L'IRCOCERVO
PRIMA RIVISTA ELETTRONICA ITALIANA DI METODOLOGIA GIURIDICA - TEORIA GENERALE DEL DIRITTO E DOTTRINA DELLO STATO
FIRST ITALIAN DIGITAL JOURNAL OF LEGAL METHODOLOGY - GENERAL THEORY OF LAW AND DOCTRINE OF THE STATE

Degenerazione e rischio creativo dell'Intelligenza Artificiale "forte": forme di prevenzione e tutela complementare*

Massimo Farina

Università degli Studi di Cagliari

Abstract: **Creative and "Strong" Artificial Intelligence Degeneration and Risk: Forms of Prevention and Complementary Safeguards**
The advent of artificial intelligence (AI) has catalyzed unprecedented debates regarding the philosophical, legal and neuroscientific implications of its use. One of the most crucial aspects is represented by the "degeneration" of AI, due to the complexity of autonomous systems. This paper aims to explore the implications of this emerging challenge through the prism of legal philosophy, considering the ethical, legal and scientific perspectives that intertwine in this context.

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

Sommario: 1. L'emergenza dell'Intelligenza Artificiale e le prospettive di degenerazione – 2. Genesi e fenomenologia della degenerazione tecnica: la vocazione "dato-centrica" – 3. La dimensione informativa – 4. La dimensione strutturale – 5. Degenerazione morale e giuridica: dalle azioni preventive generali a quelle complementari – 6. Considerazioni conclusive.

1. L'emergenza dell'Intelligenza Artificiale e le prospettive di degenerazione

L'odierna società risulta sempre più caratterizzata da continui ed evolutivi mutamenti socio-economici, entro i quali la tecnologia assume un ruolo di assoluto rilievo. L'innovazione della cd. "Quarta rivoluzione industriale"[1], infatti, è fortemente connotata da un uso massiccio delle "macchine"[2].

* 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 approfondire 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

In tale scenario, l'Intelligenza Artificiale (in seguito "IA") si propone come un fenomeno poliedrico[3], che manifesta interessanti frontiere evolutive in innumerevoli settori[4] grazie alla sua capacità di innovare tanto il *modus vivendi*[5], quanto l'*ars pensandi*[6], generando così consequenziali riflessioni di ordine etico-giuridico[7]. A differenza dei precedenti sistemi esperti, basati su *set* predefiniti di regole "*if-then*"[8], le nuove forme di IA tentano di mitigare le pregresse limitazioni operative e di apprendimento, per supportare in modo più efficace le attività

"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. 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*".

decisionali umane[9], attraverso l'individuazione rapida ed ampiamente fruibile delle informazioni rilevanti[10].

Per l'effetto, il dibattito sull'impiego dell'IA si è incrementato nel corso dell'ultimo ventennio[11], in relazione alle potenzialità operative dello strumento tecnologico rispetto all'attività umana[12], nonché al rapporto tra pensare "umano" e tecnica[13] frammentandosi tra "apocalittici" ed "integrati"[14].

Per un verso, l'ascesa dell'intelligenza artificiale ha aperto la strada a una nuova era di progresso tecnologico, in cui sistemi intelligenti possono analizzare dati complessi, automatizzare processi e assistere gli esseri umani in una vasta gamma di compiti[15].

Dal punto di vista funzionale, l'IA si propone come 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 e presentato "quesiti che l'uomo non si era mai posto"[16]. Parallelamente, il confronto tra l'intelligenza umana e artificiale si è intensificato, fino a voler comprendere come quest'ultima possa

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

[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.

[15] S. 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.

essere "integrata" nel tessuto legale e sociale, e come possa contribuire al bene comune senza compromettere i diritti individuali[17]. Così, ecco che le promesse di innovazione ed efficienza si affiancano a nuove sfide e rischi[18], tra cui spiccano quelli connessi all'impiego dell'intelligenza artificiale "generativa", denso di criticità[19] se esteso fino al suo innegabile risvolto "degenerativo". Il fenomeno è complesso, multiforme e ancora in fase di esplorazione, ma sintetizzabile nella tendenza dei modelli di IA a manifestare comportamenti indesiderati, dal punto di vista tecnico e/o giuridico, che ne compromettono progressivamente l'efficienza[20].

Dal punto di vista 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*. Come per la degenerazione delle capacità cognitive umane anche quelle di un sistema IA tendono a deteriorarsi nel tempo a causa di fenomeni progressivi. L'"anomalia", dovuta ad autoreplicazione di errori o di malfunzionamenti, si ripercuote sul risultato, comportando rischi di decisioni non conformi agli obiettivi prefissati, o addirittura 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 l'esigenza di interrogarsi su quale "cura" praticare all'IA per prevenire fenomeni degenerativi e per garantire che i sistemi intelligenti operino in modo lecito, etico e responsabile, rispettando i principi di giustizia e i diritti umani. Per affrontare questa sfida, è fondamentale innanzitutto identificare le potenziali cause di degenerazione dell'IA. Questo permetterà di comprendere le principali ricadute etico-giuridiche del suo utilizzo e, conseguentemente, di sviluppare le necessarie misure preventive.

[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.

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

2. Genesi e fenomenologia della degenerazione tecnica: la vocazione "dato-centrica"

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. Il fenomeno della degenerazione dell'IA è tra quelli che oggi preoccupa maggiormente, se non altro perché collegato ad ogni forma patologica, tecnica e giuridica, che la investe. La degenerazione dell'IA richiama l'idea di un declino progressivo, simile a quello osservato nelle malattie neurodegenerative, che colpiscono gli esseri umani (e più in generale anche altri esseri viventi), e costituisce una conseguenza "naturale" – per certi versi, fisiologica – e diretta della struttura del sistema dell'IA. Diverso è il discorso per la degenerazione giuridica, considerata l'attitudine "patologica" tipica della dialettica giuridica, che, coinvolgendo un assetto valoriale ben preciso, ne implica il necessario rispetto.

Procedendo, dunque, lungo la prima direttrice, si osserva che i sistemi di IA si connotano per una latente e sempre maggiore vocazione "dato-centrica"[21]: in realtà, l'esigenza di apprendere una massa critica di dati si propone come indefettibile premessa per la funzionalità della tecnologia stessa, ingenerando così una vera e propria "dipendenza dal" dato, anziché una mera "vocazione al" dato.

Nella pratica, i sistemi di intelligenza artificiale sono fortemente dipendenti sia dai dati usati nella fase iniziale – di addestramento del sistema (cd. *training*) –, sia da quelli che il sistema stesso raccoglie, in maniera autonoma, nella fase di interazione con l'esterno. Ecco, dunque, che le radici della degenerazione artificiale si collegano indiscutibilmente ad un problema di carattere epistemologico[22], che trova nei dati il suo punto di massima esplicitazione[23].

[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, [Data di consultazione: 03/05/2024]. 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-ism*", 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, [Data di consultazione: 03/05/2024].

[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:

Dal punto di vista teleologico, parimenti, assicurare la qualità e la sicurezza dei dati si manifesta come momento cruciale per il funzionamento dell'intelligenza artificiale: ogniqualvolta essi siano difettosi o mancanti, o vi sia un errore nella comunicazione degli stessi, anche il funzionamento dell'intelligenza artificiale verrà di fatto compromesso.

3. La dimensione informativa

Considerando il trinomio dato-informazione-conoscenza[24], la degenerazione assume una dimensione informativa[25] e – insistendo sul presupposto e parallelamente sul processo, di elaborazione e diffusione dell'informazione – tende ad assumere il carattere della "disinformazione"[26].

In questo senso, una delle principali cause di degenerazione dell'IA risiede nell'utilizzo di dati di addestramento contaminati o non rappresentativi, i quali possono introdurre *bias* e distorsioni nei modelli risultanti[27]. Sono, peraltro, sempre più frequenti i casi di addestramento malevolo volontario[28] – veri e propri attacchi

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 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.

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

Ma il fenomeno disinformativo presenta anche un'ulteriore declinazione, la cui causa risiede nella necessità di un continuo aumento del volume di dati disponibili per il processo di addestramento, allo scopo di ottenere modelli sempre più performanti (sempreché i dati immessi siano di qualità). Tuttavia, questa crescente ricerca di dati, che tende all'infinito, presenta un dilemma fondamentale: giungerà il 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 è di gran lunga superiore alla produzione di informazioni accessibili sul *web*, ciò comporterà l'inevitabile rallentamento del ritmo di addestramento dei modelli e la conseguente limitazione delle potenzialità dell'intelligenza artificiale 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 e ad adattarsi per rispondere alle crescenti esigenze delle applicazioni reali.

Uno dei tentativi più frequenti di soluzione è quello di ricorrere all'utilizzo dei cosiddetti dati sintetici[30], ossia ai dati generati dal modello IA sulla scorta di quelli originali. Il loro utilizzo amplia certamente il *dataset*, ma trattandosi di informazioni non reali potrebbero presentarsi delle disfunzioni informative. Un recente studio[31], effettuato da un gruppo di ricercatori della Rice University in collaborazione con altri ricercatori 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 ciclo autofagico o semplicemente di autoconsumo. Il gruppo di ricercatori, evocando l'analogia con la malattia della mucca pazza, hanno osservato che dopo un certo numero di cicli di autoconsumo, diversi da modello a modello, il sistema impazzisce e comincia a restituire

Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models", in *arXiv*, 2024 (arXiv:2310.13828).

[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.

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

output errati, talvolta discriminatori, finendo dunque per svilire il risultato pratico perseguito.

Tale criticità si acuisce, poi, 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 attributi indispensabili per l'indagine medica, rendendo i dati così "creati in vitro" non più idonei (né utili) allo scopo per cui sono stati inizialmente raccolti.

Quanto affermato non persegue affatto l'obbiettivo di mettere al bando i dati sintetici. In alcuni casi, addirittura, la loro produzione può essere 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, essendo fittizi (ma comunque sia realistici), non possono essere classificati come personali (perché non identificano o non rendono identificabile, direttamente o indirettamente, una persona fisica) ed è per questa ragione che il rischio di illecito trattamento è minimo anche negli usi secondari che vengono fatti dei dati raccolti. Nondimeno, 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 intelligenza artificiale 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].

[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 (2023), "Less is More? Reducing Biases and Overfitting in Machine Learning Return Predictions", in *SSRN*. Recuperato da www.ssrn.com, [Data di consultazione: 03/05/2024].

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

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].

I due aspetti patologici descritti sono 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 vettore silente di disinformazione, con evidenti implicazioni etiche.

Le "allucinazioni" (*hallucination*), talvolta simili ai deliri paranoici umani[37], si manifestano in diverse forme. Si pensi, ad esempio, ai sistemi di riconoscimento di immagini, largamente adoperati nel settore sanitario, laddove potrebbe verificarsi l'identificazione di oggetti che non sono presenti nella scena o la generazione di oggetti incoerenti con le caratteristiche reali.

4. La dimensione strutturale

Il fenomeno degenerativo, fin qui considerato dal punto di vista prevalentemente informativo, può ricondursi anche ad aspetti strutturali dell'intero processo di estrazione ed elaborazione dell'informazione. Si pensi ai cosiddetti scenari *black box*, in cui si inseriscono i dati per ottenere un risultato specifico senza alcuna

[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 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, [Data di consultazione: 03/05/2024].

[37] B.A. Huberman, S. Mukherjee (2023), "Hallucinations and Emergence in Large Language Models", in *SSRN*, September 18, 2023. Recuperato da www.ssrn.com, [Data di consultazione: 03/05/2024].

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 circostanza genera, tra gli altri, alcuni limiti di natura applicativa[40]. La maggior parte dei timori legati all'applicazione di tali tecnologie sono connesse alla scarsa prevedibilità dei risultati che, in virtù dell'autonomia[41] delle macchine, possono determinare decisioni non pienamente governate dal controllo umano[42]. In

[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, 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, [Data di consultazione: 03/05/2024]; 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

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 determinano la decisione finale (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*. Tale "opacità" degli algoritmi e dei sistemi di intelligenza artificiale inevitabilmente aumenta in misura proporzionale rispetto al numero delle variabili e dei dati oggetto di analisi[43].

Questa breve disamina delle più note forme degenerative tecniche dell'IA facilita la comprensione della già citata degenerazione giuridica, con lo scopo ultimo di individuare possibili rimedi e azioni preventive.

5. Degenerazione morale e giuridica: dalle azioni preventive generali a quelle complementari

La degenerazione dell'IA, intesa come declino progressivo delle prestazioni, rappresenta un'emergente problematica con significative implicazioni anche sul piano etico e giuridico[44].

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*".

[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, [Data di consultazione: 03/05/2024]; 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, [Data di consultazione: 03/05/2024]; 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.

In tal senso, per lo meno in linea generale, può affermarsi che qualunque impiego o sviluppo dell'IA, che causi danni o violi i diritti umani, è da considerare nel novero della degenerazione.

Nel dibattito scientifico, le criticità connesse al fenomeno di discriminazione algoritmica risultano sicuramente ricorrenti e pervasive, stimolando riflessioni sul degrado nella performance dell'IA e delle conseguenze – individuali e collettive – scaturenti da risultanze decisionali errate (il dibattito, infatti, si spinge sino a verificare in che modo tali pregiudizi e discriminazioni possano arrecare pregiudizio verso determinate popolazioni, contravvenendo ai principi di equità e giustizia).

Parimenti, ulteriori questioni insistono sulla violazione (o uso non corretto) dei dati personali, e sulle conseguenze, scaturenti dalla perdita di controllo delle informazioni di carattere personale, per l'individuo e per la collettività (esponendo gli individui a violazioni, sorveglianza abusiva, compressione delle libertà civili).

La complessità del fenomeno, insieme con l'indistricabile – ed evidente – connessione tecnico-etica-giuridica, richiede un approccio olistico, allo scopo di comprendere le connessioni e le interazioni tra i principali fattori che stanno alla base della degenerazione.

Al fine di evitare punti di osservazione "miopi", risulta dunque necessaria una strategia certamente "preventiva", ma integrata – tra cybersicurezza, gestione dei rischi, consapevolezza e governance – e costantemente orientata a verificare quei risvolti etici che, se non valutati attentamente, 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 benessere generale degli esseri umani[46] (antropocentrismo).

Indubbiamente, un'azione preventiva efficace si connette intimamente con interventi regolatori (aventi anch'essi carattere "generale") che, definendo e

[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.

[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".

perimetrando i confini di utilizzo di questa tecnologia, mirano alla tutela dei diritti e alla sicurezza dei cittadini.

Tuttavia, la stesura di testi normativi – nazionali e sovranazionali[47] – dedicati all'IA non risulta 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.

Orbene, le 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.

In tal senso, occorrerebbe anzitutto procedere all'implementazione di *policy* adeguate alla manutenzione regolare e l'aggiornamento dei sistemi IA: in questo senso, 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 principi di dignità e rispetto dei diritti umani[49].

La recente diffusione delle IA generative, peraltro, nello stimolare molteplici riflessioni di carattere etico-giuridico, al contempo offre l'occasione per richiamare, sul piano metodologico, superate 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 e il pubblico fruitore della stessa.

[47] Il riferimento va alla recente Proposta di Regolamento europeo sull'IA (cd. "*AI Act*"), *Proposta di Regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione*, COM/2021/206 final, consultabile nel sito ufficiale dell'Unione Europea https://eur-lex.europa.eu. A livello nazionale, invece, si segnalano le recenti *Disposizioni e delega al Governo in materia di intelligenza artificiale (disegno di legge)*, approvato nel corso del Consiglio dei Ministri n. 78 del 23 aprile 2024 (consultabile all'indirizzo www.governo.it).

[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.

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[50]. 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[51].

Rispetto al dibattito ingeneratosi sulla tutela del *software*, i 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 costituzione del *dataset* per nutrire e istruire l'IA, imponenti l'utilizzo lecito delle *fonti* di addestramento.

In passato, invero, 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[52].

All'indomani 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

[50] 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.
[51] 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.
[52] Per una panoramica sull'argomento, si veda, S. Dell'Arte, *Fondamenti di diritto d'autore nell'era digitale*, ed. 2, Key Editore, Milano, 2023.

quale le opere dell'ingegno digitali non esistevano) e si invocano riforme adeguate all'attuale contesto[53].

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 ripianare le criticità connesse alla composizione e 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.

Se è pur vero 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 altresì vero che si rischia, come tuttora accade in Rete, che la tutela venga confinata in una dimensione totalmente teorica. Il rischio è, dunque, che tali previsioni risultino, nella sostanza, incapaci di fronteggiare le concrete esigenze di tutela autorale.

Si rammenta, invero, 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 snodi una direttrice "complementare" che coglie la dimensione negoziale del rapporto intersoggettivo.

Tale combinazione potrebbe dotare quella tutela dell'efficienza pratica di cui necessita e, considerato che i rimedi complementari[54], 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

[53] 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).

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

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 lecitamente soltanto di quelle licenziate e, viceversa, scarterebbe quelle prive di autorizzazione[55].

Siffatta soluzione potrebbe dar vita ad una nuova era delle IA generative di "sana alimentazione" e far tramontare quelle "malnutrite".

Inoltre, nondimeno, 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: parallelamente a carattere generale e complementare.

6. Considerazioni conclusive

Le riflessioni finali sul tema della degenerazione dell'Intelligenza Artificiale (IA) ci portano a riconoscere l'impatto profondo che le nuove tecnologie hanno sul tessuto delle relazioni socio-economiche globali.

Considerando le connessioni inscindibili tra aspetti tecnici e giuridici, si evidenzia l'utilità di un approccio interdisciplinare e multilivello per affrontare con efficacia le sfide poste dalla degenerazione dell'IA. Un dialogo costruttivo tra filosofia del diritto, neuroscienze cognitive e scienze dell'informazione potrebbe stabilire le fondamenta per una *governance* responsabile dell'IA. Tale *governance* dovrebbe garantire non solo l'efficacia e l'efficienza dei sistemi autonomi, ma

[55] 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*".

anche il rispetto dei valori etici e dei principi giuridici che sono alla base delle nostre società.

L'IA, in particolare, si presenta come una sfida inedita che interpella sia il diritto che la tecnologia, sollevando questioni etico-giuridiche che non possono essere risolte esclusivamente attraverso strumenti di tutela a carattere legale.

Per un verso, infatti, la degenerazione dell'IA, come emerso dai paragrafi precedenti, si manifesta spesso come una conseguenza quasi naturale dell'uso degli strumenti tecnologici, legata alla loro intrinseca ricerca del risultato. Questo fenomeno, osservato sotto la lente dell'informazione, rivela una tendenza all'artificialità che richiede un monitoraggio umano costante per assicurare l'integrità dei risultati prodotti. La necessità di tale controllo umano diventa quindi una componente strutturale, essenziale per mitigare la fallibilità potenziale dei sistemi di IA.

Inoltre, problemi come le disuguaglianze, la mancanza di informazioni e la scarsa trasparenza sono aspetti critici che, ad un'analisi più attenta, rientrano nella fenomenologia della degenerazione dell'IA. Anche il rischio di nuove forme di discriminazione è estremamente pertinente, gettando le basi per quello che viene definito il "divario di potenziamento umano".

Per altro verso, invece, la tematica ricade altresì nel solco dell'autodeterminazione negoziale, e, pertanto, ingenera consequenziali riflessioni sul rapporto tra poteri pubblici e poteri privati, segnando i limiti esterni – e, per certi versi, fisiologici – della tutela legale.

Lo strumento negoziale, invero, è l'unico strumento attualmente legittimato a modulare a monte i rapporti tra le parti e, a valle, la circolazione e veicolazione dei flussi informativi.

Peraltro, ammantare di pari importanza e dignità il segmento negoziale consente di contribuire in misura più concreta allo sviluppo "antropocentrico" dell'IA.

La consapevolezza delle potenzialità della tutela negoziale, nondimeno, ben potrebbe contribuire alla realizzazione degli obiettivi postulati a livello generale, tesi a rafforzare la fiducia nell'IA, che dovrebbe essere vista non come un fine in sé, ma come uno strumento al servizio dell'umanità, progettato per migliorare il benessere umano.

In conclusione, si ritiene che una simile prospettiva ed approccio etico e centrato sull'autodeterminazione negoziale dell'individuo sia fondamentale per assicurare che l'evoluzione dell'IA proceda, senza degenerazione, in armonia con i bisogni e i diritti individuali.