati come assistenti giuridicamente capaci, in modo che il comportamento della macchina stessa (e non solo la condotta delle aziende che la gestiscono) costituisca una violazione del dovere per la quale le aziende devono essere ritenute responsabili.

Il *rischio di associazione* riguarda la partecipazione degli agenti software a "associazioni umano-macchina". Beckers e Teubner suggeriscono che una possibile soluzione sia attribuire agli agenti software lo *status* giuridico di membri di tali associazioni. In base a una soluzione ottimale (*de lege ferenda*), le azioni, i diritti e gli obblighi verrebbero attribuiti all'associazione stessa, aprendo nuove prospettive nel diritto privato. In una soluzione minimale (*de lege lata*), potrebbe essere sfruttato il concetto giuridico di "scopo dell'associazione" [*association purpose*], che dovrebbe guidare l'interpretazione delle dichiarazioni prodotte dall'intelligenza artificiale e la determinazione dei diritti e degli obblighi dei partecipanti.

Il *rischio di rete riguarda* il flusso di informazioni digitali. Teubner suggerisce la creazione di un "*pool* di rischio" che de-

terminerebbe lo *status* giuridico degli algoritmi nell'ambito di un flusso di informazioni digitali e ne stabilirebbe la responsabilità.

In relazione alle forme di rischio sopra elencate, che emergono dalla combinazione di fattori sociali con fattori tecnologici Teubner e Beckers distinguono tre forme di *socio-digital institutions*: l'assistenza digitale [*digital assistance*] che è alla base del rischio di autonomia, l'associazione uomo macchina [*human-machine association*] alla base del rischio di associazione, l'esposizione alla interconnettività digitale [*exposure to digital interconnectivity*] alla base del rischio di rete.

2.1.1. L'assistenza digitale come istituzione socio-digitale

Una *digital assistance* nel lessico di Teubner e Beckers è un'istituzione socio-digitale che consiste in sistemi informatici o applicazioni *software*, spesso basati sull'intelligenza artificiale e sulle tecnologie di elaborazione del linguaggio naturale, progettati per fornire supporto, aiuto o informazioni agli utenti e rispondere alle richieste degli utenti formulate in linguaggio naturale.

Secondo Teubner e Beckers la *digital assistance* trasforma i processi digitali in "*actants*", ovvero entità attive all'interno di una situazione. Gli attanti non debbono essere intesi come agenti, poiché si basano su istruzioni predefinite e sono privi di consapevolezza. Tuttavia, tali sistemi possono prendere decisioni non prevedibili o andare al di là delle istruzioni previste, esponendo i loro utenti a rischi legati alla carenza di un controllo sul processo di elaborazione dell'*output*.

2.1.2. L'associazione uomo macchina come istituzione socio-digitale

Una *human-machine association* (associazione uomo-macchina) è una relazione o un'integrazione stretta tra esseri umani e sistemi informatici o algoritmi. Si tratta di una forma di collaborazione che va oltre il semplice supporto digitale fornito dalle assistenze virtuali. Questa associazione può assumere diverse forme e manifestarsi in diversi contesti, come nel giornalismo computazionale, nella governance aziendale, nel trading finan-

ziario o nei servizi di traduzione. Ad esempio, scrivono Beckers e Teubner

> [i]l giornalismo computazionale è un chiaro esempio di questo; gli attori umani e gli attanti non umani sono connessi in reti e il loro flusso di lavoro è iteratamente riprogettato[15].

Le *human-machine associations* si distinguono dall'assistenza virtuale in quanto coinvolgono un grado maggiore di interazione e integrazione tra esseri umani e macchine. Mentre l'assistenza virtuale di solito comporta l'uso di assistenti digitali o chatbot per fornire supporto o informazioni agli utenti, le *human-machine associations* implicano una collaborazione più stretta e una condivisione di responsabilità e decisioni tra umani e macchine.

Tra i diversi esempi descritti da Teubner e Beckers in relazione all'istituzione socio-digitale *human-machine associations*, oltre agli algoritmi utilizzati nella produzione giornalistica, vi sono la presenza di algoritmi come membri autonomi nei consigli di amministrazione aziendali e, più in generale, l'interazione tra umani e macchine nei processi decisionali collettivi. Questi esempi mostrano come la collaborazione tra umani e macchine possa portare a effetti estremamente rilevanti sul sistema sociale.

2.1.3. L'esposizione all'interconnettività digitale come istituzione digitale

L'espressione "*exposure to interconnectivity*" si riferisce alla situazione in cui le persone, le organizzazioni o la società nel suo complesso sono esposte agli effetti della connettività tra diversi agenti elettronici autonomi o sistemi informatici. Questa connettività può manifestarsi in varie forme, come la comunicazione e lo scambio di dati tra dispositivi, sensori o algoritmi autonomi.

Secondo Beckers e Teubner:

> L'esposizione all'interconnettività algoritmica è una istituzione socio-digitale indipendente [dalle altre due tipologie].

[15] *Ivi*, p. 111. Ecco il testo originale: "'Computational journalism' is a clear example of this; human actors and non-human actants are assembled in a newswork, and their work flows are iteratively re-engineered". La traduzione è mia.

> Non può essere ridotta ad un processo decisionale individuale nell'assistenza digitale né a decisioni collettive derivanti dall'associazione umano-macchina. L'interconnettività è una configurazione a sé stante, la sua relazione con la società produce una specifica istituzione socio digitale[16].

Esempi di esposizione alla interconnessione possono aver luogo in reti di comunicazione e social media, nell'ambito delle piattaforme di trading, o anche in altre infrastrutture gestite da sistemi digitali in modo non sorvegliato (ad esempio una rete elettrica, o una rete di trasporto pubblico) che possono essere esposte a rischi legati sia alla sicurezza pubblica e sia alla stabilità economica.

L'interconnessione tra piattaforme di social media e reti di comunicazione può influenzare la diffusione di false informazioni e consentire una manipolazione dell'opinione pubblica, nonché agevolare la violazione della privacy degli utenti. Infine, l'interconnessione tra piattaforme di trading, istituzioni finanziarie e reti di comunicazione, può aumentare il rischio di volatilità e destabilizzazione nei mercati finanziari.

L'"*exposure to interconnectivity*" è considerata come una istituzione socio-digitale autonoma. Questo significa che la connettività tra sistemi informatici autonomi non può essere ridotta né alla decisione individuale che caratterizza l'assistenza digitale né alle decisioni collettive nelle associazioni uomo-macchina.

La connettività tra macchine non può essere personificata, poiché non ha capacità comunicative in senso stretto. Tuttavia, influisce significativamente sulla società, portando con sé rischi legati all'imprevedibilità, all'incomprensibilità e all'invisibilità delle operazioni interconnesse.

[16] *Ivi*, p. 90. "The exposure to algorithmic interconnectivity is a self-standing socio-digital institution. It can neither be reduced to individual decision-making in digital assistance nor to collective decisions by human-machine associations. Interconnectivity is a configuration in its own right, and its relation to society results in a specific socio-digital institution". La traduzione è mia.

3. Il ruolo dell'istituzione digitale nel sistema sociale (Discussione)

Se inquadriamo il ruolo dell'intelligenza artificiale nella cornice della teoria dei sistemi, ci rendiamo conto della sua potenziale rilevanza (in verità sempre più attuale) nei processi di comunicazione interna ai sistemi sociali.

Quando gli elementi di un sistema interagiscono attraverso la comunicazione, emergono nuovi significati e strutture che non esistevano prima delle interazioni. In questo senso, la comunicazione può essere intesa come una realtà emergente che è prodotta dagli agenti e dagli attanti che interagiscono in un sistema. Chiaramente questa considerazione vale anche nel caso di comunicazioni con agenti digitali. La comunicazione incide sulle dinamiche del sistema nel suo complesso e crea "nuove realtà" derivanti dalle interazioni e dalla costruzione di significato condivisa tra gli attori del sistema.

3.1. Il ruolo della comunicazione

Poiché l'intelligenza artificiale può incidere pesantemente sui processi di comunicazione, essa può contribuire in modo decisivo ad alterare le strutture dei sistemi sociali.

Più precisamente, l'intelligenza artificiale può partecipare alle interazioni comunicative all'interno di un sistema, elaborando informazioni, generando risposte e contribuendo alla costruzione e all'interpretazione di un significato condiviso.

In questo discorso è importante tener presente il senso attribuito al termine "comunicazione" [*Kommunikation*] nel lessico della teoria dei sistemi di Luhmann. Infatti, secondo Luhmann, l'atto del comunicare [*Mitteilung*] non deve essere inteso come un trasferimento di contenuti da un mittente ad un destinatario, la metafora della trasmissione è fuorviante e non chiarisce il ruolo del destinatario nel processo di comunicazione. Nelle parole di Luhmann:

> L'atto del comunicare [*Mitteilung*] non è altro che una proposta di selezione, un suggerimento. È solo quando questo suggerimento viene preso in considerazione, quando l'impul-

> so viene processato, che la comunicazione [*Kommunikation*] ha luogo[17].

Un altro elemento importante da considerare è che l'informazione rappresenta il prodotto di una selezione tra un repertorio noto o ignoto di possibilità[18] e non resta invariata nell'ambito del processo di comunicazione. Anche per questo la comunicazione contribuisce alla trasformazione del sistema sociale.

L'intelligenza artificiale può svolgere diversi ruoli all'interno del contesto della comunicazione all'interno di un sistema (ad esempio, elaborazione delle informazioni, supporto decisionale, automazione dei compiti, assistenza nella comunicazione). Inoltre, può essere utilizzata per analizzare grandi quantità di dati ed estrarre informazioni rilevanti per il sistema, che possono poi essere utilizzate per prendere decisioni o per formulare risposte alle richieste dei partecipanti al sistema.

Come evidenzia Luhmann "*information is nothing but selection exclusively internal to the system – conditioned by the system's selection 'horizon'*"[19]. Quindi la capacità dell'intelligenza artificiale di incidere sul sistema sociale è strettamente connessa alla struttura effettiva del sistema, ai dati contenuti nel sistema, ai suoi significati e alle sue norme.

Un aspetto fondamentale della visione di Beckers e Teubner riguarda il modo in cui l'intelligenza artificiale può intervenire nei processi di comunicazione e influenzare le istituzioni sociali. Questo processo può avvenire in modo diretto o in modo indiretto. L'influenza dell'intelligenza artificiale sui processi di comu-

[17] N. Luhmann, *Soziale Systeme: Grundriß einer allgemeinen Theorie*, Suhrkamp, Frankfurt am Main, 1984, p. 194. Ecco il testo originale: "Die Mitteilung ist aber nichts weiter als ein Selektionsvorschlag, eine Anregung. Erst dadurch, daß diese Anregung aufgegriffen, daß die Erregung prozessiert wird, kommt Kommunikation zustande". La presente traduzione è stata rivista da me per esplicitare la distinzione tra atto del comunicare "*Mitteilung*" e comunicazione "*Kommunikation*", tale distinzione è evidenziata nel glossario della traduzione curata da Alberto Febbrajo (N. Luhmann, *Sistemi sociali*, cit., p. 750) che resta un'irrinunciabile guida.

[18] *Ivi*, p. 195.

[19] N. Luhmann, "Closure and Openness: On Reality in the World of Law", in G. Teubner (ed.), *Autopoietic Law: a New Approach to Law and Society*, De Gruyter, Berlin, 1988, p. 343.

nicazione è diretta quando gli agenti digitali compiono dei veri e propri atti comunicativi. Tale influenza può essere esercitata in modo indiretto quando le istituzioni socio-digitali caratterizzate da *exposure to interconnectivity* hanno un'influenza sulla società basata su contatti non-comunicativi [*non communicative contacts*].

3.2. L'influenza sulla società caratterizzata da contatti non-comunicativi

In base alla ricostruzione proposta da Teubner e Beckers l'interazione tra uomo e macchina avviene in due spazi distinti: nello *spazio comunicativo* e nell'ambito di *contatti non-comunicativi*.

Esiste un'area relativamente limitata in cui la comunicazione tra esseri umani e computer è possibile. Questo spazio è popolato da attori algoritmici e ibridi uomo-macchina, che possono comunicare attraverso interfacce. Tuttavia, nella maggior parte delle operazioni algoritmiche interne, le dinamiche non sono accessibili né alla comunicazione sociale né alla coscienza umana. Secondo Beckers e Teubner:

> C'è un vasto segmento delle operazioni interne degli algoritmi che sono "macchine invisibili" per gli esseri umani; la società non ha affatto contatti comunicativi con essi. Tuttavia, essi influenzano fortemente la società, anche se in modo indiretto[20].

Qui non si tratta più di interazioni uomo-macchina, ma di relazioni *non-comunicative uomo-macchina*. Infatti, mentre gli esseri umani possono comunicare con i computer attraverso un'interfaccia, gran parte delle operazioni algoritmiche interne non sono accessibili alla comunicazione o alla coscienza umana. In altri termini, le operazioni interne delle macchine sono formalizzate in processi binari elettronici che non sono accessibili attraverso la comunicazione. Come sottolineano Beckers e Teubner:

[20] A. Beckers, G. Teubner, *op. cit.*, p. 27. "There is a large segment of internal operations of algorithms that are 'invisible machines' for humans; society has no communicative contact at all to them, but they influence society in a massive, albeit indirect way". La traduzione è mia.

> Tipicamente, le macchine interconnesse esercitano un'influenza indiretta ma enorme sulla società, che è difficile da comprendere con gli strumenti delle scienze sociali […] I sistemi tecnologici non operano per mezzo del significato [*in the medium of meaning*] ma attraverso le operazioni elettroniche e le loro interconnessioni. Questi processi binari formalizzati non sono accessibili attraverso la comunicazione. Tuttavia, attraverso le interfacce di superficie si determina un impatto sostanziale sugli eventi comunicativi, sulle strutture sociali e i loro canali [*media*][21].

In breve, mentre l'interazione diretta tra uomini e computer avviene in uno spazio comunicativo limitato, gran parte delle operazioni algoritmiche interne non sono accessibili alla comunicazione umana e influenzano la società in modo indiretto e spesso invisibile.

4. Conclusioni

Secondo Luhmann nei sistemi sociali le variabili che aumentano la complessità e quelle che la limitano possono essere molto simili[22]. Anche le istituzioni socio-digitali possono essere intese come una variabile che, declinata in modo diverso, può portare sia alla riduzione sia all'incremento della complessità intesa come eccesso di possibilità rispetto a quelle attuabili[23]. Da un lato le istituzioni socio-digitali possono essere uno strumento di

[21] *Ivi*, pp. 112-13. Ecco il testo originale: "Characteristically, interconnected machines exert a highly indirect but enormous influence on society, which is difficult to grasp with the instruments of the social sciences [...] Technological systems do not operate in the medium of meaning but through electronic operations and their interconnections. These formalised binary processes are not accessible via communication. Nonetheless, a substantial impact on communicative events, social structures, and their media is exercised through surface interfaces". La traduzione è mia.

[22] Si veda N. Luhmann, "Systemtheoretische Argumentationen. Eine Entgegnung auf Jürgen Habermas", in J. Habermas, N. Luhmann, *Theorie der Gesellschaft oder Sozialtechnologie – Was leistet die Systemforschung?*, Suhrkamp, Frankfurt am Main, 1971, p. 296. (N. Luhmann, "Argomentazioni teoretico-sistematiche. Una replica a Jürgen Habermas", in J. Habermas, N. Luhmann, *Teoria della società o tecnologia sociale*, trad. it., ETAS, Milano, 1973 p. 200).

[23] N. Luhmann, *Sistemi sociali*, cit., p. 111.

semplificazione consentendo l'automatizzazione di processi anche attraverso l'analisi di grandi quantità di dati, dall'altro lato, possono incrementare la complessità proprio in ragione del fatto che parte dell'influenza che esse esercitano opera solo indirettamente sui processi di comunicazione.

La diminuzione della complessità può portare ad una diminuzione della libertà. La riduzione della libertà è, in questo caso, un riflesso dell'assenza di strumenti di controllo affidabili in relazione alle operazioni algoritmiche interne alle istituzioni socio-digitali e della possibilità di incidere sulla società in uno spazio "non-comunicativo".

Questa capacità di agire entro uno spazio non-comunicativo, evidenziata da Beckers e Teubner, è ad un tempo interessante e paradossale; pur essendo delle entità proposizionali[24] che trovano espressione nei linguaggi di programmazione, le istituzioni socio-digitali possono dare origine a forme di regolazione non-proposizionale. Per regolazione non proposizionale si intende una forma di regolazione priva di significato normativo, che può condizionare il comportamento umano, pur non essendo adempibile in senso stretto[25].

Del resto, il meccanismo che caratterizza alcune forme indirette di incidenza delle istituzioni socio-digitali sui sistemi sociali[26] non è in sé molto dissimile da quello che caratterizza alcune forme di regolazione non-proposizionale che contraddistinguono le tecniche di architettura ostile. Secondo Lorini e Moroni:

> Se consideriamo una panchina anti-sonno; non ha senso chiedersi se qualcuno l'abbia "violata". Potrebbe esserci un

[24] Gli algoritmi che compongono le istituzioni socio-digitali trovano espressione nel linguaggio di programmazione. Pare interessante notare che un medesimo algoritmo può trovare espressione in diversi linguaggi di programmazione (con alcuni accorgimenti dovuti alle caratteristiche specifiche di ogni singolo linguaggio). Tuttavia, indipendentemente dalla loro natura "proposizionale" esse possono essere utilizzate in forme di regolazione non proposizionale.

[25] G. Lorini, S. Moroni, "Non-Propositional Regulation", in *Philosophical Investigations*, 45 (2022), n. 4, pp. 512-527.

[26] O.G. Loddo, "Artefatti regolativi digitali", in *Studi economico-giuridici*, 65 (2024), p. 113-130.

> comportamento che non si conforma all'ideale che il regolatore aveva in mente quando ha introdotto l'artefatto regolativo. […] parlare di 'violazione di un dosso' o 'violazione di una panchina anti-sonno', sarebbe usare un'espressione grammaticalmente scorretta[27].

Forme di regolazione non proposizionale possono emergere da interazioni non-comunicative tra l'uomo e la macchina nell'ambito di diversi sottosistemi sociali.

Un possibile esempio, non troppo futuristico, potrebbe riguardare l'ambito della regolazione dei flussi migratori. Si pensi, ad esempio, a degli agenti *software* programmati per stabilire la provenienza di un migrante in base ai tratti biometrici al fine di verificare la veridicità delle sue dichiarazioni e di bloccarne l'accesso in un determinato paese. In questo senso è interessante il caso degli e-gates, installati in alcuni aeroporti, che sono in grado di controllare se il numero di identificazione personale di un cittadino figura sui mandati nazionali o dell'Interpol, se i dati nel passaporto sono autentici e di bloccare l'accesso a determinate aree dell'aeroporto[28].

Le istituzioni socio-digitali possono essere progettate per agire in conformità con obiettivi specifici e regole predeterminate, ma non hanno comprensione completa del contesto o delle implicazioni delle loro azioni, né è possibile ricostruire integralmente il percorso che li ha portati alla decisione finale. Pertanto, devono essere oggetto non solo di una progettazione informatica, ma anche di un'adeguata progettazione istituzionale che tenga conto dei loro possibili errori e delle conseguenze delle loro azioni sul sistema sociale nel suo complesso e sui singoli sottosistemi sociali.

[27] G. Lorini, S. Moroni, *op. cit.*, p. 526. La traduzione è mia.

[28] O.G. Loddo, A. Addis, "Tre funzioni dell'IA nella gestione dei flussi migratori", in *ADiM Blog, Analisi & Opinioni*, (2023), p. 7.

Paolo Moro

Persona elettronica.
Una finzione giuridica per l'intelligenza artificiale

Abstract: **Electronic Person. A Legal Fiction for Artificial Intelligence.** In the premise, it is observed that the technological advancement of artificial intelligence is pushing advanced mechatronic systems to transform into sophisticated robotic forms, which now lack an android morphology but emulate typically human faculties and resemble artificial persons.
The introductory part of the study critically explores the foundational aspect of robotic subjectivity and re-examines the epistemological parallels, commonly accepted in the modern era, between the calculating machine and human consciousness: the traditional comparison between the digital device and human reasoning capability; the view of software as agents, carriers of ethical and legal responsibility; the definition of mechanical entities capable of recognising and emulating emotions.
In the next part of the essay, the tendency to superficially humanise machines, through a naive imitation of human behaviour, is criticised, and the theoretical and practical difficulties of establishing an acceptable comparison between artificial intelligence and human consciousness are highlighted, still presupposed on the basis of a reductive conception of human subjectivity.
In the conclusion, the question is raised as to whether and to what extent, based on the epistemological premises outlined in the first part, sophisticated mechatronic systems and advanced robotic entities might attain the qualification of "electronic person", a legal fiction indicated by the Resolution of the European Parliament of 16 February 2017 and currently overlooked by the recent AI Act of the European Union.

Keywords: Robotic Subjectivity, Electronic Person, Legal Fiction, Mechatronic Systems and Law, Moral Machines.

Sommario: 1. Persone artificiali – 2. Congegni intelligenti – 3. Congegni morali – 4. Congegni emotivi – 5. Congegni inconsapevoli – 6. Persone elettroniche.

1. Persone artificiali

La realizzazione di sistemi di intelligenza artificiale altamente organizzati, dotati di elevata automazione e capaci di emulare alcune facoltà tipicamente umane, non è più il frutto della fervida immaginazione letteraria della fantascienza più brillante e ri-

flessiva del Novecento[1], ma un fenomeno sociale ed economico che ha provocato in Europa molteplici studi sui limiti etici dei sistemi meccatronici e cibernetici e ha condotto all'approvazione del nuovo AI Act dell'Unione Europea sull'intelligenza artificiale, che nell'intero pianeta si presenta come la prima legge che disciplina il fenomeno.

Come è noto, l'ampia e progressiva diffusione dell'intelligenza artificiale, alimentata dai rapidi progressi della tecnologia informatica, ha portato l'Unione Europea a considerare l'importanza del tecnodiritto[2] e stabilire una nuova legislazione cogente al fine di garantire lo sviluppo continuo del settore industriale e di proteggere simultaneamente il valore intrinseco della persona in base a princìpi generali di tutela della dignità umana[3].

Pertanto, dopo una serie articolata di lavori preparatori, il Parlamento Europeo ha approvato definitivamente il 13 marzo 2024 il regolamento sull'intelligenza artificiale (*artificial intelligence act*), che è un atto normativo di portata generale e, quindi, è obbligatorio in tutti i suoi elementi e direttamente applicabile in ciascuno degli Stati membri (in base all'art. 298 del Trattato sul Funzionamento dell'Unione Europea).

Già una decina d'anni orsono si avvertiva la possibilità che, attraverso l'utilizzo più efficace dell'intelligenza artificiale, i nuovi saperi del terzo millennio, come l'informatica, la robotica e le neuroscienze, avrebbero condotto alla costruzione di sistemi meccatronici complessi, che appaiono quasi come persone artificiali, in quanto imitano differenti facoltà umane: non soltanto ragionamenti, come la comprensione del linguaggio naturale, l'apprendimento automatico o la rappresentazione della conoscenza, ma anche comportamenti, come la capacità di decidere o la reazione a stimoli esterni, e in taluni casi anche emozioni,

[1] P.K. Dick, *Do Androids Dream of Electric Sheep?*, Doubleday and Company, New York, 1968.

[2] P. Moro, C. Sarra (a cura di), *Tecnodiritto. Temi e problemi di informatica e robotica giuridica*, FrancoAngeli, Milano, 2017.

[3] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", in *Minds & Machines*, 28 (2018), n. 4, pp. 689-707.

attraverso le espressioni del volto sintetico o altre azioni comunicative[4].

È arrivata con inattesa sollecitudine l'era delle macchine come noi[5]: come è già stato illustrato in un precedente lavoro[6], l'informatica e la robotica permettono oggi di costruire *robot* anche immateriali che possiedono non solo l'abilità di calcolare, ossia di elaborare dati in base al modello computazionale come ogni *computer*, ma anche la capacità di decidere, attraverso la retroazione (*feedback*) o l'apprendimento automatico (*machine learning*), e anche, in certe condizioni, la facoltà di manifestare percezioni simili alle emozioni.

Tuttavia, i problemi etici e giuridici posti dall'intelligenza artificiale, talora erroneamente vista come un apparato univoco o un dispositivo indipendente, riguardano principalmente l'interazione umana con i sistemi robotici evoluti, i quali appaiono sempre più capaci di emulazione autonoma delle facoltà proprie del soggetto umano, del quale influenzano il modo di vivere ma dal cui controllo restano sempre più emancipati[7].

Questa rappresentazione meccanica dell'anima umana, riflessa come in uno specchio da un sistema non soltanto materiale (*hardware*) ma anche immateriale (*software*), sempre più simile alla persona e sempre più capace di interagire con essa, ripropone la questione filosofica del discusso paragone tra coscienza umana e intelligenza artificiale e ripropone temi e problemi di etica, legalità e tecnologia, che sono studiati e discussi nell'informatica giuridica contemporanea[8].

[4] P. Moro, "Biorobotics and fundamental rights. Issues and limits of artificial intelligence", in D. Provolo, S. Riondato, F. Yenisey (eds.), *Genetics Robotics Law Punishment*, Padova University Press, Padova, 2014, pp. 517-531.

[5] I. McEwan, *Machines like me*, Jonathan Cape, London, 2019.

[6] P. Moro, "Macchine come noi. Natura e limiti della soggettività robotica", in U. Ruffolo (a cura di), *Intelligenza artificiale. Il diritto, i diritti, l'etica*, Giuffrè, Milano, 2020, pp. 45-62.

[7] Q. Yang, M. Wooldridge (ed.), *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 25-31 July 2015, AAAI Press, Palo Alto, California, USA, in http://ijcai.org/papers15/contents.php

[8] P. Moro (a cura di), *Etica, diritto e tecnologia. Percorsi dell'informatica giuridica contemporanea*, FrancoAngeli, Milano, 2021.

In particolare, non è peregrino ribadire che la macchina intelligente, sempre che sia dotata di facoltà artificiali analoghe, se non superiori, a quelle umane, possa essere titolare di una sorta di autodeterminazione individuale e, quindi, anche di posizioni giuridiche soggettive simili a diritti e doveri: infatti, la capacità di apprendimento del sistema evoluto di intelligenza artificiale è interattiva e aumenta in modo progressivo quando opera con l'essere umano[9].

Un esempio plastico di questa interazione è l'architettura tecnologica dei più evoluti *chatbot*, ossia di programmi cibernetici che realizzano assistenti virtuali progettati per simulare conversazioni umane e utilizzati non solo per supporto tecnico o servizi informativi (come quelli basati su regole predefinite), ma anche per generare risposte più flessibili e contestualmente adeguate, al fine di rendere le interazioni più naturali e coinvolgenti per gli utenti.

Il più noto e diffuso sistema di questo tipo è stato prodotto e reso accessibile al pubblico della rete Internet dalla società commerciale statunitense Open AI con il nome di *ChatGPT* (acronimo di Generative Pre-trained Transformer), sistema esperto di apprendimento automatico (*machine learning*), allenato su vaste quantità di testo e fondato su un ampio modello di linguaggio (*Large Language Model*) capace di produrre testo autonomo e sofisticato, con l'obiettivo di comprendere ed elaborare le complesse strutture semantiche della lingua.

Un elemento innovativo di questa tipologia di *chatbot* risiede nella modalità con cui i sistemi di intelligenza artificiale gestiscono il trattamento del linguaggio naturale, includendo funzioni quali la risposta a interrogativi, la traduzione automatica, l'interpretazione dei testi e la loro sintesi. Queste capacità non sono sviluppate mediante un processo di apprendimento supervisionato basato su collezioni di dati specifiche per ciascuna attività. Piuttosto, questi sistemi acquisiscono autonomamente tali abilità, senza bisogno di direttive specifiche, attraverso un pro-

[9] P. Moro, "Libertà del robot? Sull'etica delle macchine intelligenti", in R. Brighi, S. Zullo (a cura di), *Filosofia del diritto e nuove tecnologie. Prospettive di ricerca tra teoria e pratica*, Atti del XXIV Congresso della Società Italiana di Filosofia del Diritto, Aracne, Roma, 2015, pp. 525-544.

cesso di formazione ininterrotto e graduale su volumi crescenti di informazioni, attingendo da milioni di pagine disponibili sulla rete Internet[10].

Queste domande, che toccano anche il dibattito scientifico sull'interazione tra soggetto umano e persona digitale, possono essere esplorate attraverso un'analisi critica che oltrepassa la pura tecnologia dei sistemi cibernetici, attualmente definiti in modo generico "intelligenze artificiali", in base a princìpi di logica matematica e informatica.

Invece, si dovrebbe considerare la visione antropologica che giustifica questo modo di intendere la tecnologia. Molti ricercatori nelle discipline naturali, etiche e sociali continuano a confrontare l'essere umano con la macchina, spesso seguendo le teorie del post-umanesimo contemporaneo[11]: tali concezioni promuovono una comprensione limitata e controversa della natura umana, secondo un'idea che affonda le radici nel razionalismo moderno e che è ancora prevalente nel pensiero scientifico attuale.

Infatti, attraverso l'uso di tecnologie avanzate e scienze sempre più raffinate, l'uomo moderno cerca da tempo di creare una sorta di "anima artificiale" nelle macchine, che non soltanto non possiedono un corpo biologico, ma ne sono sempre più spesso prive e sono gestite soltanto da *software*. Questo atteggiamento epistemologico indica una chiara direzione verso un obiettivo ben definito che la scienza cerca di raggiungere nei sistemi meccatronici, cercando di emulare le principali funzioni psichiche umane che, all'origine della filosofia occidentale, Platone suddivide nella nota tripartizione dell'anima tra ragione, volontà e passione[12].

Platone sostiene che l'equilibrio di queste tre facoltà spirituali riflette la giustizia nelle relazioni sociali e raccomanda di mantenere uniti questi aspetti dell'anima per comprendere appieno la natura umana.

[10] A. Radford, J. Wu, R. Child., D. Luan, D. Amodei, I. Sutskever, "Language Models are Unsupervised Multitask Learners", in *OpenAI blog*, 2019.

[11] R. Peppered, "The Posthuman Manifesto", in *Kritikos*, (2005), n. 2, pp. 1-16.

[12] Platone, *Repubblica*, 439 a-441 c, in *Opere,* vol. II, Laterza, Bari, 1967, pp. 262-265.

In contrasto con questa visione olistica, per ragioni di opportunità politica o di profitto economico, l'era tecnologica e postmoderna si sta concentrando sull'amplificazione nelle macchine di intelligenza, volontà ed emozione in una sequenza esponenziale, seguendo un approccio individualistico e frammentato che riduce l'uomo ad un'unica dimensione[13].

2. Congegni intelligenti

L'intelligenza artificiale non è un fenomeno nuovo. La locuzione nasce dopo che alcuni matematici del secolo scorso considerano appropriato raffrontare la natura dei processi mentali di elaborazione e apprendimento delle informazioni alla struttura biologica dei processi naturali e sociali dell'attività fisiopsichica, al fine di realizzare il sogno di costruire una "macchina per pensare".

Nel 1936, Alan Turing progetta tale dispositivo, analizzando il processo razionale di calcolo e riferendosi agli "stati mentali" per definire le configurazioni interne della macchina e alla costruzione di un "cervello" per delineare l'esecuzione del suo progetto[14].

Nel 1948, Norbert Wiener collega espressamente la ricerca sugli esseri viventi a quella sulle macchine per garantire un adeguato controllo su entrambi e sostiene l'impossibilità teorica di distinguere il sistema nervoso umano da un sistema di comunicazione meccanico come quello telefonico[15].

In questo contesto storico e scientifico nasce la metafora "intelligenza artificiale".

Il 31 agosto 1955 negli Stati Uniti, durante gli studi della scienza dei calcolatori elettronici (*computer science*) e della cibernetica (*cybernethics*), lo scienziato nordamericano John McCarthy propone ad alcuni colleghi (come Marvin Minsky,

[13] H. Marcuse, *One-Dimensional Man. Studies in the Ideology of Advanced Industrial Society*, Beacon Press, Boston, 1964.

[14] A.M. Turing, "On computable numbers with an application to the Entscheidungsproblem", in *Proceedings of the London Mathematical Society*, 42 (1936), pp. 230-265.

[15] N. Wiener, *Cybernetics, or control and communication in the animal and the machine*, Hermann, J. Wiley, New York-Paris, 1948.

Nathaniel Rochester e Claude Shannon) di designare con il sintagma *artificial intelligence* il titolo di un progetto estivo di ricerca relativo ad un modello computerizzato della mente umana, che sviluppava i presupposti della cibernetica di Wiener[16].

Nel *paper proposal* del 1955 sono già presenti i tratti essenziali dell'intelligenza artificiale contemporanea: automazione e simulazione delle funzioni del cervello umano; utilizzo del linguaggio naturale; creazione di reti neuronali; teoria della dimensione del calcolo; auto-miglioramento delle macchine; astrazioni dai dati sensoriali; casualità e creatività.

Come afferma successivamente Pat Hayes nel suo *naive physics manifesto*[17], ciò che si vuole rappresentare con le applicazioni dell'intelligenza artificiale è il modo in cui le persone vedono il mondo, non il mondo come viene visto dalla fisica, con la conseguente necessità di elaborare modelli che riguardano non solo oggetti od eventi, ma anche percezioni e azioni inerenti alle modalità di comprensione del mondo da parte delle persone.

Oggigiorno, questa comparazione scientifica tra intelligenza artificiale e ragione umana presuppone l'adesione alla tesi dell'intelligenza artificiale "forte" ed appare inevitabilmente fondata sull'assioma concettuale della cibernetica, secondo cui la macchina per pensare emula la ragione che calcola. Infatti, lo sforzo di innalzare il dispositivo tecnologico dotato di intelligenza artificiale al livello delle facoltà umane si realizza attraverso operazioni di formalizzazione che, in sostanza, si traducono nella costruzione di algoritmi informatici che governano ed accelerano il presunto ragionamento e la conseguente azione del sistema di intelligenza artificiale.

Il procedimento calcolante custodisce la sua base computante anche nei sistemi informatizzati, che costituiscono elaboratori evoluti della tecnica combinatoria sviluppata dai logici più importanti dell'epoca barocca, come Pascal e Leibniz, per con-

[16] J. McCarthy, M.L. Minsky, N. Rochester, C.E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955", in *AI Magazine*, 27 (2006), n. 4, p. 12.

[17] P.J. Hayes, "The naive physics manifesto", in *Expert Systems in the Micro-Electronic Age*, Edinburgh University Press, Edinburgh, 1978.

nettere aritmetica e meccanica e fondare così nell'aurora della scienza moderna un'autentica *mathesis universalis*[18].

Questa concezione si mostra nell'alveo teoretico del razionalismo, secondo il quale l'uomo sarebbe capace di conoscere ciò che è bene e ciò che è male attraverso l'uso della propria ragione, e ha contribuito a fondare i programmi di intelligenza artificiale sul metodo scientifico[19], riconducendo a schemi comuni di origine algoritmica le macchine e i sistemi biologici capaci di esibire un comportamento finalizzato, confermando il presupposto matematico della cibernetica[20].

Tuttavia, la rappresentazione informatizzata della conoscenza si sta realizzando in determinati sistemi di intelligenza artificiale soprattutto con l'apprendimento automatico (*machine learning*), favorito dagli studi di statistica e delle neuroscienze, perché basato su molteplici esempi (denominati *big data*) tratti dal grande archivio di Internet e sull'elaborazione di sistemi di intelligenza artificiale (denominati reti neurali) simili al cervello umano.

Questo approccio statistico, basato sui dati elaborati con le tecnologie di *machine learning* e di ottimizzazione stocastica, ha sostituito la logica deduttiva, trasformando l'inferenza intelligente nel modello inverso, risolto dalla massimizzazione induttiva di una certa quantità probabilistica[21].

Intelligenza artificiale e rete Internet sono così diventate le più importanti tecnologie convergenti dell'epoca contemporanea, aumentando il loro predominio in via esponenziale, perché interagiscono tra loro e si implicano reciprocamente, pur sempre in dipendenza di algoritmi di apprendimento automatico: questa soggezione permane sia nel caso di algoritmo supervisionato, ossia quando il sistema (per esempio, la rete neurale) apprende

[18] W. Schmidt-Biggemann, *Topica universalis. Eine Modellgeschichte humanistischer und barocker Wissenschaft*, Meiner, Hamburg, 1983, pp. 156 ss.

[19] D.A. Gillies, *Artificial intelligence and scientific method*, Oxford University Press, Oxford, 1996, trad. it., *Intelligenza artificiale e metodo scientifico*, Cortina, Milano, 1998.

[20] G. Tamburrini, *I matematici e le macchine intelligenti. Spiegazione e unificazione nella scienza cognitiva*, Mondadori, Milano, 2002.

[21] N. Cristianini, "On the current paradigm in artificial intelligence", in *AI Communications* 27 (2011), pp. 37-43.

il nesso che unisce dati particolari e ipotesi generali, imparando a classificare esempi simili, sia nel caso di algoritmo per rinforzo, ossia quando il sistema interagisce con l'ambiente ed esegue gli esempi di comportamenti più vicini e simili al risultato, che quindi migliorano l'azione da eseguire, evitando quelli più distanti e diversi, che sono considerati errati.

3. Congegni morali

I sistemi dotati di intelligenza artificiale non sono più soltanto macchine calcolanti o *robot* che sostituiscono parti ripetitive del lavoro umano, come indica l'etimologia del termine, che deriva dal ceco *robota* e che è stato introdotto dallo scrittore Karel Čapek nel dramma fantascientifico R.U.R.[22].

Infatti, i dispositivi più sofisticati ed evoluti governati dall'intelligenza artificiale appaiono anche entità agenti o congegni morali, in grado di deliberare e di assumere decisioni che, osservate dall'esterno, appaiono simili a manifestazioni di volontà soggettiva, soggette a regole di comportamento elaborate dalla roboetica e fondate su valori scientifici, culturali e tecnici universalmente condivise e dirette alla tutela del benessere della persona umana[23].

È difficile ammettere l'attribuzione di diritti e doveri alle macchine intelligenti, sebbene queste operino con sempre maggiore autonomia in analogia con il modo di pensare e di agire dell'uomo, tanto che (come documenta l'approvazione in Europa del nuovo AI Act), appare ormai in corso di evoluzione la regolazione della specifica responsabilità dei sistemi cibernetici, richiesta dalla discussione avviata negli ultimi anni nella robotica e nella bionica[24] e considerata urgente ed indifferibile nella tecnoetica, soprattutto al fine di preservare la tutela e l'integrità

[22] K. Čapek, *Rossumovi univerzální robot*, Praha, 1920, trad. it., *R.U.R. Rossum's Universal Robots*, a cura di A. Catalano, Venezia, Marsilio, 2015.

[23] G. Veruggio, "The birth of roboethics", in *Institute of Electrical and Electronics Engineers International Conference on Robotics and Automation,* Workshop on Roboethics, Barcellona, 2005.

[24] A. Montanari, "Questioni di tecnoetica in intelligenza artificiale, robotica e bionica", in P. Moro (a cura di), *Etica informatica diritto*, FrancoAngeli, Milano, 2008, pp. 33-50.

delle persone attraverso l'imposizione di limiti etici e giuridici alle applicazioni della ricerca tecnologica[25].

Peraltro, il *robot* che interagisce con l'uomo in modo autosufficiente deve affrontare una serie di dilemmi morali che impongono alla macchina intelligente la necessità di deliberare una scelta fra più alternative, come, per esempio, costringere un anziano ad assumere un farmaco.

Introducendo il concetto di "macchine morali" per designare questi sistemi particolarmente sviluppati, è stato osservato che l'implementazione di regole morali o giuridiche che permettono al *robot* di prendere decisioni consapevoli e comportarsi in modo giusto oppure sbagliato può essere programmata con algoritmi di istruzione ma anche di apprendimento dalla condotta che l'automa pone in essere, ma si è precisato giustamente che tale questione supera i confini di una forma controllabile di intelligenza artificiale[26].

Inoltre, allo stato attuale, la possibilità di stabilire un'analogia tra stato mentale e logica della macchina avviene solo sul piano dell'accettabilità dei risultati e non della identità delle procedure di ragionamento e di decisione: tali metodologie sono propriamente umane e sono radicalmente differenti nel calcolatore che, tuttavia, può arrivare ai medesimi esiti dimostrativi[27].

In ogni caso, se il *robot* diventa sempre più autonomo nell'esecuzione dei propri compiti a causa dell'accelerazione dei processi autonomi di *machine learning*, come può accadere ad una macchina cibernetica impiegata per scopi militari oppure per attività di polizia, esso può diventare centro d'imputazione, del tutto analogo all'uomo, di atti responsabili per gli eventuali danni provocati dal loro comportamento, con la conseguenza che gli agenti intelligenti (che sono composti di *hardware* e di *software*) potrebbero diventare titolari di diritti e doveri simili a quelli delle persone e, dunque, entità penalmente imputabili.

[25] A. Fabris, G. Tamburrini, "Robot al di qua del bene e del male", in *Vita e Pensiero*, (2006), n. 3, pp. 85-93.

[26] W. Wallach, C. Allen, *Moral machines. Teaching robots right from wrong*, Oxford University Press, Oxford, 2009.

[27] S. Crafa, "Artificial Intelligence and Human Dialogue", in *Journal of Ethics and Legal Technologies*, (2019), n. 1, pp. 4456.

Invero, è evidente la problematicità semantica del linguaggio che si utilizza per tentare di spiegare questa dimensione "soggettiva" dell'intelligenza artificiale con categorie antropologiche ormai cristallizzate nella storia e nel pensiero occidentale, quali quelle di autonomia, volontà, responsabilità: si tratta di un lessico del tutto insufficiente ad illustrare adeguatamente una presunta libertà di agire di una entità meccanica. La responsabilità delle macchine intelligenti deve essere ripensata con le nuove categorie del tecnodiritto, come è già stato notato nella dottrina italiana maggiormente attenta al fenomeno, sia nel settore della responsabilità civile[28] che in quello della responsabilità penale[29].

La costruzione del cosiddetto *imitation game* da parte di Alan Turing, secondo il quale un sistema artificiale può essere definito intelligente se riesce a convincere una persona che il suo comportamento intellettuale è quello di un essere umano medio, è indubbiamente condizionata da un'interpretazione della coscienza e della volontà della macchina basata sugli effetti della condotta reattiva agli impulsi esterni: tale interpretazione deriva tuttora dalla psicologia comportamentale di John Watson, secondo cui la condotta esplicita sarebbe l'unica unità di analisi scientificamente studiabile, in quanto direttamente osservabile, mentre l'introspezione non potrebbe fornire alcun dato affidabile[30].

La costruzione antropomorfica della scienza informatica odierna, che assegna alle operazioni effettuate da sistemi informatici o cibernetici la nozione di volontà come reazione della macchina (*output*) ad un impulso dell'utente (*input*), condiziona l'approccio riduzionista della "società della mente" di Marvin Minsky, secondo il quale il cervello è una macchina altamente complessa paragonabile ad una società organizzata da una rete di interconnessioni gerarchiche e competitive tra una molteplici-

[28] U. Ruffolo, "Per i fondamenti di un diritto della robotica self-learning, dalla machinery produttiva all'auto driverless: verso una responsabilità da algoritmo?", in Id., *Intelligenza Artificiale e Responsabilità*, Giuffrè, Milano, 2018, pp. 1-30.

[29] S. Riondato, "Robot: talune implicazioni di diritto penale", in P. Moro, C. Sarra (a cura di), *op. cit.*, pp. 85-98.

[30] J.B. Watson, "Psychology as a behaviorist view", in *Psychological review*, 20 (1913), n. 2, pp. 158-177.

tà di componenti fortemente diversi fra loro, denominati appunto "agenti della mente"[31].

Occorre superare un'interpretazione semplificata e positivista dell'essere umano, riconoscendo che la volontà umana che il *robot* riproduce non è una decisione di agire basata sulla reazione ad impulsi esterni, ma una tensione verso un fine indisponibile che indica una qualità sostanziale dell'uomo nel suo intero. Dunque, prima di implementare la volontà umana nella macchina e sempre che tale operazione possa essere ammissibile e anche possibile, sarebbe necessario definire più compiutamente che cosa la volontà sia effettivamente, senza limitare l'indagine dell'essenza del soggetto all'analisi biologica delle sue facoltà psicologiche, come sembra accadere nella dominante prospettiva neuroscientifica.

Il pensiero e l'azione, così come l'intelligenza e la volontà, sono forme di vita, ma tali forme sono originarie e trovano il proprio fine nel loro stesso manifestarsi, sicché dimostrano di possedere una struttura anche soltanto biologica profondamente diversa da quelle che sembrano oggi ricostruibili con i sistemi cibernetici, come si desume dal sistema metabolizzante del corpo vivente, nel quale il metodo di produzione e di scambio di energia è anche il risultato dell'attività metabolica[32].

4. Congegni emotivi

In un terzo aspetto emulativo delle facoltà psichiche umane, gli studi e le applicazioni dell'intelligenza artificiale hanno già da tempo cercato di progettare e realizzare congegni emotivi, governati da software in grado di riconoscere ed esprimere emozioni (*affective computing*)[33].

L'interazione umana con i sistemi di intelligenza artificiale sta portando allo sviluppo di macchine affettive (*sentiment ma-*

[31] M.L. Minsky, *The Society of Mind*, Simon and Schuster, New York, 1986.

[32] H. Jonas, "Ist Gott ein Mathematiker? Vom Sinn des Stoffwechsels", in *Organismus und Freiheit. Ansätze zu einer philosophischen Biologie*, Vandenhoeck & Ruprecht, Göttingen, 1973.

[33] R.W. Picard, *Affective computing*, MIT Press, Cambridge (MA), 1997.

chines) che valutano la reazione emotiva dell'utente e interagiscono con il medesimo, sollevando alcune problematiche etiche e giuridiche in materia di riservatezza e di autodeterminazione del singolo.

Le *sentiment machines* rappresentano l'avanguardia tecnologica dell'intelligenza artificiale, che esplora le intersezioni tra macchine calcolanti ed emozioni umane, e sono il risultato di una ricerca che mira a sviluppare sistemi cibernetici capaci di riconoscere, interpretare e, in alcuni casi, replicare le emozioni umane.

Un aspetto fondamentale degli studi sui computer affettivi è la capacità delle macchine di percepire e reagire alle emozioni umane in modo appropriato, con l'esigenza di realizzare avanzamenti tecnici in termini di riconoscimento facciale o elaborazione del linguaggio naturale ma anche con la necessità di affrontare sfide etiche e filosofiche su natura e limiti degli interventi sulle esperienze emotive umane.

Per esempio, nelle applicazioni di marketing finanziario e di comunicazione politica, ha assunto peculiare importanza l'analisi computazionale delle emozioni (*sentiment analysis*) basata sulla ricerca ed esame delle opinioni (*opinion mining*) pubblicate sulla rete Internet e reperibili principalmente sui maggiori social network[34].

Inoltre, la percezione emotiva degli esseri umani è studiata nel campo della robotica sociale e dell'assistenza sanitaria per programmare "robot di compagnia" in modo che possano simulare le emozioni nelle loro interazioni con gli esseri umani, con evidenti conseguenze non solo etiche e sociali, ma anche giuridiche in materia di dignità umana e diritti fondamentali[35].

Gli esempi dimostrano la rilevanza di "macchine emotive" particolarmente progredite e l'impatto giuridico che potrà avere su tali sistemi il nuovo AI Act nel territorio dell'Unione Europea. Infatti, basti rammentare che l'articolo 5 del nuovo rego-

[34] B. Liu, "Sentiment analysis and opinion mining", in *Synthesis Lectures on Human Language Technologies*, 5 (2012) n. 1, pp. 1-167.

[35] L. Devillers, "Human-Robot Interactions and Affective Computing: The Ethical Implications", in J.S. von Braun, M. Archer, G.M. Reichberg, M. Sánchez Sorondo (eds.), *Robotics, AI, and Humanity*, Springer, Cham, 2021, pp. 205-211.

lamento europeo considera vietate pratiche di manipolazione del comportamento attraverso tecniche subliminali o di sfruttamento delle vulnerabilità di minori o disabili o di profilazione illegittima; attribuzione di un punteggio sociale basato sull'AI con trattamento sfavorevole per le persone da parte di autorità pubbliche; ricorso a sistemi di identificazione biometrica remota "in tempo reale" in spazi accessibili al pubblico a fini di attività di contrasto, salvo talune eccezioni.

La capacità degli attuali sistemi di intelligenza artificiale di esprimere emozioni si fonda sull'abilità di imitare le espressioni del volto umano attraverso algoritmi di apprendimento automatico e tecniche di *deep learning*. Questi "computer affettivi" possono analizzare le espressioni facciali elaborando rapidamente miliardi di immagini, grazie all'accesso a enormi quantità di dati provenienti da diverse fonti, e si affiancano a modelli computazionali avanzati che imitano aspetti dell'intelligenza emotiva attraverso la creazione di reti neurali artificiali che replicano alcuni aspetti delle emozioni, come se fossero capaci di "sentire".

Tuttavia, nonostante questi progressi, è noto che ogni sistema di intelligenza artificiale resta confinato ad elaborare ed eseguire compiti specifici, come la combinazione di parole e il riconoscimento di immagini o suoni, ma non riesce ancora a simulare una forma di intelligenza generale che richiederebbe un'interazione complessa di funzioni cognitive coordinate e dirette alla gestione di più azioni nello stesso momento.

Infatti, le ricerche neuroscientifiche documentano che anche alle macchine affettive manca il senso dell'intero, come dimostrano gli studi sulle relazioni tra intelligenza artificiale e le dinamiche *whole brain*, che riguardano l'interazione tra aree dell'intero cervello che non sono indipendenti ma che comunicano tra loro e partecipano a un sistema di relazioni complesso e articolato[36].

In ogni caso, anche le macchine emotive fondano la propria struttura tecnologica sulle acquisizioni dell'intelligenza artificiale ed appaiono influenzate dalla visione critica della natura

[36] H. Mizutani, M. Ueno, N. Arakawa, H. Yamakawa, "Whole brain connectomic architecture to develop general artificial intelligence", in *Procedia Computer Science*, 123 (2018), pp. 308-313.

umana che le neuroscienze hanno proposto nel dibattito degli ultimi anni, opponendosi alla tradizionale concezione dualistica del soggetto, di origine cartesiana, che contrappone le facoltà intellettive alla struttura della materia e che appiattisce il soggetto umano nel campo dei fenomeni[37].

5. Congegni inconsapevoli

Nel presente lavoro si è notato che la crescita esponenziale degli studi e delle applicazioni dell'intelligenza artificiale sta percorrendo un itinerario emulativo delle facoltà umane di ragionare, di decidere e di emozionarsi.

Questa imitazione delle principali qualità psichiche da parte delle macchine, guidata da scienziati e programmatori in una sequenza quasi pedagogica, sembra l'esito narcisistico di una prospettiva antropocentrica e riduzionistica del soggetto umano, che ricompone le proprie funzioni con la fabbricazione di frammenti digitali del soggetto.

È il risultato della dogmatica e mai adeguatamente superata rappresentazione meccanica dell'intelligenza artificiale, frutto del positivismo cibernetico del secolo scorso, con la conseguenza che l'evoluzione dei *robot* continua ad essere un processo di sviluppo di alcune facoltà soggettive che sono ritenute univoche, come la ragione calcolante, la volontà reattiva o l'emozione oggettivata, ma che non costituiscono gli autentici e compiuti modi di definire queste caratteristiche della natura umana.

In effetti, la coscienza non è riducibile ad una sola delle sue funzioni e, dunque, il tentativo di paragonare la mente umana all'intelligenza artificiale è destinato a fallire: la coscienza è una proprietà dell'intera persona e deve essere considerata nella totalità che contraddistingue l'intelligenza umana, che non può essere riprodotta, quantomeno nella sua completezza, attraverso modelli computazionali anche molto evoluti, giacché essa genera un problema che non è computabile[38].

[37] S. Fuselli, *Diritto, neuroscienze, filosofia. Un itinerario*, FrancoAngeli, Milano, 2014.

[38] F. Faggin, *Irriducibile. La coscienza, la vita, i computer e la nostra natura*, Mondadori, Milano, 2022.

Forse l'intelligenza artificiale è un modo di pensare, ma resta una forma della ragione calcolante essenzialmente ridotta al codice binario che governa tuttora la scienza informatica. Invece, l'uomo è dotato di una mente colorata, come ricorda Omero a proposito del multiforme ingegno di Ulisse[39], e vive di plurime dimensioni della soggettività che, come ha accertato la psicanalisi del Novecento, provengono anche dall'inconscio: come osserva Freud, la vita psichica si svolge in modalità performative anche nelle sue forme secondarie – *lapsus*, dimenticanze, sogni – che indicano un'organizzazione alternativa della razionalità[40] e che sono del tutto ignote all'intelligenza meccanica che, quando sbaglia, esegue un'operazione che è giudicata falsa in base all'algebra di Boole e, dunque, non commette affatto un *lapsus*.

Dunque, l'equivalenza tra mente umana e cervello elettronico è inammissibile, perché le macchine calcolanti sono inconsapevoli. Infatti, applicando gli algoritmi di programmazione che ogni *computer* deve comunque eseguire, il dispositivo cibernetico elabora un ragionamento meccanico senza comprendere ciò che dice o ciò che fa.

Tutte le logiche di tipo formale, che sono largamente utilizzate nelle macchine intelligenti, si manifestano tramite un lessico e una sintassi di tipo rigidamente simbolico, analizzabile seguendo regole di inferenza che hanno natura algoritmica, deduttiva o induttiva. La comprensione del linguaggio e l'interpretazione delle regole inferenziali da parte dell'operatore che ne fa uso è indipendente dallo sviluppo e dall'esito del procedimento inferenziale.

Nel suo famoso esperimento mentale, noto come "stanza cinese" ed illustrato per criticare i sostenitori dell'intelligenza artificiale "forte", John Searle ha contestato che la mente possa essere un programma, perché il *computer* non capisce ciò che sta facendo quando elabora un'operazione mentale attraverso il linguaggio simbolico, essendovi una distanza incolmabile tra sin-

[39] P. Citati, *La mente colorata. Ulisse e l'Odissea*, Mondadori, Milano, 2002.

[40] S. Freud, *Vorlesungen zur Einführung in die Psychoanalyse*, trad. it., *Introduzione alla psicoanalisi* [1915], in *Opere di Sigmund Freud*, VIII, Bollati Boringhieri, Torino, 2000.

tassi e semantica, ossia tra struttura grammaticale del linguaggio e comprensione significante delle operazioni combinatorie[41].

Le argomentazioni di Searle colgono il punto nevralgico di ogni teoria sull'intelligenza artificiale e, pur essendo tuttora oggetto di dibattito scientifico[42], consentono indubbiamente di escludere che la logica proposizionale booleana, spesso considerata un postulato del ragionamento informatizzato, possa realizzare un qualsiasi sistema complesso in grado di superare la morfologia sintattica del linguaggio per giungere alla sua piena comprensione semantica.

Nel dominio dell'informatica, l'incapacità di un sistema di intelligenza artificiale di riconoscere e superare i propri limiti è provata dalla presenza di problemi non computabili (o indecidibili) che non sono risolvibili dall'elaboratore elettronico. Tale scenario si manifesta nell'impossibilità di sviluppare un algoritmo di verifica per programmi che possa determinare, data una serie specifica di premesse, se un dato software terminerà il suo processo con un risultato logicamente coerente o se, al contrario, continuerà all'infinito a produrre un esito incoerente[43].

Pertanto, l'approccio computazionale, che rappresenta la base teorica per la ricerca e lo sviluppo nel campo dell'intelligenza artificiale, deve affrontare la specifica capacità del pensiero umano di autotrascendersi e di ridefinirsi, che rappresenta una delle caratteristiche fondamentali e irriducibili dell'intelligenza umana: questa facoltà soggettiva è inesauribile e imprevedibile, sicché appare difficilmente afferrabile dalla computazione[44].

6. Persone elettroniche

La possibilità di attribuire una vera e propria "soggettività robotica" ad una macchina particolarmente sofisticata, che pre-

[41] J.R. Searle, "Minds, brains, and programs", in *Behavioral and Brain Sciences*, (1980), n. 3, pp. 417-457.

[42] S. Harnad, "What's Wrong and Right About Searle's Chinese Room Argument", in M. Bishop, J. Preston (eds.), *Essays on Searle's chinese room argument*, Oxford University Press, Oxford, 2011.

[43] D. Harel, *Computers Ltd.: What They Really Can't Do*, Oxford University Press, Oxford-New York, 2000.

[44] F. Chiereghin, *L'eco della caverna. Ricerche di filosofia della logica e della mente*, Il Poligrafo, Padova, 2004, pp. 200-201.

senta forma materiale (*hardware*) o immateriale (*software*) e che agisce anche senza controllo umano, sta attraendo su di sé nozioni e metafore delle scienze matematiche ed informatiche del tutto differenti tra loro (algoritmo, intelligenza artificiale, *chat bot*) in base ad una visione individualista, che è tipica dell'epoca contemporanea e che sta influenzando non solo la cultura e la società, ma anche la più tipiche categorie dell'etica e del diritto dell'era moderna, come la responsabilità.

Poiché le nozioni di uomo e di macchina sono incommensurabili, è opportuno evitare di applicare ad un software cibernetico particolarmente evoluto l'idea antropocentrica della responsabilità soggettiva, talora presupposta nel pensiero tecnologico della postmodernità, valutando attentamente la rilevanza giuridica del passaggio di un'entità robotica da *res* a *persona*[45].

Tuttavia, la macchina dotata di intelligenza artificiale avanzata opera ormai attivamente nell'odierna "infosfera", nella quale il tutto viene prima del soggetto agente e l'uomo non è l'unico soggetto cui attribuire una dignità morale, diventando così vero e proprio "organismo informazionale" (*inforg*), le cui condotte, intenzionali o meno, possono produrre effetti sulle entità interconnesse[46].

Soprattutto nel suo approccio odierno, il sistema di intelligenza artificiale ottiene un'acquisizione di cognizioni e condotte che restano indipendenti dalla programmazione informatica e, quindi, dall'intervento umano[47], e costruisce automaticamente un modello dall'analisi dei dati su cui è addestrato, generando valutazioni e previsioni sui nuovi casi ad esso sottoposti in base al continuo aggiornamento dei dati medesimi, migliorando il modello predittivo attraverso l'ottimizzazione stocastica e la massimizzazione induttiva di una certa quantità probabilistica[48].

È indubbio che l'apprendimento automatico delle macchine può forse puntare a raggiungere l'autentico "algoritmo dominan-

[45] U. Ruffolo, "Il problema della 'personalità elettronica'", in *Journal of Ethics and Legal Technologies*, (2020), n. 1, pp. 75-88.

[46] L. Floridi, *La quarta Rivoluzione*, Raffaello Cortina, Milano, 2017.

[47] S.J. Russell, P. Norvig, *Artificial Intelligence. A Modern Approach*, Prentice Hall, Englewood Cliffs (N.J.), 2010.

[48] N. Cristianini, "On the current paradigm in artificial intelligence", in *AI Communications*, 27 (2011), pp. 37-43.

te"[49] e che gli esiti del modello creato dall'intelligenza artificiale e le ragioni che ne giustificano le decisioni non possono essere spiegati dal codice sorgente o dai risultati osservabili della procedura impostata, giacché il funzionamento interno del sistema resta oscuro agli stessi programmatori e può essere considerato una *black box*[50].

Infatti, tenendo presente il progressivo aumento di autosufficienza di sistemi sempre più prossimi a compiere attività sempre meno condizionata, l'articolo 3 numero 1) del nuovo AI Act dell'Unione Europea designa il "sistema di intelligenza artificiale" come "un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali".

Nel *considerandum* 12 del regolamento, si precisa che "la capacità inferenziale di un sistema di IA trascende l'elaborazione di base dei dati e consente l'apprendimento, il ragionamento o la modellizzazione" e che "l'adattabilità *che un* sistema di IA potrebbe presentare dopo la diffusione si riferisce alle capacità di autoapprendimento, che consentono al sistema di cambiare durante l'uso".

Peraltro, occorre considerare che la soggettività umana si distingue dal sistema di intelligenza artificiale, che può animare un dispositivo immateriale oppure un corpo sintetico, anche e soprattutto per la condizione biologica, che caratterizza specificamente l'esistenza umana[51].

Inoltre, il tratto originario e irripetibile della soggettività umana, che la distingue dall'intelligenza meccanica, è la rela-

[49] P. Domingos, *The Master Algorithm. How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, New York, 2015.

[50] F. Pasquale, *The black box society: The secret algorithms that control money and information*, Harvard University Press, Cambridge (MA), 2015.

[51] P. Benanti, *The Cyborg: corpo e corporeità nell'epoca del post-umano. Prospettive antropologiche e riflessioni etiche per un discernimento morale*, Cittadella Editrice, Assisi, 2012.

zione con l'altro da sé[52], la quale costituisce la finalità e la metodologia del diritto, che è un'attività di tutela e promozione della dignità umana[53]. Anzi, è stato giustamente osservato che, nell'ipercomunicazione dell'era tecnologica, la scomparsa del senso dell'Altro è il motivo ontologico per cui l'utilizzo continuo del dispositivo digitale accresce il vuoto e aggrava la solitudine dell'umanità contemporanea[54].

Tuttavia, è indubbio che sono sempre più rilevanti le interazioni sociali ed autonome degli agenti software dotati di intelligenza artificiale[55] e che la capacità di apprendimento del robot evoluto è interattiva (*interactive learning*) e incrementa progressivamente la propria autonomia non soltanto apprendendo dall'ambiente e dagli oggetti che incontra, ma anche cooperando con l'essere umano, come avviene nella robotica collaborativa[56]. L'impatto della regolazione algoritmica della società (per esempio, attraverso attività di *crowdsourcing* o di gestione della reputazione) ha anche suggerito la nozione di "macchina sociale", che genera evidenti conseguenze etiche e giuridiche per la tutela dell'autonomia individuale e dell'ordine pubblico[57].

Pertanto, non è più azzardato ipotizzare che la macchina intelligente, presuntivamente dotata di facoltà artificiali analoghe, se non superiori, a quelle umane, possa essere titolare di una specie di autodeterminazione individuale e, quindi, di una sorta

[52] G. De Bona, "Verso la tutela giuridica dei sistemi intelligenti. Prospettive critiche della soggettività robotica", in *Journal of Ethics and Legal Technologies*, 4 (2022) n. 2, pp. 51-104.

[53] A.C. Amato Mangiameli, "IA e diritto. In luogo di una introduzione", in *Journal of Ethics and Legal Technologies*, 5 (2023), n. 2, pp. 81-94.

[54] Byung-Chul Han, *Undinge: Umbrüche der Lebenswelt*, Ullstein Verlag, Berlin, 2021, trad.it., *Le non cose. Come abbiamo smesso di vivere il reale*, Einaudi, Torino, 2022, p. 39.

[55] G. Teubner, "Die anonyme Matrix: Zu Menschenrechtsverletzungen durch ‚private' transnationale Akteur", in *Der Staat*, 45 (2006), n. 2, pp. 161-187.

[56] M.A. Goodrich, A.C. Schultz, "Human-Robot Interaction: A Survey", in *Foundations and Trends in Human-Computer Interaction*, (2007), n. 3, pp. 203-275.

[57] N. Cristianini, T. Scantamburlo, "On social machines for algorithmic regulation", in *AI & Society,* Springer, Verlag, 35 (2020), pp. 645-662.

di libero arbitrio, come ha sostenuto John McCarthy[58]. Infatti, se tale qualificazione fosse accettabile, occorrerebbe rivedere l'idea antropocentrica dell'intelligenza artificiale e prendere atto che l'agente robotico particolarmente sofisticato potrebbe acquisire una certa forma di libertà soggettiva e, quindi, una pur imprecisata titolarità di diritti e di doveri[59] oppure di una certa capacità di agire, sempre che si ammetta l'applicabilità di tali categorie ad una macchina[60].

È da notare che la programmazione di decisioni consapevoli e l'applicazione di regole condivise può essere imposta al sistema robotico da algoritmi di istruzione e di apprendimento, ma tale questione supera i confini di una forma controllabile di intelligenza artificiale[61]. Difatti, le procedure formali di apprendimento automatico restano inferenze ipotetiche e non possono interpretare o risolvere problemi complessi della vita reale, come i conflitti di valore, che manifestano costante variabilità ed incertezza e che sono difficili da codificare attraverso misurazioni o funzioni numeriche[62].

In definitiva, quanto sopra esaminato suggerisce di riflettere senza pregiudizi sulla questione della soggettività robotica, che non è paragonabile alla soggettività umana, ma che potrebbe essere attribuita a sistemi evoluti e complessi di intelligenza artificiale, le cui limitazioni e tutele dovrebbero essere indicate non dalle categorie giuridiche tradizionali, ma dalla costituzione di una nuova finzione, utilizzando il linguaggio visivo della retorica.

In sostanza, per designare un sistema robotico di intelligenza artificiale che opera in modo simile a quello umano, pare del

[58] J. McCarthy "Free Will-even for Robots", in *Journal of Experimental and Theoretical Artificial Intelligence*, 12 (2000), n. 3, pp. 341-352.

[59] P. Moro, "Libertà del robot? Sull'etica delle macchine intelligenti", cit., pp. 525-544.

[60] E. Morotti, "Una soggettività a geometrie variabili per lo statuto giuridico dei *robot*", in F. Bilotta, F. Raimondi (a cura di), *Il soggetto di diritto. Storia ed evoluzione di un concetto nel diritto privato*, Jovene Editore, Napoli, 2020, pp. 291-306.

[61] W. Wallach, C. Allen, *op cit.*

[62] Z. Lipton, "The mythos of model interpretability", in *ICML Workshop on Human Interpretability*, 2016, arXiv:1511.03677.

tutto appropriato introdurre nel tecnodiritto contemporaneo una metafora simile a quella della persona giuridica (in inglese, *legal entity*), ben nota al linguaggio dei giuristi e inventata per risolvere problemi fondamentali del traffico commerciale e della responsabilità risarcitoria[63].

Infatti, ci si deve domandare se, come ha fatto per il momento il nuovo AI Act, potrà davvero essere trascurata dal legislatore europeo la proposta, contenuta nella Risoluzione del Parlamento europeo del 16 febbraio 2017, di costituire per i robot evoluti una innovativa finzione giuridica, assegnando la condizione di "persone elettroniche responsabili" a determinati sistemi di intelligenza artificiale.

La citata Risoluzione, recante raccomandazioni alla Commissione concernenti norme di diritto civile sulla robotica e basata sulla logica della sicurezza del prodotto commerciale e sulla gestione del rischio[64], contiene un esplicito invito al massimo organo legislativo europeo, e indirettamente gli Stati membri dell'Unione, "a esplorare, esaminare e valutare, nell'ambito della valutazione d'impatto del suo futuro strumento legislativo, le implicazioni di tutte le soluzioni giuridiche possibili, tra cui: (…) l'istituzione di uno status giuridico specifico per i robot nel lungo termine, di modo che almeno i robot autonomi più sofisticati possano essere considerati come persone elettroniche responsabili di risarcire qualsiasi danno da loro causato, nonché eventualmente il riconoscimento della personalità elettronica dei robot che prendono decisioni autonome o che interagiscono in modo indipendente con terzi".

Il documento legislativo del Parlamento Europeo, cui ha dato origine il percorso che ha portato all'approvazione del nuovo regolamento sull'intelligenza artificiale, è di evidente interesse per individuare il principio giuridico, peraltro tratto da una citazione letteraria, che risolva le probabili controversie tra uomini e robot: nelle premesse giustificative, infatti, la Risoluzione considera formalmente le quattro leggi di Asimov, tra le quali

[63] F. Galgano, *Le insidie del linguaggio giuridico. Saggio sulle metafore nel diritto*, il Mulino, Bologna, 2010.

[64] N. Busto, "La personalità elettronica dei *robot*: logiche di gestione del rischio tra trasparenza e fiducia", in *Ciberspazio e diritto*, 18 (2017) n. 3, pp. 499-511.

la fondamentale "legge zero" ("Un robot non può recar danno all'umanità e non può permettere che, a causa di un suo mancato intervento, l'umanità riceva danno")[65], "come rivolte ai progettisti, ai fabbricanti e agli utilizzatori di robot, compresi i robot con capacità di autonomia e di autoapprendimento integrate, dal momento che tali leggi non possono essere convertite in codice macchina".

Nel silenzio della legislazione, la costruzione di una nuova *fictio iuris* denominata "persona elettronica" oppure "entità elettronica" (in inglese *electronic person* oppure *electronic entity*) per designare la soggettività robotica potrebbe essere demandata alla giurisprudenza per consentire di risolvere alcune rilevanti questioni pratiche, come i conflitti provocati dall'interazione sociale tra sistemi di intelligenza artificiale e soggetti umani.

La natura e i limiti di questa finzione, che sta aprendo nuovi ed importanti orizzonti di ricerca dell'informatica giuridica contemporanea, dovranno essere certamente soppesate da riflessioni adeguate ai princìpi di ogni ordinamento sociale[66], ma potranno essere definite anche dal progresso tecnologico, la cui evoluzione porterebbe a forme graduali di riconoscimento della personalità elettronica, in base allo stadio di autonomia dei sistemi di intelligenza artificiale[67].

In ogni ipotesi, la finzione giuridica della "persona elettronica" non è affatto peregrina e appare del tutto compatibile con la dottrina secondo cui le categorie concettuali, che tradizionalmente identificano nell'ordinamento sociale la tutela e la responsabilità del soggetto umano, siano applicabili già ai calcolatori elettronici degli anni Ottanta del Novecento[68] e *a fortiori* ai

[65] I. Asimov, "The evitable conflict", Id., in *I, robot*, Gnome Press, New York, 1950.

[66] S.M.C. Avila Negri, "Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence", in *Frontiers in robotics and AI*, 8 (2021), 789327 (https://doi.org/10.3389/frobt.2021.789327).

[67] M. Farina, "Brevi riflessioni sullo status delle 'persone elettroniche'", in *L'Ircocervo*, 20 (2021), n. 2, pp. 106-126.

[68] A. Narayan, D. Perrott, "Can computers have legal rights?", in M. Yazdani, A. Narayanan (eds.), *Artificial Intelligence. Human Effects*, Ellis Horwood Series, Halsted Pr., New York, (1984), pp. 52-61.

sistemi di intelligenza artificiale del nuovo millennio[69], giustificando così una nuova, seria ed imprevista generazione di diritti e doveri, riferibili a persone umane e soggettività artificiali, che dovranno abitare insieme nel nuovo "tecnomondo".

[69] B.J. Koops, M. Hildebrandt, D.O. Jacquet-Chiffelle, "Bridging the Accountability Gap: Rights for New Entities in the Information Society?", in *Minnesota Journal of Law, Science & Technology*, (2010), pp. 497-561.

Guido Saraceni

Intelligenza Artificiale e benessere psicologico

Abstract: **Artificial Intelligence and Psychological Well-being.** The exploitation of Artificial Intelligence (AI) by digital platforms could seriously damage the psychological well-being of users, diminishing their cognitive/attentional resources, favoring their isolation and their dependence (IAD). On the other hand, AI proffers manifold benefits in the realm of mental health, such as providing timely interventions through virtual therapists and chatbots. Moreover, machine learning algorithms can discern subtle shifts in behavior and language indicative of mental health issues, facilitating prompt diagnosis and intervention. Hence, this essay delves into the intricate relationship between AI and mental health, accentuating both the potential advantages and the ethical, societal, and technical hurdles that ensue.

Keywords: Artificial Intelligence, Mental Health, Machine Learning, Legal Informatics, Philosophy of Technology.

Sommario: 1. Introduzione. L'importanza del linguaggio: da Aristotele a Turing – 2. Le nostre (limitate) risorse cognitive – 3. L'*Internet Addiction Disorder* (IAD). Una pandemia digitale – 4. Benessere psicologico e *Social Network* – 5. *Pars construens*. L'IA al servizio della psiche – 6. Conclusioni.

1. Introduzione. L'importanza del linguaggio: da Aristotele a Turing

L'uomo è un animale sociale perché ha il *logos*[1]. A scapito della sua estrema e lineare sinteticità, questo antico insegnamento aristotelico è colmo di conseguenze e di spunti di riflessione. Non a caso, la prima parte della frase rappresenta uno dei motti più celebri dell'intera storia della filosofia e viene spesso citata per ribadire la natura comunitaria ed interpersonale dell'essere umano, ovvero, la sua congenita necessità di entrare in rapporto con i propri simili, evitando l'isolamento e la solitudine, considerate, *secundum quid*, alla stregua di un atto autolesionistico. In tal senso, il pensiero dello Stagirita può essere collocato

[1] Aristotele, *Politica*, I, 1252 a-1253 a.

all'interno di una nutrita ed eterogenea tradizione culturale – tipicamente occidentale – che riconosce alla comunità un valore effettivamente *salvifico*, dato che essa consente all'essere umano di vivere in maniera *sana*, ovvero, conforme alla propria più intima natura.

Tuttavia, l'importanza dell'insegnamento aristotelico travalica ampiamente queste primissime ed introduttive linee interpretative, poiché alla considerazione circa la natura sociale dell'essere umano aggiunge una spiegazione che ne unisce in maniera chiarissima e diretta la vocazione interpersonale al *logos*. Così facendo, Aristotele non intendeva rimarcare l'importanza della ragione, né tantomeno, affermare la natura compiutamente razionale del genere umano, quanto puntare l'attenzione sul linguaggio, considerato alla stregua di uno strumento indispensabile, in grado di distinguere gli uomini dagli animali[2].

Proprio il linguaggio, così intimamente connesso alla natura (sociale) dell'uomo, rappresenta, da più di settant'anni, il vero e proprio elemento di discrimine in tema di Intelligenza Artificiale[3], poiché su di esso si basa il celebre "Test di Turing", atto ad attribuire o negare, in maniera certa e definitiva, capacità intellettive ai computer.

Anche per questo motivo l'intelligenza artificiale di ultima generazione – cosiddetta *intelligenza artificiale generativa* – sta suscitando grande interesse ed altrettanto clamore: per la prima volta nella storia abbiamo la chiara percezione di essere vicini al superamento del "Test di Turing". Piattaforme come *ChatGPT* e *Gemini* sono infatti in grado di dialogare correttamente con gli

[2] Come sottolinea J. Dewey, *Logica, teoria dell'indagine*, trad. it., Einaudi, Torino, 1949, pp. 59 ss., che commenta questo celebre motto attribuendo particolare importanza esattamente al linguaggio e al *mondo*, all'ambiente, culturalmente trasmesso, all'interno del quale vive l'essere umano.

[3] Il riferimento, scontato ed assolutamente necessario, è al celebre saggio di A.M. Turing, "Macchine Calcolatrici e Intelligenza", in Id., *Intelligenza Meccanica*, trad. it., Bollati Boringhieri, Torino, 1994, pp. 121-158; per una critica, si veda A.C. Amato Mangiameli, *Diritto e Cyberspace*, Giappichelli, Torino, 2000, pp. 81 ss.; S. Shieber, *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, MIT, Cambridge-London, 2003; S.J. Russell, P. Norvig, *Artificial Intelligence: a modern approach*, Pearson, London, 2021; S. Natale, *Macchine Ingannevoli. Comunicazione, tecnologia, intelligenza artificiale*, trad.it., Einaudi, Torino, 2022, pp. 21 ss.

utenti, emulando in maniera parecchio precisa la conversazione umana; inoltre, mostrano non trascurabili tratti di originalità, potendo apprendere, migliorare e (ri)creare testi – modulandone lo stile in base ai *prompt* ed agli *input* ricevuti.

Il *logos* – inteso come linguaggio e, ancor di più, come dialogo – assume dunque un ruolo di fondamentale importanza per la riflessione filosofica ed informatica. Parimenti, esso rappresenta un tema centrale della ricerca psicologica – in particolar modo per gli studi di *cyberpsicologia*[4]. Basti pensare che uno dei primissimi *chatterbot*[5] divenne famoso grazie ad uno *script* di (auto)aiuto psicologico basato, essenzialmente, su riformulazioni e domande aperte, ovvero, sulle più elementari basi teoretiche del counseling rogersiano[6]. Infine, il linguaggio costituisce la cellula prima ed assiomatica di qualsiasi studio o ragionamento.

Intendo dire che la premessa di una ricerca dovrebbe contenere la definizione dei suoi termini essenziali – in questo caso: "IA" e "benessere psicologico". In ambito filosofico, questo compito si rivela spesso arduo, se non addirittura, impossibile. Non solo perché la stessa filosofia è strutturalmente costretta a (ri)definire continuamente l'oggetto del suo studio, ma anche perché i termini di cui si occupa risultano non di rado polisemici, sottraendosi ad una definizione chiara e scientificamente inoppugnabile. Constatata l'impossibilità di sintetizzarne il senso in poche e squadrate parole, la cosa migliore da fare è provare ad illustrarne la problematicità.

Con riguardo al disagio mentale, notiamo che il DSM censisce una vasta gamma di disturbi, strutturalmente eterogenei

[4] M. Aiken, *The Cyber effect: A Pioneering Cyberpsychologist Explains How Human Behaviour Changes Online*, Coronet, London, 2017; V. Benson, J. MacAlaney (eds.), *Cyber Influence and Cognitive Threats*, Academic Press, New York, 2019; A. Atrill-Smith, *The Oxford Handbook of Cyberpsychology*, OUP, Oxford, 2020.

[5] J. Weizenbaum, "Eliza. A Computer Program for the Study of Natural Language Communication Between Man and Machine", in *Communications of the ACM*, 9 (1966), n. 1, pp. 36-45; S. Natale, *op. cit.*, pp. 79 ss.; R. Ciesla, *The Book of Chatbots: from Eliza to ChatGPT*, Springer, New York, 2024.

[6] C. R. Rogers, *Potere personale. La forza interiore e il suo effetto rivoluzionario*, trad. it., Astrolabio, Roma, 1978; Id., *Un modo di essere*, trad. it., Giunti, Firenze, 2012; Id., *La terapia centrata sul cliente*, trad. it., Giunti, Firenze, 2019.

e certamente non omologabili in termini di gravità, sintomatologia ed eziologia[7]. Alcuni tra essi implicano, ad esempio, che il malato provi dolore, o meglio ancora che avverta una *sofferenza* (sono particolarmente significativi, a tal riguardo, la schizofrenia o il disturbo ossessivo compulsivo) mentre in altri e altrettanto importanti casi si tratta di disagi perfettamente ego-sintonici (come, ad esempio, i più comuni disturbi della personalità); alcuni tra essi impediscono al soggetto di svolgere le normali attività quotidiane, danneggiandone gravemente il funzionamento sociale (esemplari, in tal senso, l'*hikikomori* o l'agorafobia), mentre altri non impediscono al malato di uscire di casa, lavorare, svagarsi, non influendo in alcun modo sulle sue *performance*; alcuni, essendo caratterizzati da allucinazioni, sensazioni di de-realizzazione e de-personalizzazione, allontanano l'individuo dal mondo, danneggiandone l'esame di realtà, mentre altri non ne diminuiscono la lucidità, permettendogli di restare razionale, perfettamente consapevole e ben orientato.

La caratteristica che più di tutte sembra accomunare le diverse ed eterogene forme di disagio mentale è certamente quella dello stigma sociale, ovvero, l'antica e sempre attuale tendenza della società a disprezzare, temere ed isolare chi soffre di problemi psichici. Purtroppo, questo ostracismo finisce quasi sempre per aggravare la malattia – della quale può essere, peraltro, una (con)causa e/o un effetto.

A prescindere dallo stigma sociale – che può assumere, è bene ricordarlo, diversi gradi di intensità e di diffusione – non sembrerebbe esistere altro elemento in grado di accomunare tutte le forme di disagio psichico. L'assenza di sintomi e di segni comuni rende parecchio difficile elaborarne una definizione precisa, netta e comunemente accettata[8]. A maggior ragione, non è

[7] APA, *Manuale diagnostico e statistico dei disturbi mentali*, Raffaello Cortina, Milano, 2023. Tra i disturbi statisticamente più comuni ricordiamo: i disturbi dell'umore – come, ad esempio, la distimia, la ciclotimia o la depressione; i disturbi d'ansia, come il disturbo d'ansia generalizzata o il disturbo da stress post-traumatico; i disturbi dell'alimentazione – come l'anoressia nervosa, la pica o la bulimia – ed i disturbi della personalità.

[8] La complessità del concetto risulta a mio avviso comprovata dal fatto che il DSM-5, nel definire il disturbo mentale, ricorre a ben due requisiti formulati in negativo, sottolineando che "una risposta prevedibile o cultu-

facile stabilire cosa sia il benessere psicologico – considerato che esso fa riferimento alla generica e complessiva condizione psichica in cui versa il soggetto, a prescindere dalla diagnosi di un vero e proprio disagio[9].

Quel che è certo è che sarebbe meglio evitarne una concezione dicotomica e manichea, quasi che esso fosse una questione di *have* or *have not*, riconducibile al rassicurante schematismo del codice binario, adottando, in suo luogo, un approccio *dimensionale*, consapevole dell'esistenza di un *continuum* che unisce lo stato psichico *ottimale*, in cui l'individuo sperimenta il completo e soddisfacente benessere, sociale e relazionale, alla presenza di sintomi lievi o transitori, fino alle manifestazioni più gravi – che solitamente interferiscono con la qualità della vita.

La situazione non si presenta come meno problematica se poniamo mente al concetto di IA. In via del tutto preliminare, potremmo affermare che essa si suddivide in tre principali categorie:

1) l'Intelligenza Artificiale Ristretta (ANI – *Artificial Narrow Intelligence*), progettata per eseguire un singolo compito o un insieme limitato di compiti; di conseguenza, non possiede capacità cognitive generali e non è in grado di agire al di fuori del proprio dominio specifico. Ne sono un esempio i sistemi di

ralmente condivisa ad un evento stressante o a una perdita comune, come la morte di una persona cara, *non è* un disturbo mentale" e che il "comportamento socialmente deviante (per esempio politico, religioso o sessuale)" *non è* un disturbo mentale. La forma prescelta risulta davvero poco opportuna in ambito definitorio, considerato che essa indica cosa *non è* il disagio mentale, tuttavia, nasce dalla condivisibile urgenza di porre un confine ad un concetto, che, stante la sua indeterminatezza, si presta ad essere strumentalizzato (bio) politicamente (APA, *Manuale diagnostico e statistico dei disturbi mentali*, cit., p. 15).

[9] Il concetto di benessere psicologico è affine a quello di salutogenesi. Con quest'ultimo termine, coniato dal sociologo Aaron Antonovsky negli anni '70, si intende indicare un approccio olistico alla salute, incentrato sulla promozione ed il mantenimento della vita sana. La salutogenesi si occupa dunque di studiare ciò che contribuisce ad una migliore qualità della vita, piuttosto che analizzare esclusivamente i fattori eziologici e patogeni (A. Antonovsky, *Health, Stress and Coping*, Jossey-Bass, San Francisco, 1979; Id., *Unraveling the Mystery of Health. How People Manage Stress and Stay Well*, Jossey-Bass, San Francisco, 1987).

riconoscimento facciale, gli assistenti vocali e gli algoritmi commerciali che operano su piattaforme come Netflix o Amazon;

2) l'Intelligenza Artificiale Generale (AGI – *Artificial General Intelligence*), programmata per emulare capacità cognitive simili a quelle umane. I sistemi AGI possono comprendere, apprendere e applicare conoscenze in vari contesti e domini;

3) la Superintelligenza Artificiale (ASI – *Artificial Superintelligence*): si tratterebbe di una forma di intelligenza macchinica in grado di superare quella umana in ogni ambito, inclusa la creatività, la risoluzione di problemi e l'intelligenza emotiva. Pur essendo, ad oggi, una mera ipotesi teorica, l'ASI è spesso oggetto di agguerrite ed interminabili logomachie.

Nella consapevolezza delle difficoltà definitorie sin qui menzionate, il presente studio intende riferirsi al benessere psicologico inteso in una prospettiva ampia – come salutogenesi – ed all'IA considerata alla stregua di ANI e di AGI. Quindi come sistema informatico in grado di prendere decisioni autonome razionalmente corrette all'interno di uno specifico ambito, ovvero come piattaforma o software che si propone di emulare il ragionamento dell'uomo, equiparandone o addirittura superandone le competenze e le capacità.

Tutto ciò premesso, esistono almeno due grandi linee di ricerca che legano il tema dell'IA a quello del benessere psicologico e che ci inducono a riflettere su di essa in una prospettiva (bio) giuridica.

1) In ragione della sua capacità di dialogare con gli utenti, l'IA generativa si presenta sotto la veste di *sistema esperto*, ovvero come piattaforma in grado di imitare e sostituire i professionisti – dunque anche gli psicologi e gli psicoterapeuti[10];

2) Lo sfruttamento di ingenti asset di dati consente alle piattaforme informatiche di condizionare con sempre maggiore invasività il comportamento degli utenti, favorendone la dipendenza (IAD), attraverso una maliziosa gestione delle interazioni e delle visualizzazioni[11]; in tal modo, esse possono agevolare la

[10] Con riguardo ai sistemi esperti, A.C. Amato Mangiameli, M.N. Campagnoli, *Strategie digitali. #diritto_educazione_tecnologie*, Giappichelli, Torino, 2020.

[11] Per quanto concerne invece l'*Internet Addiction Disorder*, rimando alla bibliografia citata nel paragrafo ad esso dedicato.

diffusione della FoMO, l'isolamento ed il vero e proprio *hikikomori*[12].

Di questi argomenti mi occuperò nelle pagine che seguono, avendo cura di sottolineare i rischi e le grandi opportunità offerte dal sempre più eclatante miglioramento dei sistemi di IA.

Procediamo con ordine.

2. Le nostre (limitate) risorse cognitive

Con la parola "attenzione" facciamo normalmente riferimento alla capacità umana di isolare specifici stimoli – di carattere sensoriale, cognitivo o mentale –, separandoli dalla infinita messe di input prodotti dall'ambiente – dalla stessa coscienza o dal corpo –, al fine di concentrare il pensiero su di essi. Si tratta di una risorsa psichica di grande importanza, in mancanza della quale risulterebbe pressoché impossibile qualsivoglia attività intellettiva, come, ad esempio, prendere decisioni, riflettere sul passato o studiare la soluzione di un problema.

Secondo le scienze cognitive esistono almeno tre diversi tipi di attenzione: il primo è l'"attenzione sostenuta", ovvero la capacità di mantenere il *focus* su di un dato compito per tutto il tempo necessario alla sua esecuzione; il secondo è la cosiddetta "attenzione selettiva" e consiste nella capacità di selezionare solo gli stimoli rilevanti per l'esecuzione di un dato compito, relegando ciascuno degli altri input che comunque il nostro cervello percepisce a mero rumore di fondo[13]; il terzo è, infine, l'"attenzione

[12] M. Spitzer, *Solitudine digitale. Disadattati, isolati, capaci solo di una vita virtuale?*, trad. it., Corbaccio, Milano, 2016; Id., *Connessi e isolati. Un'epidemia silenziosa*, trad. it., Corbaccio, Milano, 2018; Id., *Demenza digitale. Come la nuova tecnologia ci rende stupidi*, trad. it., Corbaccio, Milano, 2019; Id., *Emergenza Smartphone. I pericoli per la salute, la crescita e la società*, trad. it., Corbaccio, Milano, 2019; J. Lanier, *Ten arguments for deleting your social media accounts right now*, Henry Holt, New York, 2018; S. Turkle, *Life on the Screen. Identity in the Age of the Internet*, Simon & Schuster, New York, 2011; Id., *Insieme ma soli. Perché ci attendiamo sempre più dalla tecnologia e sempre meno dagli altri*, trad. it., Einaudi, Torino, 2019.

[13] A tal riguardo si parla anche di "effetto cocktail", ovvero alla capacità di dialogare nel corso di una festa, sebbene la conversazione sia circondata dalla confusione dovuta alla contemporanea presenza di molte persone (Cfr. D. Kahneman, *Psicologia dell'attenzione*, trad. it., Giunti Barbera, Firenze,

(con)divisa", si tratta della capacità di eseguire contemporaneamente più compiti spostando le risorse attentive da uno all'altro – in termini informatici parliamo di multitasking[14].

Purtroppo, come ogni altra caratteristica, qualità o risorsa umana, anche l'attenzione è limitata. Poiché essa richiede una elevata dose di energia mentale, la sua attivazione e la sua tenuta dipendono da molte variabili, personali ed ambientali. Soprattutto, l'attenzione tende a calare con il passare del tempo. A seguito di questa inevitabile deplezione, la mente avverte il bisogno di riposare, si stacca dall'oggetto (pre)scelto ed inizia a vagare alla ricerca di altri contenuti – piacevoli, meno impegnativi, oppure, più urgenti.

Risulta parecchio significativo, a tal riguardo, che l'*attenzione sostenuta* dei nativi digitali sia sensibilmente inferiore rispetto a quella delle generazioni precedenti. I risultati ottenuti dalle nuove generazioni sono tuttavia equiparabili a quelli ottenuti da individui nati prima dell'avvento della società digitale, nel caso in cui facciano un uso costante e intensivo delle nuove tecnologie informazionali[15] – come a dire: se le persone non facessero un uso intensivo delle nuove tecnologie manterrebbero inalterate le proprie risorse attentive.

Cosa ancor più importante, la perdita di *attenzione sostenuta* sembrerebbe compensata da un aumento della capacità di lavorare in multitasking.

Eppure, l'aumento della attenzione (con)divisa – associato ad una diminuzione dell'attenzione sostenuta – retroagisce negativamente sull'*attenzione selettiva*: non è facile mantenere l'attenzione quando il contesto in cui siamo immersi stimola la nostra coscienza attraverso infiniti input, o forse sarebbe più corretto dire "infinite notifiche", che reclamano considerazione e, soprattutto, richiedono una risposta in tempo reale.

Oltre ad essere molteplici, questi stimoli cognitivi sono eterogenei con riguardo al mittente, alla forma ed al contenuto. Ciò

1981; H. Pashler, *The Psychology of Attention*, MIT, Cambridge, 1999; R.W. Proctor, K.-P.L. Vu, *Attention*: *Selection and Control in Human Information Processing*, APA, Washington, 2022).

[14] G. Riva, *Fake News*. *Vivere e sopravvivere in un mondo postverità*, il Mulino, Bologna, 2018, p. 93.

[15] *Ivi*, p. 92.

significa che dobbiamo essere in grado di cambiare rapidamente il nostro codice ermeneutico ed espressivo – spendendo, in tal modo, altre ed importanti risorse psichiche. L'uso (compulsivo) che facciamo degli smartphone incide dunque sulle nostre risorse attentive, favorendo la stanchezza mentale e riducendo la nostra capacità di prendere decisioni assennate[16].

3. L'*Internet Addiction Disorder* (IAD). Una pandemia digitale

Nel 1994, lo psicologo inglese Mark Griffiths pubblicò sullo *Psychology Forum* un saggio dedicato alle "*technological addictions*", ovvero alle dipendenze "*which involve human machine interaction*". In quel saggio l'autore censiva una variegata fenomenologia di comportamenti patologici, riservando particolare attenzione alla ludopatia (*fruit machine addiction*), alla *telephone sex addiction* ed alla dipendenza da internet.

Con riguardo a quest'ultima, Griffiths riteneva che i criteri diagnostici fossero essenzialmente sei: 1) *salience*, la situazione che si verifica quando una certa attività diviene la più importante nella vita di un individuo, finendo per dominarne i pensieri, i sentimenti ed il comportamento; 2) *euphoria*, l'esperienza soggettiva della persona dipendente, raccontata in termini di eccitazione o rilassamento; 3) *tolerance*, definita come la necessità che il soggetto avverte di incrementare progressivamente il proprio coinvolgimento pur di ottenere l'effetto sperato; 4) *withdrawal symptoms*, si tratta dei cosiddetti "*unpleasant feeling states*" o "*physical effects*" causati dall'astinenza; 5) *conflicts*, i conflitti interpersonali o intrapsichici dovuti alla dipendenza; 6) *relaps*e, ovvero, la propensione a riprendere l'attività dopo un periodo di astinenza, ovvero l'incapacità di guarire dalla dipendenza[17].

[16] Come aveva compreso benissimo Bauman: "L'adattamento alle condizioni create da Internet rende l'attenzione fragile e soprattutto incostante, incapace di concentrarsi a lungo: è allenata e abituata a navigare, ma non a spingersi in profondità. […] In breve, l'attenzione tende ad abituarsi a scivolare sulla superficie molto più rapidamente del tempo che le sarebbe necessario per farsi un'idea di ciò che si nasconde più in fondo" (Z. Bauman, *Danni collaterali*, trad. it., Laterza, Roma-Bari, 2013, p. 101).

[17] M.D. Griffiths, "Technological Addictions", in *Clinical Psychology Forum*, 76 (1995), pp. 14-19, in part. p. 15; per un commento, A. Civita, *Un*

A pochi mesi di distanza dalla pubblicazione di questo primo saggio, lo psichiatra americano Ivan Goldberg coniò l'espressione "*Internet Addiction Disorder*", modellandone i criteri diagnostici sulla base dei criteri elencati dal DSM-IV con riguardo al gioco d'azzardo patologico[18].

Circa un anno più tardi, Kimberly Young, psicologa e docente presso l'Università di Pittsburgh, fondò il COLA, un centro *on line* a pagamento, specializzato nella cura dell'IAD.

Ad avviso di Young, la dipendenza da internet avrebbe potuto essere diagnosticata nel caso in cui si fossero verificate cinque delle seguenti otto condizioni, per un periodo non inferiore ai sei mesi: 1) preoccupazione riguardo a Internet, ovvero, il soggetto pensa costantemente a ciò che ha fatto o farà *on line*; 2) crescente bisogno di trascorrere tempo *on line*; 3) ripetuti tentativi di ridurre l'uso del *world wide web*; 4) ritiro in sé, se viene imposto un minore utilizzo della rete; 5) problemi nella gestione del tempo, considerato che il soggetto tende a restare *on line* per un periodo sempre maggiore rispetto a quanto inizialmente programmato; 6) problemi relazionali, ovvero problemi lavorativi, di rendimento scolastico o interpersonali; 7) menzogne circa il tempo trascorso *on line* (il paziente nasconde ai suoi familiari o ad altre figure significative la quantità di tempo trascorso *on line*); 8) repentini cambiamenti di umore, dovuti essenzialmente al fatto che il paziente naviga per fuggire dai suoi pressanti problemi quotidiani, ovvero per mitigare la depressione, l'ansia, il senso di colpa o l'impotenza[19].

Dopo circa venti anni dalla pubblicazione di questi primi studi, possiamo definire l'*Internet Addiction Disorder* (IAD) come il disagio che affligge chi spende la maggior parte del proprio tempo e delle proprie energie *on line*, provando, di conseguenza, una "importante compromissione nelle principali aree esisten-

malessere sociale: la dipendenza da Internet, FrancoAngeli, Milano, 2014, p. 42 ss.

[18] Cfr. F. Metitieri, *Comunicazione personale e collaborazione in rete*: *vivere e lavorare tra e-mail, chat, comunità e groupware*, FrancoAngeli, Milano, 2003, p. 174.

[19] K. Young, "Internet Addiction: The Emergence of a New Clinical Disorder", in *CyberPsychology & Behavior*, 1 (1996), n. 3, pp. 237-244.

ziali, come quella personale, scolastica, professionale, famigliare e sociale"[20].

Come abbiamo avuto modo di considerare, la comunità scientifica si occupa di questo argomento sin dalla metà degli anni '90. Tuttavia, il disagio è stato censito "ufficialmente" solo nel 2013, quando il DSM-5 ha inserito nella III Sezione – riservata alle condizioni che richiedono ulteriori ricerche – una forma specifica di dipendenza da Internet, denominata Internet Gaming Disorder [21].

Non è ancora chiaro se esista un'unica dipendenza da internet, o se non sia più corretto parlare di molteplici dipendenze, incentrate, di volta in volta, sull'attività ludica, sulla ricerca di denaro, sul sesso, sul *trading on line* o sui *social network*… quel che è certo è che un numero sempre maggiore di utenti sta iniziando a sperimentare gli effetti deleteri di questa dipendenza comportamentale, andando incontro ai molteplici problemi di salute fisica e/o mentale che solitamente si accompagnano ad essa.

Sarebbe possibile ipotizzare che la sempre più preoccupante diffusione dell'IAD sia dovuta, in una certa misura, all'utilizzo a dir poco malizioso dell'intelligenza artificiale da parte delle più importanti compagnie e piattaforme informatiche. Non è difatti un mistero che queste società sfruttano l'IA per (ri)conoscere i gusti, le preferenze ed i comportamenti degli utenti. In tal modo, sperano di aumentarne l'*engagement*, ovvero, di incrementare il tempo medio speso da ciascuno di essi sulla piattaforma.

La situazione appare tanto più preoccupante quanto più viene messa in relazione al fenomeno dei big data, ovvero, alla crescente massa di dati che gli internauti danno quotidianamente "in pasto" alla rete, consegnando ad essa, in maniera più o meno

[20] D. Piacentino, R. Brugnoli, "Dipendenza da Internet", in D. Piacentino, R. Brugnoli (a cura di), *Psicopatologia delle dipendenze*, Raffaello Cortina, Milano, 2004, p. 297. Sul punto, si veda anche D. La Barbera, "Le dipendenze tecnologiche", in V. Caretti, D. La Barbera (a cura di), *Le dipendenze patologiche. Clinica e psicopatologia*, Raffaello Cortina, Milano, 2005, pp. 113-132; P. Wallace, *La psicologia di Internet*, trad. it., Raffaello Cortina, Milano, 2000, pp. 425-454.

[21] APA, *Manuale diagnostico e statistico dei disturbi mentali*, cit., p. 791, pp. 1081-1085.

avveduta e consapevole, i propri dati sanitari, il proprio stato d'animo o le proprie opinioni[22].

Da un lato, queste informazioni vengono quotidianamente ed esplicitamente condivise dagli utenti dei *social network*, sotto forma di vero e proprio contenuto, oppure, sotto forma di apprezzamento, commento o condivisione, dei contenuti generati da altri; dall'altro lato, una miriade di importanti indicazioni possono essere dedotte da comportamenti "muti", ma parecchio significativi – basti pensare, a mero titolo di esempio, all'orario, alla durata delle connessioni o ai siti visitati.

Tutti questi dati consentono di addestrare l'Intelligenza Artificiale nella gestione automatica delle piattaforme informatiche, in modo che l'esperienza di ogni singolo utente risulti il più possibile coinvolgente, tarata sulle sue caratteristiche di personalità, sul suo stato d'animo o sulle probabili evoluzioni della sua storia di vita.

4. Benessere psicologico e *Social Network*

L'avvento dei *Social Network* ha rappresentato il principale fattore evolutivo nel passaggio, avvenuto circa venti anni fa, dal *world wide web* al *web 2.0*[23]. Potremmo in un certo qual modo

[22] E. Cirone, "Big Data e tutela dei diritti fondamentali: la ricerca di un (difficile) equilibrio nell'ambito delle iniziative europee", in S. Dorigo (a cura di), *Il ragionamento giuridico nell'era dell'intelligenza artificiale*, Pacini, Firenze, 2020; B. Clegg, *Big data: how the information revolution is transforming our lives*, Icon Books, London, 2017; M. Del Mastro, A. Nicita, *Big Data. Come stanno cambiando il nostro mondo*, il Mulino, Bologna, 2019; L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Raffaello Cortina, Milano, 2014, Kindle ed.; A. Guthrie Ferguson, *The Rise of Big Data Policing*, New York University Press, New York, 2017; V. Mayer-Schönberger, K. Cukier, *Big data. Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*, trad. it., Garzanti, Milano, 2013; G. Mondini, P. Cuomo, *Titolarità e circolazione dei dati*, in Aa.Vv., *Big Data: Privacy, Gestione, Tutele*, Altalex, Milano, 2018; S. Zuboff, *Il capitalismo della sorveglianza*. *Il futuro dell'umanità nell'era dei nuovi poteri*, Luiss University Press, Roma, 2019.

[23] Sembra che questa espressione sia stata utilizzata per la prima volta dall'editore americano Tim O' Reilly all'interno di una serie di conferenze tenute nel 2005 e dedicate alla nuova generazione di siti e servizi che in quel momento si stava affermando sul *web*. Il tratto caratteristico del *web 2.0* è rappresentato dal ruolo giocato al suo interno dagli utenti, che, dismesse le vesti di passivi fruitori di informazioni, ne divengono, al tempo stesso, i prin-

affermare che i *social network* siano stati la *killer application*[24] del *web*, ovvero, la maggiore causa del suo successo e della sua diffusione. È vero che una cospicua porzione della rete ospita, ancora oggi, siti istituzionali, blog e portali di vario genere e natura. Tuttavia, nessuno ha mai avuto un successo neanche lontanamente paragonabile a quello ottenuto da Facebook, Instagram o Twitter[25].

La inarrestabile e rivoluzionaria crescita dei *social* ha stimolato molte e profonde riflessioni in tutti i campi delle cosiddette scienze umane – dalla psicologia alla sociologia, alla filosofia, al diritto[26]. Nel rispetto di un antico monito hegeliano – in base al quale il pensiero filosofico è attratto dalla crisi e dalla decadenza, dispiegando le sue ali al calare delle tenebre –, gran parte di queste riflessioni esprime una posizione critica se non addirittura allarmistica.

Jaron Lanier, ad esempio, accusa i *Social Network* di aver amplificato la solitudine degli utenti, isolandoli dal proprio ambiente naturale e contingente, ovvero, dal *contesto di prossimità*[27].

cipali produttori e consumatori (sul punto M.R. Allegri, Ubi social ibi ius. *Fondamenti costituzionali dei social network e profili giuridici della responsabilità dei provider*, FrancoAngeli, Milano, 2018, p.17).

[24] La *killer application* è un gioco o una *utility* che, tramite il suo successo, rende egemone la tecnologia *hardware* che ne consente l'esecuzione.

[25] Oggi denominato "X". Cfr. S. Bentivegna, G. Boccia Artieri, *Le teorie della comunicazione di massa e la sfida digitale*, Laterza, Roma-Bari, 2019; B-C. Han, *Nello sciame. Visioni del digitale*, trad. it., Nottetempo, Roma, 2015; K.S. Young, *Caught in the net, Presi nella rete*, trad. it., Calderini, Bologna, 2000.

[26] Senza alcuna pretesa di compiutezza, si rinvia a C.R. Sunstein, *Republic.com. Cittadini informati o consumatori di informazioni*?, il Mulino, Bologna, 2006; G. Granieri, *Blog Generation*, Laterza, Roma-Bari, 2009; Id., *Umanità accresciuta*, Laterza, Roma-Bari, 2009; Id., *La società digitale*, Laterza, Roma-Bari, 2011; A.L. Lafuente, M. Righi, *Internet e web 2.0*, UTET, Torino, 2011; G. Riva, *I social network*, il Mulino, Bologna, 2016; L. Floridi, *op. cit.*; G. Pitruzzella, "La libertà di informazione nell'era di internet", in *Media Laws. Rivista di Diritto dei Media*, (2018), n. 1, pp. 19-48; M.E. Bucalo, "La libertà di espressione in rete tra *content moderation* dei social network e regolazione dell'Unione Europea", in *Diritti Fondamentali*, (2022), n. 3, pp. 270-295.

[27] J. Lanier, *La dignità ai tempi di internet. Per un'economia digitale equa*, *Primo interludio. Un'antica anticipazione della singolarità*, trad. it., Il Saggiatore, Milano, 2014.

Detto in altre parole, l'utilizzo di queste piattaforme informatiche determinerebbe il progressivo allontanamento degli utenti dalle persone che si trovano fisicamente al loro fianco[28].

A questa scuola di pensiero si iscrive anche lo psicologo Manfred Spitzer, il quale, nella sua opera più nota, rimarca come la generazione che oggi viene etichettata con il termine *baby boomer* o più semplicemente *boomer*, avesse una concezione della società e della socialità molto diversa da quella attuale, definita, in maniera parecchio significativa, "generazione io"[29]. Questa vocazione egoistico narcisista risulterebbe confermata da alcune ricerche che dimostrerebbero come la capacità di partecipazione empatica degli esseri umani sia andata diminuendo nel corso degli ultimi decenni[30].

[28] Non vi è dubbio che questi siti influiscano in maniera determinante sulle comuni modalità interazionali, ovvero sul modo in cui le persone entrano in contatto ed interagiscono. Come scrive J. van Dijck "Gli elementi tecnologici, economici e socio-legali di una piattaforma modellano le dinamiche di una socialità guidata dalla piattaforma". Resta però da capire se la solitudine ed il distanziamento sociale siano stati *causati* dai *social network* o se non siano i *social network* ad aver avuto grande successo in una società già largamente afflitta da simili problemi (J. van Dijck, T. Poell, M. de Vaal, *Platform Society*. *Valori pubblici e società connessa*, trad. it., Guerini, Milano, 2019, p. 44).

[29] "A quanti erano più grandi di noi, il nostro comportamento sarà sembrato egocentrico e poco empatico. Noi però ci muovevamo sempre in gruppo: andavamo *insieme* alle manifestazioni (per solidarietà con gli operai, contro la guerra del Vietnam ecc., cioè generalmente per gli interessi degli altri) e organizzavamo seminari *di gruppo* per ritrovare *se stessi*" (M. Spitzer, *Connessi e isolati*, cit., p. 20). Per dirla con le parole di Floridi: "Qualche moderno Geremia lamenta che la generazione iperconsapevole di sé di Facebook, costantemente impegnata a rispondere alla domanda 'dove sei?' sulla mappa Google della vita, abbia perso contatto con la realtà. Si dolgono del fatto che questa nuova generazione viva in una sorta di bolla virtuale in cui le chiacchiere più superficiali sono la moneta corrente, che non sia capace di impegnarsi in alcunché di naturale e autentico […] che non sopporti niente che abbia una cadenza lenta e che duri più a lungo di una conferenza TED, che sia fatta di *selfie narcisistici ed egocentrici* […]" (L. Floridi, *op. cit.*, posizione 1251/5025).

[30] Questa tesi è ripresa anche da G. Riva, che, a tal riguardo, taccia i nativi digitali di "analfabetismo emotivo", *I social network*, cit.

L'interazione consentita dalle piattaforme informatiche sarebbe inoltre *priva di valore* e strutturalmente *ambigua.* Perché essa, risultando incentrata sul linguaggio scritto, metterebbe "tra parentesi" la comunicazione non-verbale e para-verbale, non consentendo agli utenti di esprimersi tramite il tono, il volume o la velocità dell'eloquio; impedendo loro di manifestare il proprio pensiero attraverso la postura del corpo, l'espressione del volto o la prossemica[31].

Oltre a quanto sin qui esposto, i *social network* favorirebbero la diffusione di notizie false[32], agevolando la radicalizzazione e la polarizzazione delle opinioni – dovuta al successo degli *hate speech* ed al lavoro di meccanismi informazionali denominati "*filter bubble*" ed "*echo chambers*"[33].

Infine, l'utilizzo compulsivo ed eccessivo di queste piattaforme viene spesso menzionato come un elemento cardine di due forme di disagio mentale denominate FoMO ed "*hikikomori*", parecchio diffuse tra le nuove generazioni. Con l'acronimo FoMO (*Fear Of Missing Out*) si fa infatti riferimento al timore di essere "tagliati fuori", ovvero di essere emarginati e rifiuta-

[31] "Con l'introduzione di un medium, il soggetto diventa *disincarnato* (*disembodied*) per il suo interlocutore: la fisicità del corpo viene sostituita da quella del medium. Più precisamente, nei *social network* la fisicità e l'immediatezza del corpo reale vengono sostituite da un corpo digitale, composto da una pluralità di immagini parziali e contestualizzate" (*Ivi*, posizione 400/3535.)

[32] R. Brotherton, *Bad News*, *Why we fall for fake news*, Bloomsbury Sigma, London, 2020.

[33] E. Irving sottolinea, ad esempio, le responsabilità avute da Facebook nelle violenze scoppiate, il 01/08/2018, in Birmania, ai danni dei Rohingya e di altre minoranze ("Suppressing Atrocity Speech on Social Media", in *American Journal of International Law*, 113 (2019), pp. 256-261). La tesi della polarizzazione sociale indotta dai *social network* è molto ben espressa da C. R. Sunstein che, a tal riguardo, usa l'espressione *cybercascades*, sottolineando come internet sia un "terreno fertile per l'estremismo", perché "persone con la stessa area di pensiero trattano tra di loro con grande frequenza e facilità e spesso senza sentire una controparte. Un'esposizione ripetuta a una posizione estrema, unita all'idea che molte altre persone condividano quella posizione, prevedibilmente porterà le persone che vi sono esposte, e forse già propense, a credere in essa": C. R. Sunstein, *op. cit.*, p. 87. Per quanto riguarda l'*hate speech* si veda, inoltre, G. Ziccardi, *L'odio on line*, Raffaello Cortina, Milano, 2016.

ti dalla (o dalle) comunità (virtuali) di riferimento. Più esattamente, esso indica

> The preoccupation of users […] with lost opportunities when they are offline or unable or unwilling to connect or communicate with others to the extent they wish […] FoMO is a main type of problematic attachment to social media and is associated with a range of negative life experience and feelings, such as lack of sleep, reduced life competency, emotional tension, negative effects on physical well being, anxiety and a lack of emotional control[34].

Dal canto suo, l'espressione "*hikikomori*" è divenuta nota al grande pubblico nel 1998, quando lo psichiatra Tamaki Saito pubblicò un libro intitolato Shakaiteki Hikikomori in cui descriveva i soggetti affetti da questa sindrome come "coloro i quali si ritirano completamente dalla società e restano nelle loro case per più di sei mesi"[35].

[34] A.K. Przybylski, K. Murayama, C.R. De Haan, V. Gladwell, "Motivational, emotional, and behavioral correlates of fear of missing out", in *Computers in Human Behavior*, 29 (2013), pp. 1841-1848; A. Alutaybi, D. Al-Thani, J. McAlaney, R. Ali, "Combating Fear of Missing Out (FoMO) on Social Media: The FoMO-R Method", in *Int. J. Environ. Res. Public Health*, 17 (2020); M. Akbari, M. Seydavi, S. Palmieri, G. Mansueto, G. Caselli, M.M. Spada, "Fear of missing out (FoMO) and internet use: A comprehensive systematic review and meta-analysis", in *Journal of Behavioral Addictions*, 10 (2021), n. 4, pp. 879-900.

[35] A partire da quel momento, la sindrome è divenuta sempre più nota, al punto che, nel 2010, il Ministero giapponese per la Salute, il Lavoro ed il Welfare dedicò una *guideline* all'*hikikomori*, riprendendo, sostanzialmente, la definizione contenuta nel libro di Saito. Circa sei anni più tardi, l'Ufficio di Gabinetto governativo giapponese divulgò uno studio in cui si stimava che il numero di cittadini affetti da vero e proprio *hikikomori* (*core hikikomori*) fosse pari a 176,000, mentre altri 540,000 sarebbero stati affetti da una sindrome meno grave, definita di *pre-hikikomori*. Con quest'ultima espressione si faceva riferimento a persone dedite a trascorrere la maggior parte del proprio tempo libero *on line*, che tuttavia non rinunciavano ad uscire di casa per fare acquisti nei negozi del quartiere (T. Saito, *Shakaiteki hikikomori: owaranai shishunki (social withdrawal: a neverending adolescence)*, PHP Shinsho, Tokyo, 1998; M. Tateno, A.R. Teo, W. Ukai, J. Kanazawa, R. Katsuki, H. Kubo, T.A. Kato, "Internet Addiction, Smartphone Addiction, and Hikikomori Trait in Japanese Young Adults: Social Isolation and Social Network", in *Frontiers in Psychiatry*, 10/07/2019, art. 455).

L'ultima versione del DSM precisa che la parola *hikikomori* è un termine composito, derivante dall'unione di "*hiku*" (tirarsi indietro) e "*moru*" (appartarsi); lo definisce come "una sindrome di ritiro sociale prolungato e grave", sottolineando come "l'immagine tipica" dell'"*hikikomori*" sia quella dell'adolescente – o del giovane adulto – che non abbandona la propria stanza e rifiuta le interazioni sociali; aggiungendo che questo comportamento, inizialmente egosintonico, non di rado, con il passare del tempo, sfocia nell'angoscia, rimarcandone infine la correlazione con altre condizioni patologiche descritte nel manuale – quali, ad esempio, il disturbo d'ansia sociale, il disturbo depressivo maggiore o il disturbo evitante di personalità.

A differenza delle primissime definizioni di "*hikikomori*", la concezione scientifica attualmente in voga non esclude le interazioni digitali, anzi, il DSM rimarca come esso sia "spesso associato a un'elevata intensità di utilizzo di Internet e di scambi sociali virtuali"[36].

Al fondo di tutte queste accuse sembra agitarsi il fantasma della *strumentalità inversa*, ovvero l'idea, a mio avviso corretta, che ogni forma di interazione tecnologicamente mediata sia in grado di retroagire in maniera negativa sull'utenza, oggettivizzando ed a suo modo "sfruttando" i soggetti che ne fanno uso.

5. *Pars construens*. L'IA al servizio della psiche

L'Intelligenza Artificiale può fornire un grande aiuto in tema di benessere psicologico, essa è infatti in grado di sfruttare gli asset di dati a disposizione delle piattaforme per diffondere messaggi di prevenzione e di cura tra particolari categorie considerate "a rischio" per specifici disturbi mentali – in base all'età, al ceto o alla professione. Pensiamo, a mero titolo di esempio, al *burn out*, del quale spesso soffrono le persone impegnate in una professione di aiuto; all'anoressia, che nella stragrande maggioranza dei casi riguarda individui di genere femminile in giovane

[36] APA, *Manuale diagnostico e statistico dei disturbi mentali*, cit., p. 1039.

età, oppure, all'*internet gaming disorder* di cui sembrano ammalarsi soprattutto gli adolescenti di genere maschile[37].

Ancor di più, se il concetto di salute viene considerato in maniera ampia, in modo da ricomprendere al suo interno anche le attività di prevenzione, appare evidente quanto di buono possiamo ottenere dall'Intelligenza Artificiale in termini di campagne informative.

Inoltre, l'IA può essere usata come strumento di tecno-regolazione e/o di auto-contenimento[38]. Nel primo caso, potrebbe intervenire autonomamente, ed in maniera tempestiva – ad esempio per bloccare l'accesso ad un sito di scommesse o oscurare un contenuto pornografico, ove riscontri nel comportamento dell'utente gli indizi di una dipendenza incipiente o conclamata. Nel secondo caso, essa agirebbe su richiesta dello stesso utente, ad esempio per interdire il traffico dati una volta superata una certa soglia temporale, oppure, per impedire il superamento di una specifica soglia di spesa all'interno dei siti dedicati al gioco d'azzardo.

Infine, poniamo mente ai *chatbot* dedicati alla salute mentale[39]. A tal riguardo, è importante notare che i sistemi esperti non

[37] Ancor di più, l'intelligenza artificiale può raggiungere, con i suoi *input*, un determinato utente che potrebbe aver bisogno, in quel preciso momento della giornata, di essere edotto circa i rischi che conseguono alla mancanza di sonno, ad una cattiva alimentazione o alla pessima abitudine di controllare il cellulare mentre si guida.

[38] Con riguardo al concetto di tecno-regolazione, R. Brownsword, "What the World Needs Now: Techno Regulation, Human Rights and Human Dignity", in *Human rights: Global Governance and the Quest for Justice*, 4 (2004), n. 1; O.W. Lembcke, "Techno-regulation and law: rule, exception or state of exception?", in *Netherlands Journal of Legal Philosophy*, (2011), n. 2; A.C. Amato Mangiameli, "Tecno-diritto e tecno-regolazione. Spunti di riflessione", in *Rivista di filosofia del diritto*, (2017), *n. speciale,* pp. 87-112; Id., "Tecno-regolazione e diritto. Brevi note su limiti e differenze", in *Il diritto dell'informazione e dell'informatica*, (2017), n. 2, pp. 147-167; R. Zaccagnino, C. Capo, A. Guarino, N. Lettieri, D. Malandrino, "Techno-regulation and intelligent safeguards", in *Multimedia Tools and Applications*, 80 (2021), pp. 15803-15824; A.C. Amato Mangiameli, voce "Tecnoregolazione", in A.C. Amato Mangiameli, G. Saraceni (a cura di), *Cento e una voce di informatica giuridica*, Giappichelli, Torino, 2023.

[39] Di seguito, alcuni esempi: 1) *Woebot*. Si tratta di un chatbot progettato per fornire supporto emotivo. Utilizza principi terapeutici basati sui princi-

possono sostituire la vera e propria consulenza professionale. Tuttavia, possono risultare estremamente utili per offrire agli utenti un supporto a breve termine, migliorandone l'autoconsapevolezza. Ad essi va inoltre riconosciuto il pregio di essere facilmente accessibili, tanto dal punto di vista economico quanto dal punto di vista della disponibilità – h 24 e a domicilio. Senza considerare che consentono di mantenere una seppur mitigata forma di anonimato e dunque di vincere le ritrosie, le reticenze e gli imbarazzi che non di rado allontanano chi soffre di un disagio mentale dall'intraprendere un percorso di cura.

Per quanto concerne, invece, i *social network*, esistono molte e delicatissime evenienze nelle quali essi offrono un valido antidoto contro l'emarginazione e l'isolamento sociale, alcune di queste riguardano il contesto sociale, altre le personali condizioni in cui versano gli utenti.

Con riguardo alle prime, è sin troppo facile, oggi, richiamare alla mente l'esperienza della pandemia, sottolineando quanto siano stati importanti i *social network* per mitigare le conseguenze negative dell'isolamento sociale; per quanto concerne invece le seconde, basterebbe porre mente alla condizione in cui versano gli anziani, i malati e le persone diversamente abili, per capire che i *social network*, a prescindere dal verificarsi di circostanze eccezionali che investano l'intera società, offrono quotidianamente a larghe fasce della popolazione una insostituibile opportunità di socializzazione e di interazione. In tal senso, essi favoriscono il benessere psicologico, combattendo lo stress e mitigando la solitudine.

pi della terapia cognitivo-comportamentale per aiutare gli utenti a gestire lo stress, l'ansia e altre forme del disagio; 2) *Wysa*: è un sistema esperto che sfrutta l'intelligenza artificiale per offrire supporto per l'ansia, la depressione, l'insonnia. Offre anche esercizi di autoaiuto e tecniche di gestione e di autocontenimento; 3) *Tess*: è stato espressamente progettato per i militari e i veterani degli Stati Uniti, ma è accessibile a tutti. Offre sostegno emotivo e strumenti per affrontare lo stress, la depressione e l'ansia; 4) *Youper*: questa piattaforma combina principi della terapia cognitivo-comportamentale e dell'intelligenza artificiale per aiutare gli utenti a monitorare il loro stato emotivo, identificare schemi di pensiero e sviluppare strategie di coping; 5) *Replika*: un chatbot di intelligenza artificiale che può essere utilizzato per conversazioni leggere ma anche per discutere di questioni personali e emozionali. Può fornire supporto e compagnia.

6. Conclusioni

Per quanto concerne il valore etico, prima ancora che giuridico, dell'intelligenza artificiale, notiamo come una postura filosofica parecchio diffusa sostenga che la tecnica non abbia e non possa avere alcun valore intrinseco, ovvero, che ogni tecnologia debba essere considerata, per la sua stessa natura strumentale, "eticamente neutra", priva di una chiara connotazione assiologica. Ciò a dire, tutti gli strumenti sarebbero, in quanto tali, privi di responsabilità, attitudine, carattere; ogni ragionamento circa la loro "bontà" dipenderebbe dal modo in cui l'uomo decide di farne concretamente uso.

Questa tesi, pur apparendo lineare e sensata, rischia tuttavia di confondere il discorso sulla *libertà* (e sulla responsabilità che ne consegue), con il diverso ragionamento sull'*essenza* o sulla natura della tecnologia.

Se è vero che gli strumenti di cui ci serviamo non sono liberi, e dunque non sono responsabili, è altrettanto vero che essi possiedono una struttura, un'architettura, una forma oggettiva che ne esplicita altrettanto oggettivamente il senso. Intendo dire che il loro meccanismo ci consente di analizzarne fenomenologicamente l'essenza e dunque di ragionare attorno a ciò che essi sono – a prescindere dall'uso che può farne l'uomo.

Il sapere giuridico è ben consapevole dello scarto che separa il discorso sulla libertà da quello sulla natura – non a caso, riconosce che anche gli oggetti e gli strumenti hanno una specifica "vocazione", considerando che alcuni tra essi sono vere e proprie armi, mentre altri lo sono solo in senso "improprio". I primi vengono qualificati come oggetti atti ad offendere, i secondi possono essere maliziosamente distratti dal loro *normale* e *corretto* utilizzo per arrecare un danno fisico ad un'altra persona. Un altro esempio concerne il materiale pedopornografico, del quale è vietata la produzione, la detenzione, la vendita, lo scambio e l'offerta per *ciò che esso è*, ovvero, a prescindere dall'uso che intende farne il soggetto; lo stesso discorso riguarda le sostanze stupefacenti.

Le norme che ho appena menzionato non sottendono alcun giudizio circa il valore etico di un oggetto, di una sostanza o di uno strumento – non è la foto, la molecola o il filmato ad essere

moralmente o giuridicamente "responsabile" – ma ne sanzionano la struttura oggettiva, perché ritenuta in sé dannosa, pericolosa e quindi illecita, a prescindere dal concreto utilizzo che intenda farne l'essere umano.

Se accettiamo di scontrarci su questo terreno, ovvero sul campo fenomenologico, dobbiamo convenire che l'intelligenza artificiale, nella sua essenza, non è un virus, una droga o un'arma, ma uno strumento eminentemente culturale: il più alto livello raggiunto dalla nostra competenza informatica, un risultato del quale dovremmo andare fieri, piuttosto che un nemico da temere e combattere.

Tuttavia, anche l'intelligenza artificiale, come tutti gli strumenti, si presta ad essere violentata e distratta verso finalità di carattere politico o crematistico – che non appaiono consustanziali alla sua natura.

Appare dunque chiaro quale sia la più grande sfida che attende il mondo del diritto negli anni a venire.

In particolare, le leggi nazionali e gli accordi internazionali dovranno: 1) evitare che i chatbot vengano percepiti dagli utenti come sostitutivi del colloquio con uno specialista – non tanto per salvaguardarne il posto di lavoro, quanto per tutelare gli stessi utenti dalla illusione di poter bypassare il dialogo interpersonale, ovvero, di potersi curare a prescindere da una relazione umana – elemento che possiede, di per sé un elevato e comprovato valore terapeutico[40]; 2) controllare lo sfruttamento dei sistemi di IA da parte delle app, dei siti e delle piattaforme informatiche, al fine di tutelare gli utenti – in particolar modo i minori – dalle

[40] Sul punto, D.J. Safran, Z.V. Segal, Il processo interpersonale in psicoterapia cognitiva, Feltrinelli, Milano, 1993; C. Sica, "Fattori prognostici relativi all'esito della terapia cognitivo-comportamentale del disturbo ossessivo compulsivo", in *Acta Comportamentalia*, (1996), n. 4, pp. 107-127; P. Gilbert, R.R. Leahy, *La relazione terapeutica nella terapia metacognitiva interpersonale*, trad. it., Eclipsi, Firenze, 2009; A. Semerari, *La relazione terapeutica. Storie, teorie, problemi*, Laterza, Roma-Bari, 2022; F.C. Casula, "La relazione che cura in psicoterapia", in A. Maniaci (a cura di), *La cura della relazione e la relazione di cura. Dialogo fra giuristi, medici e psicoterapeuti*, Pacini, Pisa, 2023, pp. 75-97.

trappole che ne aggrediscono le risorse attentive, favorendone la dipendenza[41].

Spetterà quindi ai giuristi creare un insieme di regole che sappiano limitare lo sfruttamento dell'intelligenza artificiale con esiti manipolatori e/o dannosi per la salute mentale degli utenti, consentendole di restare fedele alla sua vocazione *lato sensu* umanistica e strutturalmente culturale.

[41] A tal riguardo, risulta parecchio significativo il dibattito che si sta sviluppando nei Paesi europei circa l'opportunità di consentire agli studenti di portare uno smartphone all'interno delle aule scolastiche. In Francia, la legge n. 698/2018 vieta l'utilizzo di smartphone e di altri oggetti connessi alla rete nelle scuole elementari e medie inferiori – a meno che questi strumenti non risultino necessari per finalità didattiche e pedagogiche; in senso sostanzialmente analogo, in Grecia, tramite una circolare ministeriale datata 22 giugno 2018, è stato proibito agli studenti delle scuole elementari e medie di fare uso dei telefoni cellulari e di apparecchiature connesse alla rete; più recentemente, il governo inglese ha diramato le linee guida denominate *Mobile Phones in Schools. Guidance for schools on prohibiting the use of mobile phones throughout the school day*, adottando, a partire dal 19/02/2024, una politica assimilabile a quella dei Paesi già citati.

Maurizio Sozio

Il deficit fronetico dello "*Jus ex machina*"

Abstract: **The Phronetic Deficiency of "*Jus ex Machina*".** In 1666, Leibniz prophetically hypothesized that in the future, in the face of a dispute, the parties could sit down and resolve it with a calculation. In 2017, a study by the Universities of London and Sheffield developed an algorithm capable of predicting the decisions of the ECHR with a degree of accuracy of 79%. In the face of the realization of Leibniz's prophecy, we must ask ourselves whether a calculating entity, an artificial intelligence, could replace human's judgment. The need therefore arises to go back to the necessary conditions for the ability to judge law. Conditions that we believe are of two types: epistemic and phronetic.

Keywords: Fronesis, Embodiment, *Jus ex Machina*.

Sommario: 1. Dalla Corasmia a Dortmouth – 2. Aspettando la *Singolarità* – 3. *Jus ex machina*? – 4. Condizioni epistemiche – 4.1. Una fallacia epistemica? – 5. Condizioni fronetiche – 6. L'insostenibile leggerezza dell'incalcolabile e le forme di messa in ordine del mondo.

1. Dalla Corasmia a Dortmouth

Quando nell'825 Abu Ja'far Muhammad ebn Musa al-Khwarezmi scrisse la sua opera *al-Kitab al-mukhtasar fi hisab al-jab wa l-muqab ala*, tradotta in latino con il titolo *Algoritmi de numero indorum*, non scrisse esclusivamente il più importante trattato antico sul sistema numerale indiano, diffondendo la scienza delle procedure di calcolo espresse come operazioni matematiche, ma diede inconsapevolmente la paternità, grazie alla traduzione latina del toponimo *al-Khawrezm i* traslitterato dai traduttori latini in "*Algoritmi*", allo schema di funzionamento operativo fondamentale per un futuro da lui lontanissimo: quello dell'intelligenza artificiale.

Probabilmente quel futuro lontanissimo ha come arco temporale di riferimento il secondo dopoguerra, e segnatamente prima

l'*Imitation Game* di Turing[1] e poi la nascita del sintagma "intelligenza artificiale" nel documento per la conferenza di Dartmouth del 1956 in cui McCarthy, Minsky, Rochester e Shannon, dichiaravano che "ogni aspetto dell'apprendimento o di qualsiasi altra caratteristica dell'intelligenza può in linea di principio essere descritto così precisamente da farne una macchina in grado di simularlo"[2].

Ogni innovazione nasconde l'inquietudine della contraddizione, del passaggio dal *notum* al *novum*, ripropone antiche risonanze prometeiche, mostra l'incertezza del futuro. È sempre stato così, dall'ambivalenza del *pharmakón* della scrittura, prudenzialmente nascosto sotto il mantello di Fedro, alla rivoluzione della stampa, *ars scribendi artificialiter* autentica *meretrix* nelle invettive dei libelli del Quattrocento dopo l'innovazione di Gutenberg[3].

Probabilmente l'epoca della "quarta rivoluzione", come l'ha ribattezzata Luciano Floridi[4], genera le stesse inquietudini. Il tumultuoso sviluppo della ITC ha impattato sempre più nella vita quotidiana, tanto che James Lovelock ha parlato di una nuova era dell'umanità: il Novacene[5].

Si è giunti a parlare di Novacene, specifica Lovelock, in seguito all'avvento della *iper-intelligenza*, introdotta dallo sviluppo delle *deep learning machine* e dalla possibile simbiosi tra uomo e macchina.

L'influenza dell'intelligenza artificiale si mostra senza dubbio invasiva, numerosi sono infatti i campi di utilizzo: arte, economia, medicina, ingegneria e diritto, e soprattutto si mostra con il volto ancipite di Giano, facendo sorgere questioni sui suoi limiti di utilizzo.

[1] A.M. Turing, "Computing Machinery and Intelligence", in *Mind*, 59 (1950), pp. 433-460.

[2] J. McCarthy, M.L. Minsky, N. Rochester, C.E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 31/09/1955.

[3] F. Pierno, *Stampa Meretrix: scritti quattrocenteschi contro la stampa*, Marsilio, Venezia, 2012.

[4] L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, trad. it., Raffaello Cortina, Milano, 2017.

[5] J. Lovelock, *Novacene. L'età dell'iperintelligenza*, trad. it., Bollati Boringhieri, Torino, 2020.

Nel nostro breve lavoro intendiamo porre il focus sulla questione se il giudizio *ex machina* risponda alle condizioni minime che deve soddisfare un giudizio giuridico: la conoscenza dei fatti e delle norme e la capacità e la saggezza di giudicare. Condizioni che chiameremo *epistemiche* e *fronetiche*. Condizioni che riguardano non solo la scienza "calcolabile" di fatti, fattispecie e precedenti giudiziari ma anche la capacità di esercitare la saggezza nel giudizio e aprirsi all'incalcolabilità del caso concreto.

Prima, però, occorre innanzitutto definire cosa si intende per *Intelligenza* dal momento che per l'*Oxford Companion to the Mind*: "sono disponibili innumerevoli test per misurare l'intelligenza, ma nessuno sa con sicurezza che cosa misurino i test disponibili"[6].

Si può, tuttavia, azzardare, nell'ottica di valutare la capacità operativa dell'intelligenza artificiale, una definizione del tipo: L'intelligenza razionale (perché c'è anche l'intelligenza emotiva) è la capacità di inferire corrispondenze e correlazioni da un insieme di dati (*input*) e la conseguente capacità di pianificare un comportamento o una strategia, costituita da scelte o comportamenti concordanti con un fine prestabilito (*output*).

L'intelligenza, la conoscenza, sono sempre in-formatiche; tomisticamente sono ciò che determina i singoli enti e fa essere un essere quello che è, *dat esse rei*. Allo stesso modo l'intelligenza "giuridica" del giudicante rimanda alla conoscenza e scelta di dati, fatti, che danno *forma* alla fattispecie.

È un giudizio che potrebbe essere ridotto a giudizio *calcolante*[7], se non fosse che, come vedremo, non può fare a meno di un'eccedenza di senso, di un pensiero potenzialmente divergente *perché* fenomenologicamente situato.

Prima di indagare le condizioni di compatibilità tra giudizio calcolante e giudizio giuridico, può essere utile valutare lo stato dell'arte delle "*macchine*" approfondendo l'evoluzione dello sviluppo dell'intelligenza artificiale e le prospettive future.

[6] R.L. Gregory, "Intelligence", in Id., *The Oxford Companion to the Mind*, Oxford University Press, Oxford, 1987, p. 375.

[7] Sul tema della calcolabilità giuridica si veda: N. Irti, *Un diritto incalcolabile*, Giappichelli, Torino, 2016; A. Carleo (a cura di), *Calcolabilità giuridica*, il Mulino, Bologna, 2017.

2. Aspettando la *Singolarità*

Possiamo distinguere gli stadi di sviluppo dell'A.I. in tre fasi. La nascita dell'*hardware* (1), lo sviluppo del *software* (2) e il *deep learning* (3), in attesa di un ennesimo punto di svolta che i futurologi indicano imminente, una macchina dalle capacità inimmaginabili: la *Singolarità*.

1) La storia dell'intelligenza artificiale nasce negli anni Quaranta in seguito agli sforzi orientati a costruire una macchina in grado di *emulare* le attività del cervello umano. Konrad Zuse (in Germania) e Von Neumann (in USA) costruirono la prima architettura *hardware* ancora oggi alla base di tutti i *devices* elettronici. Questa architettonica è composta da tre sottosistemi interconnessi: L'unità centrale di elaborazione (CPU), la memoria (RAM), i dispostivi di ingresso e di uscita (*input*/*output*) e l'operatore di interconnessione (il BUS). Alan Turing, come accennato, mise a punto le fondamenta del calcolo universale attraverso la *Intelligent Machinery* e la celebre prova dell'*Imitation Game*.

2) La seconda fase dello sviluppo dell'intelligenza artificiale, si ebbe con lo studio di sistemi operativi e linguaggi di programmazione. Grazie ai nuovi *software* le macchine furono capaci di risolvere problemi sempre più complicati e cimentarsi in attività tipicamente umane come, ad esempio, il gioco degli scacchi. Lo sviluppo dei *software* ha dato la stura a un *upgrade* dell'operatività dell'intelligenza artificiale e all'ampliarsi della domanda filosofica attorno agli stati mentali. Fu Searle in particolare a sottolineare, attraverso il ricorso al celebre esempio della *stanza cinese*[8], come un *software* che manipola simboli possa

[8] Cfr. J.R. Searle, *Menti, cervelli e programmi. Un dibattito sull'intelligenza artificiale*, trad. it., CLUP-CLUED, Milano, 1989. L'esperimento mentale della stanza cinese consiste nel pensare a una persona che non conosce il cinese chiusa in una stanza con un libro di regole per manipolare simboli cinesi. La persona riceve dei fogli con simboli cinesi, consulta il libro per trovare la regola corrispondente e produce altri simboli in risposta. Dall'esterno, la persona sembra in grado di comprendere e rispondere al cinese, ma in realtà non coglie l'aspetto semantico. Manipola i simboli seguendo le regole senza attribuire alcun significato a ciò che fa. Searle utilizza questo esperimento per argomentare contro l'intelligenza artificiale forte e l'ipotesi che un computer possa realmente comprendere il linguaggio e avere una mente. La sintassi (l'insieme di regole per manipolare i simboli) non è sufficiente per la semantica (la comprensione del significato). Un computer può seguire le regole, ma non può capire il significato dei simboli come fa un

dare risposte corrette ma mai essere considerato una mente. Le questioni sul rapporto tra il *software* e gli stati mentali hanno preparato il terreno alla terza fase dell'evoluzione dell'intelligenza artificiale: l'intelligenza artificiale forte, le reti neurali e il *deep learning*.

3) Nata come emulazione del cervello umano, la programmazione di *software* per intelligenza artificiale, a partire dalla fine degli anni Novanta, ha una sorta di mutamento di paradigma in ragione del parallelo sviluppo delle neuroscienze e della conoscenza del funzionamento delle connessioni cerebrali.

Le nuove tecniche di indagine neuroscientifica hanno evidenziato il funzionamento del *connettoma*[9] cerebrale individuando nel particolare funzionamento dei circuiti neurali la struttura portante dell'intelligenza umana. Il tentativo di collegare le due discipline fu identificato con la nascita dell'intelligenza artificiale forte, improntata, appunto, all'emulazione strutturale del cervello umano e orientata, in un futuro remoto, a consentire alle macchine di diventare senzienti o autocoscienti. Ma non è questo che ci interessa in questa sede. La novità proposta dallo sviluppo dell'intelligenza artificiale basata sulle reti neurali ha mostrato come il passaggio dal primo paradigma delle AI, quello basato sul *simbolismo* incentrato operativamente su stringhe di dati manipolati dai software, a quello *connessionista*, abbia avviato la rivoluzione del *deep neural network* e del *machine learning*, *software* in grado di consentire alle macchine di poter imparare da se stesse e con grandi margini di sviluppo. All'orizzonte, i futurologi ipotizzano il superamento da parte delle macchine dotate di intelligenza artificiale della soglia della consapevolezza, adombrando uno snodo di passaggio nel concetto di *Singularity.*

Cos'è la Singolarità?

La Singolarità, così come ipotizzato nella seconda metà del Novecento prima da von Neumann[10] e poi da Vernor Vinge[11], è

essere umano. L'esperimento ha avuto un impatto significativo sulla filosofia dell'intelligenza artificiale e sulla linguistica cognitiva.

[9] Cfr. S. Seung, *Connettoma*. *La nuova geografia della mente*, trad. it., Codice edizioni, Torino, 2013.

[10] La profezia di John von Neumann è riportata in: S. Ulam, "John von Neumann, 1903-1957", in *Bullettin of the American Mathematical Society*, 64 (1958), n. 3, pp. 1-49.

[11] V. Vinge, "The Coming Technological Singularity: How to Survive in the Post-Human Era", in *Vision-21*: *Interdisciplinary Science and Engineer-*

una *reazione di fuga* di cicli di auto-accrescimento per cui un'intelligenza artificiale estremamente intelligente sarà così tanto intelligente da migliorarsi ricorsivamente portando ad un aumento esponenziale della sua intelligenza. Alcuni datano una simile transizione tecnologica al 2030-2050, come Ray Kurzweil nel suo *Singularity is Near*, che prevede la singolarità nel 2045[12].

Non sappiamo se il progresso tecnologico tra qualche decennio ci porterà alla transizione verso il post umanesimo, ciò che importa al momento è domandarsi se e quanto possano contribuire i sistemi di *machine learning* a migliorare il sistema giuridico, a prendere decisioni giuridiche più "intelligenti" di quelle umane.

Le potenzialità in ambito di decisione giuridica sono molteplici, grazie ai protocolli algoritmici di *riconoscimento* (*pattern matching*) e *derivazione* vi è la possibilità di gestire una casistica giuridica non solo attraverso il classico ragionamento per analogia (*pattern matching*) ma anche attraverso un sistema dotato di basi per una sorta di "apprendimento autonomo".

Questa tecnologia, in attesa di quella che potrebbe essere la rivoluzione della singolarità, intercetta le caratteristiche del "decisore giuridico". Bisogna valutarne, però, la piena compatibilità. Vediamo.

3. *Jus ex machina?*

Passata in rassegna la struttura dei sistemi di A.I., resta da indagarne più approfonditamente la funzione in ambito giuridico, la domanda giusfilosofica da porsi investe le condizioni di sostituibilità della funzione dell'organo giudicante da parte di un ente artificiale; è sostanzialmente una *Grundfrage* attorno alle questioni di validità della c.d. "giustizia predittiva", una giustizia basata sul giudizio algoritmico[13].

La logica predittiva dell'I.A. deve confrontarsi con le logiche del processo e con i suoi modelli; il *Richterrecht*, il diritto

ing in the Era of Cyberspace, NASA Conference Publication, 01/12/1993, pp. 11-22.

[12] R. Kurzweil, *La singolarità è vicina*, trad. it., Maggioli, Rimini, 2008.

[13] Sulla giustizia predittiva si veda, da ultimo, F. Corona, *Giustizia predittiva. Quando gli algoritmi pervadono il diritto*, Aracne, Roma, 2023.

dei giudici, è fondato essenzialmente sulla logica deduttiva, per questo la sua "calcolabilità" potrebbe essere simulata dall'operatività algoritmica delle "macchine" predisposte di reti neurali e *deep learning*.

La logica deduttiva, tuttavia, non esaurisce la gamma dei modelli della logica del processo. Incampo nota come accanto al modello analitico-deduttivo, ci siano almeno due modelli: il modello dialettico-retorico e quello *fuzzy*, modelli che prendono in esame rispettivamente la logica di proposizioni possibilmente vere e la logica di proposizioni che "vivono" nella *fuzziness*, possono essere, cioè, allo stesso tempo sia vere che false[14].

La decisione giuridica si articola attraverso diversi livelli di discorso, alle prese non solo con la molteplicità dei metodi interpretativi, ma soprattutto con la natura proteiforme del caso concreto, la cui manifestazione è imprevedibile e sfuggente rispetto alla razionalità dell'agente giudicante.

Una razionalità che appare limitata. Gli agenti giudicanti, per dirla con H. Simon[15], si misurano con una forma di razionalità approssimata per difetto, in cui chi giudica non riesce a darsi una raffigurazione completa della realtà, per via della non linearità e complessità dell'ambiente o per via dell'incapacità del soggetto di raccogliere, trattare e memorizzare tutti i dettagli del mondo reale. "Si tratta di una forma semi forte di razionalità, nella quale si assume che gli agenti siano razionali nelle intenzioni ma di fatto limitatamente tali"[16].

La domanda sull'utilizzo e sulla compatibilità dell'A.I. per le decisioni giuridiche ha probabilmente qui la sua ragion d'essere, il modello di risoluzione dei problemi associato all'approccio della razionalità limitata descrive un agente umano decisore inaffidabile, a tratti disfunzionale rispetto all'esigenza primaria del diritto quale *medium* sociale, vale a dire, come insegnano i sociologi[17], quella di stabilizzare le aspettative. In questo senso,

[14] A. Incampo, *Metafisica del processo*, Cacucci, Bari, 2016, pp. 24-39.

[15] H. Simon, *Causalità, razionalità, organizzazione*, il Mulino, Bologna, 1985, p. 295.

[16] O.E. Williamson, *Le istituzioni economiche del capitalismo. Imprese, mercati, rapporti contrattuali*, trad. it., FrancoAngeli, Milano, 1987, p. 127.

[17] Si pensi a Luhmann e Habermas che partendo da due impostazioni differenti (codici sistemici e linguaggio orientato all'intesa) delineano esiti

certezza e *calcolabilità* sembrano e elementi costitutivi del fenomeno giuridico.

Di fronte alle limitazioni cognitive, ogni presa di decisione diventa permeabile all'errore e aperta al rischio dell'arbitrio e dell'irrazionalità. La domanda, quindi, attorno alle condizioni di sostituibilità del giudizio del giudicante umano con quello artificiale, alla luce dell'impetuoso sviluppo tecnologico, appare legittimarsi sempre più.

Tuttavia, ci sembra opportuno verificare se alcune condizioni di validità, che si possono ritenere essenziali del giudizio giuridico, siano integrate o integrabili dal giudizio "ex machina", dalle scienze calcolatrici che, parafrasando Vico,

> espiano il difetto (*vicio*) della nostra mente e della sua estrema limitatezza, e per mezzo delle operazioni diventano simili alla scienza divina in quanto vero e fatto si convertono[18].

Una macchina che interviene a dirimere le controversie umane propone *ricorsivamente*, per citare ancora Vico, la soluzione affidata a un *Deus ex Machina*, come nelle tragedie greche, riproponibile come uno *Jus ex machina*, calato dall'alto dell'intelligenza artificiale.

Ma quali condizioni deve soddisfare lo *Jus ex machina* per proporsi come alternativa al decisore giuridico?

Come già accennato, possono individuarsi almeno due condizioni: una *epistemica* che riguarda le condizioni riguardanti scienza e conoscenza di atti, fatti, dati giuridici e precedenti e una condizione *fronetica*, che richiama la giurisprudenza della virtù[19], vale a dire il possesso della capacità di bilanciare i principi che guidano la funzione di utilità della decisione giuridica, esercitando quella particolare virtù dianoetica che Aristotele identificava nella *fronesis*[20].

simili.

[18] G.B. Vico, "De antiquissima italorum sapientia", in Id., *Opere filosofiche*, Sansoni, Firenze, 1971, p. 68.

[19] Sulla "giurisprudenza della virtù" cfr. L. Corso, "La virtù del giudice fra emozioni, giustizia particolare e ruolo istituzionale. Un abbozzo a partire da Aristotele", in *Ars Interpretandi*, 11 (2022), n. 1, pp. 53-69.

[20] Si veda specialmente il libro VI dell'*Etica Nicomachea*.

Un'altra condizione che dovrebbe essere soddisfatta dal "decisore digitale" è la *condizione di pubblicità* (la *Öffentlickeit* viene posta da Kant addirittura alla base della giuridicità stessa) del *software* che struttura l'impianto algoritmico. Le caratteristiche di progettazione e funzionamento del software alla base della decisione non possono restare inaccessibili, anche solo per rispetto del principio di motivazione deve essere sempre possibile accedere a qualsiasi *software* deputato alla decisione giuridica. Ma è una questione che si intreccia con il diritto commerciale e merita una riflessione non proponibile in questa sede.

4. Condizioni epistemiche

Gli algoritmi dell'intelligenza artificiale svolgono principalmente funzioni epistemiche, attraverso attività di classificazione, pianificazione e decisione, applicando regole preesistenti.

Allo stesso modo, la prima attività del giudicante, intesa come *prius*, è la sussunzione del fatto nella norma, attraverso la riconduzione della fattispecie concreta a quella astratta, dai fatti-dati alle norme.

Esattamente come *input* (e *output*) nel linguaggio digitale, *quaestio facti* e *quaestio iuris* sono le premesse logiche del giudizio giuridico, in cui fatto e fattispecie sono intimamente interconnessi.

Nell'universo di senso giuridico esiste (è giuridicamente rilevante) solo ciò che integra la fattispecie, la cosa interessante è che lo strato di senso epistemico della *cognizione* giuridica è sovrapponibile a quello che è lo schema dell'algoritmo incardinato nel periodo ipotetico "*If-Then*" che opera come il nesso di imputazione giuridica del *Deutungsschema* di tipo prescrittivo-sanzionatorio di Kelsen: "Se è A – allora (deve essere) B".

Sembra che fatto e fattispecie seguano la logica algoritmica delle proposizioni condizionali poggiate sullo stesso schema della protasi e dell'apodosi del periodo ipotetico.

Le analogie tra i due impianti cognitivi sono molto interessanti e operano entrambi una riduzione epistemica della realtà.

Il paradigma del giudizio algoritmico comporta una riduzione della realtà naturale, *analogica*, ontologicamente complessa, nella realtà digitalizzata. La differenza è di tutto rilievo, mentre il linguaggio analogico può assumere qualsiasi valore in qualsia-

si intervallo tra 0 e 1, il linguaggio digitale riduce tutto alla logica del valore binario 0/1 che non contiene altri valori all'interno.

Il modello di cognizione algoritmica e il modello di cognizione giuridica sembrano agire secondo la stessa logica: 0/1 – giuridicamente rilevante/giuridicamente irrilevante.

La natura analogica del reale è irriproducibile nel suo continuo e incessante divenire e per poter essere trattata attraverso metodi algoritmici l'informazione relativa ai fatti, alla realtà, deve essere necessariamente impoverita, proprio come la legge positiva che nella sua formulazione generale e astratta non può essere "tutto il diritto", non può prevedere ogni atto infinitamente orientato verso il seguire la regola, non può essere "una mappa grande quanto il territorio da rappresentare"[21].

La struttura epistemica dello *Jus ex machina*, non ha solo la capacità di sussumere il fatto nella fattispecie (*input*) e di esprimere il risultato giuridico applicabile (*output*) ma ha anche la capacità di orientare il giudizio elaborando i precedenti giudiziari.

Come si è visto, l'apprendimento automatico, il *machine learning*, consente di costruire, a partire dalle correlazioni esistenti all'interno di una grande massa di dati, alcuni modelli computazionali in grado di proporre una decisione.

La decisione secondo *precedenti* è logicamente connessa alla possibilità di una "giustizia predittiva" e si basa essenzialmente sul modello di Solomonov, vale a dire: una probabilità che si decida in futuro in una certa maniera è predeterminabile sulla base della divisione tramite algoritmo degli eventi raccolti in *cluster* differenti accomunati al loro interno da caratteristiche (*features*) comuni[22].

Il sistema di predizioni giuridiche opera su inferenze e correlazioni tra fatti e dati di tipo semantico, in modo "cieco", esclusivamente sul nesso *sintattico* tra *input* e *output* che, come schematizza Sartor,

> consisterebbero nell'inferenza logica tra premesse e conclusioni nel caso di sistemi basati sulle conoscenze e nella

[21] È una felice espressione di L. Lombardi Vallauri riportata da A. Incampo. Cfr. A. Incampo, *op. cit*., p. 13.

[22] A. Carcaterra, "*Machinae* autonome e decisione robotica", in A. Carleo, *op. cit.*, pp. 48-49.

correlazione statistica tra casi e predizioni per sistemi basati sull'apprendimento automatico[23].

Anche Carcaterra ha sottolineato, con spunti interessanti, i punti di contatto tra la decisione giuridica e quella delle macchine, secondo Carcaterra la decisione è azione collocata nel presente, si basa sull'esame dei dati passati, su uno sviluppo degli scenari futuri, sul rispetto di vincoli/regole ed è finalizzata, come ogni modello basato sulla teoria delle decisioni, sulla massimizzazione della funzione di utilità attesa (*expected utility*)[24].

Il focus decisivo sulle condizioni di sostituibilità del giudice con macchine dotate di intelligenza artificiale, non riguarda solo la capacità epistemica sussuntiva dei fatti nelle fattispecie e neanche solo la capacità elaborativa dei precedenti inferita da una massa di dati, ma riguarda anche la capacità di decidere in funzione dell'utilità attesa.

Come si può misurare l'utilità attesa? Il giudizio giuridico non riguarda solo la conoscenza di regole e precedenti, ma deve essere orientato all'interno del quadro dei principi, sono questi che teleologicamente orientano la funzione di utilità.

È il punto di maggior criticità della decisione giudiziale e, per certi versi, rinvia al più ampio tema del rapporto tra diritto e scienza. Campo su cui si gioca la questione del conflitto tra il diritto *science driven* e *science based*, riproposto nei termini di un conflitto tra diritto *data driven e* diritto *data based.* Il diritto e le decisioni connesse non possono essere guidati da un'epistemologia scientista e matematizzante, a rischio di incorrere in una fallacia epistemica in quanto diritto e scienza hanno uno statuto epistemologico differente[25].

[23] G. Sartor, *L'intelligenza artificiale e il diritto*, Giappichelli, Torino, 2022, p. 135.

[24] A. Carcaterra, *op. cit.*, p. 48.

[25] Sulla fallacia epistemica riguardo al rapporto tra diritto e scienza si veda A. Incampo, "*Extremus necessitatis casus.* Ai confini dell'ordine giuridico", in A. Incampo, L. Franzese (a cura di), *Potere e libertà al tempo delle emergenze*, Cacucci, Bari, 2021, pp. 23-43.

4.1. Una fallacia epistemica?

La questione della fallacia epistemica nel giudizio *ex machina* si manifesta sotto un duplice aspetto: rispetto al divario tra la cifra semantica del compito da eseguire e quella sintattica del programma che la sviluppa (a) e rispetto alla incalcolabilità della funzione di utilità attesa (b).

a) Notano Garapon e Lasségue come la fallacia riguardi innanzitutto il *vulnus* tra rappresentazione *semantica* del *compito* che la macchina deve portare a termine e la rappresentazione che avviene su un piano *sintattico* all'*interno* del programma da realizzare. C'è indubbiamente uno slittamento epistemico tra i due piani del linguaggio. Questo perché la scrittura informatica non possiede tutte le proprietà del linguaggio naturale, né, tantomeno, cosa non trascurabile, riesce a veicolare la performatività del linguaggio incardinato nel sistema simbolico rituale del giudizio[26].

b) La funzione dell'utilità attesa è messa in discussione quando la decisione tramite A.I. non si basa esclusivamente su identificazione di fattispecie (siano esse riguardanti norme o grandi masse di dati riguardanti precedenti) ma riflette sulla funzione di utilità stessa, sui propri principi.

È ciò che pone in evidenza, in dialogo ideale con Irti, Carcaterra quando propone una *metrica dei valori*[27] finalizzata alla soluzione di casi in cui il criterio decisorio si confronta con le *Tragic Choices*[28] o con uno scontro tra principi.

Il decisore giuridico segue una tassonomia decisoria per regole, precedenti e valori. La questione del giudizio per principi o valori può essere risolta attraverso una metrica dei valori che assegni *quantificazione* e *misurabilità*[29] ai principi valoriali in gioco. Come, ad esempio, nelle regole dei veicoli a guida autonoma il principio/valore "sicurezza" ha un "punteggio" maggiore del principio della percorrenza più rapida ed economica.

[26] A. Garapon, J. Lassègue, *La giustizia digitale. Determinismo tecnologico e libertà*, trad. it., il Mulino, Bologna, 2021, pp. 46-60.

[27] A. Carcaterra, *op. cit.*, pp. 55-61.

[28] Il riferimento è al classico di G. Calabresi, P. Bobbitt, *Tragic Choices*, Norton & Co., New York, 1978.

[29] A. Carcaterra, *op. cit.*, p. 57.

Pur ritenendo apprezzabile l'idea di giustizia procedurale di Carcaterra, utilizzare la metrica del valore nella progettazione di macchine per decisione giuridica non esaurisce la plasticità del fenomeno giuridico, né totalmente la "defettibilità" del ragionamento mediante principi.

L'argomentazione *defeasible* si differenzia dall'argomentazione deduttiva perché non è conclusiva: nell'argomentazione deduttiva la catena epistemica dell'inferenza dalla base di conoscenza alla decisione [*output*] passa linearmente e automaticamente attraverso il motore sillogistico-inferenziale, al contrario nell'argomentazione defettibile intervengono argomenti in conflitto che generano necessità di bilanciare i principi in gioco. Il bilanciamento dei principi in gioco nel caso concreto eccedono le competenze di calcolabilità algoritmica, si smarcano dalla conoscenza dimostrabile *more geometrico*, la decisione sul caso concreto apre alle *dissenting opinion*, al pensiero divergente, riporta l'eco, filtrata dal moderno costituzionalismo, delle stagioni della *Interessenjurisprudenz* e del *Freirechtsbewegung*.

Il decisore giuridico deve operare con *virtù* verso la scelta della decisione del caso concreto bilanciando i principi, proponendo talvolta soluzioni eque, disponendosi a scegliere il giusto mezzo[30], agendo con saggezza, con *fronesis*, che, per restare nel registro linguistico di Carcaterra, ha una funzione di utilità massima nella capacità di giudizio.

La decisione ottimale della macchina, che opera generando decisioni e contenuti sulla base di *cluster* delineati all'interno di una grande mole di dati immessi, non è capace di cogliere quell'eccedenza di senso che il testo giuridico veicola e che impegna pragmaticamente l'interprete alla ricerca del bilanciamento dei principi in gioco nel caso concreto, principi che sfuggono anche alla logica del precedente.

Riannodando le fila dell'analisi, sia la decisione per fattispecie che quella per precedenti condividono la logica della sussunzione in *cluster* e la logica della funzione di utilità orientata dal *cluster* formato dai dati del passato, non sembra, invece, condividere la stessa capacità elaborativa da parte di sistemi di *ma-*

[30] Aristotele, *op. cit*., II, 1106 b.

chine-learning (supervisionati e non supervisionati) la decisione per princìpi o per equità.

I principi sono criteri meta-positivi, non calcolabili: non hanno peso calcolabile negli algoritmi, hanno, per così dire, l'insostenibile leggerezza della incalcolabilità.

Il punto centrale resta la *funzione di utilità*, l'orizzonte teleologico che muove le decisioni della macchina e gli orientamenti del giudice, la capacità di orientarsi verso la decisione per valore.

Se la macchina vuole essere lo *Jus ex machina*, dirimente rispetto ai casi perplessi, tragici, della vita, deve essere capace di maneggiare la funzione di utilità propria del diritto.

La natura *defeasibility* dei principi, non sempre schematizzabili nella logica del *modus ponendo ponens*, mette a rischio le condizioni di compatibilità del giudizio macchina con il giudizio giuridico. Non è sufficiente la conoscenza epistemica del diritto, è necessaria una capacità fronetica nel giudizio.

5. Condizioni fronetiche

"Chi sa deliberare sarà saggio"[31]. E così che Aristotele ìndica icasticamente l'attributo del buon giudice. Per Aristotele non si può deliberare attorno al necessario, "non si può deliberare su ciò che non può essere diversamente", cioè sulla *scienza*, perciò la *fronesis*, la saggezza, "è una disposizione abituale dell'animo veritiero, unito a ragionamento pratico, che riguarda ciò che è bene e male per l'*uomo*"[32].

Con il richiamo alla *fronesis* nel giudizio ne va dell'essere dell'uomo.

Coglie nel segno Aldo Pisano per cui la

> *fronesis* richiama la questione secondo cui la macchina *conosce* ma non può agire secondo una propria ragion pratica […], per questo motivo la macchina soffre di un vero e proprio "deficit fronetico" anche riconducibile all'assenza della dimensione somato-percettiva e di quella emotiva[33].

[31] Aristotele, *Etica Nicomachea*, Laterza, Roma-Bari, 2005, p. 231.

[32] *ivi*, VI, 1140 b 4-6.

[33] A. Pisano, "La macchina e le forme dell'azione: deficit fronetico e autonomia artificiale", in *Mechane* (2021), n. 1, p. 46.

Si accennava alle "scelte tragiche", al conflitto tra principi giuridici fondamentali, che talvolta impegna conflitti morali non risolvibili con la logica del *calcolo* perché rientrano nella sfera esistenziale. La macchina non può farsi carico del piano *simbolico-esperienziale* perché resta ancorata a quello *simbolico-formale*.

La capacità fronetica è un elemento essenziale per la capacità di giudizio, per Aristotele è parte delle virtù dianoetiche che riguardano la parte razionale dell'anima che consente di conoscere e ragionare correttamente rispetto alle decisioni del caso concreto.

L'uomo virtuoso quando giudica, per dirla con Calvino, rispetto alla macchina conosce i "gelosi attributi dell'umano"[34].

Il giudizio giuridico sull'uomo non può che essere neurofenomenologicamente incarnato. *Fronesis* ed *embodiment*[35] sono legati nel diritto dalla sua caratteristica esperienziale, come insegna Capograssi[36].

Se il diritto è esperienza il significato delle norme trabocca dal segno linguistico, si fa esperienza sociale, incarnata nel corpo, che esperisce le relazioni giuridiche. Non è in questione il mero *Körper* ma il *Leib*, non il corpo biologico, ma quello che sente e che esperisce[37]. Non il corpo che "abbiamo" ma il corpo che "siamo". La disposizione fronetica, la saggezza del caso concreto, segue l'ordine esperienziale, è maggiormente aperta alla *dissenting opinion* à là Holmes, o a ciò che Calvino chiama pensiero divergente.

Ma non è tutto, come si è accennato, la disposizione fronetica del giudizio si articola, nel suo essere *medium* tra conoscenza razionale e conoscenza pratica, nella cognizione incarnata [*embodied cognition*]. Quello che una *machina* è impossibilitata a fare è *sperimentare* attraverso il proprio corpo un significato. Non si

[34] I. Calvino, "Cibernetica e fantasmi", in Id., *Una pietra sopra*, Mondadori, Milano, 1995, pp. 210-211.

[35] Il concetto di *necessary embodiment* proposto da C. Taylor, *Hegel*, Cambridge University Press, Cambridge, 1975, ha trovato successivo sviluppo negli studi di neurofenomenologia. Cfr. M. Cappuccio (a cura di), *Neurofenomenologia. Le scienze della mente e la sfida dell'esperienza cosciente*, Mondadori, Milano, 2006.

[36] Cfr. G. Capograssi, *Opere*, II, Giuffrè, Milano, 1959.

[37] E. Husserl, *Idee per una fenomenologia pura e per una filosofia fenomenologica*, II, trad. it., Einaudi, Torino, 2002.

tratta semplicemente di rispecchiarsi nell'altro attraverso il sistema di simulazione incarnata del corredo dei neuroni specchio[38], ma si tratta di possedere veri e propri *Embodied Schemata* di comprensione di concetti astratti, di significati di "stati" mentali che secondo la teoria bio-funzionale di Mark Johnson sono comprensibili esclusivamente da un punto di vista incarnato. E sono "stati" che per il giudizio giuridico sono importanti, ad esempio quello di buona fede, buon padre di famiglia, gli stati passionali, l'*affectio maritalis*, il concetto di reato putativo, *etc. etc*. La comprensione concettuale secondo Lakoff e Johnson[39] vive nella comprensione di stati attraverso una mappatura di schemi concettuali metaforicamente incentrati su domini esperienziali, si tratta di una capacità *enattiva* che nel giudizio è veicolata dalla *fronesis* e sfugge alla capacità calcolante degli algoritmi.

Le condizioni fronetiche, richiamate necessariamente dal processo epistemico del giudizio, sono irriducibili al sapere calcolante, sfuggenti e incarnate, hanno l'insostenibile leggerezza dell'incalcolabile.

6. L'insostenibile leggerezza dell'incalcolabile e le forme di messa in ordine del mondo

L'alba di una rivoluzione, di un cambiamento di paradigma, crea sempre inquietudine, smarrimento. Quando si tratta, per giunta, di una simulazione dell'attributo umano di giudizio, la macchina opera come *das Unheimliche*, come nota Antonio Punzi riprendendo un celebre scritto di Freud[40], come il *perturbante* che ingenera angoscia e orrore, non solo perché ci somiglia, tanto più oggi che vuole sostituirsi all'uomo, ma anche perché "l'uomo sa di essere *anche* una macchina" e percepisce la

[38] Sul sistema *mirror* della molteplicità condivisa, intesa come cognizione sociale cfr. V. Gallese, "La molteplicità condivisa. Dai neuroni *mirror* all'intersoggettività", in S. Mistura (a cura di), *Autismo: l'umanità nascosta*, Einaudi, Torino, 2006, pp. 302-335; Id., "Corpo vivo, simulazione incarnata, intersoggettività. Una prospettiva neuro-fenomenologica", in M. Cappuccio (a cura di), *op. cit.*, pp. 293-326.

[39] G. Lakoff, M. Johnson, *Philosophy in the Flash: The Embodied Mind and Its Challenge to Western Thought*, Basic Books, New York, 1999.

[40] S. Freud, "Il perturbante", in *Saggi sull'arte, la letteratura e il linguaggio*, trad. it., Torino, Boringhieri 1991.

macchina (o il robot) calcolante come qualcosa di familiare, ma allo stesso tempo "raggelante"[41], come "il suo fantasma"[42].

Certo, come si diceva, le esigenze di certezza del diritto, messe in discussione, per citare Irti, dalla "crisi della fattispecie e della decodificazione"[43], oltre a una incontrollabile e rizomatica gerarchia delle fonti e degli organi giudiziari (con il proliferare di quelli sovranazionali e del "*soft law*") quasi invocano lo *Jus ex machina*. Una sorta di intervento risolutivo della "macchina intelligente" così come nella tragedia greca l'*apò mechanés theós* compariva nella messa in scena teatrale, che tanto ricorda la *mise-en-scène* processuale, al fine di dirimere questioni e dispute irrisolvibili.

Due paradigmi tendono, in definitiva, a contrapporsi quando si guarda alle condizioni di sostituibilità del giudicante ad opera della macchina dotata di intelligenza artificiale, sono simili a quelle che Garapon e Lassègue chiamano *forme di messa in ordine del mondo*[44]: la forma "automatizzata" e la forma "autonoma". Queste, in qualche modo veicolano due tipi di efficienza processuale: quella procedurale e quella sociale.

La forma *automatizzata* di messa in ordine del mondo si regge sull'idea del principio del mondo stabile[45], e tende attraverso il sistema di calcolo a stabilizzare le aspettative facendo leva sull'estrema predicibilità delle sentenze e sull'assoluta imparzialità del giudicante, lontano dai rischi della *headache jurisprudence*, o di *bias* sempre in agguato, su tutti il *bias* della rappresentatività o della disponibilità[46].

[41] È l'espressione utilizzata da S. D'Amico, *Autori del Teatro ceco*, in *La tribuna*, 03/10/1930, p. 3. A commento della prima opera letteraria (distopica) in cui è presente il *robot*. Si tratta dell'opera K. Čapek, *Rossumovi univerzální roboti*, pubblicata nel 1920, che ovviamente narra della trasformazione dei robot da servi a dominatori.

[42] A. Punzi, "*Judge in the machine*. E se fossero le machine a restituirci l'umanità del giudicare?", in A. Carleo, *op. cit.*, p. 319

[43] N. Irti, "La crisi della fattispecie", in *Rivista di diritto processuale civile*, 1 (2014), pp. 36-44.

[44] A. Garapon, J. Lassègue, *op. cit.*, p. 158.

[45] Sul "principio del mondo stabile" cfr. G. Gigerenzer, *Perché l'intelligenza umana batte ancora gli algoritmi*, Raffaello Cortina, Milano, 2023, pp. 54-57.

[46] Per un approfondimento sul tema del ruolo dei *bias* nella decisione giudiziale si rinvia a M. Sozio, "La decisione giudiziale in condizioni di in-

La forma *autonoma* dell'applicazione del diritto affida al giudice l'efficienza sociale del giudizio, quella "incarnata" nell'esperienza, quella che si pone in consonanza con le esigenze sociali e culturali ed è capace di intercettarne i cambiamenti, anche perché li percepisce "*embodied*" su se stesso, che utilizza la propria capacità emotiva come "antidoto alla rigidità percettiva"[47]: quella che può fare affidamento su una saggezza pratica.

Un sistema giudiziario perfetto probabilmente è quello che riesce a creare la migliore sinergia tra i due tipi di efficienza, quella *procedurale* e quella *sociale*.

È questo il punto. La grande trasformazione digitale genera inquietudini, apre alla paura di un dislivello prometeico tra gli uomini e le macchine che l'avvento della Singolarità, come visto, potrebbe amplificare. Non è un caso che la letteratura distopica, come ad esempio il "racconto di fantascienza giudiziaria" *Justice Machines* di Jacques Charpentier (1954)[48], tratteggi un futuro in cui lo *Jus ex machina* amministri la società attraverso un'efficienza del calcolo, assolutamente pura, imparziale, e per questo equanime. La distopia si annida proprio nel deficit fronetico, in cui una macchina giudiziaria "a guida autonoma" dispone senza avere la virtù "incarnata" del buon giudice.

Resta praticabile la ricerca della sinergia tra efficacia calcolante-procedurale ed efficacia simbolico-sociale del processo, una sinergia che indica la strada di un processo simbiotico "uomo-macchina" e che affidi le sorti di cittadini in attesa di giustizia più che a una macchina della giustizia a guida autonoma, alla virtù di un giudice uomo assistito da una macchina[49], l'unico in grado di governare l'insostenibile leggerezza di ciò che è incalcolabile.

certezza: tra distorsioni cognitive e neuroscienze", in *Quaderni del Dipartimento Jonico*, 11 (2019), pp. 355-370.

[47] L. Corso, *op. cit.*, p. 60.

[48] J. Charpentier, *Justice Machines*. *Racconto di fantascienza giudiziaria*, trad. it., Liberilibri, Macerata, 2023.

[49] La metafora dell'auto a guida assistita è felicemente utilizzata anche in A. Punzi, *op. cit.*, p. 330.

Autori

Ali Shan	*Dottorando in Filosofia del diritto*, Università degli Studi di Padova
Amato Salvatore	*Ordinario di Filosofia del diritto*, Università degli Studi di Catania
Amato Mangiameli Agata C.	*Ordinaria di Filosofia del diritto*, Università degli Studi di Roma Tor Vergata
Campagnoli Maria Novella	*Associata di Filosofia del diritto*, Università degli Studi di Roma Tor Vergata
D'Agosto Sveva	*Dottoranda in Filosofia del diritto*, Università degli Studi di Bari Aldo Moro
D'Avanzo Wanda	*Assegnista in Filosofia del diritto*, Università degli Studi "Magna Graecia" di Catanzaro
Farina Massimo	*Associato di Filosofia del diritto*, Università degli Studi di Cagliari
Franzese Lucio	*Ordinario di Filosofia del diritto*, Università degli Studi di Napoli "Federico II"
García-Berrio Teresa María	*Associata di Filosofia del diritto*, Università Complutense di Madrid (Spagna)
Loddo Olimpia Giuliana	*Ricercatrice di Filosofia del diritto*, Università degli Studi di Cagliari
Marra Piero	*Ricercatore di Filosofia del diritto*, Università degli Studi di Bari Aldo Moro

Moro Paolo	*Ordinario di Filosofia del diritto*, Università degli Studi di Padova
Muscatiello Vincenzo Bruno	*Ordinario di Diritto penale*, Università degli Studi di Bari Aldo Moro
Pulito Lorenzo	*Ricercatore di Diritto processuale penale*, Università degli Studi di Bari Aldo Moro
Romeo Francesco	*Ordinario di Filosofia del diritto*, Università degli Studi di Napoli "Federico II"
Santamaría Ramos Francisco José	*Professore a contratto di Filosofia del diritto*, Università Complutense di Madrid (Spagna)
Saraceni Guido	*Associato di Filosofia del diritto*, Università degli Studi di Teramo
Sommaggio Paolo	*Associato di Filosofia del diritto*, Università degli Studi di Padova
Sozio Maurizio	*Aggregato di Filosofia del diritto*, Università degli Studi di Bari Aldo Moro
Strugała Radosław	*Associato in Diritto privato e Diritto privato internazionale*, Università di Wrosław (Polonia)

Indice dei nomi

Abbagnano Trione, Andrea, 245n, 246n.
Adam, George Alexandru, 48n.
Adcock, Brett, 314.
Addis, Andrea, 28n.
Agazzi, Evandro, 373.
Aghihotri, Arpita, 172n.
Agre, Philip E., 36n.
Agrusti, Stefania, 91n.
Aiken, Michael, 437n.
Akbari, Mehdi, 450n.
Alberta, Valentina, 247n.
Alemohammad, Sina, 43n, 55n.
Ali, Raian, 450n.
Ali, Shan, 6, 12, 95n, 257 e n, 475.
Al-Khalili, Jim, 322 e n.
Allegri, Maria R., 447n.
Allen, Colin, 420n, 431n.
Allen, Layman Edward, 78n.
Alò, Loredana, 2.
Alonso González, Marián, 162n.
Al-Thani, Dena, 450n.
Altman, Sam, 60.
Alutaybi, Aarif, 450n.
Alvino, Francesco, 246n.
Amarelli, Giuseppe, 245n, 247n.
Amato Mangiameli, Agata Cecilia, 5, 10, 19, 22n, 24n, 34n, 35n, 37n, 40n, 73n, 74n, 76n, 79n, 177n, 430n, 436n, 440n, 452n, 475.
Amato, Cristina, 172n.
Amato, Salvatore, 6, 13, 79n, 81n, 180n, 307, 475.
Amato, Stefania, 248n.
Amaturo, Enrica, 40n.
Ammanniti, Massimo, 335n.
Amodei, Dario, 317n, 415n.
Anders, Günther, 309 e n.
Anderson, Chris, 39n.
Andorno, Roberto, 266n, 275n, 276n, 277, 279n, 365n.
Andrews, Donald Arthur, 242n.
Andriani, Roberto, 41n.
Angwin, Julia, 234n.
Anscombe, Gertrude Elizabeth Margaret, 400 e n.
Antonelli, Mauro, 110n, 112n.
Antonovsky, Aaron, 439n.
Aonuma, Hitoshi, 256n.
Applegarth, Daniel Michael, 253n, 255n.
Aragona, Biagio, 40n.
Araiz Huarte, David Edgar, 149n.

Arakawa, Naoya, 424.
Araszkiewicz, Michal, 102n.
Archambault, Laurent, 297n.
Archer, Margaret Scotford, 423n.
Aristotele [latino: *Aristoteles*] di Stagira, 67, 78 e n, 80n, 83n, 84n, 87n, 109n, 112n, 114, 121, 326, 435 e n, 436, 464 e n, 469n, 470 e n, 471.
Arnese, Aurelio, 84n.
Asimov, Isaac, 432, 433n.
Atkinson, Ian, 173n.
Atrill-Smith, Alliso, 437n.
Attanasio, Antonino, 221n, 224n.
Ausín, Txetxu, 124n.
Avila Negri, Sergio, 433n.
Avitabile, Luisa, 62n, 90n.
Aviv, Oshrit, 191n.

Baccari, Gian Marco, 231n, 232n, 234n, 235n, 236n.
Baker, Stephanie, 173n.
Baliño, Juan Pablo Pampillo, 124n.
Bang, Yejin, 45n.
Barbaro, Clementina, 241n.
Barboni, Lucía, 240n.
Barcellona, Mario, 84n.
Barfield, Woodrow, 230n.
Barrat, James, 34n, 36n.
Barrick, Kelle, 239n.
Barron, Amy, 94n.
Barta, Paweł, 291n.
Bartel, Patrick A., 239n.
Bartoli, Roberto, 244n, 245n, 249n.
Basile, Fabio, 219n, 220n, 222n, 233n.
Bassi, Alessandra, 246n.
Bathaee, Yavar, 46n.
Battezzati, Luigi, 180n.
Bauman, Zygmunt, 347, 348, 349 e n, 350-352, 443n.
Baumgartner, Peter, 239n.
Bauzon, Stéphane, 86n.
Beauchamp, Tom Lamar, 382 e n.
Becchi, Paolo, 36n.
Beck, Ulrich, 350 e n, 351n.
Beckers, Anna 73n, 395, 396 e n, 397-399 e n, 400 e n, 401-403, 407 e n, 409.
Beck-Gernsheim, Elisabeth, 348, 351n.
Beitz, Charles R., 274n.
Beltrametti, Monica, 412n.
Beltran de Heredia Ruiz, Ignasi, 361 e n, 362.
Benanti, Paolo, 70n, 320n, 429n.
Bender, Emily, 343 e n.
Bengio, Yoshua, 46n, 313n.
Benson, Vladlena, 437n.
Bentivegna, Sara, 447n.
Benvenuti, Feliciano, 64n.
Bernstein, Jeremy, 317e n.
Bettiol, Giuseppe, 64 e n.
Bhattacharya, Saurabh, 74n, 172n.
Bianchi, Davide, 246n.
Biasiotti, Maria Angela, 52n.
Bickerton, Derek, 322n.
Bigenwald, Ariane, 264n.

Biondi, Giuseppe, 252n.
Bishop, Cristopher M., 336n.
Bishop, Mark, 427n.
Bisschop, Arno, 278n.
Bittman, Ladislav, 41n.
Blakeslee, Sandra, 28n.
Bloch, Ernst, 331n.
Bobbio, Norberto, 77n, 78n, 193n.
Bobbitt, Philip Chase, 468n.
Boccaccini, Federico, 110n, 112n.
Boccia Artieri, Giovanni, 447n.
Bochicchio, Mario Alessandro, 80n.
Boddington, Ghislaine, 173n.
Boehm, Christopher, 105n.
Boire, Richard Glen, 268n.
Bolukbasi, Tolga, 41n.
Bommasani, Rishi, 74n.
Bonavina, Luigi, 66 e n.
Bonta, James, 242n.
Boole, George, 246.
Borgia, Francesco, 36n.
Borja, Martínez-Pérez, 191n.
Borruso, Renato, 36n.
Borsari, Riccardo, 220n.
Borum, Randy, 239n.
Bostrom, Nick, 47n, 271n.
Botezatu, Alex, 269n.
Bottari, Carlo, 35n.
Bou Pastor, Mónica, 143n.
Bowart, Walter Howard, 280n.
Boyd, Robert, 105n.
Braun, Joachim von, 423n.
Brennan, Tim, 233n.
Brentano, Franz Clemens Honoratus Hermann, 101n, 110n, 112n.
Brighi, Raffaella, 181n, 414n.
Brighton, Henry, 130n.
Brittain, Brian, 239n.
Brooks, David, 39n, 40n.
Brotherton, Rob, 449n.
Brownsword, Roger, 452n.
Brugnoli, Roberto, 445n.
Brynjolfsson, Eric, 36n.
Bublitz, Jan-Christoph, 268n, 277n, 281 e n.
Bucalo, Maria Esmeralda, 447n.
Bucci, Enrico, 325n.
Buccino, Giovanni, 339 e n.
Budiansky, Stephen, 325n.
Buffardi, Annalisa, 75n.
Burchard, Christoph, 200n.
Burt, Andrew, 342n.
Buscemi, Antonino, 180n.
Busto, Nicola, 432n.
Butler, Judith, 342n.

Cabitza, Federico [Antonio Niccolò Amedeo], 80n.
Cacioppo, John Terrence, 260n.
Caianiello, Michele, 232n, 236n.
Calabresi, Guido, 468n.
Calavita, Oscar, 246n.
Calvillo-Arbizu, Jorge, 173n.
Calvino, Italo, 471 e n.
Calvo Palomares, Ricardo, 143n.
Camera, Guido, 252n.

Campagnoli, Maria Novella, 5, 15, 40, 177n, 440n, 475.
Campanale, Anna Maria, 90n.
Campara, Flavio, 199n, 207n.
Canessa, Nicola, 339n.
Canzio, Giovanni, 199n, 231n, 236n, 241n.
Čapek, Karel, 419 e n, 473n.
Capo, Carmine, 452n.
Capograssi, Giuseppe, 63n, 471n.
Capparelli, Marilu, 41n.
Cappuccio, Massimiliano, 472n.
Carcaterra, Antonio, 466n, 467 e n, 468 e n, 469.
Caretti, Vincenzo, 445n.
Carleo, Alessandra, 36n, 77n, 89n, 459n, 466n, 473n.
Carnap, Paul Rudolf, 100, 103.
Carnevale, Antonio, 94n, 243n, 254n.
Carrer, Stefania, 61n, 235n.
Cartesio, Renato [*Descartes, René*], 112n.
Casalegno, Daniele, 36n.
Casco-Rodriguez, Josue, 43n 55n.
Caselli, Gabriele, 450n.
Casonato, Carlo, 41n 187n, 236n.
Castellanos Claramunt, Jorge, 387 e n.
Castronuovo, Donato, 244n.
Casula, Federica C., 455n.
Catalano, Alessandro, 419n.
Cavaceppi, Cecilia, 202n, 221n.
Cavaliere, Armando, 188n.
Cavaliere, Antonio, 246n.
Cavalli Sforza, Luigi Luca, 105n.
Cave, Stephen, 308n.
Ceci, Michelangelo, 77n.
Čerka, Paulius, 46n.
Cesarini, David, 105n.
Chambon, Valerian, 264n.
Chang, Kai-Wai, 41n.
Charles, Wendy M., 278n.
Charpentier, Jacques, 474 e n.
Chatila, Raja, 412n.
Chazerand, Patrice, 412n.
Chen, Feng, 35n.
Chiereghin, Franco, 427n.
Child, Rewon, 415n.
Childress, James Franklin, 382 e n.
Chiriatti, Massimo, 327n.
Chittaro, Luca, 36n.
Chu, Kayla, 81n.
Cicero, Vincenzo, 93n.
Ciesla, Robert, 437n.
Cilardo, Chiara, 45n.
Cipolla, Carlo Maria, 307n.
Cirone, Enza, 446n.
Citati, Pietro, 426n.
Civita, Anna, 443n.
Clegg, Brian, 446n.
Coeckelbergh, Mark, 66n, 330n.
Cohen, Jonathan, 262n.
Comfort, Megan, 239n.
Conill Sancho, Jesús, 374.
Contaldo, Alfonso, 199n, 200n, 202n, 207n.

Conte, Amedeo Giovanni, 82n, 90n.
Contissa, Giuseppe, 225n, 236n.
Copeland, Jack, 36n.
Coplan, Amy, 340n.
Coppola, Fabio, 252n, 253n, 255n.
Corciulo, Simona, 80n.
Corea, Francesco, 37n.
Corona, Fabrizio, 89n, 462n.
Corso, Lucia, 84n, 464n, 474n.
Cortina, Adela, 126n, 347, 374, 392.
Costa, Pedro, 342n.
Costabile, Gerardo, 221n, 224n.
Cotino Hueso, Lorenzo, 387 e n.
Courville, Aaron, 46n.
Cowls, Josh, 412n.
Crafa, Silvia, 89n, 420n.
Crawford, Kate, 225n.
Cristianini, Nello, 59n, 65n, 418n, 428n, 430n.
Croce, Benedetto, 346 e n.
Crosby, Michael, 269n.
Cukier, Kenneth, 446n.
Cuomo, Paolo, 446n.
Cusano, Nicola *vel* Niccolò Cusano, 112n.
D'Acquisto, Giuseppe, 94n.
Dafoe, Allan, 25n.
D'Agostino, Lorenzo, 233n, 235n, 237n.
D'Agosto, Sveva, 312, 333, 473.
D'Aloia, Antonio, 41n, 51n.
Dallery, Jesse, 80n, 174n.
D'Amico, Silvio, 473n.
Daraio, Girolamo, 246n.
D'Avack, Lorenzo, 35n.
D'Avanzo, Wanda, 11, 171, 475.
D'Aversa, Annunziata, 77n.
Davenport, Thomas, 35n.
Dawes, Christopher T., 105n.
D'Elia, Giacomo, 2.
Deacon, Terrence William, 321n.
De Anna, Gabriele, 91n.
De Bona, Giulia, 430n.
De Collibus, Francesco Maria, 202n.
De Felice, Fabio, 326n.
De Francesco, Giovannangelo, 250n.
De Haan, Cody R., 450n.
de Kerckhove, Derrick, 75n, 312n.
De la Torre-Díez Isabel, 191n.
Dell'Arte, Salvo, 53n.
Del Mastro, Marco, 446n.
De Martino, Graziella, 77n.
DeMichele, Matthew, 239n.
Deng, Xuechao, 35n.
Dennett, Daniel C., 215n.
Desmarais, Sarah, 242n.
De Vaal, Martijn, 448n.
Devillers, Laurence, 423n.
De Vito, Roberto, 246n, 247n.
de Vries Robbé, Michiel, 239n, 240n.
Dewey, John, 436n.
de Zan, Julio, 126n.
Di Cesare, Donatella, 114n.

Dick, Philip K., 412n.
Dieterich, William, 233n.
Dignum, Virginia, 412n.
Dihal, Kanta, 308n.
Di Lucia, Paolo, 82n, 90n.
Di Michele, Matteo. 204n.
Ding, Wenxin, 41n.
Di Paolo, Gabriella, 236n.
Di Porto, Andrea, 90n.
Di Prospero, Alfonso, 40n.
Dix, Alan, 80n, 81n, 95n.
Djeffal, Christian, 182n.
Dolcini, Emilio, 246n.
Domingos, Pedro, 318n, 429n.
Donati, Filippo, 35n, 51n, 236n.
Donini, Massimo, 246n.
D'Onofrio, Cristina, 189n.
Dorati, Marco, 78n.
Dorigo, Stefano, 233n, 446n.
Douglas, Kevin S., 239n, 242n.
Dressel, Julia, 237n.
Ducrotay de Blainville, Henri Marie, 107n.
Dupuy, Jean-Pierre, 318n.
Dvoskin, Joel, 242n.
Dyoub, Abeer, 95n.
Dziri, Nouha, 45n.

Eagleman, David M., 246 e n.
Eaglin, Jessica, 233n.
Eco, Umberto, 36n, 222.
Edwards, Lilian, 289n.
Ehret, Beate, 233n.
Elhalal, Anat, 230n.
Eliot, Thomas Stearns, 326 e n.
Ellul, Jacques, 309 e n.
Engelhardt, Hugo Tristram Jr., 381n.
Errigo, Maria Chiara, 74n.
Esposito, Elena, 41n.
Es-Sbai, Hadil, 278n.
Eusebi, Luciano, 246n.
Eysenbach, Gunther, 45n.

Fabiano, Laura, 96n.
Fabris, Adriano, 420n.
Faggin, Federico, 425n.
Faigman, David L., 261n.
Farabee, David, 233n.
Farah, Martha Julia, 261n, 271n, 179n.
Faralli, Carla, 193n.
Farid, Hany, 237n.
Farina, Massimo, 5, 10, 33, 34n, 50n, 54n, 79n, 177n, 433n, 475.
Faro, Sebastiano, 82n.
Febbrajo, Alberto, 406n.
Felicioni, Paola, 231n, 232n, 234n, 235n, 236n.
Ferola, Laura, 173n.
Ferrajoli, Luigi, 90n, 102n.
Ferrari, Gianfranco A., 78n.
Ferrari, Maurizio, 37n, 52n.
Ferrari, Vincenzo, 62n.
Ferraris, Maurizio, 100 e n, 114n.
Ferro, Fulvio, 258n.
Fichte, Johann Gottlieb, 112n.

Finocchiaro, Giusella Dolores, 72n, 207n, 208n, 209n, 215n, 217n, 219n.
Fiorentin, Fabio, 246n.
Fiorinelli, Gaia, 200n, 201n.
Fiorio, Carlo, 235n.
Fiormonte, Domenico, 337n, 340n.
Fittipaldi, Edoardo, 90n.
Flach, Joost, 278n.
Flaxman, Seth, 330n.
Fleisch, Elgar, 174n.
Floridi, Luciano, 48n, 68n, 70n, 73n, 80n, 93n, 230n, 278n, 310 e n, 314n, 324 e n, 336n, 347, 365, 366 e n, 367, 391, 392, 412n, 428n, 446n, 447n, 448n, 458 e n.
Floyd, Juliet, 91n.
Forth, Adelle, 239n.
Forza, Antonio, 231n.
Foschini, Gaetano, 230n.
Fossa, Fabio, 37n.
Foucault, Paul-Michel, 367n.
Fowler, James H., 105n.
Franks, David D., 260n.
Franz, Christoph, 174n.
Franzese, Lucio, 5, 11, 59, 64n, 89n, 467n, 475.
Franzosi, Mario, 52n.
Frege, Friedrich Ludwig Gottlob, 100 e n, 101 e n, 102 e n, 103 e n, 119.
Freud, Sigmund, 426 e n, 472 e n.
Frosini, Tommaso Edoardo, 82n.
Fu, Bo, 81n.
Fuller, Lon Luvois, 82n, 83n, 90n.
Fuselli, Stefano, 84n, 182n, 425n.

Gal, Michal S., 191n.
Galgani, Benedetta, 233n.
Galimberti, Umberto, 37n.
Gallese, Vittorio, 335n, 338n, 339n, 472n.
Galstyan, Aram, 63n.
Gamberini, Luciano, 36n.
Gao, Zhiyuan, 35n.
Garapon, Antoine, 63n, 73n, 80n, 82n, 200n, 207n, 468 e n, 473 e n.
García-Berrio, María Teresa, 6, 13, 347, 475.
Gardner, Howard, 40n.
Gargani, Alberto, 244n, 245n, 246n.
Galgano, Francesco, 432.
Garrat, Chris, 126n.
Gatsby, Peter, 81n.
Gatta, Gian Luigi, 246n.
Gatzonis, Stylianos, 178n.
Gawdat, Mo, 310, 311n.
Gazzaniga, Michael S., 261n.
Gebru, Timnit, 343n.
Gehlen, Arnold, 34n.
Gellert, Raphaël, 289n.
Gentile, Francesco, 63 e n, 68n.
Georges, Leah, 239n.
Geraci, Rosa Maria, 245n.
Germani, Emiliano, 173n.
Gerola, Alessio, 258n.

Gialuz, Mitja, 252n, 253n.
Gianni, Robert, 66n.
Giannini, Alice, 200n.
Gigerenzer, Gerd, 473n.
Gilbert, Paul, 455n.
Gilbert, Stephen, 278n.
Gillies, Donald Angus, 418n.
Ginsburg, Simona, 323n.
Giordano, James, 259n, 273n.
Giova, Stefania, 90n.
Gladwell, Valerie, 450n.
Gnoli, Antonio, 335n.
Gödel, Kurt Friedrich, 91 e n.
Godfrey-Smith, Peter, 323n.
Goldberg, David Edward, 46n.
Goldberg, Ivan, 444.
Gómez Pavajeau, Carlos Arturo, 140n.
Gómez, Santiago Botero, 124n.
Goodfellow, Ian, 46n.
Goodman, Bryce, 330n.
Goodrich, Michael A., 430n.
Goretta, Renato, 194n.
Gorgoni, Guido, 66n.
Gould, Carol C., 175n.
Gould, Glenn, 322.
Grace, Katja, 25n.
Granieri, Giuseppe, 447n.
Grassi, Edmondo, 51n.
Greely, Henry T. [*vel* Hank], 266 e n, 267 e n.
Greene, Joshualand Brent, 262 e n.
Greene, Kristen, 342n.
Greengard, Mig, 319n.
Gregory, Richard Langton, 459n.
Grieco, Chiara, 95n.
Griffiths, Mark D., 443 e n.
Grigienė, Jurgita, 46n.
Guardado, Nicolas, 81n.
Guarino, Alfonso, 452n.
Guerra, Michele, 339n.
Guerrini, Roberto, 249n.
Guidi, Dario, 247n.
Guizzardi, Silvia, 52n.
Günther, Klaus, 397n.
Guo, Priscilla, 234n.
Guthrie Ferguson, Andrew, 446n.
Gwinn, Jason, 242n.

Habermas, Jürgen, 329 e n, 373, 374, 408n, 463n.
Hagan, Margaret, 192n.
Haibe-Kains, Benjamin, 48n.
Hajin, Lee [Ha-jin, Lee], 116n.
Halberstam, Judith, 343n.
Hall, Patrick, 342n.
Hamilton, Zachary, 242n, 254n.
Han, Byung-Chul, 430n, 447n.
Hao, Rui, 243n.
Harari, Yuval Noah, 40n.
Haraway, Donna J., 206n, 208n.
Hare, Richard Mervyn, 84n.
Harel, David, 247n.
Harnad, Stevan, 427n.
Hasegawa, Eisuke, 256n.
Hawking, Stephen, 321n.
Hawkins, Jeff, 26n.

Hayes, Patrick Joseph, 417 e n.
Hayles, Nancy Katherine, 324n.
Hegel, Georg Wilhelm Friedrich, 112n.
Heidegger, Martin, 365 e n.
Heisenberg, Werner Karl, 115n.
Helzel, Paola Barbara, 93n, 189n.
Herrmann, Andreas, 174n.
Hester, Rhys, 242n.
Hilbert, David, 90, 91n.
Hildebrandt, Mireille, 183n, 434n.
Hinton, Geoffrey, 313n.
Höffe, Otfried, 85n.
Hofstadter, Douglas Richard, 215n.
Hogger, Christopher John, 79n.
Holland, John Henry, 46n, 79n.
Holmes, Oliver Wendell Jr., 92, 471.
Hosny, Ahmed, 48n.
Hottois, Gilbert, 373.
Howard, Clint, 44n.
Hu, Linmei, 243n.
Huang, Yu-Ping, 344n.
Huberman, Bernardo A., 45n.
Hume, David, 112n.
Huq, Aziz Zahirul, 63n.
Husserl, Edmund Gustav Albrecht, 334 e n, 471.
Hydén, Håkan, 396n, 397n.
Hygounet, Jean Louis, 180n.

Ianulardo, Mario, 221n, 224n.
Ienca, Marcello, 266n, 274, 275n, 276n, 277, 279n, 354n.
Incampo, Antonio, 78n, 85n, 86n, 87n, 88n, 91n, 93n, 463 e n, 466n, 467n.
Ippocrate [latino: *Hippocrates*], 66.
Irti, Natalino, 36n, 62n, 459n, 468, 473 e n.
Irving, Emma, 449n.
Isgur Flores, Sarah, 264n.

Jablonka, Eva, 323n.
Jaquet-Chiffelle, David-Oliver, 434n.
Jiang, Liang, 268n.
Jiang, Yuan, 45n.
Joffe, Helene, 260n.
Johannesson, Magnus, 105n.
Johnson, Mark, 472 e n.
Jonas, Hans, 34n, 364 e n, 365, 370, 371 e n, 372, 373, 378, 391, 422n.
Jones, Owen D., 261n.
Jori, Mario, 90n.
Jotterand, Fabrice, 259n.
Jurado Almonte, José Manuel, 143n.

Kahneman, Daniel, 441n.
Kaku, Michio, 309n, 323n.
Kalai, Adam, 41n.
Kalakota, Ravi, 35n.
Kalyanaraman, Vignesh, 269n.
Kanazawa, Junichiro, 450n.
Kant, Immanuel, 93n, 112n, 465.

Kaplan, Jerry, 216n.
Karnouskos, Stamatis, 254n.
Karp, Beth, 243n.
Kasparov, Garry, 203, 319 e n.
Kato, Takahiro A., 450n.
Katsuki, Ryoko, 450n.
Katz, Daniel Martin, 254n.
Kawecki, Maciej, 291n.
Keen, Andrew, 312n.
Kehl, Danielle, 234n.
Kelsen, Hans, 88n, 102n, 119, 120, 328 e n, 465.
Kessler, Samuel, 234n.
Kigerl, Alex, 254n.
Kinsey, Libby, 230n.
Kirchner, Lauren, 234n.
Kirsten, Toralf, 278n.
Kleeven, Anne Theodora Hermina, 239n, 240n.
Kolber, Adam, 261n, 262n.
Koops, Bert-Japp, 434n.
Kostoris, Roberto, 230n.
Kottow, Miguel H., 381 e n.
Kowalski, Melissa, 254n.
Kowalski, Robert Anthony, 79n.
Kroner, Daryl G., 242n.
Kubo, Hiroaki, 450n.
Kubrick, Stanley, 207n.
Kudan, Rie, 20n.
Kudo, Tatsumi, 256n.
Kurzweil, Raymond, 462 e n.
Kymlicka, Will, 376n.

La Barbera, Daniele, 445n.
La Rosa, Emanuele, 224n.
Labatut, Benjamín, 319n.
Labrecque, Ryan, 242n.
Lachmayer, Friedrich, 102n.
Lafuente, Alberto L., 447n.
Lagioia, Francesca, 234n.
Lagravinese, Giovanna, 339n.
Lakoff, George Philip, 472 e n.
Lanier, Jaron, 441n, 447 e n.
Larriviere, Daniel, 261n.
Larson, Jeff, 234n.
Lasagni, Giulia, 199n, 236n.
Lassègue, Jean, 63n, 80n, 82n, 200n, 207n, 468 e n, 473 e n.
Latessa, Edward, 242n.
Latour, Bruno, 207n.
Lattanzi, Giorgio, 246n, 247n.
Lavagnini, Simona, 52n.
Leahy, Robert R., 455n.
Le Doux, Joseph, 322n.
Leibniz, Gottfried Wilhelm von, 14, 112n, 417, 457.
Lembcke, Oliver W., 452n.
Lemoine, Blake, 319n.
Leonetti, Michele, 93n.
Lerman, Kristina, 63n.
Lettieri, Nicola, 230n, 253n, 452n.
Lewis, Raven A., 253n, 255n.
Li, Guiqin, 35n.
Lichtenstein, Paul, 105n.
Licklider, Joseph Carl Robnett, 71 e n, 72n, 76, 94, 96 e n, 254n.

Linneo, Carlo [latino: *Linnaeus, Carl Nilsson*], 114.
Lipovetsky, Gilles, 348.
Lippmann, Walter, 341n.
Lipton, Zachary Chase, 431n.
Lisi, Francesca Alessandra, 95n, 243n, 254n.
Litwiński, Paweł, 291n.
Liu, Bing, 423n.
Liu, Dianbo, 243n.
Locke, John, 112n.
Loddo, Olimpia Giuliana, 6, 13, 73n, 75n, 92n, 94n, 395, 409n, 410n, 476.
Lodewijks, Henny, 239n.
Lohr, Steve, 40n.
Lombardi Vallauri, Luigi, 84n, 85 e n, 89n, 92n, 466n.
Lombardi, Antonio, 95n, 243n, 254n.
Longo, Alessandro, 34n.
López Estupiñán, Andrés Mauricio, 163n.
López-Koronado, Miguel, 191n.
Lord, Sarah E., 80n, 174n.
Loreggia, Andrea, 37n.
Lorini, Giuseppe, 395n, 409 e n, 410n.
Lorusso, Sergio, 230n.
Losco, Eugenio, 248n.
Lotto, Beau, 108n, 117 e n.
Lovelock, James Ephraim, 458 e n.
Lovins, Brian, 242n.
Luan, David, 415n.
Lucci, Giuliana, 178n.
Luetge, Christoph, 412n.
Luhmann, Niklas, 94n, 396n, 405, 406 e n, 408 e n, 463n.
Lui, Fausta, 339n.
Lullo, Raimondo [latino: *Raimundus Lullus*], 112n.
Lupária Donati, Luca, 236n, 241n.
Luther King, Martin Jr. [King, Michael Jr.], 370, 371.
Lux, Jennifer, 242n.
Luzi, Lorenzo, 43n, 55n.
Luzzati, Claudio, 84n.
Lynch, Gary, 272n.
Lynskey, Orla, 289n.

MacAlaney, John, 437n.
Maceratini, Arianna, 187n.
Machnikowski, Piotr, 299n.
MachInnis, Deborah, 179n.
MacIntyre, Alasdair, 378 e n.
Macioce, Fabio, 22n.
Madelin, Robert, 412n.
Madia, Nicola, 247n.
Malandrino, Delfina, 452n.
Maldonato, Lucia, 236n, 237n, 241n, 243n, 252n, 255n, 256n.
Malizia, Alessio, 81n, 95n.
Mamoshina, Polina, 269n.
Manes, Vittorio, 220n, 232n.
Maniaci, Arturo, 455n.
Mannelli, Chiara, 55n.
Mansueto, Giovanni, 450n.
Manzin, Maurizio, 78n.
Marani, Roberto, 180n.

Marcus, Gary, 38n.
Marcuse, Herbert, 416n.
Marinelli, Vincenzo, 89n.
Marini, Gaetano, 64n.
Marone, Beatrice, 20n.
Márquez Domínguez, Juan Antonio, 143n.
Marra, Piero, 5, 11, 15, 71, 77n, 82n, 83n, 90n, 94n, 476.
Marsch, Lisa A., 80n, 174n.
Martin, Christian, 91n.
Martin, Jim, 239n.
Martino, Antonio Anselmo, 78n.
Martoni, Michele, 93n, 192n.
Marx, Karl, 307, 337n.
Maslen, Hannah, 271n, 273n.
Massaro, Antonella, 232n.
Mattu, Surya, 234n.
Matwyshyn, Andrea, 172n.
Maugeri, Anna Maria, 240n.
May, Hope Elizabeth, 84n.
Mayer, Verena E., 103n.
Mayer-Schönberger, Viktor, 446n.
Mazza, Giuseppe, 232n.
Mazzarese, Tecla, 78n.
Mazzocca, Marco, 258n, 273n.
McAfee, Andrew, 36n.
McCarthy, John, 72n, 130n, 416, 417n, 431n, 458 e n.
McEwan, Ian, 413n.
McFadden, Johnjoe, 322n.
McKinney, Scott M., 48n.
McLeod, Christopher J, 278n.
McMillan-Major, Angelina, 343n.
McNeese, Nathan, 94n.
Mehrabi, Ninareh, 63n.
Mei, Xiaohan, 254n.
Mercuri, Ilenia, 221n.
Merenda, Ilaria, 247n.
Metitieri, Fabio, 444n.
Meynen, Gerben, 266n.
Milton, Sivan, 45n.
Minsky, Marvin Lee, 72n, 416, 417n, 421, 422n, 458 e n.
Minssen, Timo, 278n.
Miotti, Andrea, 313n.
Mistura, Stefano, 472n.
Mitcham, Carl, 373.
Mittelstadt, Brent, 278n.
Mizutani, Haruo, 424n.
Moerel, Lokke, 183n.
Molinari, Enrico, 194n.
Mondini, Giorgio, 446n.
Mongioì, Costanza Rosa, 220n.
Monod, Jacques, 321.
Montagna, Mariangela, 234n, 235n, 241n.
Montagnani, Maria Lillà, 34n.
Montanari, Angelo, 178n, 419n.
Montella, Pietro, 189n.
Moon, Youngme, 343 e n.
Moradi, Bonnie, 344n.
More, Max, 260n.
Morelli, Ugo, 335n.
Morgante, Gaetana, 200n, 201n.
Morley, Jessica, 230n.
Moro, Paolo, 6, 14, 73n, 76n, 81n, 89n, 96n, 178n, 411, 412n, 413n, 414n, 419n, 421n, 431n, 476.

Moroni, Stefano, 409 e n, 410n.
Morotti, Emanuela, 431n.
Morse, Stephen J., 262 e n, 263n, 280n.
Morstatter, Fred, 63n.
Moscovici, Serge, 337n.
Mui, Anita, 202n.
Mukherjee, Sayandev, 45n.
Mulder, Trix, 191n.
Mulet Trobat, Bartomeu, 146n.
Müller, Berndt, 81n.
Muller, Sam, 264n.
Murayama, Kou, 450n.
Muscatiello, Vincenzo Bruno, 6, 12, 95n, 197, 200n, 476.
Mustil, Tom, 325n.

Nachiappan, 269n.
Nadelhoffer, Thomas A., 262n.
Nagera, Humberto, 265n.
Narayanan, Ajit, 433n.
Nass, Clifford, 343 e n.
Natale, Simone, 320n, 436n, 437n.
Nathan, Oz, 269 e n.
Negri, Daniele, 244n.
Nicita, Antonio, 446n.
Nida-Rümelin, Juliane, 196n.
Nielsen, Emily, 81n.
Nitsch, Carlo, 107n.
Nitti, Renato, 200n.
North, John David, 71n.
Norvig, Peter, 37n, 75n, 130n, 428n, 436n.
Notaro, Laura, 232n.
Noto La Diega, Guido, 52n.
Novaco, Raymond, 242n.
Nowotny, Helga, 196n.
Numerico, Teresa, 337n, 340n.
Nussbaum, Martha [Craven], 347, 375 e n, 376, 377n, 378, 392.

O'Neill, Thomas, 94n.
Ocone, Corrado, 114n.
O'Connor, Cliodhna, 260n.
Ojomoko, Lucy, 269n.
Ostermann, Max, 278n.
Ostrovski, Alex, 269n.
Otto, Randy K., 239n.

Pacella, Cecilia, 143n.
Padua, Giorgia, 235n.
Pagallo, Ugo, 36n, 412n.
Pajno, Alessandro, 35n, 51n.
Palazzani, Laura, 76n, 178n, 182n.
Palazzo, Francesco, 246n, 247n, 249n.
Palmieri, Sara, 450n.
Palmirani, Monica, 41n, 192n.
Papa, Michele, 222n, 228.
Pardo, Michael S., 262n, 265n.
Parisi, Domenico, 36n.
Parodi, Cesare, 235n, 246n.
Pascal, Blaise, 417.
Pasceri, Giovanni, 34n.
Pasquale, Frank, 46n, 318n, 429n.
Pasquinelli, Matteo, 314n.

Passananti, Josephine, 41n.
Passerini Glazel, Lorenzo, 82n.
Paternò, Fabio, 36n.
Patrick, Kate, 239n.
Patrikelis, Panayiotis, 178n.
Pattanayak, Pradan, 269n.
Patteri, Ilaria, 339n.
Patterson, Dennis, 262n, 165n.
Patti, Salvatore, 84n.
Paulesu, Pier Paolo, 235n, 237n, 241n.
Pavčnik, Marijan, 102n.
Pavlik, Greg, 27n.
Pązik, Adam, 292n.
Pazos García, Francisco Josè, 143n.
Pecchioli, Giulia, 236n.
Pellicani, Antonio, 77n.
Peluso, Maria Grazia, 42n.
Peña Mesa, Laura, 163n.
Pentland, Alex Sandy, 269n.
Peppered, Robert, 415n.
Perelman, Chaïm, 78n.
Perine, Lori, 342n.
Perlingieri, Pietro, 90n.
Perri, Anna Gina, 180n.
Perrott, David A., 433n.
Perrucci, Antonio, 35n, 51n.
Perucchietti, Enrica, 36n.
Peruginelli, Ginevra, 82n.
Philips, Kathleen, 195n.
Piacentino, Daria, 445n.
Picard, Rosalind Wright, 319n, 422n.
Piergallini, Carlo, 219n.
Pierno, Franco, 458n.
Piñeyro, Sebastián, 240n.
Pinotti, Giulia, 20n.
Pio, Gianvito, 77n.
Pisano, Aldo, 470 e n.
Pisati, Michele, 241n.
Pitruzzella, Giovanni, 447n.
Pizzetti, Franco, 185n.
Platone [latino: *Plato*] di Atene, 44n, 67, 83n, 415 e n.
Poell, Thomas, 448n.
Popma, Arne, 239n, 240n.
Porro, Carlo, 339n.
Portet Tiebas, Jaume J., 143n.
Prandini, Riccardo, 396n.
Pressacco, Luca, 236n.
Preston, John, 427n.
Price, Linda, 179n.
Prisco, Immacolata, 90n.
Proctor, Robert W., 442n.
Provolo, Debora, 413n.
Przybylski, Andrew K., 450n.
Pulito, Lorenzo, 6, 12, 77n, 94n, 95n, 229, 476.
Punzi, Antonio, 196n, 472, 473n, 474n.

Quattrocolo, Serena, 229n, 233n, 235n, 237n, 238n, 239n, 241n, 255n.
Quintarelli, Stefano, 37n.

Rabelais, François, 89n.
Race, Roberto, 239n, 326n.

Radbruch, Gustav, 90 e n.
Radford, Alec, 415n.
Raichle, Marcus E., 261n.
Raimondi, Fabio, 431n.
Raine, Adrian, 265n.
Ramalho, Quintela Ribeiro Neves, 52n.
Ramos, Paulina, 267n.
Ranieri, Enrico, 248n.
Rauccio, Chiara, 192n.
Recanati, Francesco, 89n.
Reddy, Siva, 45n.
Rees, Geraint, 260n.
Regazzoni, Simone, 114n.
Reichberg, Gregory M., 423n.
Reijers, Wessel, 66n.
Reina-Tosina, Javier, 173n.
Ren, Giovanni, 220n.
Resta, Giorgio, 75n.
Ribas, Luisa, 342n.
Ricci, Stefano, 72n.
Riccobono, Francesco, 102n.
Richerson, Peter J., 105n.
Ricœur, Paul, 373.
Rief, Rachael, 253n, 255n.
Righi, Marco, 447n.
Rijcken, Claudia, 174n.
Riondato, Silvio, 413n, 421n.
Riva, Giuseppe, 442n, 447n, 448n.
Rizzo, Maria Livia, 189n.
Rizzolatti, Giacomo, 335n, 339n.
Roach, Matt, 81n, 95n.
Roberts, Robert E.L., 233n.
Robinson, Dave, 126n.
Rochester, Nathaniel, 72n, 417 e n, 458 e n.
Rodotà, Stefano, 190n.
Rogers, Carl R., 437n.
Rollnert Liern, Göran, 363n.
Román-Martínez, Isabel, 173n.
Romano, Bruno, 65n, 67n, 88n.
Romano, Francesco, 52n.
Romanò, Leonardo, 241n.
Romele, Alberto, 66n.
Romeo, Francesco, 5, 11, 75n, 90n, 97, 102n, 103n, 105n, 108n, 111n, 121n, 476.
Rosen, Jeffrey, 277n.
Roskies, Adina, 258n.
Rossetti, Andrea, 72n.
Rossi, Arianna, 192n.
Rossi, Francesca, 412n.
Rovatti, Riccardo, 234n.
Rovelli, Carlo, 322n.
Ruffolo, Ugo, 35n, 37n, 232n, 413n, 421n, 428n.
Ruggieri, Fabrizio, 249n.
Ruiz, Sergio, 267n.
Russell, Bertrand Arthur William, 100.
Russell, Stuart Jonathan, 37n, 313n, 428n, 436n.
Russo, Stefano, 36n.
Sadin, Eric, 313n.
Safran, Jeremy D., 455n.
Sagri, Maria Teresa, 52n.
Saito, Tamaki, 450 e n.
Salerno, Angelo, 249n.

Saligrama, Venkatesh, 41n.
Salvadori, Ivan, 222n.
Sandberg, Anders, 271n.
Sandel, Michael Joseph, 382 e n.
Sanseverino, Giuseppe, 52n.
Santamaría Ramos, Francisco José, 5, 11, 74n, 123, 124n, 131n, 136n, 476.
Santosuosso, Amedeo, 20n.
Sapienza, Salvatore, 37n.
Saraceni, Guido, 6, 14, 22n, 24n, 75n, 435, 452n, 476.
Sarra, Claudio, 73n, 76n, 412n, 421n.
Sartor, Giovanni, 51n, 52n, 182n, 225n, 234n, 236n, 466, 467n.
Savarese, Paolo, 83n.
Savater, Fernando, 373 e n.
Savulescu, Julian, 265n.
Saxena, Nripsuta, 63n.
Scalfati, Adolfo, 86n, 245n.
Scantamburlo, Teresa, 430n.
Schafer, Burkhard, 412n.
Schelble, Beau, 94n.
Schelling, Friedrich Wilhelm Joseph, 112n.
Schmidt-Biggemann, Wilhelm, 418n.
Schmitt, Francis Otto, 177 e n.
Schrödinger, Erwin Rudolf Josef Alexander, 115n.
Schuett, Jonas, 69n.
Schultz, Alan C., 430n.
Schwab, Klaus, 123, 124n.
Schwarz, Reva, 342n.
Schweighofer, Erich, 102n.
Scorza, Guido, 34n.
Scott, Darren, 81n.
Searle, John Rogers, 34n, 82 e n, 118n, 207n, 400 e n, 426, 427 e n, 460 e n.
Sedol, Lee [Se-dol, Lee], 203, 204, 319.
Segal, Zindel V., 455n.
Sejnowski, Terrence Joseph, 323n.
Selina, Howard, 130n.
Sellaroli, Valentina, 235n.
Semerari, Antonio, 455n.
Senabre, Isabel, 240n.
Sententia, Wrye, 260n, 268n, 270n.
Sepúlveda, Magdalena, 281n.
Sequeri, Pierangelo, 182n.
Sergio, Aquilina, 189n.
Seth, Anil, 323n.
Seung, Sebastian, 461n.
Seydavi, Mohammad, 450n.
Sgubbi, Filippo, 222n.
Shan, Shawan, 41n.
Shannon, Claude Elwood, 72n, 417 e n, 458 e n.
Shapiro, Stuart Charles, 79n.
Shieber, Stuart, 436n.
Shneiderman, Ben, 329n.
Sica, Claudio, 455n.
Signorelli, Andrea, 47n.
Simon, Herbert Alexander, 463n.
Simoncini, Andrea, 67n, 233n, 236n.

Singer, Peter Albert David, 382 e n.
Singh, Ilina, 265n.
Singh, Jay P., 242n.
Sinigaglia, Corrado, 335n.
Siniscalchi, Guglielmo, 79n, 96n.
Sinnott-Armstrong, Walter P., 265n.
Sirbikytė, Gintaré, 46n.
Sitaram, Ranganatha, 267n.
Skeem, Jennifer, 242n.
Sloterdijk, Peter, 367 e n.
Smith, Hannah Louise, 278n.
Snarski, Tomasz, 84n.
Soler Martínez, José Antonio, 138n.
Sommaggio, Paolo, 6, 12, 73n, 95n, 257 e n, 273n.
Soriano, Angelo Ryan, 81n.
Soro, Antonello, 192n.
Sorondo, Marcelo Sánchez, 423n.
Sozio, Maurizio, 6, 14, 96n, 457, 473n.
Spada, Marcantonio M., 450n.
Spangher, Giorgio, 248n.
Spitzer, Manfred, 441n.
Sreedharan, Sarath, 336n.
Stern, Günther Siegmund: *vide* Anders, Günther.
Steven, Megan S., 261n.
Stiegler, Bernard, 347, 367 e n, 368, 369, 370, 391, 392.
Stimson, Charles, 242n.
Storm, Marijn, 183n.
Straini, Mauro, 248n.
Strugała, Radosław, 6, 12, 95n, 285.
Summerfield, Christopher, 320n, 324n.
Sunstein, Cass R., 447n, 449.
Sutskever, Ilya, 415n.

Taddei Elmi, Giancarlo, 200n.
Taddeo, Mariarosaria, 93n.
Talamo, Gregorio, 24n.
Tamanaha, Brian Z., 263n.
Tamburrini, Guglielmo, 51n, 418n, 420n.
Tapscott, Alex, 270n.
Tapscott, Don, 270n.
Tateno, Masaru, 450n.
Taylor, Charles, 471n.
Taylor, Kathleen, 280n.
Telesca, Mariangela, 246n.
Teo, Alan R., 450n.
Teofrasto [latino: *Theophrastus*] di Ereso, 113.
Teubner, Gunther, 13, 73n, 207n, 395, 396 e n, 397 e n, 398, 399 e n, 400 e n, 401, 402, 403, 406 e n, 407 e n, 409, 430n.
Thagard, Paul, 80n, 81n, 317n, 320n.
Tiberi, Carlo, 36n.
Tognazzi, Sonia, 232n.
Tomasi, Francesca, 337n, 340n.
Tommaso d'Aquino [latino: *Thomas Aquinas*], santo, 109n, 110n, 112n.
Torre, Valeria, 201n.
Tosi, Emilio, 288n.

Travaglia Ciciriello, Teresa, 244n, 245n, 247n.
Trezza, Remo, 36n, 201n.
Triolo, Paul, 199n.
Trockij, Lev, 311.
Trujillo Perez, Isabel [Ascension], 84n.
Tsamados, Andreas, 93n.
Turchi, Tommaso, 81n, 95n.
Turing, Alan Mathison, 79 e n, 90 e n, 91n, 124 e n, 177n, 207n, 215n, 230n, 323n, 343, 416 e n, 421, 435, 436 e n, 458 e n, 460.
Turkle, Sherry, 441n.

Ubertis, Giulio, 199n, 254n, 255n.
Ueno, Michihiko, 424n.
Ukai, Wataru, 450n.
Ulam, Stanisław Marcin, 461n.
Urbina Carreño, Alfredo, 165n.

Valcke, Peggy, 412n.
Valera, Luca, 267n.
Van Alsenoy, Brendan, 293n.
van Banning, Theo, 281n.
van der Waal, Mark B., 278n.
van der Waal, Raymond X., 278n.
van Dijck, José, 448n.
van Genugten, Willem J.M., 281n.
Varraso, Gianluca, 247n.
Vassalli, Giuliano, 90n.
Vassilev, Apostol, 342n.
Vayena, Effy, 412n.
Veale, Michael, 289n.
Velluzzi, Vito, 84n.
Verma, Sanjeev, 269n.
Veruggio, Gianmarco, 419n.
Vespigani, Alessandro, 35n.
Vestoso, Margherita, 253n.
Vico, Giambattista, 464 e n.
Vincent, Nicole A., 261n.
Vinge, Vernor, 461 e n.
Vita-More, Natasha, 260n.
Volpato, Chiara, 344n.
von Hagen, Alexa, 240n.
Vozza, Lisa, 335n.
Vu, Kim-Phuong L., 442n.

Wachter, Sandra, 278n.
Wagner, Anthony D., 261n.
Wallace, Björn, 105n.
Wallace, Patricia, 445n.
Wallach, Wendell, 420n, 431n.
Washington, Anne, 233n.
Watanabe, Saori, 256n.
Watson, John Broadus, 421 e n.
Weidenfeld, Nathalie, 196n.
Weizenbaum, Joseph, 23n, 437n.
Welzel, Cindy, 278n.
Welzel, Hans, 331n.
Wenger, Michael, 239n.
Werthner Hannes, 196n.
Wiener, Norbert, 416 e n, 417.
Wiggers, Kyle, 48n.
Williams, Michael A., 261n.
Williams, Rhiannon, 48n.
Williamson, Oliver Eaton, 463n.

Wilson, David Sloan, 105n.
Wilson, Edward O., 105n.
Wilson, Ben, 81n, 95n.
Wittes, Benjamin, 277n.
Wittgenstein, Ludwig [Josef Johann], 91 e n, 340 e n.
Wooldridge, Michael, 413n.
Woolf, Virginia, 341n.
Worden Frederic, 177n.
Wormith, Stephen, 242n.
Wu, Jeffrey, 415n.

Xiang, Wie, 173n.

Yamakawa, Hiroshi, 424n.
Yang, Qiang, 413n.
Yang, Yaling, 265 e n.
Yanovich, Yury, 269n.
Yazdani, Masoud, 433n.
Yenduri, Gokul, 23n.
Yenisey, Feridun, 413n.
Yoshimura, Jin, 256n.
Young, Kimberley S., 444 e n, 447n.
Yu, Mo, 45n.

Zaccagnino, Rocco, 452n.
Zaccaria, Giuseppe, 63n.
Zadeh, Lofti Asker, 78n.
Zagrebelsky, Gustavo, 77n, 83n.
Zaiane, Osmar, 45n.
Zanesco, Giacomo, 220n.
Żełaniec, Wojciech, 85n.
Zeno-Zencovich, Vincenzo, 82n.
Zhang, Haotian, 45n.
Zhang, Lin, 45n.
Zhang, Sheldon, 233n.
Zhao, Ben Y., 41n.
Zheng, Haitao, 41n.
Zhok, Andrea, 334n.
Ziccardi, Giovanni, 42n.
Zimmer, Carl, 325n.
Zimmermann, Lars, 297n.
Zingaro, Leonardo, 2.
Ziviani, Bianca, 220n.
Zorzetto, Silvia, 72n., 182n.
Zou, James, 41n.
Zuboff, Shoshana, 196n.
Zullo, Silvia, 414n.
Zyskind, Guy, 269 e n, 278n.

Unità del sapere giuridico

Quaderni di scienze penalistiche e filosofico-giuridiche

Serie I. Editore A. Giuffré, Milano

1. Vincenzo Garofoli (a cura di), *L'unità del sapere giuridico tra diritto penale e processo.* Atti del Convegno (Bari, 21-22 maggio 2004, Facoltà di Giurisprudenza), 2005.
2. Angiola Filipponio (a cura di), *Verità e normatività*, 2005.
3. Vincenzo Garofoli (a cura di), *Unità del sapere giuridico ed eterogeneità dei saperi del giudice*, 2005.
4. Vincenzo Garofoli (a cura di), *Problematiche tradizionali e incaute innovazioni legislative*, 2006.
5. Giuseppe Spagnolo (a cura di), *La responsabilità da reato degli enti collettivi. Cinque anni di applicazione del d.lgs. 8 giugno 2001, n. 231.* Atti del Convegno (Bari, 26-27 maggio 2006-2007, Facoltà di Giurisprudenza), 2007.
6. Vincenzo Garofoli (a cura di), *La sentenza della Corte Costituzionale 6 febbraio 2007 n. 26: un energico richiamo al metodo della giurisdizione.* Atti del Convegno (Trani, 2-3 febbraio 2007, Hotel S. Paolo al Convento), 2007.
7. Miranda Zerlotin, *Diritto originario in Erik Wolf*, (con la traduzione italiana dell'opera di Erik Wolf, *Recht des Nächsten. Ein rechtsthelogischer Entwurf*), 2008.
8. Vincenzo Garofoli (a cura di), *Unità del sapere giuridico e poliformismo normativo*, 2008.
9. Vincenzo Garofoli (a cura di), *L'udienza preliminare e il contenimento dei tempi processuali*, 2008.
10. Vincenzo Garofoli (a cura di), *L'azione penale tra obbligatorietà e discrezionalità.* Atti del Convegno (Bari, 29 novembre 2008, Villa Romanazzi Carducci), 2009.
11. Angiola Filipponio/Aldo Regina (a cura di), *In ricordo di Aldo Moro.* Atti del Convegno (Bari, 20 giugno 2008, Facoltà di Giurisprudenza), 2010.
12. Angiola Filipponio/Vincenzo Garofoli (a cura di), *In ricordo di Franco Cipriani*, 2010.
13. Vincenzo Garofoli/Antonio Incampo (a cura di), *Verità e processo penale*, 2012.

Serie II. Editore Cacucci, Bari

14. Antonio lncampo/Wojciech Żełaniec (eds.), *Universality of Punishment*, 2015.

Serie III. Editore Cacucci, Bari

15. Antonio Incampo/Adolfo Scalfati (a cura di), *Giudizio penale e ragionevole dubbio*, 2017.
16. Lucio Franzese/Antonio Incampo (a cura di), *Potere e libertà al tempo delle emergenze*/Power and Freedom in the Time of Emergencies, 2021.
17. Piero Marra, *Per una moralità procedurale del diritto. Considerazioni attuali a partire da Lon L. Fuller*, 2022.
18. Domenico Siciliano, *Della violenza nel diritto penale fascista. Il caso della legittima difesa a tutela del patrimonio*, 2023.
19. Antonio Incampo/Maria Antonella Pasculli (a cura di), *Partecipazione, rappresentanza e crimini di genere. L'obiettivo 5 dell'Agenda 2030 delle Nazioni Unite,* 2024.
20. Leonardo Di Carlo, *La validità tra identificazione,applicatività e concetto di diritto*, 2024.
21. Maurizio Sozio, *Dall'opinione pubblica all'infosfera. Linguaggio, diritto e pubblicità in J. Habermas*, 2025.
22. Torquato G. Tasso, *Rapsodia del Diritto. Percorsi diacronici di Filosofia e Teoria Generale del Diritto,* 2026.
23. Maria Novella Campagnoli/Piero Marra (a cura di), *Intelligenza artificiale e metamorfosi del diritto*/Artificial Intelligence and Metamorphosis of Law, 2026.