theFuture ofScience andEthics

Rivista scientifica a cura del Comitato Etico della Fondazione Umberto Veronesi

Volume 9 ■ 2024 ■ ISSN 2421-3039

Fondazione VERONESI


theFuture
ofScience
andEthics


Fondazione
VERONESI

theFuture ofScience andEthics

Rivista scientifica
del Comitato Etico
della Fondazione Umberto Veronesi
ISSN 2421-3039
ethics.journal@fondazioneveronesi.it
Via Solferino, 19
20121, Milano

Volume 9 ■ 2024

SOMMARIO

ARTICOLI

ARTICOLI / PROSPETTIVE

DOCUMENTI DI ETICA E BIOETICA

RECENSIONI

theFuture ofScience andEthics

Articoli

Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

Realtà e risvolti bioetici, biogiuridici e sociali dell'intelligenza artificiale

Reality and bioethical, biolaw, and social implications of artificial intelligence

ALESSANDRO VOLPE[1]
volpe.Alessandro1@hsr.it
MARCO ANNONI[2]
marco.annoni@cnr.it

AFFILIAZIONE
1. Università Vita-Salute San Raffaele; European Centre for Social Ethics
2. Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca (CID-Ethics), Consiglio Nazionale delle Ricerche; Fondazione Umberto Veronesi

L'accelerazione delle tecnologie ad intelligenza artificiale (IA), segnata in particolare dalla recente diffusione su larghissima scala dell'intelligenza artificiale generativa, è andata di pari passo con la voluminosa crescita dei dibattiti sul tema IA, di carattere scientifico, etico, politico e giuridico. La percezione è quella di un discorso già per certi aspetti saturo – già solo per il numero di pubblicazioni scientifiche dedicate[1] – ma che non può certo rallentare o del tutto arrestarsi, considerate le continue novità che provengono dalla ricerca applicata e di conseguenza dal mercato tecnologico.

La riflessione a tutto campo sull'intelligenza artificiale è forse e sempre più destinata ad assumere i contorni di una vera e propria 'ontologia dell'attualità', un modo specifico di parlare di noi esseri umani al presente, pensando già al nostro immediato futuro. Ciò rende sempre più l'IA non semplicemente un oggetto di dibattito tra gli altri, bensì uno sfondo complessivo a partire dal quale la maggior parte delle questioni contemporanee vanno affrontate.

Il nuovo numero di *The Future of Science and Ethics* (vol. 9/2024), dedicato alle prospettive bioetiche, biogiuridiche e sociali dell'intelligenza artificiale, intende precisamente provare a riflettere sull'intelligenza artificiale non già come evento o serie di eventi eccezionali, bensì come realtà ormai consolidata. Molti dei contributi ospitati presentano l'ecosistema algoritmico ormai come un vero e proprio *habitus*, per nulla neutro e oggettivo, incubatore di contraddizioni, bias, deficit di riflessività, non dimenticando però gli aspetti estremamente promettenti che l'IA restituisce nell'ambito dello sviluppo delle scienze e delle tecniche. In questa realtà, si tratta di ripensare il ruolo di esperti di IA come 'vedette etiche'[2], capaci di osservare le trasformazioni in atto, considerandone i molteplici risvolti e prevedendone anche direzioni ed esiti possibili.

Il numero, infatti, indaga a partire da diverse prospettive e in tono fortemente multidisciplinare i numerosi aspetti di influenza e problematicità dell'intelligenza artificiale: dalle questioni di giustizia epistemica alla formazione del soggetto, dagli interrogativi sugli usi in campo bellico a quello giuridico e scientifico. Simona Tiribelli discute criticamente dei risvolti ormai evidenti dell'opacità epistemica degli algoritmi, e di come quest'ultima stia assumendo gradualmente la forma di un'ingiustizia sistemica; Guglielmo Tamburrini si misura con un tema particolarmente controverso e spesso sottostimato se non poco conosciuto, vale a dire l'utilizzo dell'IA nell'ambito militare tramite la progettazione e il dispiego di armi autonome; l'articolo di Giuseppe De Ruvo, a partire dalla questione dell'accelerazione sociale, propone un modello di *critical digital literacy*, quanto mai necessario per non soccombere alla rimozione di complessità e riflessività indotta dagli algoritmi; Enrico Radaelli affronta nel suo contributo un aspetto cruciale nel rapporto uomo-macchina, ovvero la ridefinizione della soggettività e di facoltà tipicamente considerate 'umane' di fronte all'IA generativa; Tommaso Ropelato affronta le prospettive ma soprattutto i limiti dell'impiego dell'IA in ambito sanitario, in particolare nel counseling genetico; Ilaria De Gasperis affronta il problema giuridico della brevettabilità delle invenzioni connesse all'IA; in conclusione della sezione monografica Ludovica Marinucci si confronta con la questione di com definire e prevenire le condotte di ricerca discutibili e irresponsabili nello sviluppo di sistemi di IA.

Il tema della *call for papers* prosegue poi idealmente anche nella successiva sezione dedicata ai documenti di etica e bioetica. Il volume ospita il documento del Comitato Etico di Fondazione Umberto Veronesi dedicato al ruolo dell'intelligenza artificiale nella ricerca, dal titolo "Intelligenze Future. La ricerca scientifica nell'era dell'intelligenza artificiale". Questo *position paper* auspica una maggiore integrazione tra intelligenza umana e artificiale per fini di ricerca scientifica al fine di inaugurare una nuova era di scoperte e di progresso i cui benefici devono però essere equamente condivisi a vantaggio di tutti.

A seguire, viene pubblicata anche la nuova versione, rivista e aggiornata, della "Dichiarazione in materia di integrità nella ricerca" della Fondazione Veronesi, la quale include ora una sezione dedicata proprio all'uso dell'IA per fini di ricerca. Questo documento si sofferma in maniera innovativa su una possibile e auspicabile deontologia della ricerca scientifica alla luce dell'utilizzo massiccio dell'IA nel mondo della scienza. Come atto di applicazione concreta di queste linee guida, la Fondazione chiede alle ricercatrici e ricercatori che svolgono attività di ricerca finanziate dalla Fondazione stessa o condotte sotto la sua egida, di condividerne e rispettarne i contenuti.

Inoltre, con la diffusione di questi due documenti, la nostra rivista intende avviare una riflessione condivisa sul futuro della ricerca scientifica, ripensandone presupposti, modalità e finalità, in considerazione dei cambiamenti in atto e degli strumenti ormai disponibili.

Accanto al tema principale della *call for papers*, il volume ospita nella sezione "Prospettive" anche un articolo di Giulia Perri dedicato al rapporto tra genomica e discriminazione, nonché il *Manifesto per un'etica procedurale* redatto dall'Istituto di Bioetica per la Veterinaria e l'Agroalimentare (CBV-A), accompagnato dai commenti di approfondimento a firma di Laura Palazzani e Vito Tenore.

Infine, come di consueto, il volume ospita una ricca sezione dedicata alla recensione di libri di recente pubblicazione attinenti all'intelligenza artificiale e ai nuovi sviluppi della bioetica e dell'etica normativa – volumi di giovani ricercatori nonché di già affermati studiosi.

Ci auguriamo che la lettura di questo nuovo numero possa suscitare attenzione e sollecitare un ampio dibattito pubblico, che non si limiti a quello degli addetti ai lavori, come da propositi originari e ancora attuali di *The Future of Science and Ethics*.

NOTE

1. Center for Security and Emerging Technology (2024) – processed by Our World in Data. "Annual scholarly publications on artificial intelligence". Center for Security and Emerging Technology, "Country Activity Tracker: Artificial Intelligence". Disponibile al collegamento: https://ourworldindata.org/grapher/annual-scholarly-publications-on-artificial-intelligence

2. Guglielmo Tamburrini, *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Carocci, Roma 2020, p. 77.

Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

Ingiustizia epistemica, intelligenze artificiali e sfide morali

Epistemic Injustice, Artificial Intelligence, and Moral Threats

SIMONA TIRIBELLI
simona.tiribelli@unimc.it

AFFILIAZIONE
Università degli Studi di Macerata, Macerata, Italia
Institute for Technology & Global Health, Boston, US

SOMMARIO

Nelle società dell'informazione odierne, i sistemi di intelligenza artificiale (IA) appaiono detenere un vantaggio epistemico sugli individui, poiché capaci di processare, dare senso e inferire informazioni preziose dalla realtà datificata in cui viviamo, noi inclusi. Tale vantaggio ha funto da sprone al loro uso pervasivo al punto che oggi co-partecipano e rimodellano gran parte dei nostri compiti epistemici e delle nostre decisioni. Se tale ruolo e impatto sono riconosciuti, se tali sistemi siano vantaggiosi o dannosi per gli individui dal punto di vista epistemico è materia di indagine. Questo articolo si propone di chiarire i benefici e, in modo particolare, i danni producibili dalle tecnologie algoritmiche sugli individui quali soggetti epistemici, traendo spunti dalle teorie sull'ingiustizia epistemica elaborate nell'ambito della filosofia morale e dall'epistemologia sociale. In tal modo, l'articolo mira anche a chiarire le sfide morali intrinseche a tali danni e, infine, sottolineare il ruolo della filosofia morale nell'etica dell'IA per affrontarli.

PAROLE CHIAVE

Ingiustizia epistemica

Intelligenza artificiale

Etica dell'intelligenza artificiale

Agency epistemica

Danni epistemici

ABSTRACT

Artificial intelligence (AI) systems sound to hold an epistemic advantage on individuals in nowadays mature information societies, insofar as they are capable of navigating, making sense, and inferring valuable knowledge from the datafied reality we live in, ourselves included. Such advantage spurred their pervasive use so that today AI systems co-participate and reshape the majority of epistemic tasks and decisions we form and take. While such a role and impact are recognized, whether they benefit or harm individuals epistemically ought to be investigated. This paper aims to elucidate the epistemic challenges posed by algorithmic technology to individuals in their epistemic standing, by drawing on theories on epistemic injustice developed in moral philosophy and social epistemology. By doing so, the paper aims also to clarify the moral threats entrenched with the AI-based epistemic harms outlined and to finally point out the role of moral philosophy for AI ethics to tackle them.

KEYWORDS

Epistemic Injustice

Artificial Intelligence

AI Ethics

Epistemic Agency

Epistemic Harms

DOI: 10.53267/20240101

1. INTRODUZIONE

Sistemi di intelligenza artificiale (IA) come algoritmi di *machine learning* e *deep learning* governano oggi in modo pervasivo gran parte delle tecnologie digitali con cui facciamo esperienza e conosciamo il mondo, diamo senso alle nostre esperienze, apprendiamo e ci esprimiamo, comunichiamo con gli altri e persino con noi stessi. In altre parole, le tecnologie basate su tali sistemi co-partecipano in vari ruoli a innumerevoli nostre attività di natura *epistemica*. Secondo alcuni, questi sistemi costituiscono un valido *supporto epistemico*, date le nostre comprovate limitate capacità cognitive, soprattutto negli ambienti delle società mature dell'informazione, caratterizzati da sovraccarico informativo e scarse risorse temporali[1]. Altri, invece, li descrivono come strumenti di *oppressione epistemica* e ingiustizia sociale, essendone stata dimostrata la tendenza a incorporare pregiudizi iniqui storicamente radicati (quali i *bias* etnici e di genere), difficili da scovare a causa della relativa *opacità intrinseca*, e così di perpetuare *modelli epistemici* di *discriminazione sistemica*, influenzando in modo strutturale il modo in cui conosciamo, interpretiamo esperienze, formiamo credenze, comunichiamo e agiamo, perlopiù a discapito di alcuni gruppi minoritari[2].

Tuttavia, se il rapporto tra sistemi di IA e ingiustizia sociale è oggi al centro di un'ampia letteratura nell'ambito dell'etica delle IA, soprattutto negli studi sull'equità algoritmica, la relazione tra IA e ingiustizia epistemica è trattata in modo parziale e frammentario.

Questo articolo intende mostrare come le lenti teoriche elaborate in materia di ingiustizia epistemica ci consentano di elaborare un quadro sistematico e chiaro di quali siano le sfide di matrice epistemica che i sistemi di IA possono inaugurare, perpetuare ed esacerbare, nonché il portato morale a queste intrinseco. A tal fine, la prima sezione mostra la rilevanza del tema a partire dal dibattito in materia; la seconda sezione introduce la tassonomia concettuale di analisi delle sfide e dei possibili danni epistemici introdotti o riprodotti dalle IA, e li discute, traendo spunti dagli studi più dirimenti sull'ingiustizia epistemica elaborati in filosofia morale e nell'epistemologia sociale; la terza sezione mostra la rilevanza morale di tali sfide e conclude sottolineando il ruolo chiave della filosofia ed epistemologia morale nell'etica dell'IA al fine di affrontarle.

2. INGIUSTIZIA EPISTEMICA NELL'ETICA DELL'IA

Il tema dell'ingiustizia è uno tra più discussi negli studi sull'etica dell'IA. Quivi il dibattito si è concentrato principalmente su questioni di ingiustizia sociale; questioni innescate e perlopiù connesse alla tendenza dei sistemi di IA a incorporare *bias* iniqui nei set di dati utilizzati per svilupparli (bias di *minoranza* inclusi: tendenza all'uso di set di dati mancanti di una rappresentazione adeguata di gruppi vulnerabili e minoranze) e, quindi, a riprodurre ed esacerbare nei relativi risultati (decisioni, punteggi, raccomandazioni) asimmetrie di potere, oppressione sociale e diseguaglianze inique socialmente stratificate e storicamente radicate. Rare sono le indagini filosofiche di matrice etico-applicata sul particolare tipo di ingiustizia nota, invece, come *ingiustizia epistemica* in relazione alle tecnologie in questione[3]. Chiariamo, in primo luogo, perché una tale indagine appare rilevante.

Nel suo celebre lavoro "Epistemic Injustice: Power and the Ethics of Knowing", Miranda Fricker definisce le ingiustizie epistemiche come quei torti o danni che minano gli individui quali *soggetti epistemici (knowers)*, cioè, nelle proprie capacità epistemiche (di conoscere, interpretare, comunicare, nonché essere ascoltati e creduti)[4]. L'impatto dei sistemi algoritmici sugli individui come soggetti epistemici (conoscitivi e decisionali) appare oggi evidente. Si pensi, ad esempio, al ruolo mai neutrale che esercitano quali *gatekeepers* dell'informazione, già centrale nelle piattaforme digitali quali motori di ricerca e *social networking services* e ora ancora più critico considerati i progressi nell'ambito dei *large language models* (e delle IA generative in senso ampio). Inoltre, se, come è noto, le tecnologie algoritmiche sono le sole, oggi, a esibire le capacità necessarie per processare e dare senso agli enormi flussi di dati risultanti dai processi di digitalizzazione e datificazione del reale (tutto è datificabile: incorporabile o traducibile in dati) e, da ciò, produrre informazioni, anche di carattere predittivo, di grande valore nei processi economici e politici odierni, appare altresì evidente il *vantaggio epistemico (epistemic advantage), o posizione epistemica privilegiata*, di tali sistemi e di chi li possiede e ne trae profitto su scala[5]. Tuttavia, se questo vantaggio epistemico detenuto e prodotto dalle IA produca dei benefici o danni sulle capacità epistemiche umane (conoscitive e decisionali), per cui si possa parlare

di ingiustizie, e con quali implicazioni etiche, richiede una riflessione sistematica ben precisa.

Vari studi nell'etica digitale e dell'IA hanno descritto tali sistemi, incluse le tecniche da questi dispiegate (tecniche di personalizzazione, dalla profilazione algoritmica, filtro e classificazione ai sistemi di raccomandazione), come *abilitatori cognitivi ed epistemici*[6]. Ciò appare ragionevole almeno per due ragioni. In primo luogo, consentono di far fronte alle nostre ridotte capacità cognitive di processare e scegliere in contesti caratterizzati da un numero elevato di opzioni (cioè informazioni) e alle limitate risorse umane temporali ed epistemiche (assenza di tempo e informazioni rilevanti) nei contesti quotidiani, presentandoci, filtrando, a battuta di click, ciò che cerchiamo e potrebbe servirci, potenzialmente prevenendo l'utilizzo di scorciatoie cognitive o euristiche errate a cui siamo, invece, avvezzi. A tal riguardo, un numero di studi valorizza come tali sistemi permettano di rilevare noti e/o nuovi *bias* cognitivi, fallacie epistemiche e *pregiudizi iniqui* umani, difficili da identificare e rilevare come *patterns epistemici* ricorrenti, consentendoci una maggiore *vigilanza epistemica*: un'attenzione critica sui nostri processi epistemici interni ed esterni, cioè di acquisizione, processazione e condivisione della conoscenza, ad esempio, nella formazione delle credenze e relativa comunicazione e condivisione nei processi di dialogo, argomentazione e giustificazione reciproca interpersonale. In secondo luogo, tali sistemi consentono di leggere e rappresentare il reale (invisibile) su scala micro e macro e scoprire conoscenza (risoluzione di problemi inclusa) fondamentale al progresso scientifico e sociale in molti ambiti (si pensi all'ambito medico) attraverso *operazioni epistemiche complesse*, dall'analisi dati all'individuazione di modelli, passando per processi di inferenza e scoperta di evidenza, impensabili per le capacità cognitive ed epistemiche umane naturali.

Altri studiosi sostengono che tali sistemi siano più di meri abilitatori cognitivi; sono tecnologie *intrinsecamente epistemiche*: sistemi progettati per funzionare in ambienti *epistemici*, come gli ambienti informazionali, utilizzati su *contenuti epistemici*, come dati e informazioni, e implementati per operare attraverso *operazioni epistemiche* come analisi, inferenze e predizioni, senza necessariamente migliorare le nostre capacità epistemiche (possono generare conoscenza inesatta in modo opaco)[7].

In sintesi, le IA hanno capacità fondamentalmente epistemiche, in ultima istanza, analizzano dati, e generano benefici e danni che sono, quindi, *in primis*, di natura epistemica. È all'analisi di questi danni che dedichiamo la prossima sezione. Utilizzando le lenti teoriche elaborate nell'ambito degli studi sull'ingiustizia epistemica, vagliamo se le tecnologie algoritmiche non solo possano promuovere le nostre capacità epistemiche, come emerge *prima facie*, ma pure danneggiare la nostra agency epistemica, generando danni epistemici ingiusti.

Capire la natura dei danni riprodotti e generati dai sistemi di IA è di fondamentale importanza: ci consente non solo di chiarire perché (giustificazione morale) i sistemi che li producono dovrebbero essere regolati; ci consente soprattutto di elaborare in modo realmente efficace *come* farlo e, nello specifico, attraverso quali strumenti teorici e pratici, per prevenire o mitigare i rischi in questione.

3. IA E DANNI EPISTEMICI

Nei casi di ingiustizia epistemica è sempre coinvolto un certo tipo di danno epistemico, cioè un danno arrecato a qualcuno quale agente epistemico. Fricker, il cui merito è *in primis* quello di aver distinto nel suo lavoro questa particolare ingiustizia, definisce due tipologie di ingiustizia epistemica: *l'ingiustizia testimoniale* e l'*ingiustizia ermeneutica*. L'*ingiustizia testimoniale* si verifica quando a un parlante è dato un inferiore livello di credibilità a causa di un pregiudizio identitario; si pensi ai molti esempi di casi in cui la parola di un testimone di un crimine non è presa sul serio a causa di un *bias* etnico. L'*ingiustizia ermeneutica* si verifica, invece, «quando una lacuna nelle risorse interpretative collettive mette qualcuno in una posizione di svantaggio ingiusto quando si tratta di dare un senso alle proprie esperienze sociali»[8]. Uno degli esempi più noti di ingiustizia ermeneutica è il caso di Carmita Wood: una vittima di molestia sessuale nel 1970, incapace di comprendere e, quindi, comunicare la propria esperienza a causa dell'assenza di un concetto adeguato per interpretare e comunicare l'ingiustizia subita. Fricker distingue inoltre tra *danni epistemici primari* (intrinseci) e *secondari* (estrinseci). Il danno primario dell'ingiustizia epistemica consiste nell'essere danneggiati *qua knower* e quindi simbolicamente *qua personae*, in altre parole, quali agenti epistemici capaci di partecipare ai processi di produzione di conoscenza;

quivi, secondo Fricker, vi è una riduzione ingiusta del *soggetto* da *agente* epistemico *attivo* a *oggetto* epistemico *passivo*, da cui esclusivamente estrarre conoscenza (processo che Fricker chiama di *oggettificazione epistemica*). Non tutti concordano sul fatto che risieda nell'oggettificazione il danno primario dell'ingiustizia epistemica; altri sostengono che sia invece nell'essere visti come soggetto inaffidabile, incapace di contribuire alle pratiche epistemiche in modo *unico*, cioè a partire dalle proprie *particolari* esperienze vissute[9]. I danni secondari dell'ingiustizia epistemica si distinguono in a. danni secondari pratici e b. danni secondari epistemici. I primi si riferiscono ai costi pratici causati dall'ingiustizia in questione. Ad esempio, l'assenza di risarcimento per un danno subito a causa di una sentenza negativa dovuta a un pregiudizio etnico o, in parte, a causa dell'incapacità della persona di interpretare e poi comunicare l'esperienza subita. I secondi si riferiscono alla perdita di fiducia nelle proprie convinzioni e competenze epistemiche, ad esempio, sulla percezione e peso di ciò che hanno vissuto, successiva all'ingiustizia epistemica subita.

Chiarita, in breve, la tassonomia di analisi di cui ci avvaliamo, è possibile elaborare alcune considerazioni. In effetti, è possibile interpretare molti dei rischi generati e riprodotti dai sistemi di IA come sfide o danni di natura epistemica. Ad esempio, i sistemi di IA appaiono perpetuare una condizione di *ingiustizia epistemica ermeneutica by default* a causa dell'*opacità epistemica* a essi intrinseca; in altre parole: la loro complessità e seguente non intellegibilità priva gli individui degli strumenti concettuali necessari per comprendere le proprie esperienze come mediate, informate o rimodellate dagli algoritmi. In effetti, il primo limite epistemico di tali sistemi è nella relativa opacità a causa della complessità del processo probabilistico in atto. Ciò costituisce una delle preoccupazioni maggiori essendo utilizzati in misura crescente per compiti e decisioni prima esclusivamente umani in settori fondamentali quali l'educazione e la sanità. Come è noto, il 'comportamento' algoritmico, cioè il modo attraverso il quale un sistema elabora un certo *output* (il quale può assumere varie forme: punteggio, risultato, decisione, raccomandazione, a seconda del sistema considerato), processando probabilisticamente enormi quantità di dati, non solo è spesso inaccessibile per ragioni quali la proprietà intellettuale, ma – anche nei casi di modelli *open source* – risulta incomprensibile per natura probabilistica ai suoi stessi designer; in altre parole, è una 'scatola nera'[10]. Questa opacità mina *by default* gli individui come agenti *epistemici* (*knowers*), cioè nella relativa possibilità (*epistemic standing*) di conoscere e comprendere il modo, le ragioni e gli aspetti (e, nella terminologia specifica, i *patterns* e le correlazioni) sulla base dei quali sono formulate certe decisioni e le relative conseguenze a cui sono soggetti. In questo senso, se da un lato tali sistemi appaiono *prima facie* agevolare le nostre scelte svolgendo alcune operazioni epistemiche quali la processazione delle informazioni, i processi algoritmici sulla base dei quali questi operano fornendo conoscenza, suggerendoci certe informazioni invece di altre, o plasmando i nostri contesti informazionali, sono opachi. Tale opacità si aggrava poiché non solo riguarda i processi per ragioni di complessità probabilistica (opacità *tecnica*), ma anche i fini che li orientano. In altre parole, nonostante il sistema si comporti in modo opaco, il fine verso cui tale comportamento si dispiega (il sistema apprende probabilisticamente come raggiungerlo) è, invece, definito da alcuni (*provider*), secondo le relative agende, in modo opaco agli utenti. La conoscenza dei fini in questione mitigherebbe l'opacità tecnica in questione, ma è preclusa agli utenti, perlopiù lasciati *out of the loop* nella definizione delle agende economiche dei *provider*, a loro discapito: si tratta, dunque, di un'opacità anche *politica*. In tal senso, l'utente è privato *by design* della conoscenza sia per comprendere che cosa (e perché) sta contribuendo a plasmare le proprie esperienze sociali e credenze personali e per partecipare a tali processi, sia per riconoscere possibili fenomeni di ingiustizia a cui è soggetto – ad esempio, l'esclusione nell'esposizione a certe informazioni (beni epistemici) o a opportunità sociali ed economiche che tali informazioni incorporano (ad esempio, una posizione professionale) a causa di un'errata profilazione basata su di un pregiudizio identitario[11].

Questa duplice opacità produce danni secondari epistemici e pratici rilevanti. Infatti, mina altresì gli individui come agenti *pratici*, nella possibilità di agire a riguardo di tali decisioni, *sociali* e *politici*, nella possibilità di contestarle e sovvertirle, e *morali*, nella possibilità – adeguatamente informata – di avallarne il portato valoriale intrinseco. Ciò, oggi, è di estrema rilevanza: basti pensare agli esempi di usi dei sistemi di IA per

prevedere la possibilità di recidiva di un incriminato in base a cui decidere sulla relativa libertà vigilata o per determinarne l'accesso prioritario a cure mediche. L'impossibilità di conoscere e contestare tali processi può risultare in esclusioni o privazioni pratiche sostanziali e ingiuste (ad esempio, la possibilità di un risarcimento se erroneamente valutati da un algoritmo in un caso giudiziario, si consideri il caso COMPAS, o l'accesso a una posizione professionale). Inoltre, l'incapacità di interpretare in modo adeguato e dimostrare l'ingiustizia subita può provocare danni secondari epistemici problematici, dalla perdita di fiducia rispetto alle proprie convinzioni (es. di aver ricevuto un'inesatta prevalutazione algoritmica) a un generale senso di impotenza in conoscenza e azione. Questi danni rinforzano in modo particolare quei fenomeni di marginalizzazione sociale che affliggono soprattutto le minoranze e coloro già vulnerabili, privandoli ulteriormente degli strumenti concettuali e pratici per riconoscere, interpretare e avvalorare, prima, e poi manifestare e rivendicare, le proprie particolari esperienze di ingiustizia e oppressione vissute. L'opacità duplice analizzata, infatti, crea, acuisce e cela effetti ingiusti di natura ermeneutica che, per l'appunto, tendono a colpire soprattutto gruppi storicamente marginalizzati (minoranze etniche e di genere). Tale opacità maschera una progettazione tecnologica dei sistemi in questione non inclusiva bensì storicamente escludente di gruppi socialmente marginalizzati, basata perlopiù su approcci, agende politiche ed economiche, priorità e preferenze orientate alla massimizzazione del profitto e di stampo esclusivamente occidentale, risultando nello sviluppo di sistemi che generano – come prodotto diretto – l'inevitabile rafforzamento di quei fenomeni di marginalizzazione ed esclusione sociale che ne hanno improntato le pratiche di sviluppo[12].

Tali considerazioni consentono di mettere in luce ulteriori danni epistemici riconducibili a ingiustizie testimoniali. Come ampiamente mostrato da Zuboff nel suo lavoro "The Age of Surveillance Capitalism", gli algoritmi trattano gli individui (e i loro dati) come materia grezza da cui ricavare o estrarre prodotti di previsione comportamentale da vendere al miglior inserzionista[13]. Ciò avviene in modo particolare nei processi di profilazione individuale attraverso cui gli algoritmi inferiscono caratteristiche simili tra individui (*collaborative filtering*) al fine di assegnare loro profili (processo di etichettatura) utili al loro inserimento all'interno di categorie predefinite (a cui corrispondono stime di identità e comportamenti) basate su macro-generalizzazioni. In questi processi, l'individuo non ha quasi mai un ruolo attivo, cioè non è posto nelle condizioni di partecipare in modo proattivo ai meccanismi attraverso cui è 'identificato' algoritmicamente e sulla base di cui (quali correlazioni e *patterns*) risulta esposto ad alcune informazioni e opportunità e soggetto a certe decisioni, rispetto ad altre. In questo senso, utilizzando la terminologia di Fricker, appare verificarsi un vero e proprio processo di *oggettificazione algoritmica* dell'individuo che risulta minato nella sua capacità agenziale epistemica di fornire conoscenza (agente epistemico attivo) e trattato, invece, quale – o ridotto a – *oggetto epistemico passivo* da cui ricavare conoscenza utile ai profitti di parti terze. Si tratta di un danno epistemico primario al centro di un'ingiustizia testimoniale perpetrata algoritmicamente, dove l'individuo viene danneggiato nella sua capacità come agente epistemico di fornire conoscenza su sé stesso e sulla propria identità epistemica e pratica. Infatti, nonostante possibili sforzi dell'individuo al fine di farsi percepire nella propria singolarità e irriducibilità alle categorie algoritmiche citate, qualora informato del *modus operandi* descritto, le tecnologie algoritmiche, per come ora progettate, finirebbero comunque per ignorare o fornire minor peso alla prospettiva del singolo in virtù della relativa rappresentazione stereotipata (*biased*) poiché quantitativamente più ricorrente. Infatti, come emerso nei casi di discriminazioni algoritmiche discussi nel dibattito in materia, i sistemi algoritmici finiscono per stereotipare l'individuo, trattandolo come un mero aggregato di associazioni ricorrenti inferibili dai dati a disposizione, e privandolo, in parallelo, della possibilità di contribuire alle pratiche epistemiche che pure lo coinvolgono (quali la ridefinizione dell'ambiente epistemico a cui è esposto e di elementi epistemici personali quali le credenze) in modo *univoco*, cioè sulla base della persona particolare (informata da relazioni, affiliazioni, valori, affetti, impegni e progetti particolari) che è.

Le tecnologie algoritmiche, allo stato attuale, appaiono oggi dunque capaci anche di danneggiare gli individui in quanto soggetti epistemici nella misura in cui privano loro di partecipare attivamente ai processi di produzione di conoscenza, o alle pratiche epistemiche che li riguardano in senso ampio, in modo duplice.

In primo luogo, gli individui risultano danneggiati *by design* nella loro posizione epistemica in quanto posti in condizioni di mancanza di risorse epistemiche collettive (svantaggio epistemico) per comprendere i processi e gli annessi fenomeni in cui sono coinvolti e le decisioni di matrice algoritmica a cui sono soggetti (*ingiustizia ermeneutica*). In secondo luogo, in quanto sistematicamente declassati o esclusi quali valide fonti di conoscenza nei processi algoritmici in questione a favore di una loro rappresentazione e trattamento da parte di tali sistemi perlopiù basati su correlazioni semplificanti, distorte o errate, cioè stereotipi e pregiudizi identitari ingiusti (*ingiustizia testimoniale*).

4. TECNOLOGIE EPISTEMICHE E SFIDE MORALI: IL RUOLO DELLA FILOSOFIA MORALE NELL'ETICA DELL'IA

L'analisi etica condotta ha mostrato come le tecnologie algoritmiche *qua* tecnologie intrinsecamente epistemiche possano non solo promuovere la nostra agency epistemica ma anche danneggiarla, creando danni epistemici *ingiusti* riconducibili a vere e proprie forme di ingiustizia epistemica testimoniale ed ermeneutica. Questi danni sono moralmente carichi: pongono *sfide morali serie*, già ravvisabili nei costi pratici discussi sopra, che appare qui fondamentale almeno menzionare.

L'opacità (su due livelli: tecnica e politica) nella progettazione di tali sistemi non consente all'individuo di verificare i processi algoritmici a cui è soggetto, richiederne giustificazione, eventualmente contestarli, o chiederne la modifica; in tal senso, riducendo la nostra agency epistemica, minacciano anche la nostra autonomia morale e agency *pratica*, la nostra capacità di avallare criticamente gli *output* algoritmici e il portato epistemico e valoriale che questi riflettono e/o incorporano. A essere sempre più esiguo è, dunque, anche il controllo che possiamo più o meno esercitare su ciò che informa, plasma o persino determina i nostri processi decisionali, le nostre scelte e azioni, e degli elementi che li motivano, quali preferenze, convinzioni, valori, relazioni, che appaiono sempre più plasmati dai sistemi algoritmici; a essere messa in discussione è anche la relativa autenticità, la pluralità epistemica e assiologica e la riflessione critica di cui necessitano per una formazione e adesione personale realmente genuina. A essere sfidata è anche la nostra possibilità di autorialità nei processi di formazione delle nostre identità personali come persone particolari, e la libertà di orientarle o improntarne lo sviluppo sulla base di ciò che per noi *qua* agenti epistemici e morali singolari è significativo e realmente (non algoritmicamente) conta.

L'analisi elaborata evidenzia, quindi, il ruolo fondamentale della filosofia e dell'epistemologia morale all'interno della teoria etica applicata alle IA come capaci di fornire gli strumenti concettuali necessari all'indagine non solo delle ingiustizie sociali riproducibili dai sistemi di IA, su cui vi è già un dibattito fervente, ma anche dei danni epistemici connessi a forme particolari di ingiustizia epistemica concatenati alle caratteristiche dei sistemi in questione come attualmente progettati. Tali strumenti appaiono altresì cruciali per facilitare lo sviluppo di criteri di *ethics by design*, cioè di quei requisiti etici che debbono orientare la progettazione dei sistemi di IA tenendo conto dei rischi messi in luce, per la realizzazione di tecnologie *epistemicamente* (al fine di essere anche socialmente) giuste, cioè capaci di rispettare gli individui *qua personae* a partire dal loro rispetto quali agenti epistemici.

NOTE

1. Julian Savulescu e Hannah Maslen, "Moral Enhancement and Artificial Intelligence: Moral AI?", in J. Romportl, E. Zackova, J. Kelemen, *Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics*, vol. 9, Springer, Cham, 2015.

2. Safiya Umoja Noble, *Algorithms of oppression: How search engines reinforce racism* (New York: New York University Press, 2018).

3. Per un'introduzione al tema soprattutto in relazione a questioni di privacy mentale si veda Fiorella Battaglia, "Algoritmi predittivi e ingiustizia epistemica", in M. Galletti, S. Zipoli Caiani, *Filosofia dell'Intelligenza Artificiale. Sfide etiche e teoriche*, Bologna: Il Mulino, 2024, 63-82.

4. Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford: Oxford University Press, 2007).

5. Simona Tiribelli, *Moral Freedom in the Age of Artificial Intelligence*, Milan-London: Mimesis International, 2022.

6. Paul Humphreys, *Extending ourselves: Computational science, empiricism, and scientific method*, Oxford: Oxford University Press, 2004.

7. Ramon Alvarado, "AI as an Epistemic Technology", *Science and Engineering Ethics* 29, 32 (2023).

8. Fricker 2007, *Epistemic Injustice*, cit., p. 1.

9. Gayle Pohlhaus, "Discerning the Primary Epistemic Harm in Cases of Testimonial Injustice", *Social Epistemology* 28(2), 9, 2014.

10. Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge (MA): Harvard University Press, 2015.

11. Simona Tiribelli, *Identità personale e algoritmi. Una questione di filosofia morale*, Roma: Carocci editore, 2023.

12. Sábëlo Mhlambi, Simona Tiribelli, *"Decolonizing AI Ethics. Relational Autonomy to Counter AI Harms"*, *Topoi*, 2023.

13. Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future and The New Frontier of Power*, New York: Public Affairs, 2019.

Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

L'IA nell'azione bellica: problemi e sfide per il controllo umano della guerra

AI in warfare: problems and challenges for human control of war

GUGLIELMO TAMBURRINI
guglielmo.tamburrini@unina.it

AFFILIAZIONE
Università di Napoli Federico II

SOMMARIO

Conservare il controllo umano sull'azione bellica è una questione di grande rilevanza etica e giuridica. Non ricadono sulle macchine le responsabilità per lo svolgimento di un'azione bellica e per violazioni del Diritto internazionale umanitario. La vigilanza dei controllori umani può inoltre contribuire a evitare errori nei quali la macchina potrebbe incorrere operando su campi di battaglia dinamici e poco prevedibili. Questa esigenza normativa si scontra con i rapidi sviluppi tecnologici nel settore delle armi autonome e con le condizioni materiali di impiego dei sistemi dell'IA per il supporto alle decisioni belliche. Lo scarto crescente tra richiesta normativa e realtà dello sviluppo tecnologico mette impietosamente in luce la lentezza delle attuali iniziative diplomatiche, intergovernative e della società civile sulla regolamentazione dell'IA in ambito bellico, insieme all'urgenza di affrontare con maggiore efficacia le sfide etiche e giuridiche poste dalle applicazioni militari dell'IA.

PAROLE CHIAVE

Armi autonome

Sistemi di supporto alle decisioni belliche

Teoria della guerra giusta

Diritto internazionale umanitario

Intelligenza artificiale.

ABSTRACT

Preserving human control on warfare is a major ethically and legally motivated demand. Machines cannot be held responsible for the outcomes of warfare actions or for violations of moral and legal norms that combatants are required to respect. Human control may additionally help one to avoid errors made by machines in the fog of war. However, these normative needs increasingly jar with both developments of AI technologies enabling autonomous weapons and circumstances of use for AI decision-support systems in warfare scenarios. These tensions shed light on the inadequacy of current actions on the regulation of AI systems in warfare, and urgently call for more effective societal, diplomatic and intergovernmental initiatives.

KEYWORDS

Autonomous weapons systems

Decision-support systems in warfare

Just war theory

International humanitarian law

Artificial intelligence.

DOI: 10.53267/20240102

1. IA, ARMI AUTONOME E SUPPORTO ALLE DECISIONI IN AMBITO BELLICO

Le tecnologie dell'Intelligenza Artificiale (IA) si stanno ampiamente diffondendo nel settore militare. In questo ambito, le applicazioni dell'IA sono utilizzate anche per lo svolgimento di funzioni specifiche delle forze armate in ambito bellico. Vi sono sistemi dell'IA che forniscono un supporto alle decisioni necessarie per pianificare e svolgere le azioni belliche. E i cosiddetti sistemi di arma *autonomi* (che indicheremo qui più brevemente come 'armi autonome') sono in grado di individuare un obiettivo e di attaccarlo senza che sia necessario l'intervento di un operatore dopo la loro attivazione. Lo sviluppo delle armi autonome più avanzate si basa su tecnologie dell'IA per la percezione artificiale, la pianificazione e l'esecuzione delle azioni[1].

Lo sviluppo delle armi autonome si trova all'origine di un problema nuovo nella storia dell'interazione tra esseri umani e macchine in ambito bellico. La diffusione delle armi autonome potrebbe rendere il controllo umano sull'azione bellica tecnicamente superfluo su una scala difficilmente concepibile in precedenza, perché nessun operatore umano è inserito nel processo di elaborazione e decisione che conduce dalla selezione di un possibile obiettivo, alla pianificazione dell'attacco contro di esso e infine alla messa in atto dell'attacco. A quali condizioni sia possibile mantenere nella mano dell'uomo il controllo sulle armi autonome e perché dovrebbe essere mantenuto è una questione ampiamente dibattuta in ambito militare, accademico, diplomatico e intergovernativo. Ma il problema del controllo umano sull'IA impiegata a scopi bellici è balzato più recentemente alla ribalta in un altro contesto di interazione umano-macchina, che non riguarda l'autonomia operativa nell'azione bellica, bensì il supporto che l'IA offre alla scelta delle azioni da compiere sul campo di battaglia. Il problema è emerso in tutta evidenza da un'inchiesta giornalistica condotta delle riviste "+972" e "Local Call" a proposito dei sistemi informatici Habsora e Lavender, che sono stati sviluppati utilizzando anche tecniche di IA. Questi sistemi sono stati utilizzati nel 2023 dalle Forze armate israeliane (IDF) per la pianificazione di bombardamenti aerei da mettere in atto nella Striscia di Gaza[2].

Le liste di obiettivi da bombardare compilate da questi sistemi devono essere vagliate da un operatore umano, che ha il compito di accettare, modificare oppure rifiutare ciascuna delle proposte in lista. La configurazione dell'interazione umano-macchina prevede costitutivamente la subordinazione decisionale della macchina all'essere umano. Al contrario di quanto accade nel caso delle armi autonome, l'operatore umano svolge una funzione di controllo essenziale, e più specificamente di filtro sugli obiettivi selezionati dalla macchina. E tuttavia, come vedremo, le reali condizioni al contorno per lo svolgimento di una tale funzione di filtraggio non bastano a garantire un controllo umano efficace.

Sotto l'ipotesi che alle testimonianze riportate dall'inchiesta giornalistica corrispondano elementi di verità, il caso di Habsora e Lavender illustra drammaticamente come il problema del controllo umano sulle decisioni e sulle azioni messe in atto da macchine 'intelligenti' dell'IA in ambito bellico abbia una portata più ampia di quanto sia stato comunemente assunto o rilevato nel dibattito etico e giuridico. Infatti, l'attenzione alle questioni normative si è finora concentrata sull'autonomia operativa dei sistemi d'arma che possono selezionare e attaccare un obiettivo senza che sia necessario un intervento umano.

Oltre a determinare un'estensione del problema del controllo umano sull'azione bellica, gli sviluppi tecnologici recenti dell'IA contribuiscono ad aggravarne ulteriormente la portata anche all'interno del suo perimetro originario, e cioè in riferimento alle armi autonome. La recente sperimentazione di un caccia da combattimento manovrato in volo da un sistema dell'IA ci pone di fronte alla concreta realizzazione di un velivolo autonomo da combattimento capace di ingaggiare un confronto ravvicinato con altri velivoli da caccia, senza richiedere in nessuna fase del confronto l'intervento da parte di un pilota o di un qualsiasi altro operatore umano[3].

Conservare il controllo umano sull'azione bellica ha grande rilevanza etica e giuridica. Non possono ricadere sulle macchine coinvolte le responsabilità per lo svolgimento di un'azione bellica e per eventuali violazioni delle norme morali e giuridiche che i combattenti devono rispettare.[4] Gli attuali sistemi dell'IA, per quanto intelligenti possano essere ritenuti, non sono agenti morali e non possono dare conto delle azioni nelle quali sono causalmente coinvolti. La vigi-

lanza dei controllori umani può inoltre contribuire a evitare errori che la macchina potrebbe commettere negli ambienti dinamici e poco prevedibili che spesso caratterizzano i teatri bellici. Bisogna però constatare che questa esigenza normativa non mantiene il passo con i rapidi sviluppi tecnologici nel settore delle armi autonome e con le condizioni materiali di impiego dei sistemi dell'IA per il supporto alle decisioni belliche. Lo scarto crescente tra esigenza normativa e realtà tecnologica mette impietosamente in luce la lentezza e l'inadeguatezza delle attuali iniziative diplomatiche e intergovernative sulla regolamentazione dell'IA in ambito bellico, ponendo con urgenza il problema di mettere rapidamente in campo iniziative più tempestive ed efficaci.

Ma procediamo con ordine, a partire da un esame del problema del controllo umano nel caso di sistemi per il supporto all'azione bellica abilitati dalle tecnologie dell'IA.

2. IA E SELEZIONE DI OBIETTIVI MILITARI

Secondo l'inchiesta giornalistica condotta dalle riviste "+972" e "Local Call", l'IDF avrebbe impiegato due sistemi di supporto alle decisioni basati sull'IA – noti con i nomi di Habsora e Lavender – per generare liste di obiettivi potenziali da bombardare nella striscia di Gaza. Questi sistemi sarebbero stati utilizzati sistematicamente nei primi mesi dopo il massacro di civili innocenti perpetrato il 7 ottobre 2023 in territorio israeliano da Hamas e il concomitante sequestro di ostaggi. Poiché si tratta di sistemi di supporto alle decisioni, le scelte operate da tali sistemi sono solo condizionalmente accettate ed eseguite. Spetta infatti a operatori militari competenti il compito di passare in rassegna le proposte avanzate dalla macchina per poi accettarle, rivederle o respingerle.

Si possono addurre rilevanti motivazioni di carattere etico e giuridico a giustificazione di una siffatta subordinazione della macchina al giudizio di un operatore umano. La teoria della guerra giusta ammette che in determinate situazioni il ricorso alle armi sia moralmente giustificato. Nondimeno, essa prescrive che tutte le parti coinvolte in un conflitto bellico debbano ottemperare a determinati vincoli morali nella conduzione delle operazioni belliche. In particolare, bisogna rispettare l'immunità dei non combattenti e bisogna astenersi dall'infliggere danni sproporzionati in relazione agli obiettivi militari da conseguire[5]. Questi elementi della teoria della guerra giusta sono incorporati nel principio di distinzione e nel principio di proporzionalità del Diritto Internazionale Umanitario (DIU)[6]. Il giudizio degli operatori deve servire a controllare, tra l'altro, che l'attacco agli obiettivi designati dalla macchina sia compatibile con questi e altri principi del DIU.

Per l'inchiesta di "+972" e "Local Call", il controllo effettuato dagli operatori umani non sarebbe stato all'altezza di questo gravoso compito. Essi avrebbero invece esercitato un controllo affrettato e in definitiva del tutto nominale, svolto in condizioni tali da rendere la loro attività assimilabile a quella di semplici passacarte, che avallano meccanicamente le indicazioni fornite dalla macchina. Fonti anonime dell'IDF menzionate nell'inchiesta giornalistica hanno affermato che l'introduzione dei sistemi di supporto alla scelta degli obiettivi da bombardare ha comportato un chiaro incremento del numero di possibili obiettivi per unità di tempo: «da 50 obiettivi all'anno» elaborati manualmente in situazioni precedenti si è arrivati «fino a 100 obiettivi al giorno». Questo incremento avrebbe innescato una forte pressione psicologica da parte dei superiori militari, con la richiesta rivolta ai controllori di mantenere il passo con la macchina: «Prepariamo gli obiettivi automaticamente e lavoriamo in base a una lista di controllo (checklist) … È proprio come in una fabbrica. Operiamo con rapidità e non c'è tempo per un approfondimento sull'obiettivo. L'idea è che siamo giudicati in base al numero di obiettivi che riusciamo a generare». Emerge qui chiaramente la pressione psicologica percepita dagli operatori umani di allineare il ritmo delle loro valutazioni all'incremento di 'produttività' impressa dalla macchina e la denuncia della conseguente difficoltà di dare un giudizio ponderato sulle azioni belliche suggerite dal sistema di supporto alle decisioni.

È possibile che la ricerca di un rapido e decisivo vantaggio militare sul campo sia stata la fonte principale della pressione psicologica esperita dalle fonti interne all'IDF citate nell'inchiesta giornalistica. Gli operatori umani sarebbero diventati il principale collo di bottiglia per il conseguimento degli obiettivi militari desiderati. Il tempo richiesto da un controllo scrupoloso sulle liste di obiettivi generati dalla macchina avrebbe indotto una coda sempre più lunga di obiettivi in attesa di convalida. Una lezione fondamentale emerge da questo resoconto,

anche indipendentemente da quale sia realmente la motivazione militare alla base della pressione psicologica esperita dagli operatori e da quali siano le eventuali negligenze dei superiori militari relative al rispetto del DIU. La lezione riguarda il peso da dare al fattore tempo nella progettazione del controllo umano sui sistemi di supporto alle decisioni per l'azione bellica. Una configurazione delle interazioni umano-macchina che preveda la presenza di uno o più operatori come filtro non basta a garantire che questi ultimi riescano a valutare con cura le proposte avanzate dalla macchina. Finestre temporali troppo ristrette per l'esercizio del controllo umano sono una fonte significativa di perturbazioni del processo decisionale affidato agli operatori.

Daniel Kahneman ha esposto in forma semplificata e comunicativamente efficace questo tipo di problema, facendo riferimento a due sistemi cognitivi distinti che generalmente sono all'opera nei processi decisionali[7]. Il Sistema 1 consiste di processi decisionali euristici: veloci, per lo più automatici e connotati emotivamente. Il Sistema 2 consiste di un insieme di processi decisionali più lenti e meno prontamente attivabili, riflessivi e analitici. I due sistemi operano in maniera concorrente, ma non sempre cooperativa. Le opzioni più prontamente indicate dal Sistema 1 vengono spesso accettate per *default*, soprattutto quando bisogna decidere sul da farsi avendo poco tempo a disposizione. Le opzioni di attacco selezionate da Habsora o Lavender sono prontamente disponibili, mentre l'esplorazione di altre opzioni può rivelarsi molto più dispendiosa in termini di tempo e di altre risorse cognitive.

Un procedimento mentale euristico – che induce a preferire opzioni prontamente disponibili ignorando alternative mentalmente più faticose da esplorare, è stato designato da Kahneman e Amos Tversky con l'acronimo WYSIATI (What-You-See-Is-All-There-Is). Questa euristica riflette, per Kahneman, una «grande asimmetria tra i modi in cui la nostra mente tratta le informazioni immediatamente disponibili e quelle che non lo sono». Le persone tendono a focalizzare su ciò che si trova innanzi ai loro occhi, trascurando informazioni rilevanti che potrebbero essere rintracciate o recuperate dalla memoria. WYSIATY accelera il processo decisionale ma porta con sé un pregiudizio (*bias*) che può indurre in errore: «Le informazioni che non sono recuperate (nemmeno inconsciamente) dalla memoria potrebbero anche non esistere»[8]. E così la pressione psicologica a prendere decisioni a ritmo serrato può indurre gli operatori ad accettare acriticamente i suggerimenti resi prontamente disponibili dalla macchina.

È stato stimato un margine di errore del 10% per le liste di obiettivi generate dai sistemi di supporto alle decisioni utilizzate dall'IDF. Le implicazioni etiche e giuridiche di una finestra temporale insufficiente ad esprimere un giudizio ponderato risultano ancora più gravi alla luce di questa stima di errore. Un giudizio ponderato su liste di obiettivi gravate da una notevole percentuale di errore è cruciale per risparmiare la vita di persone fuori combattimento, di civili innocenti, del personale sanitario e di altre persone protette dal principio di distinzione codificato nel DIU. Sui comandanti militari ricade la responsabilità di fare il possibile per garantire che ci siano le condizioni per arrivare a un tale giudizio ponderato. E nello scenario descritto tali condizioni evidentemente non sussistono.

Insieme alla disponibilità di tempo sufficiente, altre condizioni devono essere rispettate per esprimere un giudizio umano ponderato sui suggerimenti forniti da una macchina. Evidentemente è necessario che gli operatori abbiano competenze sufficienti e la capacità di resistere al cosiddetto *bias* da automazione («Chi sono io, con le mie limitate capacità cognitive, per contraddire le scelte operate da una macchina che è stata addestrata con quantità gigantesche di dati, dei quali potrò padroneggiare ed elaborare solo una minima frazione nel corso della mia vita?»)[9]. Gli operatori dovranno inoltre convivere con l'opacità dei processi di elaborazione degli attuali sistemi dell'IA; con una mancanza di trasparenza che ostacola l'interpretazione e la spiegazione delle ragioni che si trovano alla base delle decisioni proposte dalla macchina.

In definitiva, i sistemi di supporto alle decisioni belliche basati sull'IA sono evidentemente dei sistemi socio-tecnici. Bisogna affrontare e coordinare tra loro formidabili sfide tecniche e scientifiche, sociali, organizzative e psicologiche per mitigarne le limitazioni ed evitarne usi eticamente e giuridicamente ingiustificabili. Nelle narrazioni più comuni sull'innovazione tecnologica – in ambito bellico, ma non solo – lo sguardo viene troppo spesso distolto da questi problemi e dalle loro implicazioni per la società.

3. CONTROLLO DELLE ARMI AUTONOME

Nel caso di un'arma autonoma, come è stato sottolineato già all'inizio di questo articolo, non si richiede che le decisioni della macchina siano vagliate da un operatore umano prima di essere messe in pratica. Si pone dunque il problema se sia ancora possibile garantire un controllo umano efficace in queste circostanze e, in caso contrario, se l'uso di un'arma autonoma non soggetta a un controllo umano in corso d'opera sia compatibile con i vincoli etici e giuridici che sovrintendono alla conduzione delle attività belliche.

Una risposta influente a questo problema è stata fornita dal Comitato Internazionale della Croce Rossa (CICR). Nel 2021, il CICR ha elaborato uno schema generale per arrivare a un trattato internazionale vincolante che introduca forti restrizioni sullo sviluppo e l'uso delle armi autonome[10]. Lo schema, sostenuto da varie argomentazioni etiche e giuridiche, si basa su due richieste di divieto e su una richiesta di regolamentazione. La proposta del CICR si inserisce in una famiglia più ampia di proposte di regolamentazione delle armi autonome, che sono dette 'a due livelli' perché comprendono sia divieti sia restrizioni. Le proposte di regolamentazione a due livelli riscuotono attualmente un consenso molto ampio nel dibattito internazionale sulle armi autonome. Un consenso molto più ampio rispetto a proposte più restrittive che prevedono la messa al bando di qualsiasi tipo di arma autonoma. Le richieste di divieto avanzate dal CICR riguardano le armi autonome che siano progettate per attaccare direttamente le persone oppure che abbiano effetti imprevedibili rispetto al DIU. Le armi autonome che non sono vietate su queste basi dovranno essere regolamentate nel loro impiego in base a una serie di vincoli che articolano in varie forme la richiesta di una supervisione umana efficace. Ecco, in sintesi, le tipologie di vincoli individuate dal CICR:

1. Bisogna proibire le armi autonome progettate o utilizzate in maniera tale che i loro effetti non possono essere previsti, compresi e spiegati allo scopo di prevenire effetti indiscriminati che possono scaturire dal loro uso e che risultano essere incompatibili con il principio di distinzione del DIU.
2. Bisogna proibire le armi autonome che sono progettate o utilizzate per attaccare direttamente le persone. Il motivo principale di questa proibizione è la salvaguardia della dignità delle potenziali vittime di un'azione bellica. Queste ultime non dovrebbero mai essere soggette a decisioni da parte di una macchina che riguardano la loro incolumità fisica.
3. Bisogna regolamentare mediante opportuni vincoli la progettazione e l'utilizzazione delle armi autonome che non ricadono nei casi 1-2. In particolare, bisogna limitare a oggetti di natura esclusivamente militare il tipo di obiettivi da attaccare; introdurre limitazioni sulla durata, sull'area geografica e sulla portata dell'azione dell'arma autonoma, al fine di rendere possibile la supervisione umana su ogni specifico attacco; introdurre vincoli sull'interazione umano-macchina, per garantire che vi sia una supervisione umana efficace in corso d'opera, che comprenda la possibilità di intervenire tempestivamente e disattivare l'arma autonoma[11].

Delle considerazioni etiche e giuridiche a sostegno del punto 1 si è già detto in riferimento alla teoria della guerra giusta e al DIU. Anche il punto 2 si basa sul rispetto del DIU e fa inoltre appello al rispetto della dignità umana[12]. A questo proposito, il filosofo Peter Asaro ha sostenuto che dal rispetto della dignità umana discende il diritto di ogni essere umano a non essere deprivato arbitrariamente della vita. E affinché una decisione di vita o di morte soddisfi questo requisito, ha affermato Asaro, essa deve essere presa da un altro essere umano, da un *qualcuno* che possa provare empatia e avere compassione per chi è oggetto di una decisione di vita o di morte. Un'arma autonoma è invece un *qualcosa*, una macchina che non ha la capacità di apprezzare il valore della vita umana e di valutare adeguatamente il significato della sua perdita. Per questo motivo, affidare a una macchina la decisione di togliere la vita a un essere umano, conclude Asaro, costituisce una violazione della dignità umana[13].

Il punto 3 si basa sull'ipotesi che armi autonome sufficientemente prevedibili e comprensibili nei loro effetti esistano o possano essere progettate, in modo tale da consentire un'adeguata forma di controllo umano sul loro impiego, da indirizzarne l'uso in conformità al DIU e da permettere l'attribuzione chiara di responsabilità a persone inserite nella catena di comando e controllo, qualora si verifichino violazioni di norme morali o

giuridiche sulla condotta delle azioni belliche.

Un sistema d'arma autonomo che sembra conformarsi al punto 3 è il sistema mobile antimissile *Iron Dome*, utilizzato in Israele per monitorare e neutralizzare con il lancio di missili intercettori i razzi e altri proiettili balistici diretti verso il territorio israeliano.[14] Gli operatori posizionano sul terreno *Iron Dome* e circoscrivono l'area che esso deve monitorare e proteggere. Dopo aver compiuto queste operazioni preliminari, *Iron Dome* viene abilitato a rispondere in piena autonomia, lasciando però agli operatori la facoltà di disabilitare il sistema in corso d'opera. Dotato di analoghe capacità di intercettazione, anche il sistema NBS (*Nächstbereichschutzsystem*) *MANTIS* è utilizzato dalle forze armate tedesche per proteggere i soldati e le installazioni militari da proiettili balistici in arrivo.[15]

Gli attuali sviluppi tecnologici di armi autonome (e di loro precursori) non sembrano accordarsi sempre con il punto 3 e con le altre richieste di proibizione totale avanzate dal CICR. Come è stato già ricordato, nel 2023 l'aviazione militare statunitense ha condotto con successo delle prove sperimentali su un aereo da caccia la cui navigazione aerea è interamente controllata da un sistema dell'IA. Il caccia è stato testato in uno scenario di confronto aereo a distanza ravvicinata (*dogfight*) con un altro caccia, eseguendo manovre di attacco o di evasione. I piloti che erano presenti nella cabina di pilotaggio avevano facoltà di escludere il sistema di IA e di assumere essi stessi il controllo della navigazione aerea. Ma non c'è stato bisogno di fare ciò. Dai test condotti risulta perciò evidente la possibilità tecnologica di trasformare questo aereo in un'arma autonoma, dotandolo delle capacità di selezionare e attaccare un obiettivo, in aggiunta alla capacità acquisita di navigazione autonoma e di *dogfighting* avvalorata dalle prove sperimentali.

Dalla prospettiva assunta nel punto 1 più sopra, è opportuno chiedersi se il comportamento di un caccia di questo genere sia sufficientemente prevedibile, allo scopo di prevenire effetti indiscriminati proibiti dal DIU. Non è ovvio che questa condizione sia invariabilmente soddisfatta in confronti aerei ravvicinati che potrebbero anche coinvolgere un numero elevato di velivoli in interazione cooperativa o competitiva tra loro. Inoltre, i vincoli enunciati al punto 3 prevedono che vi sia la garanzia di un intervento umano tempestivo, per disattivare l'arma autonoma o correggerne l'azione. A causa della latenza prolungata di segnali di controllo impartiti a grande distanza dal velivolo – come accade nel caso di droni che sono pilotati da una stazione a terra situata a migliaia di chilometri di distanza – questa richiesta impone che i segnali di controllo per correggere l'azione del caccia autonomo siano impartiti da una postazione collocata a distanza più ravvicinata. Nelle missioni che comportano il sorvolo di territorio nemico o di zone militarmente contestate, questa possibilità di controllo sembra restringersi a postazioni collocate su altri velivoli sufficientemente vicini, con tutte le limitazioni del caso, che comprendono sia la perturbazione delle comunicazioni nello spettro elettromagnetico da parte del nemico sia la neutralizzazione delle postazioni ravvicinate di controllo. Come verrà interpretato il principio di necessità militare incorporato nel DIU se il controllo ravvicinato verrà meno? Si rinuncerà a concludere la missione oppure si lascerà piena autonomia all'aereo da caccia senza pilota a bordo? È difficile conciliare la seconda opzione con la proposta di regolamentazione a due livelli avanzata dal CICR.

Anche senza elaborare ulteriormente sui problemi tecnologici collegati all'esercizio del controllo umano su un aereo da caccia autonomo abilitato a confronti ravvicinati, la sperimentazione di un tale velivolo entra in tensione con il rispetto di vincoli eticamente e giuridicamente motivati che sono contenuti nella piattaforma 'a due livelli' per la regolamentazione delle armi autonome proposta dal CICR. La possibilità tecnologica di un *dogfight* 'autonomo' non era stata ancora avvalorata sperimentalmente nel 2021, all'epoca in cui fu resa pubblica la proposta del CICR. Ma tale proposta è stata reiterata pressoché *verbatim* anche nel 2024 nel contributo del CICR al Rapporto del Segretario Generale dell'ONU sulle cosiddette 'armi autonome letali'.[16] La persistente assenza di una regolamentazione internazionale non può che aggravare il problema posto da uno sviluppo incalzante di applicazioni militari dell'IA che erodono il controllo degli esseri umani sulle attività belliche, con le conseguenti minacce per il rispetto del DIU e per la protezione della pace. Lo scarto tra esigenza normativa e realtà tecnologica in evoluzione mette impietosamente in luce i ritardi delle attuali iniziative della società civile, degli organismi intergovernativi e della

diplomazia sulla regolamentazione dell'IA in ambito bellico, la persistente assenza di concreti esiti normativi, insieme alla necessità, sempre più pressante, di fronteggiare i rischi crescenti derivanti da una perdita di controllo umano sull'azione bellica.

NOTE

1. Per una analisi critica recente del dibattito sulle armi autonome in lingua italiana e per un approfondimento sui principi etici e giuridici soggiacenti, si veda F. Farruggia (a cura di), *Dai droni alle armi autonome*, Roma: FrancoAngeli, 2023. Il volume *open access* è liberamente scaricabile all'indirizzo https://www.francoangeli.it/Libro/Dai-droni-alle-armi-autonome?Id=28524. Si vedano inoltre i capp. 4 e 5 di G. Tamburrini, *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Roma: Carocci, 2020, e, in inglese, D. Amoroso, *Autonomous weapons systems and international law. A study on human-machine interactions in ethically and legally sensitive domains*, Napoli / Baden-Baden: Edizioni scientifiche italiane / Nomos, 2020, nonché i capp. 9 e 13 di G. Mecacci et al. (a cura di) *Research handbook on meaningful human control of artificial intelligence systems*, Cheltenham, UK: Edward Elgar Publishing, 2024.

2. Si vedano gli articoli "A mass assassination factory: Inside Israel's calculated bombing of Gaza" (accessibile all'indirizzo https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/) e "Lavender: The AI machine directing Israel's bombing spree in Gaza" (accessibile all'indirizzo https://www.972mag.com/lavender-ai-israeli-army-gaza/) I contenuti di questi articoli sono stati ripresi da autorevoli testate giornalistiche. Si veda per esempio l'articolo apparso su The Guardian il 1° dicembre 2023 (https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets) . Per un'analisi condotta dalla prospettiva del Diritto Internazionale Umanitario (DIU) delle procedure documentate in questi articoli giornalistici, si vedano D. Amoroso, "Sistemi di supporto alle decisioni basati sull'IA e crimini di guerra: alcune riflessioni alla luce di una recente inchiesta giornalistica", *Diritti Umani e Diritto Internazionale* 2 (2024): 347-368 e D. Mauri, "Numeri, persone, umanità: sistemi di supporto alle decisioni umane in campo militare da parte dell'IDF e diritto internazionale umanitario", *Diritti Umani e Diritto Internazionale* 2, (2024): 329-346.

3. I test sperimentali sono stati condotti nella seconda metà del 2023 e i risultati conseguiti sono stati resi pubblici nell'aprile del 2024. Si veda https://www.defensenews.com/air/2024/04/19/us-air-force-stages-dogfights-with-ai-flown-fighter-jet/

4. Daniele Amoroso e Guglielmo Tamburrini, "Toward a Normative Model of Meaningful Human Control over Weapons Systems", *Ethics and International Affairs* 35 (2021): 245-272.

5. Michael Walzer, *Guerre giuste e ingiuste. Un discorso morale con esemplificazioni storiche*, Napoli: Liguori Editore, 1990, 175-216.

6. Entrambi i principi sono codificati nei protocolli aggiuntivi alle Convenzioni di Ginevra del 1949, consultabili in rete all'indirizzo https://www.icrc.org/eng/assets/files/other/icrc_002_0321.pdf

7. Daniel Kahneman, *Pensieri lenti e veloci*, Milano: Mondadori, 2021.

8. Kahneman, *Pensieri lenti e veloci*, 96.

9. Per approfondimenti, si vedano A. Coco, "Exploring the Impact of Automation Bias and Complacency on Individual Criminal Responsibility for War Crimes", *Journal of International Criminal Justice* 21 (2023): 1077–1096; M. L. Cummings, "Revising human-systems engineering principles for embedded AI applications", *Frontiers in Neuroergonomics* 4 (2023): 1-6 (doi: 10.3389/fnrgo.2023.1102165); M. L. Cummings, "Rethinking the maturity of artificial intelligence in safety-critical settings", *Artificial Intelligence Magazine* 42 (2021): 6-15.

10. CICR (Comitato Internazionale della Croce Rossa), *ICRC Position on Autonomous Weapons Systems*, Geneva: ICRC, May 12, 2021 https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems.

11. CICR, ICRC *Position on Autonomous Weapons Systems*, 2.

12. Daniele Amoroso, Frank Sauer, Noel Sharkey, Lucy Suchman, Guglielmo Tamburrini, *Autonomy in Weapon Systems. The Military Application of Artificial Intelligence as a Lit-*

mus Test for Germany's New Foreign and Security Policy, Berlino: Heinrich Böll Foundation, 2018.

13. P. Asaro, "On banning autonomous weapon systems: human rights, automation, and the de-humanization of lethal decision-making", *International Review of the Red Cross*, 94 (2012): 687-709.

14. https://www.army-technology.com/projects/irondomeairdefencemi/ . Si veda Amoroso e Tamburrini, cit., per un'analisi della compatibilità con il DIU della forma di autonomia "difensiva" esercitata da tali sistemi sotto la supervisione umana.

15. https://www.army-technology.com/projects/mantis/

16. UN Secretary General, Lethal Autonomous Weapons Systems, Report to the UN General Assembly, 79th Session, 1 July 2024. https://digitallibrary.un.org/record/4059475?v=pdf

Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

Per un'etica critica dell'accelerazione digitale

For a critical ethics of digital acceleration

GIUSEPPE DE RUVO
giuseppederuvo1@gmail.com

AFFILIAZIONE
Università Vita-Salute San Raffaele (Milano),
European Centre for Social Ethics

SOMMARIO

Partendo dalla teoria di Rosa sulla modernità come processo di accelerazione, mostreremo come l'accelerazione digitale produca dinamiche di potere che legittimano il dominio delle piattaforme sulla vita delle persone. Sosteniamo che l'accelerazione digitale installi nei soggetti un deficit strutturale di riflessività, ovvero un'attitudine in virtù della quale i soggetti non colgono l'inaccettabilità riflessiva della loro esistenza online. Mostreremo come tale deficit strutturale di riflessività sia prodotto 1) dal fatto che gli algoritmi si comportano come meccanismi di sincronizzazione che non vengono discussi e 2) dal fatto che, per affrontare l'accelerazione, i soggetti assumono una identità situazionale, in virtù della quale la lotta per il riconoscimento, etica e riflessiva, è sostituita da un'extramorale corsa per la reputazione. In conclusione, l'articolo analizza le strategie della critica etica e della critica immanente e sostiene la necessità di implementare una *critical digital literacy* per rinforzare la riflessività e favorire un atteggiamento critico.

PAROLE CHIAVE

Accelerazione

Algoritmi

Teoria Critica

Riflessività

Reputazione

ABSTRACT

Drawing on Rosa's theory of 'modernity as a process of acceleration', we will show how digital acceleration produces dynamics of power that legitimise platforms' domination over people's lives. We claim that digital acceleration imbues subjects with a structural reflexivity deficit, i. e. an uncritical attitude by virtue of which users do not grasp the reflexive unacceptability of their online existence. We will show that such a structural reflexivity deficit is produced 1) by the fact that algorithms behave as synchronization tools that are not discussed and 2) by the fact that, in order to face acceleration, subjects assume a 'situational identity', by virtue of which the ethical and reflexive struggle for recognition is replaced with an extra-moral race for reputation. In the end, the article analyses the strategies of ethical and immanent critique and argues that a model of critical digital literacy should be implemented in order to reinforce reflexivity and favour a critical attitude.

KEYWORDS

Acceleration

Algorithms

Critical Theory

Reflexivity

Reputation

DOI: 10.53267/20240103

1. MODERNITY AS A PROCESS OF ACCELERATION

Defining modernity is a very complex issue. Studies concerning the concepts of *rationalisation*, *differentiation* and *individualisation* have therefore multiplied in recent decades. The basic idea is that modernity is characterised by processes of rationalisation and secularisation of the social order, within which relations of functional differentiation are established and subjects are individualised as such. This analysis is confirmed not only by objective data, but also by how modern society has represented itself: just think of Balzac's or Kafka's novels about the power of bureaucracies, which embody the spirit of rationalisation and social differentiation, or of the 'discovery' of the human subject's role (individualisation) in eighteenth- and nineteenth-century novels[1]. In short, one of the characteristics of modernity – conceived not as a static sociological category but as a *form of life* – is that it is not simply endowed with certain characteristics, but is also capable of *reflecting on itself*.

It is from these premises that Hartmut Rosa, an exponent of the new generation of the Frankfurt School, proposes to add a new characteristic to the classical definitions of modernity: in his interpretation, modernity is characterised primarily by the «acceleration of social life and, concretely, by the rapid transformation of the material, social and spiritual world»[2]. According to Rosa, the heart of modernity lies in the «logic of social acceleration»[3]: in the increased speed that characterises communicative, political and existential processes. The experience of acceleration is not only objectively measurable in various spheres of social life[4], but can also be traced in the «cultural self-observations of modernity»[5]: that is, in the forms of reflexive expression in which modernity represents itself[6]. According to Rosa, the acceleration that characterises modernity takes three interrelated forms. Firstly, there is an evident acceleration in the development of «*end-oriented* processes in transport, communication and production, which can be called technological *acceleration*»[7]. In modernity, technological innovations have become more frequent and are capable of completely changing subjects' representations of the world, insofar as they compress space-time and favour the movement of people and communication between them, giving rise to a process that culminates in globalisation[8].

Secondly, in modernity we are witnessing an acceleration of social change. This means that «the rhythms of change themselves are changing»[9]. Changes that, in the past, required several generations are now seen as *intragenerational*[10]. There is therefore a social and political «contraction of the present», characterised by «an increasingly rapid decline in the reliability of experiences and expectations»[11]. Cultural standards and models of political legitimation tend to change rapidly; as a result, according to Rosa, the very logic of social change is extremely difficult to govern[12].

Finally, the logic of acceleration also generates an increase in the pace of life. This may seem paradoxical: technological acceleration should relieve human subjects of many tasks, leaving them with more free time. In reality, Rosa notes, the acceleration of social changes and the myriad activities and experiences that new technologies make available to subjects exceed the reduction in complexity that they generate. There is therefore an «increase in the number of individual actions or experiences [that occur] per unit of time»[13]. In such a context, the growth rates of possible experiences «exceed the rates of acceleration, and this is why time is becoming increasingly scarce […] *despite* the remarkable pace of technological acceleration»[14]. The consequence is that, as the number of possible experiences increases, «the time required in order to make rational and informed choices, and to coordinate and synchronise actions, steadily increases»[15]. However, we lack the time we need because the pace of life is structurally accelerated. The consequence is that subjects increasingly rely on tools providing collective synchronisation, and on «some external instance»[16], to ensure that they do not remain desynchronised and therefore isolated from the rest of the world.

In this sense, we are witnessing a contraction of the present, not so much at the socio-political level, but at the 'existential' level. This, Rosa states, transforms «the forms of human subjectivity, and also our being-in-the-world»[17]. Subjects' existence is characterised by a fading experience of the present and regulated by «the *silent normative force* of temporal *laws*»[18]. This is why Rosa speaks of a *totalitarianism of acceleration*, in which «the progression of social acceleration […] can be […] defined as all-pervasive and all-inclusive: it exerts its pressure by inducing a

permanent fear of losing the battle and of no longer being able to keep pace»[19]. According to Rosa, such totalitarianism does not spring from propaganda or violence[20], but from *silent* temporal mechanisms that make the subject «regulated, dominated and oppressed by a temporal regime that is mostly invisible, depoliticised, undisputed»[21]. In short, subjects accept the temporal organisation that is characterised by acceleration without reflecting on it, on its genesis or on the power and alienation to which it gives effect: «these dictates are hardly recognised or perceived as a social construction»[22].

The temporal regime also achieves this because subjects assume a *situational identity,* by virtue of which, instead of deciding on their life projects autonomously, they prefer to adapt and «follow the flow»[23]. Due to social acceleration, the increase in possible experiences and the fact that their «awareness of contingency is unavoidably heightened»[24], human beings assume attitudes that favour reducing complexity; following the flow and adapting to it seems to them to be a useful mechanism for synchronisation. The price to pay, however, is the renunciation of any reflective attitude that is capable of grasping how, in reality, such an existence is profoundly alienated, governed by heteronomous laws and ultimately modelled on the needs of the capitalist system of production, which demands performance, competition and an ability to multiply the number of productive actions[25].

In this sense, Rosa's theory of modernity as acceleration is a *critical* theory. Its aim is not only to analyse the vicissitudes of modernisation, but also to criticise them *immanently*[26]. In fact, Rosa's critical theory does not exclusively aim to 'denounce' the pathologies that social acceleration generates. Rather, its aim is to show how acceleration, though a structural feature of modernity, violates «the promise of autonomy and reflexivity that lies at the heart of modernity itself»[27]. The temporal norms of acceleration must therefore be criticised because, as Rosa writes, «if the project of modernity and the Enlightenment culminate in the idea […] of individual and collective autonomy, social philosophy must certainly pay attention to this phenomenon of automation [of acceleration], which so far has gone unnoticed»[28].

Having clarified Rosa's perspective, this article aims to show how these dynamics are further reinforced in our digital existence, while still inextricably linked to capitalist accumulation. We will therefore attempt to show how synchronisation mechanisms operate (§2) and situational identities are reshaped (§3) in a digital context. In the conclusion (§4), we will attempt to develop what can be called a *critical ethics of digital acceleration.*

2. DIGITAL ACCELERATION: ALGORITHMS AS SYNCHRONISATION TOOLS

It is beyond doubt that, in the onlife experience, the number of possible experiences per unit of time has increased and continues to increase. The volume of data created, consumed, copied and captured on the Internet has increased from 2 zettabytes in 2010 to 147 in 2024[29]. At the same time, the number of emails sent every day rose from approximately 300 billion (2020) to 380 billion (2024)[30]. From 2011 to 2022, the number of WhatsApp notifications increased from a few million per day to 125 billion[31]. Time spent on social media is also increasing, as are the number of users and the amount of content posted on the various platforms[32]. It is estimated that users now deal with around 10,000 digital advertisements per day, a figure that has been growing steadily since 2010[33].

In short, if – following Rosa – we define the acceleration of the pace of life as the increase in possible experiences per unit of time, then it is quite evident that *onlife* existence, to use Luciano Floridi's fortunate neologism, is characterised by a certain acceleration. As Rosa writes, quoting Kenneth Gergen's work, acceleration transforms 'everyday life into a sea that floods us with requests'[34], and it is easy to adapt this metaphor to the digital world, within which requests for friendship, sponsorship, email and so on are becoming increasingly frequent. All these issues, moreover, must be understood in relation to the peculiar ontological structure of the Internet, which is configured 'as an infinite and constantly moving object'[35], absolutely impossible to represent or treat as a mere repository of information. On the contrary, the Internet is ontologically constituted by data's *reciprocal action*: the entry of each new bit into the system implies its reorganisation at ever higher levels of complexity, according to a feedback process that cannot be *slowed down*[36].

It is precisely for these reasons that the web must be constantly *cut*.

For users to have a satisfying digital experience – one that does not consist of a mere (and very rapid) succession of unconnected information – it is necessary to find a way to synchronise their experiences, especially in social media. This operation is carried out by the various algorithms that govern our digital experience, without which the Internet as we know it could not exist. From this point of view, algorithms are first and foremost the fundamental tools of digital synchronisation, useful for ordering the acceleration that characterises online existence. For this reason, as David Beer has recently noted, the narrative of big digital companies is based on their algorithms' ability to organise the user experience quickly. Indeed, in the words of Big Tech itself, algorithms allow subjects 'to be *in the moment* and to react without delay or hesitation'[37]. By guaranteeing synchronisation and by being able to compress the Internet's plethora of information computationally, algorithms are thus supposed to be able to synchronise and simplify users' online lives.

Yet the problem of acceleration remains far from being solved. Algorithms are indeed an extremely useful innovation, but – as Rosa has already noted – technological innovations are not always able to reduce the acceleration of everyday life: the fact that we see a synchronisation of experiences does not imply a decrease in acceleration. It is somewhat *organised*, but not stopped. The question to be thematised, and potentially criticised, therefore becomes not only that of acceleration as such, but also that of how it is organised: if the increasing pace of life risks impeding reflexivity and reducing autonomy, does the algorithmic organisation of digital acceleration mitigate or radicalise such social pathologies?

To answer this question, one must focus on the temporal dimension of algorithmic practices. In fact, while they serve to reduce the complexity of subjects' lives *in the present*[38], this is not achieved through a 'decompression' of the present, but through practices that anticipate the future. The point is not to encourage reflective or resonant practices[39] by slowing down the digital experience, but to offer synchronised and, above all, rapid algorithmic suggestions and predictions that, in any case, leave subjects no time to reflect on them: «the analytic industry is tapping into a wider rationality, in which speed and agility are seen to be crucial»[40]. Due to the exponential increase in the number of possible experiences per unit of time, subjects lack time for reflection, and on social media platforms they encounter predetermined and anticipated futures. These are not decided autonomously, but calculated algorithmically. To an increasing extent, especially among the youngest users[41], they are unreflectively accepted as useful mechanisms for synchronisation and simplification.

Big Tech, then, presents its algorithms as neutral[42], and therefore «the lack of time for reflection is presented as holding no risk»[43]. By insisting on the neutrality and omnipotence of their algorithms' predictions, platforms legitimise their domination over the future of human beings. According to this narrative, there is no need for subjects to reflect autonomously on their courses of action, because 1) the speed and quantity of information would lead them to make mistakes anyway, and 2) there are *algorithms that do it for them*, with a degree of certainty, neutrality and impartiality that no human will ever achieve[44]. Consequently, now that we live in an era in which 'we have little space for critical reflection outside of the flow of information to which we are exposed'[45], it is better to replace human rationality, which is epistemically fallacious, with algorithmic rationality, which is «a rationality that promotes quick and accessible know how that enables all-seeing predicting and smart decision making»[46]. Above all, it is configured «as a potential solution to the need to keep up»[47] with the accelerating pace of life online.

The problem is that, as the literature has amply shown by now, these instruments are neither neutral nor objective. On the contrary, they tend to reproduce the stereotypes and discrimination that are already present in society, thus radicalising political and social problems that should be radically and reflectively addressed[48]. Moreover, as several scholars have pointed out, algorithmic predictions, far from being able to totalise subjects' experiences in order to guide them towards their most authentic and personal desires, tend to push them towards whatever actions platforms judge to be 'optimal', that is those which allow the platforms to maximise their profits through data collection (e.g. by encouraging users to continue engaging with an app[49]) or by increasing advertisements' conversion rate, leading to greater investment in digital advertising[50].

From this point of view, then, algorithms are not neutral synchronising

agents that, by skimming the surface of the Internet's complexities, enable the decompression of the present by helping users to adopt reflective attitudes. On the contrary, thanks to their capacity to order and organise subjects' online experience they behave as a seemingly objective external influence that continually solicits users, leading them to follow algorithmic suggestions unreflectively. Algorithms are presented as the only tools through which it is possible to follow the rapid flow of digital existence: «The [Big Data] analytics industry cultivates and nurtures the risk of being left behind if you choose to take the slow route and not adopt the speediness of these analytics»[51].

The goal is above all economic: this rapid and continuous algorithmic solicitation, together with narratives that emphasise the objectivity of algorithms, means that subjects, already exposed to the accelerating pace of digital life, do not act reflectively when connected, preferring to go with the flow of algorithmic suggestions, and thus legitimising the practices of surveillance, profiling and *digital nudging* necessary for the maintenance of the political economy of surveillance capitalism. As Shoshana Zuboff has noted, this system necessitates passive subjects – ones unable to reflect critically on algorithmic operations and thus to challenge them – since any «unpredictable behaviour is the equivalent of lost revenue»[52].

3. SITUATIONAL DIGITAL IDENTITIES, RECOGNITION/REPUTATION AND THE (IM)POSSIBILITY OF REFLEXIVE CRITICISM

The importance of algorithms as synchronisation tools is not the only consequence of the acceleration of the pace of life that is taking place online. On the contrary, following Rosa's work, it is possible to show how social acceleration impedes reflection even by reshaping the structure of subjectivity that previously developed in modernity. In an accelerated context, «the modern, "classical" sense of identity, which was based on an individual "life project" and on self-determination […], tends to be replaced by new forms of "situational identity" and flexibility, which accept the precariousness of all definitions of the self and of identity parameters, and no longer attempts to follow a life project, but tends instead to ride the wave»[53].

What emerges is a transition from a reflexive and active identity – one which asks itself, in a Kantian fashion, *what ought I do?* [54] – to a subjectivity which, faced with increasingly many possible experiences, renounces asking this question and prefers to go with the flow of novelties, on the basis of which it defines itself *a posteriori*. As Gergen writes: «it is the difference between swimming to reach a certain point in the ocean – taming the waves to achieve a goal – and surrendering in harmony to the unpredictable movements of the waves»[55].

The subject, therefore, does not act reflectively by setting long-term goals and aims. Indeed, social acceleration makes such a project immediately «anachronistic and doomed to failure»[56]: in an accelerated context, characterised by the contraction of the present, the subject is deprived of the time one needs to reflect critically on one's situation and thus on one's future. Social change occurs at such an accelerating rate that any long-term project is in vain, since the conditions that today make it possible and desirable could fail at any moment. Consequently, in a context marked by increasing uncertainty and the «contraction of time units that are definable as the present»[57], it is normal that «forms of identity based on flexibility and on a readiness to change are systematically favoured»[58].

In the digital world, this process is unprecedentedly radical. Indeed, given the ubiquitous presence of new information and communication technologies (ICTs), subjects themselves are beginning «to conceptualize the whole reality […] in ICT-friendly terms»[59], even from a temporal point of view. In the face of increasing numbers of external stimuli and the social pressure of what we can call «digital gazes»[60], digital existence is characterised by subjects' growing tendency to *constantly update* their digital identities in such a way as to synchronise themselves 1) with the speed of the information flow and 2) with the expectations of the digital echo chambers into which they find themselves thrown. From this point of view, the assumption of situational identities in the digital world does not imply, as Rosa seems to suggest, *adaptation* to the pace of acceleration. Rather, in a context characterised by the omnipresence of «reputation metric systems»[61] (likes, retweets, shares and so on), assuming situational identities allows subjects to feel that they are actually appreciated and esteemed, immediately and individually: «To them, it seems most natural to wonder about their person-

al identities online, treat them as a serious work-in-progress, and to toil daily to shape and update them»[62].

This is particularly important from a critical and normative point of view, because it means that the struggle for recognition does not have increasing social freedom as its goal[63], but – as Rosa had already noted – turns into a merely performative exhibition[64]. Being recognised no longer means acquiring a social *right.* In an accelerated context, and especially a digital one, recognition becomes something that «must be reconquered every day. […] It is no longer accumulated, but is always in danger of being completely devalued by the constant flow of events and the shifting of social landscapes. One's position is important [as a means] to increase one's chances of maintaining or gaining social esteem, but it is not certain that one will retain it forever»[65]. In short, the struggle for recognition is disengaged from its normative and reflective dimension. It takes the form of a continuous struggle to maintain one's reputation, which is not maintained by acting reflectively in accordance with a state of affairs that is deemed desirable, but by adapting – in an uncritical and extra-moral manner – to the order of meanings that is 'trendy' (albeit for a very short period) in a digital context. And this, as Rosa notes, is «one of the tragedies of the modern individual: feeling imprisoned in a hamster wheel, while his hunger for life and the world is never satisfied, but instead is increasingly frustrated»[66].

The ethical–political problem that emerges, then, is not only that of heteronomy or the pressure exerted by *filter bubbles* and *echo chambers*. Rather, the question concerns the conditions for any possibility of criticism. The use of algorithmic tools as heteronomous synchronisation mechanisms, which aim at platforms' profit and not at users' wellbeing, and the continuous assumption of situational identities, dictated by the demand for content and the pressure of the digital gaze, together place subjects in a state in which they do not grasp the *reflective unacceptability*[67] of digital platforms' practices of power, surveillance and manipulation. In fact, the mechanisms platforms have put in place to organise digital acceleration obstruct social learning, generating a structural reflexivity deficit that does not allow ordinary agents to criticise the present state of affairs. This also happens because, as we have seen, the normative need for recognition and social freedom is somehow replaced by a spasmodic search for approval and reputation, which leads subjects not to think about how to transcend the present state of affairs, but about how to adapt to it without questioning the practices of power that characterise it[68].

For these reasons, in the conclusion we will outline a critical ethics of digital acceleration, whose primary goal should be to lead subjects to grasp the reflexive unacceptability of such power practices.

4. CONCLUSION: FOR A CRITICAL ETHICS OF DIGITAL ACCELERATION RELATIONS

Thinking about a critical theory of digital acceleration is extremely difficult because, as we have seen, the power of such acceleration lies in how it makes the emergence of practices of reflexivity extremely complex for ordinary agents. Social criticism therefore risks presupposing an asymmetry between subjects, subjected to overwhelming power, and critics, who are able to unveil the practices of power that oppress the subjects. Without entering into this meta-theoretical debate[69], it is possible to show how, in reality, ordinary agents are able to assume, and actually have assumed, critical attitudes towards digital acceleration. For instance, as Judy Wajcman has shown[70], the feeling of being *pressed for time* is widely recognised – and with negative connotations. Similarly, users' trust in Big Tech's practices is very low throughout the West and, not surprisingly, calls for more regulation are the order of the day[71].

These reflective attitudes are usually awakened by what we can call 'ethical' forms of criticism, i.e. forms that emphasise the impossibility of a good life under a regime of power that generates oppressive forms of life. For instance, after the Snowden revelations or the Cambridge Analytica case, many have become aware of digital capitalism's practices of surveillance or of its manipulative nature, emphasising how digital capitalism behaves as an oppressive institutionalized social order. The strategy of ethical critique, notably of structural ethical critique, has an enormous advantage, clearly identified by Rahel Jaeggi and Nancy Fraser: «the ethical perspective is certainly thought-provoking and informative»[72] because it clearly shows how «capitalism's institutional structure predefines some fundamental contours of our form of life, and it does so in a

way that deprives us of our collective capacity to design the modes of living we want»[73]. Structural ethical critique prompts reflection by denouncing the *reflexive unacceptability* of the current state of affairs; it is an extremely useful tool for enacting social learning and consciousness-raising processes. However, this form of criticism tells us nothing about how to *change* the current constellation of power. This is especially true in an accelerated context, in which each technological innovation requires a different and specific ethical critique, and this takes time to emerge and to be reflectively deployed by ordinary agents.

Such an ethical critique must therefore be flanked by a different critique, capable of organising and contextualising the reflexive unacceptability that the former generates. Such a critique can be defined as immanent, and consists in showing how the current practices of power not only are morally unacceptable but, on the basis of their own normative premises, generate forms of life that are *uninhabitable* and are based on *practical contradictions*. This should generate a «tension within a formation that will drive it beyond itself»[74]. From this point of view, immanent critique is characterised as a form of critique that derives its criteria from within the practices being criticised, without presupposing any external criteria or circumstances that, after all, may or may not emerge. According to Rahel Jaeggi, when these practical contradictions are grasped reflectively, an immanent process of social learning is activated. The aim is to bring out the new from the old – to release those emancipatory and normative forces which the structure of a life force does not promote but obstructs once it has become sclerotized and oppressive[75].

According to this scheme, an accelerated digital existence is not only *bad* (a moral critique) but is also in contradiction with itself from a practical point of view: it is in fact impossible to want one thing and its opposite at the same time. To be brief, we can say that, in such a context, we have a practical contradiction between the idea of reflexive autonomy – as embedded into modernity as well as digital practices – and the algorithmic organization of online reality[76]. This practical contradiction, if grasped reflectively (and ethical criticism serves precisely this purpose), should generate a conflict immanent to the digital form of life. This form of life should be considered *uninhabitable*, by ordinary subjects *first and foremost*, insofar as its normative claims – freedom, reflexivity, a reduction of complexity – «cannot be realized without contradiction»[77].

However, even if immanent critique seems to offer a much more structured framework than ethical critique, in a context such as the one we have described it encounters structural limitations too. Indeed, given the speed of the digital world and the relative slowness of the public's deliberative capacity, it is not clear how it is possible to organise and politically represent any collective learning process, which requires a long process of public and discursive deliberation[78]. This depends, to a large extent, on the fact that subjects' reflexivity can certainly be reactivated by an ethical critique; but this in no way implies that it takes the form of immanent critique. In short: it is easier for rebellious, catastrophist and short-term practices of contestation to emerge than *radically reformist* practices[79], i.e. those capable of denying and contesting, in a determinate manner and on the basis of an immanent–transformative perspective, what actually obstructs the processes of learning and the practical appropriation of new technologies.

It is obviously not possible to resolve this dilemma in this paper. What does seem evident, however, is that more and more young people are entering the digital world without being aware of the power dynamics that run through it, of the technical limitations of IT infrastructure, and of the interests of Big Tech. The Internet, even now, is considered a sort of *locus amoenus* where it is possible to exercise freedom and creativity, all while achieving – thanks to algorithms – sensible reductions in complexity.

As things stand, what is lacking is a process of *digital literacy* that goes beyond the attention, fundamental though it is, which is already paid to the use and practical applications of new technologies[80]. In order to give rise to situated and immanent practices of contestation, it would therefore be necessary to implement what we might provisionally call *critical digital literacy*: an educational process that makes young people aware of the contradictions, power dynamics and risks that digital capitalism and the acceleration of digital existence pose to values such as autonomy and reflexivity[81].

As people become aware of the reflexive unacceptability of power prac-

tices, such an awareness, stimulated by a true critical *Bildung*[82], should replace the contingency and the external nature of moral criticism, which fails to identify – except in a vague and sentimentalistic manner – the practical and immanent contradictions inherent in digital capitalism. The aim is to create subjectivities that, aware of the nature of the digital world (and not only from a pragmatic point of view), can reflectively grasp these contradictions, not limiting themselves to denouncing them (ethical criticism), but generating reflectively and communicatively mediated practices of political organisation that can actually generate consensus. Such a consensus should not be oriented towards limiting the use of new technologies, but rather towards separating what is good about them from what is pathological, determinedly negating the practices of digital power without resorting to catastrophising. After all, «the challenge for critical theory is to extricate such potentialities from the logics of domination within which they are situated, and which, to a great extent, they currently serve to perpetuate»[83].

NOTE

1. A. Giddens, *Modernity and Self-Identity*, London: Polity Press, 1991

2. H. Rosa, *Acceleration and Alienation*, Torino: Einaudi 2019, p. 6. From now on, AA. My translation.

3. *Ivi*, p. VIII.

4. Id., *Social Acceleration. A new theory of modernity*, New York: Columbia University Press, 2015, pp. 63–67. From now on, *SA*.

5. Id., *AA*. pp. 6-7.

6. Rosa mainly focuses on the works of Shakespeare, Goethe, Marinetti and Proust. See Id., *SA*, cit., pp. 31-55.

7. Id., *AA*, cit., p. 9.

8. *Ivi*, 10.

9. *Ivi*, p. 11.

10. Id., *SA*, pp. 110-111.

11. Id., *AA*, cit., p. 13.

12. See Id., *SA*, cit., pp. 251–276.

13. Id., *AA*, cit., pp. 15-16.

14. *Ivi*, p. 21.

15. Id., *SA*, cit., p. 125.

16. Id., *AA*, cit., p. 86.

17. *Ivi*, p. 45.

18. *Ivi*, p. 44.

19. *Ivi*, p. 71.

20. H. Arendt, *The origins of totalitarianism* (New York: Penguin, 2019).

21. H. Rosa, *AA*, cit., p. VIII.

22. *Ivi*, p. 71.

23. *Ivi*, pp. 47-48.

24. Id., *SA*, p. 232.

25. Id., *AA*, cit.

26. We will return to this concept in §4.

27. *Ivi*, p. 88.

28. *Ivi*, p. 46.

29. P. Taylor, 'Amount of data created, consumed, and stored 2010–2020, with forecasts to 2025', *Statista*, 16 November 2023.

30. S. Singh, 'How many emails are sent per day in 2024?', *Demandsage*, 21 May 2024.

31. 'WhatsApp Statistics, Users, Demographics as of 2024', *What's the big data*, 12 December 2023.

32. For an in-depth analysis of each individual platform, see 'Global social media statistics', *Data Reportal*, last update 24 July 2024.

33. 'How many ads do we see a day', *Siteefy*, 25 April 2024.

34. K. Gergen, *The saturated self*, New York: Basic Books, 2000, p. 75.

35. Floyd, S., & Paxson, V. (1997). 'Why we don't know how to simulate the Internet', In Andradòttir, S., Healy, K. J., Nelson, B. L. & Whiters, D. H. (Eds.). *Proceedings of the 1997 Winter Simulation Conference*, p. 1037.

36. L. Barabasi, *Linked. The new science of networks*, Cambridge: Perseus, 2002.

37. D. Beer, *The Data Gaze*, London: SAGE, 2019, p. 22.

38. J. Cohn, *The burden of choice*,

New Brunswick: Rutgers University Press, 2019.

39. H. Rosa, *Resonance*, New York: Columbia University Press, 2019.

40. D. Beer, *op. cit.*, p. 39.

41. V. Barassi, *Child, Data, Citizen*, Cambridge: MIT Press, 2020.

42. T. Gillespie, *Algorithms*, in Peters, E. (ed), *Digital Keywords*, Princeton: Princeton University Press, 2016, pp. 18-30.

43. D. Beer, *op. cit.*, p. 48.

44. See, G. De Ruvo, 'Algorithmic objectivity as ideology: toward a critical ethics of digital capitalism', *Topoi*, 3 (2024), pp. 1-12.

45. *Ivi*, p. 40.

46. *Ivi*, p. 32.

47. *Ivi*, p. 38.

48. K. Crawford, *Atlas of AI*, New Haven: Yale University Press 2021.

49. N. Srnicek, *Platform Capitalism*, London: Polity Press, 2016.

50. P. Langley, A. Leyshon, 'Platform Capitalism: The Intermediation and Capitalization of Digital Economic Circulation', *Finance and Society*, 1 (2017) pp. 11-31.

51. D. Beer, *op. cit.*, p. 46.

52. S. Zuboff , *The age of surveillance capitalism*, London: Polity Press, 2019, p. 151.

53. H Rosa, *AA*, cit., pp. 47-48.

54. I. Kant, *Critique of practical reason*, London: Hackett, 2002.

55. K. Gergen, *op. cit.*, p. XVIII.

56. H. Rosa, *SA*, cit., p. 243.

57. Id., *AA*, cit., p. 13.

58. Id., *SA*, cit., p. 243.

59. L. Floridi, *The fourth revolution*, Oxford: Oxford University Press, 2014, p. 44.

60. *Ivi*, p. 73.

61. D. Cardon, *Che cosa sognano gli algoritmi*, Milano: Mondadori, 2015, p. 25. My translation.

62. L. Floridi, *op. cit.*, p. 60.

63. A. Honneth, *Freedom's right*, New York: Columbia University Press, 2015.

64. H. Rosa, *AA*, cit., p. 69.

65. *Ivi*, p. 68.

66. *Ivi*, p. 28.

67. R. Celikates, *Critique as social practice*, London: Rowman & Littlefield, 2018, p. 158.

68. G. Origgi, *Reputation*, Princeton: Princeton University Press, 2018.

69. For a reconstruction, see Celikates, *op. cit.*, chapter 1.

70. J. Wajcman, *Pressed for time*, Chicago: University of Chicago Press, 2014.

71. A. Bradford, *Digital Empires*, Oxford: Oxford University Press, 2023.

72. N. Fraser, *Capitalism. A conversation with Rahel Jaeggi*, London: Polity Press 2018, p. 178.

73. *Ivi*, pp. 180-181.

74. R. Jaeggi, *Critique of Forms of Life*, Oxford: Oxford University Press, 2018, p. 249.

75. See T. Stahl, 'Oppressive forms of life', *Critical Horizons*, 2 (2024), pp. 77-93.

76. This approach, of course, understands the concepts of freedom and reflexivity to be transcendental conditions of any possibility of a good life, so it is also a *transcendental* critique. On how transcendental critique can be implemented in a model of immanent critique, see R. Mordacci, *Critica e Utopia*, Rome: Castelvecchi, 2022.

77. R. Jaeggi, *op. cit.*, p. 285

78. See T. Stahl, *Immanent Critique*, London: Rowman & Littlefield, 2024.

79. N. Fraser, A. Honneth, *Redistribution or Recognition?:A Political–Philosophical Exchange*, London: Verso, 2003.

80. See, for an overview of this approach, D. Ng *et alia*, 'Conceptualizing AI literacy: An exploratory review', *Computers and Education: Artificial Intelligence*, 2 (2021).

81. See, although in a journalistic fashion, D. Buckingham, *The media education. A manifesto* (London: Polity Press, 2019).

82. N. Dreamson, *Critical Understandings of Digital Technologies in Education*, New York: Routledge, 2019.

83. G. Delanty, N. Harris, 'Critical theory and the question of technology: The Frankfurt School revisited', *Thesis Eleven*, 1 (2021), pp. 88–108.

Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

Automatismo e innovazione. L'etica e la formazione del soggetto nell'epoca dell'intelligenza artificiale

Automatism and innovation. Ethics and the formation of the subject in the age of artificial intelligence

ENRICO REDAELLI
enrico.redaelli@univr.it

AFFILIAZIONE
Università degli Studi di Verona

SOMMARIO

In questo articolo provo a delineare il tipo di impatto che la progettazione e l'utilizzo dei sistemi di intelligenza artificiale avranno sul nostro modo di concepire la soggettività umana. Ovvero, come si modificheranno quelle nozioni cui di solito ci riferiamo quando intendiamo sottolineare la peculiarità della mente umana rispetto al comportamento di macchine e animali. Nozioni come "intelligenza", "ragionamento", "creatività", "innovazione", spesso indicate come "facoltà" tipicamente o esclusivamente umane, vengono qui considerate nelle loro trasformazioni storiche alla luce della loro stretta relazione con la tecnica, dalla tecnologia alfabetica sino ai sistemi di intelligenza artificiale generativa.

PAROLE CHIAVE

Intelligenza artificiale

Innovazione

Creazione

Soggettivazione

Ragionamento

ABSTRACT

In this article, I attempt to outline the kind of impact that the design and use of artificial intelligence systems will have on our understanding of human subjectivity. That is, how certain notions will change. In particular, those notions we usually refer to when we want to emphasise the peculiarity of the human mind compared to the behaviour of machines and animals. Notions such as 'intelligence', 'reasoning', 'creativity', 'innovation', often referred to as typically or exclusively human 'faculties', are here considered in their historical transformations in the light of their close relationship with technology, from alphabetic technology to generative artificial intelligence systems.

KEYWORDS

Artificial intelligence

Innovation

Creation

Subjectification

Reasoning

DOI: 10.53267/20240104

1. INTRODUZIONE

Che tipo di impatto avranno la progettazione e l'utilizzo dei sistemi di intelligenza artificiale sul nostro modo di concepire la soggettività umana? 'Intelligenza', 'ragionamento', 'creatività', 'innovazione' sono spesso indicate come 'facoltà' tipicamente o esclusivamente umane. Sono cioè alcune delle nozioni che di solito chiamiamo in causa quando intendiamo sottolineare la peculiarità della mente umana rispetto alle modalità di azione di macchine e animali. Queste nozioni si stanno già trasformando con l'uso e la diffusione dei software di IA. Provo qui a considerare, in particolare, creatività e innovazione alla luce della loro stretta relazione con la tecnica. Come ogni tecnica, infatti, anche l'intelligenza artificiale produce una significativa variazione di ciò che si intende per 'creativo' e 'innovativo' e dunque anche di ciò che si intende per 'umano'.

Mi soffermerò in un primo momento sulla relazione che lega innovazione e creazione, da una parte, e tecnica, dall'altra. Metterò poi in luce quali tipologie di attività intellettuali – un tempo attribuite alle capacità umane di creazione e innovazione – vengono già oggi assorbite nel campo degli automatismi replicabili dagli algoritmi dell'intelligenza artificiale, delineando alcune delle conseguenze che ciò comporta nel modo di concepire l'etica e la formazione dei soggetti umani.

2. AUTOMATISMO E CREATIVITÀ

La facoltà di innovare e creare è da sempre legata alla tecnica e cambia con essa. Che cosa avviene in ogni trasformazione tecnica? Un salto. La trasformazione fa compiere all'attività creativa e innovativa un salto di gradino. Essa, infatti, introduce un blocco di operazioni automatiche (da immaginare proprio come un blocco di marmo che costituisce il gradino di una scala) svolte dall'artefatto tecnico e sposta l'attività creativa e innovativa a un altro livello, ossia al di sopra di quel gradino. Per così dire, dal livello *n*, che è la base del gradino, al livello *n*+1 che è l'altezza del gradino.

Si prenda ad esempio l'introduzione dell'aratro nelle pratiche agricole: essa genera un blocco di operazioni automatiche, che saranno compiute dall'artefatto tecnico, per cui il lavoro creativo e innovativo si sposta dal livello *n* (arare con la zappa) al livello *n*+1 (arare con l'aratro). A livello *n* il contadino arava con la zappa e questa pratica apriva un campo di automatismi (operazioni svolte dall'artefatto tecnico che sostituiscono operazioni prima svolte dall'uomo) e, contestualmente, anche un campo di potenziali non-automatismi, ossia un campo di possibilità creative e innovative: ad esempio, trovare nuovi nodi di impugnare la zappa che garantisca maggiore efficacia oppure trovare nuove posture del corpo che permettano minore fatica. Una volta introdotto l'aratro si genera un gradino, un livello *n*+1, al di sotto del quale il campo di possibili operazioni con la zappa è assorbito dagli automatismi dell'aratro e dunque in quel range la libertà di innovare, creare, pensare nuove possibilità e sperimentarle è azzerata: tutto quello che accade tra *n* e *n*+1 è automatizzato, ossia svolto da quel nuovo artefatto tecnico che è l'aratro. Il pensiero creativo e innovativo non è cancellato *in toto*, ma si trova ora dislocato al livello *n*+1: ad esempio, pensare nuovi modi di ammaestrare i buoi che spingeranno l'aratro o creare nuove forme di aratro che perfezionino quelle esistenti. Tutto il campo di possibilità creative e innovative che si estende tra *n* e *n*+1 (tutti i possibili usi creativi e innovativi della zappa) non solo non esiste più, essendo sostituito da un blocco di automatismi, ma non serve più, non importa più, scivola nell'irrilevanza. Laddove è sostituita dall'aratro, la zappa cade nell'oblio portandosi dietro tutte le possibilità di innovazione e creazione a essa legate. In conclusione, ogni nuovo artefatto tecnico introduce un salto di gradino che non si limita a generare un blocco di automatismi e, contestualmente, un campo di non-automatismi (un campo di potenzialità che si apre solo ora, a livello *n*+1), ma produce anche l'effetto di cancellare retroattivamente un intero mondo di possibilità creative e innovative, le quali si eclissano con la caduta in disuso della tecnica precedentemente in uso e delle relative pratiche.

Creatività e innovazione sono dunque, come si diceva, strettamente intrecciate alla tecnica. Esse si determinano a partire da – e per differenza da – un regime di automatismi: costituiscono il campo potenziale dei non-automatismi. Ma tale campo potenziale non esiste in assoluto, bensì sempre e soltanto in relazione a un determinato gradino, ossia a una specifica soglia di automatismi generata da una specifica novità tecnica: questa apre un campo di possibilità creative e innovative cancellando simultaneamente un

altro campo di possibilità creative e innovative. Ogni nuovo gradino, ogni novità tecnica, cioè, non seleziona soltanto, all'interno del campo di azioni attuali, la serie di azioni che andranno a costituire un blocco di automatismi (le azioni che saranno ora svolte in automatico dall'artefatto tecnico) ma seleziona anche un campo di potenzialità (il bacino delle potenziali azioni non-automatizzate). Per dirla col linguaggio di Deleuze, ogni gradino non seleziona soltanto un campo di attualità, ma anche un campo di virtualità[1].

Ora, se questo è quanto accade con l'introduzione di ogni novità tecnica, anche con l'ingresso dell'intelligenza artificiale ci troviamo di fronte a uno scenario simile: essa genera un blocco di automatismi che non azzera creatività e innovazione ma le sposta a un altro livello, selezionando un campo di potenziali non-automatismi (tutti gli usi liberi e creativi che l'intelligenza artificiale rende possibile) e cancellando al contempo un altro campo di potenziali non-automatismi (tutti gli usi liberi e creativi che l'intelligenza artificiale rende obsoleti o irrilevanti). Dunque, se ora guardiamo in particolare all'intelligenza artificiale generativa, così come si è venuta configurando in tempi recenti, e osserviamo quali tipologie di attività verranno assorbite dagli automatismi dell'algoritmo, come si verrà riconfigurando il campo potenziale delle attività creative e innovative?

3. RAGIONAMENTO E PENSIERO CRITICO

L'intelligenza artificiale generativa rende automatiche tutta una serie di operazioni che in precedenza erano attribuite alla nostra creatività, capacità di giudizio e di pensiero critico. Ad esempio, scrivere articoli accademici o tesi di laurea. Come noto, ChatGPT ha già sfornato articoli accademici passando anche la *peer review* di una rivista scientifica[2], ha dimostrato apprezzabili doti come revisore di articoli di riviste accademiche[3] e ha già realizzato la tesi di laurea di diversi studenti universitari[4]. Tutto questo, anziché dirci qualcosa dei sistemi di intelligenza artificiale generativa (se siano davvero intelligenti, se siano in grado di pensare, se ragionino come un essere umano, ecc.), dice semmai qualcosa di noi. Ci fa da specchio. Come gli operai all'inizio della rivoluzione industriale, anche noi scopriamo che molte delle nostre attività sono automatizzabili, con la sorpresa che si tratta ora di attività intellettuali e non più soltanto manuali. Apprendiamo, non senza una certa inquietudine, che siamo noi, prima ancora dei nostri sistemi di intelligenza artificiale, a essere dei «pappagalli stocastici»[5]. Molto di ciò che pensiamo e di ciò che scriviamo è stocasticamente prevedibile, come l'intelligenza artificiale dei nostri dispositivi non smette di ricordarci ogni volta che ci suggerisce la parola successiva da scrivere il più delle volte indovinandola.

In un certo senso i sistemi di intelligenza artificiale generativa offrono una sorprendente conferma empirica di quanto la filosofia e le scienze umane degli ultimi due secoli avevano ampiamente teorizzato in merito al soggetto, alle sue modalità di pensiero e di comportamento in quanto modalità storicamente determinate. Basti pensare a come, in modi assai diversi ma ampiamente convergenti, il pragmatismo americano, da una parte, e lo strutturalismo francese, dall'altra, hanno affrontato il tema della costituzione del soggetto a partire dalle «pratiche (per dirla con Foucault[6]) o dagli «*habits*» (per dirla con Peirce, James e Wright[7]), nozioni oggi al crocevia tra scienze cognitive, antropologia, semiotica e sociologia. Come si costituisce il soggetto, come si forgia il suo comportamento, come si plasma il suo modo di pensare? Entro alcune cornici, a partire da alcuni *frames*. Ovvero, entro pratiche e abiti di condotta – direbbero pragmatisti e strutturalisti – che sono già dati, che non dipendono cioè dai singoli soggetti, dal libero arbitrio di ciascuno, bensì dal contesto sociale, culturale, storico, tecnologico, ecc. Sia le pratiche sia agli *habits* sono infatti definibili come mobili strutture che guidano l'azione umana, *frames* pratico-operativi dotati di un loro funzionamento relativamente autonomo e passibili di evoluzione in base a contingenze empiriche storicamente determinate. Ed è entro queste strutture, in continua trasformazione, che i soggetti umani forgiano i loro comportamenti e i loro modi di pensare[8]. Sicché, per dirla con Foucault, «non si può in qualunque epoca parlare di qualunque cosa»[9]. Queste strutture sono cioè, per usare un altro termine foucaultiano, dispositivi di soggettivazione: producono un certo modo di essere soggetti[10].

L'intelligenza artificiale oggi ci mostra in modo tangibile il nostro essere iscritti entro queste pratiche, ossia entro una serie di abiti di condotta e di pensiero largamente comuni: molto di quello che facciamo e diciamo è codificabile in procedure standardizzate ed è dunque oggi immedia-

tamente codificato dall'intelligenza artificiale e tradotto in automatismi. L'omologazione prodotta da pratiche di vita comuni (viviamo nella stessa epoca, usiamo gli stessi strumenti, ecc.), una volta tradotta in *big data* sottoposti a calcoli statistici, ci viene restituita dall'IA sotto forma di prevedibilità con lievi margini di errore rendendo platealmente visibile quanto siamo dipendenti dal *milieu* in cui viviamo. Ovvero, mostrandoci fino a che punto siamo il prodotto dei dispositivi in cui siamo immersi. Sicché, ciò che stupisce (e un po' inquieta) dei *chatbot* stile ChatGPT non è il loro rendere evidente quanto la macchina sia ormai diventata 'umana', semmai il loro rendere evidente quanto noi umani siamo sempre stati 'macchine'.

In particolare, quali tipologie di attività intellettuali, un tempo attribuite al libero pensiero di ciascuno, vengono già oggi assorbite nel campo degli automatismi replicabili dall'algoritmo? Limitandoci all'ambito delle competenze di area umanistica, molto di quello che viene solitamente incluso sotto le etichette di 'ragionamento' e 'pensiero critico'. Ovvero, tutta una serie di capacità logiche e analitiche legate alle pratiche di lettura e scrittura di testi: la capacità di riassumere un testo o il pensiero di un autore ricavato da più testi, la capacità di confrontare più testi o più autori tra loro individuando analogie e differenze, la capacità di argomentare una tesi tramite esempi e casi specifici, la capacità di applicare in un nuovo ambito di riferimento una tesi sostenuta da un autore in un ambito diverso ovvero di applicare in un nuovo contesto gli strumenti concettuali utilizzati da un autore in un contesto diverso. Tutte queste sono attività automatizzabili e già automatizzate dall'intelligenza artificiale generativa, con risultati non sempre soddisfacenti ma con ampi margini di miglioramento (siamo solo all'inizio)[11]. Ci troviamo cioè ora di fronte a queste attività intellettuali nello stesso modo in cui ci troviamo di fronte alle attività di calcolo da quando è stata messa sul mercato la calcolatrice portatile: possiamo farle 'a mano' o 'a macchina' o in modo misto ricorrendo a entrambe le modalità. E non c'è da stupirsi se l'algoritmo è in grado di prelevare una tesi o un concetto da un determinato contesto e applicarli automaticamente a un altro contesto (un tipo di attività che è sempre stata ritenuta esclusiva della nostra umana capacità di giudizio critico): la nuova intelligenza artificiale è definita 'generativa' perché fa esattamente questo sui dati tradotti in numeri, ovvero non solo applica deduttivamente regole a nuovi contesti di dati diversi da quelli di partenza, ma "genera" nuove regole individuando *pattern* (regolarità stocastiche) nel bacino dei dati di partenza e le applica generando così nuovi dati[12]. Guardati dal lato della macchina, tesi e concetti non sono che parole o agglomerati di parole tradotti in cifre da cui è possibile ricavare regolarità e dunque leggi probabilistiche. Detto in breve: queste macchine "pensano" (se "pensare" è saper cogliere un concetto in un contesto e applicarlo a un nuovo contesto) pur senza capire nulla di quello che "pensano" (essendo quella operazione nient'altro che un calcolo statistico).

Questo, a grandi linee, è il gradino di automatismi che si è venuto configurando nell'ambito delle attività intellettuali un tempo svolte unicamente "a mano" (ossia in modo umano). Come ogni gradino di automatismi, anche quello introdotto dall'intelligenza artificiale generativa sposta il campo potenziale di attività creative e innovative a un altro livello. Ovvero, genera un nuovo campo potenziale (a livello n+1) e allo stesso tempo cancella retroattivamente un campo di possibilità creative e innovative (a livello n). Questa trasformazione genera importanti effetti di soggettivazione.

Guardiamo allora alle conseguenze nella costituzione del soggetto.

4. RISOGGETTIVAZIONE

Con l'introduzione dei sistemi di intelligenza artificiale generative alcune potenzialità creative e innovative non saranno più rilevanti in relazione a una certa attività intellettuale una volta che questa sia svolta in automatico dalla macchina (si pensi a tutte le professioni 'umane' che verranno a cadere con i loro bagagli di *know how*, *expertise* e relativi campi di possibilità). L'attività intellettuale in questione (qualunque essa sia), nella misura in cui viene sostituita dall'operatività della macchina, non è cancellata come tale, è solo esternalizzata. Tutte le attività intellettuali automatizzabili, nel momento in cui vengono almeno in parte esternalizzate (affidate alla macchina), generano nel corpo umano un'atrofia. È quanto già diceva Platone a proposito dell'introduzione e della diffusione della scrittura: essa produce un'atrofia della memoria, che infatti non è più necessaria per i contenuti esternalizzati su un supporto non corporeo (dalla tavoletta di cera alla

carta). L'atrofia avviene dunque in concomitanza con ciò che Bernard Stiegler definisce un'*exosomatizzazione*, ossia la produzione di un organo esterno (l'utensile tecnico) che svolge funzioni prima appannaggio di un organo interno[13]. Ma non si tratta di una semplice dislocazione di alcune funzioni, che prima erano corporee e che ora sono extracorporee, sicché l'atrofia corporea verrebbe compensata da un organo tecnico extracorporeo. In questo passaggio avviene infatti una risoggettivazione complessiva. È quanto si può osservare proprio con l'introduzione della scrittura alfabetica, per riprendere l'esempio suggerito da Platone. Essa genera una nuova forma di soggettività, che in parte ha eclissato la soggettività tipica della cultura delle civiltà orali e che a sua volta sarà almeno in parte destinata ad eclissarsi con la diffusione di tecniche e pratiche basate sull'uso dell'intelligenza artificiale generativa. Proviamo qui a ricostruire brevemente il tipo di soggettività generato dalla pratica di scrittura alfabetica, poiché è proprio il tipo di soggettività (o, quanto meno, uno dei tipi di soggettività più significativi) che verrà appunto riplasmato dai sistemi di intelligenza artificiale generativa.

5. L'ALGORITMO ALFABETICO

La nostra civiltà occidentale ha osannato una serie di attitudini quali la capacità argomentativa, la capacità di analisi e il senso critico, ossia la possibilità di analizzare i discorsi con un certo distacco emotivo, confrontando un'opinione con un'altra opinione, individuando analogie e differenze, ecc. Ha fatto di queste attitudini i vessilli dell'autonomia e della libertà, di contro alle civiltà arcaiche e tribali in cui l'omologazione dei soggetti ai pensieri, agli usi e ai costumi della comunità di appartenenza è più spiccata, dovuta a un minor distacco critico del soggetto e a un suo maggior coinvolgimento emotivo e sociale nel *milieu* di riferimento. Ma la scuola di Toronto (Harold Innis, Eric A. Havelock, Walter J. Ong, Marshall McLuhan), nei vari studi dedicati ai *media* e in particolare al passaggio dalla civiltà orale alla civiltà della scrittura, ha mostrato come queste attitudini non siano il frutto della 'libertà' e dell''autonomia' dell'individuo intese come facoltà soggettive ascrivibili alla mente umana. Paradossalmente, capacità di analisi e senso critico sono generate dalla tecnica e perciò dipendono dalla tecnica. In particolare, da quel particolare dispositivo tecnico che è la pratica di scrittura alfabetica. Questo tipo di scrittura ha prodotto trasformazioni radicali poiché ha anzitutto innescato una diversa modalità di fruizione della parola e, di conseguenza, una differente collocazione del soggetto[14]. E il senso di questo profondo mutamento della soggettività diviene comprensibile solo se guardiamo alle diverse modalità di fruizione delle scritture prealfabetiche rispetto a quelle alfabetiche, cioè ai differenti dispositivi che esse incarnano e ai diversi effetti di soggettivazione che producono.

Le scritture prealfabetiche, come i sistemi ideografici e i sillabari, erano 'scritture sacre' ampiamente dipendenti dalla cultura orale: il loro contenuto, per poter essere letto, doveva prima essere stato recitato e memorizzato, perciò tali scritture erano semplicemente una traccia del discorso parlato. I segni sillabici e logografici non sono cioè altro che uno stimolo per la bocca e per l'orecchio aventi il fine di riportare alla memoria un discorso già fatto, come le formule rituali, o già organizzato in una serie di moduli mnemonici, come nel caso della narrazione mitologica. Queste tipologie di scritture sono quindi efficienti nella conservazione di testi, di discorsi e di formule già conosciuti a memoria ma non altrettanto nella creazione di nuovi enunciati. La mentalità ripetitiva e conservatrice delle civiltà prealfabetiche non è dovuta al fatto che tali popolazioni fossero 'arcaiche' mentre noi saremmo 'moderni' ma è un effetto concreto e una precisa conseguenza dell'utilizzo della comunicazione orale o di un sistema di scrittura che dipende ancora largamente dall'uso della voce e della memoria orale, come il caso delle scritture logografiche e sillabiche. Ben diversa è la situazione per la scrittura alfabetica.

La notazione alfabetica, che fa la sua comparsa nella Grecia del IX-VIII secolo a. C., determina anzitutto una nuova modalità di lettura e di fruizione del testo scritto: altro dispositivo, altri effetti di soggettivazione. Se il lettore di un testo sillabico è totalmente immerso, con la mente e con il corpo, nel flusso elocutivo che va rianimando, il lettore alfabetico si trova invece a una costitutiva distanza dall'oggetto del discorso: il flusso elocutivo non è da lui incarnato nelle proprie membra, ma è interamente oggettivato su un supporto materiale che gli sta di fronte. *Interamente*, ovvero: lettera per lettera. Mentre il lettore sillabico si trova davanti solo delle tracce, stimoli utili a riattivare dei percorsi mnemonici, il lettore al-

fabetico si trova davanti un oggetto del tutto autonomo e indipendente dalla memoria e dalla tradizione orale, separato dal proprio corpo, svincolato cioè dalla necessità di una rievocazione acustica e musicale. Si trova insomma di fronte a un muto e semplice 'manufatto', che, richiedendo unicamente l'intervento della vista, cancella dall'atto della lettura tutta la componente multisensoriale. Da una lettura sinestetica, empatica e partecipativa, si passa così a una lettura silenziosa, analitica, 'distaccata'. In pratica, si potrebbe dire, dall'intonazione di un canto si passa alla 'presa visione' di un testo.

La distanza che separa il lettore alfabetico dal testo che gli sta di fronte gli permette tutta una serie di operazioni che erano precedentemente impossibili: apre un campo di potenzialità. Il lettore può ora fermare il flusso discorsivo in ogni momento, può rileggere alcune parti del testo e saltare delle righe. Può cioè bloccarsi, tornare indietro e andare avanti, come se avesse tra le mani il telecomando di un videoregistratore. Può insomma meditare su ciò che sta leggendo, ragionare freddamente sul contenuto e prenderne le distanze. Questo tipo di operazioni rendono possibile e, alla lunga, producono un atteggiamento critico e distaccato nei confronti del testo e del suo contenuto. Producono, cioè, un *soggetto critico*. Tale distacco risulta impossibile a chi, come un lettore sillabico, incarna il contenuto e il pensiero di un testo nelle proprie membra, rivivendolo in un fiume ipnotico di parole dall'andamento ritmico e formulaico. Con la scrittura alfabetica il contesto sinestetico ed empatico, musicale e gestuale, non è invece più necessario e diventa anzi ridondante. Viene così a cadere proprio il rito, con i suoi gesti, le sue formule e la sua coreografia. La *cancellazione del corpo*, operata dall'alfabeto, produce una *cancellazione del rito*[15]. Con la diffusione della scrittura alfabetica – a causa delle sue modalità di fruizione e del tipo di operazioni che essa innesca – l'uomo esce dalla dimensione sacrale e rituale e si avvia a divenire quell'uomo logico e analitico che oggi conosciamo, protagonista di una cultura critica e razionale quale quella che caratterizza la storia dell'Occidente alfabetizzato.

Da allora molte delle operazioni rese possibili dall'alfabeto (sguardo «distaccato» sui contenuti di un testo, capacità di analisi e di confronto tra testi, e dunque tra opinioni, applicazione di un concetto a un contesto diverso da quello di partenza, ecc.) sono divenute patrimonio importante della formazione culturale del soggetto occidentale e per questo sono state insegnate nel corso dei secoli per oltre due millenni. Sono cioè diventate degli *habits* che, tra XIX e XX secolo, con l'istituzione dei sistemi scolastici obbligatori, hanno raggiunto un'ampia diffusione, ben al di là dei confini occidentali, sino a divenire sostanzialmente globali. Non si tratta di semplici tecniche e abilità, ma di nuovi 'valori' e principi sulla base dei quali plasmare l'educazione umana: valori e principi – come l'atteggiamento analitico e il distacco critico nei confronti del sapere e della tradizione – che hanno soppiantato quelli precedenti, legati alla conservazione e alla rimemorazione della tradizione in cui si era immersi. La traduzione algoritmica di questi *habits* e la loro trasformazione in operazioni automatiche svolte dai sistemi di IA non può che generare un'atrofia dovuta alla parziale exosomatizzazione (esattamente come i calcoli affidati alla calcolatrice hanno reso meno necessario, se non inutile, l'esercizio 'a mano') e una riconfigurazione delle abilità specificatamente umane (le quali non riguarderanno più lo svolgimento di quelle determinate operazioni ora svolte dalla macchina, semmai l'uso libero, creativo e critico della macchina stessa). Con quali conseguenze sul tipo di soggettività che così si verrà plasmando e sui modelli pedagogici che si verranno imponendo?

6. L'ALGORITMO GENERATIVO

L'atrofia rispetto a un *habit* (prima incorporato e ora scorporato) non significa la scomparsa assoluta di tale *habit*, ma la sua esternalizzazione (quello che svolgeva il corpo umano è ora svolto dalla macchina). Si tratta di quel processo che Bernard Stiegler definisce una exosomatizzazione: viene prodotto un organo artificiale esterno che ora viene a svolgere la funzione prima svolta dal corpo umano. Dunque, l'*habit* non scompare ma la sua dislocazione all'esterno (nello strumento artificiale) non è senza conseguenze per il soggetto. L'atrofia e l'exosomatizzazione di quegli *habits* e di quelle funzioni che oggi definiamo in termini di analisi e confronto 'critico' non possono che condurre a una profonda risoggettivazione, ovvero a una rivoluzione dei modi di essere soggetti: una trasformazione di portata non meno imponente di quella avvenuta nel passaggio dalle culture orali alla civiltà alfabetica. Infatti, che significa affidare agli automatismi della mac-

china le pratiche di analisi, confronto, applicazione di concetti, ossia ciò che generalmente intendiamo (o sin qui abbiamo inteso) con 'ragionare'? Detto altrimenti: quale altro senso viene ad assumere ciò che sin qui abbiamo chiamato 'ragionare' una volta che questo è (almeno in parte) tradotto in operazioni automatiche esternalizzate?

Va anzitutto osservato che si tratta di un processo in parte già avvenuto, in modalità non dissimili, nel recente passato. Con l'avvento dei calcolatori una parte di ciò che si intendeva per 'ragionare' è stato assorbito dagli automatismi della macchina col risultato che oggi quella parte è considerata la 'meno nobile' del ragionare (ovvero, la meno nobile tra le varie accezioni che questo verbo può assumere). Mi riferisco ai procedimenti deduttivi. La logica deduttiva, resa possibile dalla tecnologia alfabetica e tenuta in grande considerazione da Aristotele che ne fa addirittura un tratto distintivo dell'umanità in quanto tale («l'uomo è un animale razionale»), è oggi considerata poco più di una serva: i procedimenti deduttivi, infatti, non sono altro che la mera applicazione di una regola. Niente meno e niente più di un algoritmo. Sono stati proprio l'ingresso e la diffusione dei calcolatori nella seconda metà del secolo scorso a portare a una parziale 'svalutazione' (in ambito umanistico) della logica deduttiva (direttamente proporzionale al suo uso sempre più massiccio in ambito tecnico, nei software e nella programmazione informatica) e a una rivalutazione della logica abduttiva. Delineata nel suo funzionamento da Charles Sanders Peirce alla fine del XIX secolo ma ampiamente ignorata sia prima sia dopo, la logica abduttiva ha iniziato a essere considerata soltanto negli ultimi 50 anni come la 'vera' logica della scoperta scientifica e la più 'creativa' tra le modalità di ragionamento[16]. Il ragionamento abduttivo contempla infatti un lato innovativo nella misura in cui non si limita ad applicare a nuovi casi una regola già nota (il modo di procedere della logica deduttiva) ma avanza una 'scommessa', ossia propone un'ipotesi da mettere alla prova.

Ora, con i sistemi di intelligenza artificiale, la logica abduttiva è destinata ad andare incontro a un destino simile a quello della logica deduttiva: finirà, cioè, un gradino più in basso nella considerazione della sua importanza in relazione all'idea di 'umanità' (ovvero, in relazione a ciò che distinguerebbe la menta umana dalla macchina) dato che anche la macchina è ora in grado di procedere in modo abduttivo. I sistemi di intelligenza artificiale, anche prima dell'introduzione dell'intelligenza artificiale generativa, facevano già ricorso a procedimenti abduttivi (si pensi, ad esempio, a Google Maps quando propone il percorso in auto più breve, tra i molti possibili, avanzando una 'scommessa' sulla base di un calcolo statistico ricavato dai dati del traffico in corso). Ciò che era creativo e innovativo – come la possibilità di avanzare un'ipotesi 'scommettendo' sulla sua veridicità – è oggi oggetto di calcolo statistico e, come tale, inserito in un processo di automazione. È quanto i recenti sistemi di IA generativa hanno reso ancora più evidente: anche la macchina è in grado di avanzare ipotesi generando nuovi 'casi' (ad esempio, disegnare un nuovo esemplare di essere umano o un nuovo esemplare di albero) con contenuti margini di errore statistico (che si traduce nella generazione di un 'caso' non verosimile, come accadeva ad esempio ai primi sistemi di IA grafica nel disegnare la disposizione delle dita di una mano). Proprio l'occorrenza di errori dimostra che non si tratta di un semplice procedimento deduttivo, in cui l'errore è impossibile, ma di un procedimento abduttivo, in cui l'errore è sempre possibile.

La definizione di umano e di ciò che sarebbero le abilità specificatamente umane (creative e innovative) si modifica col modificarsi delle innovazioni tecniche: nel momento in cui la macchina svolge funzioni prima svolte dall'uomo, quelle stesse funzioni cessano di definire la specificità umana e finiscono con l'essere svalutate. Creatività e innovazione si dislocano su un terreno inedito a un nuovo livello (n+1) ogni volta che gli automatismi della tecnica assorbono il livello n sottostante. Con l'esternalizzazione di alcune delle pratiche legate alla logica abduttiva assisteremo dunque a una loro atrofia nel corpo umano e a una loro simultanea 'svalutazione' nella misura in cui tali pratiche sono traducibili in automatismi e riproducibili dalla macchina. Ci riferiamo a tutto ciò che riguarda l'estrazione di una regola da uno o più contesti e la sua applicazione "creativa" a nuovi contesti. Dunque, quali pratiche intellettuali? Tutte quelle introdotte dalla tecnologia alfabetica, quali l'analisi di un testo, il confronto tra testi ed opinioni individuando analogie e differenze, l'applicazione di concetti ad altri contesti, ecc. Se questo è quanto accade a livello n, simultaneamente, a livello n+1, acquisirà più

importanza la capacità di governare questi automatismi (ossia, la capacità di "giostrarsi" tra i software e le loro inesauribili possibilità di applicazione) in modo inedito e creativo. Detto in figura, potremo finalmente dire addio alle tesi universitarie compilatorie – dato che quel tipo di lavoro è ora svolto in automatico dalla macchina – e avere solo tesi davvero innovative.

In realtà l'uso creativo e innovativo degli strumenti basati sull'IA generativa apre un campo di potenzialità del tutto inedito e di cui è ancora difficile individuare i contorni. L'unica certezza è che tale uso produrrà una risoggettivazione radicale almeno quanto lo è stata quella generata dall'introduzione delle tecniche di scrittura. Dunque, non solo nuove modalità di agire e di pensare, ma anche nuovi 'valori' e principi sulla cui base educare la soggettività umana. Ragionare non sarà più sufficiente, come già oggi non lo è il semplice dedurre (da quando esistono i calcolatori, applicare una regola nota a un nuovo caso non è più segno di particolare 'genialità' del soggetto umano). A ragionare – nel senso di analizzare e confrontare testi e opinioni – saranno uno o più software. Creativo e innovativo sarà semmai l'uso combinato da parte dell'uomo di tali software. I nuovi valori e principi su cui basare l'educazione umana si modelleranno su questi usi, così come, per più di due millenni, la pedagogia occidentale, oggi globale, è stata forgiata dalla tecnologia alfabetica.

NOTE

1. Sul rapporto tra attuale e virtuale in Deleuze, cfr. Gilles Deleuze, *Differenza e ripetizione*, Milano: Cortina, 1993

2. Cfr. Michael DePeau-Wilson, "Peer-Reviewed Journal Publishes Paper Written Almost Entirely by ChatGPT", *MedPage Today*, 03.02.2023

3. Cfr. Som Biswas, Dushyant Dobaria, Harris L. Cohen, "ChatGPT and the Future of Journal Reviews: A Feasibility Study", *Yale Journal of Biology and Medicine*, 29.09.2023, doi: 10.59249/SKDH9286.

4. Cfr. Andrea Vivaldi, "Le tesi con l'aiuto di Chat Gpt, primi casi sospetti nell'ateneo di Firenze", *La Repubblica*, 12.03.2023, https://firenze.repubblica.it/cronaca/2023/03/12/news/firenze_universita_teso_chat_gpt_allarme_sistemi_sospetti_studenti_elaborati-391651240/.

5. Cfr. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922.

6. Cfr. Michel Foucault, *Archeologia del sapere*, Milano: RCS, 1998.

7. Cfr. Charles Sanders Peirce, Collected Papers, Cambridge Mass.: Harvard University Press, 1931-1960; C. Wright, *L'evoluzione dell'autocoscienza*, Milano: Spirali, 1990.

8. Cfr. Carlo Sini, *Gli abiti, le pratiche, i saperi*, Milano: Jaca Book, 1996.

9. Cfr. Michel Foucault, *Archeologia del sapere*, cit., p. 61.

10. Cfr. Michel Foucault, *Sorvegliare e punire*, Torino: Einaudi, 2014.

11. Cfr. Priyanka Sharma, Monika Jyotiyana, A.V. Senthil Kumar, a cura di, *Applications, Challenges, and the Future of ChatGPT*, Hershey PA: IGI Global, 2024.

12. Cfr. Anil Ananthaswamy, *Why Machines Learn: The Elegant Maths Behind Modern AI*, Penguin, 2024.

13. Cfr. i tre volumi della trilogia *La Technique et le temps* ora raccolti in Bernard Stiegler, *La Technique et le temps. 1. La faute d'Épiméthée — 2. La désorientation — 3. Le temps du cinéma et la question du mal-être. Suivis de Le nouveau conflit des facultés et des fonctions dans l'Anthropocène*, Parigi: Fayard, 2018.

14. Il tema è molto vasto ed è stato affrontato in un gran numero di studi, di cui ci limitiamo a riportare qui i principali: Walter J. Ong, *Oralità e scrittura. Le tecnologie della parola*, Bologna: il Mulino, 1986; Marshall McLuhan, Dall'occhio all'orecchio, Roma: Armando, 1982; Marshall McLuhan, *La galassia Gutenberg. Nascita dell'uomo tipografico*, Roma: Armando, 1976; Eric. A. Havelock, *Dalla A alla Z. Le origini della civiltà della scrittura in Occidente*, Genova: il Melangolo, 1987; Eric A. Havelock, *La musa impara a scrivere.*

Riflessioni sull'oralità e l'alfabetismo dall'antichità al giorno d'oggi, Bari: Laterza, 1995.

15.Cfr. Carlo Sini, *Filosofia e scrittura*, Roma-Bari: Laterza, 1994.

16. Cfr. Lorenzo Magnani, *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*, Heidelberg: Springer-Verlag Berlin, 2009; John R. Josephson, and Susan G. Josephson, *Abductive Inference: Computation, Philosophy, Technology*, Cambridge: Cambridge University Press, 1995.

Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

GenGPT can understand your DNA, but can it handle your decisions?

GenGPT può interpretare il tuo DNA, ma come se la cava con la comprensione delle tue scelte?

TOMMASO ROPELATO
tommaso.ropelato@unito.it

AFFILIAZIONE
Fondazione Bruno Kessler - Centro per le Scienze Religiose (ISR)
Università degli studi di Torino

SOMMARIO

L'intelligenza artificiale (IA) sta rivoluzionando il settore medico, offrendo strumenti innovativi che potrebbero trasformare vari aspetti della sanità. Tra questi, spiccano per versatilità e popolarità i Large Language Models (LLMs) come ChatGPT. Questo articolo propone l'applicazione di una specifica variante allenata su database scientifici e specialistici, GenGPT, al settore del counseling genetico nel contesto riproduttivo. Nonostante il potenziale di questo strumento nel migliorare l'efficienza e il supporto ai pazienti emergono preoccupazioni riguardo alla loro capacità di gestire decisioni complesse e delicate, come quelle legate ai test genetici prenatali e preimpianto. Soffermandosi su due elementi spesso sottovalutati dalla letteratura tradizionale sul consenso informato, vale a dire i concetti di affidabilità e trasparenza dei valori, l'articolo conclude che, sebbene questi strumenti possono essere di supporto nella pratica medica, essi non possono sostituire completamente il giudizio umano e l'interazione personalizzata necessari in ambiti così sensibili.

PAROLE CHIAVE

Counseling genetico

Consenso informato

Trasparenza dei valori

Test genetici riproduttivi

Autonomia Riproduttiva

ABSTRACT

Artificial intelligence (AI) is revolutionizing healthcare with innovative tools that promise to transform various aspects of the field. Among these, Large Language Models (LLMs) like ChatGPT stand out for their versatility and popularity. This paper explores the application of GenGPT, a variant specifically trained on scientific and specialized databases, to genetic counseling in the reproductive context. Despite its potential to enhance efficiency and patient support, concerns arise about its ability to handle complex and sensitive decisions, such as those involving prenatal and preimplantation genetic testing. By focusing on often-overlooked elements in informed consent literature—namely, trustworthiness and value transparency—the paper concludes that while such tools can support healthcare, they cannot fully replace the human judgment and personalized interaction essential in such sensitive areas.

KEYWORDS

Genetic Counseling

Informed Consent

Value Transparency

Reproductive Genetic Testing

Reproductive Autonomy

DOI: 10.53267/20240105

1. INTRODUCTION

Artificial intelligence (AI) is rapidly transforming the medical field, introducing groundbreaking tools that are reshaping various aspects of healthcare. Among these innovations, generative large language models (LLMs) like ChatGPT are gaining increasing popularity. These models can engage in conversational exchanges and produce diverse textual content, ranging from emails and articles to computer code, sparking considerable excitement about their potential applications in the clinical setting[1]. It is possible to assume, as confirmed by ChatGPT itself[2], that they could soon be used to facilitate the documentation of patient reports, improve diagnostic accuracy[3], and assist in various clinical care[4]. However, there are also important concerns regarding hallucinations, biases[5], stereotype fabrication[6], and risks to patient privacy[7].

Yet, it is crucial to recognize that current models are not specifically designed for healthcare use. The following paper explores the potential application of a variant specifically designed for medical applications, such as the recently proposed BioGPT[8], a domain-specific generative pre-trained transformer language model for biomedical text generation and mining[9], in one of the most emerging, high-information and ethically complex field: genetic counseling in the procreative context. Indeed, one of the most promising application for LLMs in terms of streamlining healthcare resources and optimizing hospital human capital is patient support[10]. However, regarding this specific use case, as already highlighted in some papers on the use of ChatGPT in psychoterapy[11], experiments show that while ChatGPT is a good causal interpreter[12], it is not a good causal reasoner[13]. This raises significant concerns about its ability to effectively assist in making informed decisions, particularly in critical and sensitive areas such as genome-driven reproductive decision-making.

2. LLMS AND DATA-DRIVEN MEDICINE

The majority of us have now encountered ChatGPT, witnessing firsthand its remarkable ability to analyze, process, and reinterpret vast amounts of data with unprecedented speed and precision. Consequently, LLMs hold great potential within the increasingly data-driven medical field[14]. Their applications range from identifying research priorities to supporting healthcare professionals in clinical and laboratory diagnostics. Additionally, they can assist medical students, doctors, nurses, and other healthcare providers in staying updated on advancements in their fields. One of the most peculiar and critical application of LLMs in medicine could be their role in developing virtual assistants aimed at helping patients manage their health[15]. Such applications have the potential not only to offer cost-effective, scalable, and inclusive solutions but also to drive healthcare toward more personalized digital health ecosystems. These advancements are particularly crucial in clinical genomics, where the effective governance, interpretation, and communication of extensive data from genome sequencing and population-level studies pose significant challenges.

Therefore, this paper delves into the potential application of a specialized variant of ChatGPT, which we will refer to as GenGPT, trained specifically on medical specialized literature, within the domain of genetic and genomic screening and testing (GSTs)[16]. While ChatGPT has been already trialed in medical education, primarily for manuscript writing and standardized exams[17], and garnered interest for streamlining workflows, and educating patients in various specialties, including ophthalmology[18], radiology[19], rheumatology[20], and cardiology[21], assessments regarding its utility in clinical genetics remain limited. A recent study[22], published in the “American Journal of Medical Genetics”, surveyed 118 genetic counselors (GCs) in North America about the integration of ChatGPT into their profession. Among the 92 GCs who spend some of their time in a clinical role, 29.3% (27) report using it for some aspect of their work. The most commonly stated use is drafting clinical documentation including consult notes and result letters. More specifically, GCs said that ChatGPT is helpful in providing patient-friendly language suggestions, generating text for informational files, and finding support resources. Of the 35 GCs who spend some part of their time doing research, 37.1% (13) say that the most commonly use of ChatGPT in this setting is to help draft a literature review by pulling citations and references and summarizing papers. Other uses included assisting with data analysis by providing guidance on type of hypothesis testing, writing code for statistical software, and developing themes for interview codebook. Many of these participants also

use ChatGPT to write research documents such as grant applications, IRB protocols, survey questions and interview scripts. More generally, the ability to save time on administrative tasks was the most frequently reported benefit (74; 62.7%), which could help alleviate burnout, an issue exacerbated by the significant time GCs spend on non-clinical duties[23].

We seek to extend the current discussion by questioning whether a tool like GenGPT could enable GCs and clinical professionals to delegate more than just administrative tasks to AI. By identifying four levels of medical services that AI tools could potentially provide[24], namely, 'information' (e.g., using voice assistants, chatbots, and dialogue-based applications to initiate self-care guidance), 'assistance' (e.g., setting reminders for medication or self-therapy), 'assessment' (e.g., identification, detection, prediction with digital biomarkers, and management), and 'support' (prescribing, substituting, or supplementing medication and therapy tools), LLMs have the potential to significantly expand the range of virtual assistant applications toward the 'assessment' and 'support' levels. Focusing on this paradigm shift allows us to underscore several traditionally overlooked issues regarding the use of LLMs in clinical medicine. Alongside the well-known ethical and legal considerations associated with LLMs, such as avoiding biases and hallucinations, and preventing misinformation, which could potentially be mitigated by developing a model specifically pre-trained and designed for the proposed tasks, this paper addresses a central concern specific to GenGPT: its capability to establish meaningful relationships with prospective parents to support informed and consensual reproductive choices.

3. WHAT SHOULD I CONSENT TO, AND WHAT INFORMATION SHOULD GUIDE MY DECISION?

In medicine and research, consent occurs when A (who could be a patient or research participant) agrees to B (who could be a physician or researcher) performing an action on A (such as conducting a medical test). Consent is considered informed when A has been provided with relevant information and possesses sufficient decision-making capacity. Consent is deemed fully informed when a capacitated (or competent) patient or research participant, having received complete disclosures and comprehended all information disclosed, voluntarily agrees to treatment or participation[25].

Achieving this standard is challenging, if not impossible, in the context of GSTs due to the inherent complexity and ambiguity surrounding the interpretation of 'genomic results'. Additionally, accurately assessing a patient's level of understanding, satisfaction with, and perceived utility of such information presents an additional hurdle, further complicating efforts to ensure truly informed consent[26]. In the reproductive context these challenges are even greater[27] not only because genomic sequencing is carried out with reference to a subject C (the future child), but also because of the characteristics of emotional distress and responsibility that emerge from the shift to an offspring-determinant test[28]. Firstly, despite the inherent uncertainties in interpreting and communicating genomic results, the information derived from these results leads prospective parents to make significant and often binary decisions: whether to continue a pregnancy or not, or whether to implant a specific embryo. Secondly, as rapid technological advancements in repro-genetics, coupled with deterministic narratives around genomics[29], tempt prospective parents to believe they should be able to search for, understand, and operationalize the implications of each genetic variant, a new sense of parental obligation is emerging to incorporate this knowledge into reproductive decision-making[30]. Moreover, despite the possibility of selecting complex traits remains distant, the knowledge gap between specific genotypes and complex phenotypes is gradually narrowing: genome-wide association studies and AI tools, such as machine learning, are indeed advancing knowledge by providing information about what particular genes do and also how they interact to shape polygenic traits[31].

It is for this reason that the traditional informed consent model, currently inadequate for addressing the ethical and practical challenges posed by the growing integration of genetic biotechnologies in clinical settings, particularly in reproductive healthcare, requires serious scrutiny. By analyzing and reformulating the elements and principles of informed consent, we can not only emphasize the essential aspects needed to ensure ethical and effective communication between healthcare providers and prospective parents in this rapidly evolving field, but also assess whether a tool like GenGPT could be

a viable option for taking on certain tasks in genetic counseling. These include offering overviews of potential risks or outcomes associated with specific genetic variants, explaining available genetic tests and even recommending them, and providing patient-friendly explanations of complex genetic information and test results.

4. REVISING INFORMED CONSENT: MOVING BEYOND AUTONOMY

We argue that a revised and adequate model of informed consent for the reprogenetic context requires two additional principles to the traditionally recognized elements of comprehension, disclosure and voluntariness[32]: trustworthiness and value transparency.

The first one refers to the 'public and social sphere' of genomic data collecting, usage purposes, and sharing. Trustworthy systems for genomic data governance emphasize a culture of patient safety, a special attention to the instance of privacy and data allocation[33] and the making explicit of all the stakeholders involved, both direct and indirect[34]. On the other hand, value transparency principle (VTP) refers to an approach between clinical professionals and prospective parents that diverges from the traditional principle of non-directiveness (NDP), which is foundational to the practice of genetic counseling[35]. VTP involves not only fostering open dialogue between clinicians and prospective parents about the implications of genomic data but also acknowledging that the various stakeholders (clinicians, parents, and actors within broader social systems) may prioritize or interpret the importance of genomic information differently. By aiming to make these influences explicit and transparent, VTP ensures that prospective parents are fully informed about the reasoning behind which conditions are screened, including the implicit values underlying the selection of specific tests over others. There could also be strictly deontological reasons why such a principle is necessary to address the limitations of NDP, especially in the context of reproductive decision-making. Rooted in a commitment to respecting patient autonomy through non-interference, NDP may inadvertently undermine the professional responsibilities of genetic counseling by functioning as a defensive tool to protect clinicians from social criticism or litigation. This defensive posture may, in turn, fail to fully support prospective parents and, in some cases, may ultimately work against the best interests of both the parents and the future child.

At its core, VTP addresses these shortcomings by recognizing that ethical decision-making, especially in sensitive contexts like the one under examination, is rarely an isolated act of personal autonomy. As underlined by Rehmann-Sutter, while genetic data should primarily be treated as private[36], genetic knowledge and agency inherently extend beyond the individual, encompassing a social dimension in three significant ways: backward, as it reveals information about ancestors; forward, as it anticipates characteristics of future descendants; and laterally, as it affects other family members. Furthermore, genetic data is not 'raw' or neutral; it is interpreted within a context of symbols, narratives, and discourses about genes, embodiment, and identity, making it deeply meaningful and socially constructed[37].

It is important to emphasize that NDP is not inherently 'wrong' or 'bad' but is insufficient as a standalone framework. As the number of disorders screened for varies widely between countries and clinics, creating a broad constellation of 'gene-worlds'[38] in which clinicians and prospective parents face the challenge of interpreting information that is more voluminous, complex, granular, and sometimes of unknown significance, NDP must be integrated into a broader and more comprehensive ethical model. As mentioned, while respect for autonomy as a negative obligation is a critically important value in medical ethics and has a strong tradition also in the context of reproductive rights[39], we argue that it does not fully capture the moral significance of a 'meaningful informed consent'. In this model, consent is not merely a legal formality or transactional event, but rather a process aimed at promoting thoughtful and responsible decision-making, particularly within the context of emerging repro-genetic decision-making pathways. Therefore, the goal of respecting autonomy must be complemented by the goal of promoting autonomy, which involves not only providing information but also ensuring that prospective parents fully understand it, along with its short- and long-term consequences. This approach may enable a form of professional selective paternalism when the exchange of value-sensitive information reaches points of tension or conflict. In such instances, physicians may, at times,

assert their position, not as an exercise of decision-making authority or manipulative dominance, but as a form of discursive relational persuasion[40] rooted in what they believe to be in the best interests of the patient or, in this case, the future child, even when, and indeed precisely because, prospective parents are capable of making decisions themselves. By reaffirming the fundamental importance of the principle of beneficence, which is often mistakenly viewed as an alternative or surrogate for autonomy, particularly in contexts of profound existential significance, such as procreation or end-of-life care, we can, as noted by Savulescu et al.[41], restrict the label 'respect for autonomy' to refer to the negative duty of refraining from interference with autonomous choices; the element of trustworthiness we proposed is essential for fulfilling and protecting this principle. Conversely, we can adopt the term 'promotion of autonomy' to describe the positive duty to assist in decision-making, for which value transparency is a key principle. Although these two interpretations of the same principle have been recognized since the foundational work of Beauchamp and Childress[42], their practical implementation, particularly in a manner that ensures their coexistence and the harmonization of the values they represent, remains largely unexplored.

It should now be clear why a digital tool like GenGPT would be inadequate in addressing the questions we have raised. As noted by Verbeek, repro-genetic biotechnologies, by granting a form of contact with the fetus that goes beyond a mere ultrasonographic 'peek into the womb'[43], shape new interpretive frameworks in which prospective parents' agency becomes morally more relevant. Indeed, as technology expand the scope of actionability in the procreative process, effectively 'broadening biological contingency'[44], prospective parents may increasingly face clashing preferences or desires that intersect with, or even conflict with the well-being of their future child. It is therefore crucial to emphasize that such tensions cannot be disregarded when evaluating the use of AI tools designed to assist and guide decision-making in such sensitive and complex clinical contexts. As these systems become more advanced by collecting ever-increasing amounts of data and gaining deeper insights into our lives, the risk increases that they might reach 'existential' conclusions about what would be 'rational' for us 'to screen or not to screen'[45], which may significantly diverge from our genuine desires, shaped by our cognitive skills, emotions, and a priori beliefs[46].

Humans are often less rational, less consistently aware of their true desires, and less motivated to act in ways that promote their own or their future child's well-being than intelligent machines (or the engineers designing these tools) might assume. Consider, for example, the delicate yet unavoidable question of how to manage the broad spectrum of 'incidentalome'[47]. Imagine a scenario in which prenatal screening reveals an actionable incidental finding. From the perspective of a tool like GenGPT, disregarding such information might appear irrational, as it would be deemed undesirable from a purely outcome-focused standpoint. However, prospective parents, driven by fear or other complex motivations, might prefer not to know this information. More immediate and, in some ways, radical examples arise when individuals hold deep personal or ideological beliefs that make it challenging to confront certain existential scenarios, particularly in decisions surrounding the beginning and end of life. Consider a couple undergoing in vitro fertilization who face the decision of whether to use genetic screening on embryos before implantation. While emotionally invested in the hope of having a genetically healthy child, they may also feel uncomfortable selecting embryos based on genetic traits. This unease might stem from moral dilemmas or ethical concerns about the idea of 'designer babies'. GenGPT might recommend screening embryos to maximize the probability of favorable health outcomes, presenting this as the most 'rational' choice based on medical probabilities and potential health risks. However, this recommendation may overlook the couple's deeper emotional and ethical reservations, exemplifying the tension between the AI's probability-driven conclusions and the couple's value-based considerations. Similarly, a pregnant woman might be offered prenatal testing to screen for spina bifida, a condition associated with mobility challenges and potential cognitive delays. Coming from a family that has a history of overcoming physical challenges, she holds a strong belief that individuals with disabilities deserve support, dignity, and inclusion in society. While an AI system might suggest that prenatal testing is 'rational', given the potential for

medical complications and long-term care, the woman's values may lead her to reject this perspective. If the AI recommends termination upon detecting the condition, its suggestion would conflict with her moral belief in the inherent worth and dignity of all lives. Her decision to proceed with the pregnancy would be guided not by medical probabilities but by her "philosophical" stance on disability.

In such cases, the necessity of embracing the principle of value transparency becomes clear. This principle involves openly discussing and critically evaluating the (legitimate) decision of prospective parents to 'opt out' of knowing such findings, as well as their decision regarding whether or not to act upon them. It requires a careful examination of their evolving narrative experiences, acknowledging that their perspectives may shift over time. Appreciating and engaging with them is crucial because human autonomy is deeply rooted in personhood — a nuanced concept that encompasses consciousness, subjectivity, and free will. Genuine respect for autonomy, and even more so its active promotion, can only occur through interactions with another autonomous agent within a framework of mutual recognition[48].

5. CONCLUSION

Artificial intelligence, particularly generative large language models like ChatGPT, holds immense promise in revolutionizing healthcare. These models have demonstrated capabilities in processing vast amounts of data and generating content that can potentially guide clinical decision-making. However, their application, especially in sensitive areas such as genetic counseling, requires careful consideration of the complex dynamics of patient-provider relationships.

While LLMs can perform many tasks much more efficiently and rapidly than humans, processing and elaborating preferences through quantitative analysis and probabilistic reasoning, they still have significant gaps in their ability to reason causally and ethically, particularly in contexts where negotiating values and goals is crucial. As a result, these tools lack the capacity to genuinely understand, respect, and balance the intricate and evolving dynamics underlying human decisions. Without the ability to engage in authentic interpersonal recognition, LLMs fail to support a dynamic and evolving sense of autonomy, ultimately diminishing their capacity to fulfill their ethical responsibilities effectively. This limitation undermines the dual and non-mutually exclusive obligations of beneficence and autonomy. Therefore, there is a clear boundary, at least today, within which we can, and perhaps should, benefit from these tools, but beyond which we cannot venture.

NOTE

1. Lee P, Bubeck S, Petro J., "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine", *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184; Patel, Sajan B, and Kyle Lam, "ChatGPT: the future of discharge summaries?" *The Lancet. Digital Health* vol. 5,3 (2023): e107-e108. doi:10.1016/S2589-7500(23)00021-3; Ali, Stephen R et al. "Using ChatGPT to write patient clinic letters." *The Lancet. Digital Health*, vol. 5,4 (2023): e179-e181. doi:10.1016/S2589-7500(23)00048-1.

2. Patrinos, G. P. et al. "Using ChatGPT to predict the future of personalized medicine." *The pharmacogenomics journal*, vol. 23,6 (2023): 178-184. doi:10.1038/s41397-023-00316-9.

3. Hirosawa, T. et al. "Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study." *International journal of environmental research and public health*, vol. 20,4 3378. 15 Feb. 2023, doi:10.3390/ijerph20043378.

4. Grünebaum, A. et al. "The exciting potential for ChatGPT in obstetrics and gynecology." *American journal of obstetrics and gynecology*, vol. 228,6 (2023): 696-705. doi:10.1016/j.ajog.2023.03.009; Cascella, M. et al. "Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios." *Journal of medical systems*, vol. 47,1 33. 4 Mar. 2023, doi:10.1007/s10916-023-01925-4.

5. Straw, I., and Callison-Burch, C. "Artificial Intelligence in mental health and the biases of language based models." *PloS one*, vol. 15,12 e0240376. 17 Dec. 2020, doi:10.1371/journal.pone.0240376.

6. Azamfirei, R. et al. "Large language models and the perils of their hallucinations", *Critical care* (London, England) vol. 27,1 120. 21 Mar. 2023, doi:10.1186/s13054-023-04393-x.

7. Li, H. et al. "Ethics of large language models in medicine and medical research", *The Lancet. Digital Health*, vol. 5,6 (2023): e333-e335. doi:10.1016/S2589-7500(23)00083-3.

8. Peng, C. et al. "A study of generative large language model for medical research and healthcare", *NPJ digital medicine*, vol. 6,1 210. 16 Nov. 2023, doi:10.1038/s41746-023-00958-w.

9. In the specific case of BioGPT, the Transformer language model is pre-trained on 15 million PubMed abstracts.

10. Ahimaz, P. et al. "Genetic counselors' utilization of ChatGPT in professional practice: A cross-sectional study", *American journal of medical genetics*. Part A vol. 194,4 (2024): e63493. doi:10.1002/ajmg.a.63493; Blease, C., and John T. "ChatGPT and mental healthcare: balancing benefits with risks of harms", *BMJ mental health*, vol. 26,1 (2023): e300884. doi:10.1136/bmjment-2023-300884.

11. Cheng, S. et al. "The now and future of ChatGPT and GPT in psychiatry", *Psychiatry and clinical neurosciences*, vol. 77,11 (2023): 592-596. doi:10.1111/pcn.13588.

12. Emmert-Streib, F. "Can ChatGPT understand genetics?", *European journal of human genetics: EJHG*, vol. 32,4 (2024): 371-372. doi:10.1038/s41431-023-01419-4.

13. Jinglong, G., et al. "Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation". In *Findings of the Association for Computational Linguistics: EMNLP* (2023), Singapore. Association for Computational Linguistics.

14. Dave, T. et al. "ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations", *Frontiers in artificial intelligence*, vol. 6 1169595. 4 May. 2023, doi:10.3389/frai.2023.1169595.

15. Sezgin, E. "Redefining Virtual Assistants in Health Care: The Future With Large Language Models", *Journal of Medical Internet Research* (2024), doi: 10.2196/53225.

16. GSTs include both whole genome and whole exome sequencing, as well as chromosomal microarray. As the reference is to the reproductive context, GSTs also includes preimplantation genetic testing, genetic testing during pregnancy (such as amniocentesis and chorionic villus sampling), and more recent innovations such as non□invasive prenatal testing, or the use of polygenic scores. Although each of these techniques raises some peculiar problems, which will be addressed during the writing of the thesis, the research question allows them to be grouped under the same heading.

17. Giannos, P., Delardas, O. "Performance of ChatGPT on UK standardized admission tests: Insights from the BMAT, TMUA, LNAT, andTSA examinations", *JMIR Medical Education*, 9 (2023), doi.org/10.2196/47737; Prasad S. S., Manohar N. "Genital and Extragenital Lichen Sclerosus et Atrophicus: A Case Series Written Using ChatGPT", *Cureus* 15(5) (2023) doi:10.7759/cureus.38987.

18. Rojas-Carabali, W. et al. "Chatbots Vs. Human Experts: Evaluating Diagnostic Performance of Chatbots in Uveitis and the Perspectives on AI Adoption in Ophthalmology", *Ocular Immunology and Inflammation*, 32(8), pp. 1591–1598 (2023), doi: 10.1080/09273948.2023.2266730.

19. Mese, I., et al., "Improving radiology workflow using ChatGPT and artificial intelligence", *Clinical Imaging*, 103 (2023), doi:10.1016/j.clinimag.2023.109993

20. Krusche, M., et al. "Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4", *Rheumatol International*, (2024), doi.org/10.1007/s00296-023-05464-6.

21. Krittanawong, C., Rodriguez, M., Kaplin, S., & Tang, W. H. W. "Assessing the potential of ChatGPT for patient education in the cardiology clinic", *Progress in Cardiovascular Diseases* (2023), doi.org/10.1016/j.pcad.2023.10.002.

22. Ahimaz, P., et al. "Genetic counselors' utilization of ChatGPT in professional practice: A cross-sectional study", *American Journal of Medical Genetics* (2023), doi.org/10.1002/ajmg.a.63493.

23. Caleshu, C., et al. "Contributors to and consequences of burnout among clinical genetic counselors

in the United States", *Journal of Genetic Counseling*, 31, (2022), doi.org/10.1002/jgc4.1485.

24. Yang, S, et al. "Clinical Advice by Voice Assistants on Postpartum Depression: Cross-Sectional Investigation Using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana", *JMIR Mhealth and Uhealth* (2021), doi:10.2196/24045.

25. Koplin, J. J et al. "Moving from 'fully' to 'appropriately' informed consent in genomics: The PROMICE framework", *Bioethics*, vol. 36,6 (2022): 655-665. doi:10.1111/bioe.13027.

26. Roberts, J. S., et al. "Patient understanding of, satisfaction with, and perceived utility of whole-genome sequencing: findings from the MedSeq Project", *Genetics in medicine: official journal of the American College of Medical Genetics* (2018), doi.org/10.1038/gim.2017.223.

27. Shkedi-Rafid, Shiri et al. "What is the meaning of a 'genomic result' in the context of pregnancy?" European journal of human genetics: *EJHG* vol. 29,2 (2021): 225-230. doi:10.1038/s41431-020-00722-8.

28. Verbeek P.P., *Moralizing Technology: Understanding and Designing the Morality of Thing*s, University of Chicago Press, 2013; Rueda, J. "Value change, reprogenetic technologies, and the axiological underpinnings of reproductive choice", *Bioethics*, 10.1111/bioe.13287. 8 May. 2024, doi:10.1111/bioe.13287.

29. Nelkin, D. "Molecular metaphors: the gene in popular discourse." *Nature reviews. Genetics* vol. 2,7 (2001): 555-9. doi:10.1038/35080583; Harden, K. P. "Genetic determinism, essentialism and reductionism: semantic clarity for contested science", *Nature reviews. Genetics* vol. 24,3 (2023): 197-204. doi:10.1038/s41576-022-00537-x.

30. Battisti D., *Procreative Responsibility and Assisted Reproductive Technologies*, Routledge, 2024.

31. Sigala, Rafaella E et al. "Machine Learning to Advance Human Genome-Wide Association Studies", *Genes* vol. 15,1 34. 25 Dec. 2023, doi:10.3390/genes15010034.

32. Minor J., *Informed Consent in Predictive Genetic Testing*, Springer Cham, 2016.

33. Wan, Zhiyu et al. "Sociotechnical safeguards for genomic data privacy", *Nature reviews. Genetics* vol. 23,7 (2022): 429-445. doi:10.1038/s41576-022-00455-y.

34. Steven U., "Designing Genetic Engineering Technologies for Human Values", *Etica & Politica / Ethics & Politic*s XXIV/2", EUT Edizioni Università di Trieste, Trieste, 2022.

35. Schupmann, Will et al. "Re-examining the Ethics of Genetic Counselling in the Genomic Era", *Journal of bioethical inquiry* vol. 17,3 (2020): 325-335. doi:10.1007/s11673-020-09983-w.

36. Brassington, I. *The Private Life of the Genome: Genetic Information and the Right to Privacy* (2023), Routledge, London.

37. Rehmann-Sutter, C "Why Non-Directiveness is Insufficient: Ethics of Genetic Decision Making and a Model of Agency", *Medicine Studies* (1), pp. 113–129 (2009). https://doi.org/10.1007/s12376-009-0023-7.

38. Timmermans, S., Shostak, S. "Gene worlds", *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine* (2016), 20(1), doi:10.1177/1363459315615394.

39. In the landmark 1973 ruling of Roe v. Wade, to provide a readily understandable example, the U.S. Supreme Court anchored the right to terminate a pregnancy in the 'right to privacy.'

40. Rubinelli, S., "Argumentation as Rational Persuasion in Doctor-Patient Communication", *Philosophy & Rhetoric* (2013), doi.org/10.5325/philrhet.46.4.0550.

41. Koplin, Julian J et al., cit.

42. Beauchamp, T., Childress, J. *Principles of Biomedical Ethics* (1979), Oxford University Press, New York.

43. Verbeek., P. P. *Moralizing technology: understanding and designing the morality of things*, Chicago University Press (2011).

44. Habermas, J. *The future of Human Nature* (2003) Polity Press, Cambridge.

45. Petrova, D., et al. "To screen or not to screen: What factors influence complex screening decisions?", *Journal of Experimental Psychology: Applied*, 22(2), (2016) https://doi.org/10.1037/xap0000086.

46. Rueda, J. "Value change, repro-

genetic technologies, and the axiological underpinnings of reproductive choice", *Bioethics*, (2024), doi.org/10.1111/bioe.13287.

47. Roche, M. I., Berg, J. S. "Incidental Findings with Genomic Testing: Implications for Genetic Counseling Practice", *Current genetic medicine reports* (2015) 3(4), pp. 166–176, doi.org/10.1007/s40142-015-0075-9.

48. Pereira, G. "Reciprocal Recognition Autonomy as a Decentred Autonomy" (2013). In: *Elements of a Critical Theory of Justice*, Palgrave Macmillan, London.

Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

Le questioni giuridiche poste dalle invenzioni connesse all'Intelligenza Artificiale

Legal issues deriving from inventions related to Artificial Intelligence

ILARIA DE GASPERIS
ilaria.degasperis@cnr.it

AFFILIAZIONE
Consiglio Nazionale delle Ricerche,
Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca (CID-Ethics)

SOMMARIO

Il notevole sviluppo delle tecnologie connesse all'Intelligenza Artificiale (IA) e la loro suscettibilità ad essere applicate ad ogni ambito della scienza e della tecnica, sono stati accompagnati da un crescente interesse per lo sviluppo e lo sfruttamento commerciale di tali sistemi, anche attraverso i diritti di proprietà intellettuale. Si ritiene difatti che la ricerca e l'innovazione nel campo delle invenzioni ad alto contenuto tecnologico sia generalmente incoraggiata dalla protezione giuridica offerta dal brevetto, come nel caso della produzione di farmaci e dell'impiego delle biotecnologie. Secondo le teorie economiche che possono dirsi prevalenti nei sistemi giuridici contemporanei il monopolio temporaneo riconosciuto in capo all'inventore ed i conseguenti ritorni economici derivanti dal brevetto, consentirebbero di recuperare i capitali impiegati e di incoraggiare l'attività inventiva e l'innovazione a beneficio della società nel suo complesso. Con riguardo alla brevettabilità delle invenzioni correlate all'IA, si pongono due ordini di questioni riguardanti, da un lato, lo status giuridico delle invenzioni che integrano sistemi di IA e, dall'altro, l'attribuzione della paternità delle invenzioni generate autonomamente dalla macchina. Il contributo si prefigge, dunque, di esaminare gli aspetti più rilevanti del diritto dei brevetti, nonché della più recente giurisprudenza in materia di invenzioni connesse all'IA, segnalando talune criticità nel quadro giuridico vigente e possibili meccanismi correttivi.

PAROLE CHIAVE

Brevetti

Intelligenza Artificiale

Tecnologia

ABSTRACT

Technological advancements in Artificial Intelligence (AI) techniques and its potential of being applied to all domains of science and technology, have been flanked by a growing interest for its commercial exploitation worldwide, also by the means of IP rights. As a matter of fact, patents may encourage research and innovation, especially when it comes to inventions which feature highly innovative technologies, as it has been experienced in the pharmaceutical industry and biotechnology. According to economic theories prevailing in modern legal systems, temporary monopolies associated to patents would provide revenues to inventors, allowing them to recoup initial expenses and, ultimately, to spur inventiveness and innovation, to the benefit of the whole society. Inventions related to AI may pose two legal issues concerning, on the one hand, the legal status of inventions implementing AI systems and, on the other hand, the recognition of inventorship when it comes to inventions autonomously generated by a machine. The article has thus the aim of examining the most pertinent aspects of patent law with regards to AI inventions, together with recent judicial precedents on the subject matter, also underlying some gaps in the present legal framework and possible solutions.

KEYWORDS

Patents

Artifical Intelligence

Technology

DOI: 10.53267/20240106

1. INTRODUZIONE

Il notevole sviluppo delle tecnologie connesse all'Intelligenza Artificiale (IA), suscettibili di molteplici impieghi in ogni ambito della scienza e della tecnica, è stato accompagnato da un crescente interesse per l'espansione e lo sfruttamento commerciale di tali sistemi, anche attraverso il ricorso ai diritti di proprietà intellettuale[1]. Si ritiene difatti che la ricerca e l'innovazione nel campo delle invenzioni ad alto contenuto tecnologico sia generalmente incoraggiata dalla protezione giuridica offerta dal brevetto, come nel caso della produzione di farmaci e delle biotecnologie[2].

Secondo le teorie economiche che possono ritenersi prevalenti nei sistemi giuridici contemporanei, il monopolio temporaneo riconosciuto in capo all'inventore e i conseguenti ritorni economici derivanti dal brevetto consentirebbero di recuperare i capitali impiegati e di incoraggiare l'attività inventiva e l'innovazione, a beneficio della società nel suo complesso[3]. Con riguardo alla brevettabilità delle invenzioni correlate all'IA si possono individuare un duplice ordine di questioni giuridiche riguardanti, da un lato, lo *status* delle invenzioni che integrano sistemi di IA e, dall'altro, l'attribuzione della paternità delle invenzioni generate autonomamente dalla macchina.

Per quel che riguarda le invenzioni aventi a oggetto sistemi di IA, come ad esempio i *software*, queste potrebbero essere escluse dalla tutela offerta dal brevetto in quanto riconducibili a meri sistemi di calcolo o algoritmi, la cui brevettabilità è esclusa dall'art. 52 della Convenzione sul Brevetto Europeo (CBE)[4]. Più recentemente il divieto è stato in parte superato dagli organi decisori dell'Ufficio Europeo dei Brevetti, ammettendo che le invenzioni realizzate tramite *computer*, cui l'IA è assimilata, possano essere brevettate purché il loro impiego contribuisca al «carattere tecnico» dell'invenzione[5]. Inoltre, il 22 marzo 2023 il Consiglio d'Amministrazione dell'Ufficio Europeo dei Brevetti ha approvato un documento diretto a fornire principi orientativi aggiornati per l'esame delle domande di brevetto aventi ad oggetto invenzioni costituite dai sistemi di IA, in cui è stato chiarito come queste ultime non possano essere escluse *a priori* dalla brevettabilità, ma debba valutarsi caso per caso il loro contributo effettivo agli aspetti tecnici propri dell'invenzione[6]. Un approccio analogo può rinvenirsi anche nel sistema giuridico statunitense ove, per costante orientamento della giurisprudenza, è esclusa la brevettabilità di idee astratte e formule matematiche, tra cui i *software*.

Al contrario, le invenzioni che riguardano *computer* (*Computer Related Inventions-CRI*) sono brevettabili a condizione che queste presentino dei concreti avanzamenti tecnici che vadano al di là del modo in cui un computer è solito funzionare[7].

La segnalata evoluzione della giurisprudenza e della prassi consolidata degli uffici brevettuali appare del resto coerente con i principi sottesi al sistema dell'Organizzazione Mondiale del Commercio (OMC) e, in particolare, con l'art. 27 dell'Accordo sui *TRIPs* del 1994, secondo cui gli Stati contraenti devono ammettere la brevettabilità di tutte le invenzioni, in ogni ambito della scienza e della tecnica, senza operare discriminazioni in base al settore tecnologico di appartenenza[8]. Chiarita la possibilità di riconoscere, a talune condizioni, brevetti per invenzioni riguardanti sistemi di IA, si deve notare come sia invece diverso il caso delle invenzioni generate autonomamente dall'IA, quale risultato del processo di apprendimento e di conseguente elaborazione dei dati acquisiti dalla macchina a seguito di *training* e *machine learning*.

Per tale seconda ipotesi si pone la questione giuridica dell'attribuzione della paternità dell'invenzione (*inventorship*), nonché della titolarità dei diritti di privativa brevettuale. Se è pur vero che i diritti morali connessi all'invenzione e spettanti all'inventore possano essere disgiunti dalla titolarità dei diritti di privativa, come nell'ipotesi dell'invenzione realizzata nell'ambito del rapporto di lavoro, per la quale la titolarità spetta al datore di lavoro (persona fisica o giuridica), in ogni caso, la qualifica di inventore può attribuirsi solo ad una persona fisica[9]. E ciò in quanto l'IA difetterebbe, per il fatto di essere una macchina e non un essere umano, della capacità giuridica propria della persona fisica, da cui discende altresì la titolarità di diritti e di obblighi e la facoltà di disporre dei medesimi, con la conseguenza che in una domanda di brevetto non possa validamente designarsi una macchina quale inventore[10].

2. LA PATERNITÀ DELLE INVENZIONI GENERATE DALL'IA ALLA LUCE DEL CASO DABUS. LA SOLUZIONE DELL'ORDINAMENTO EUROPEO DEI BREVETTI

La questione dell'attribuzione della paternità delle invenzioni generate

dall'IA è stata recentemente oggetto di un vivo dibattito suscitato dalla intensa attività dell'imprenditore statunitense Stephen Thaler, volta a far riconoscere la qualifica di inventore in capo a un sistema di IA denominato *DABUS* (acronimo di *Device for the Autonomous Bootstrapping of Unified Science*). Secondo quanto affermato da Thaler, proprietario e gestore del sistema operativo di *DABUS*, quest'ultimo sarebbe una «macchina creativa» la cui attività principale consisterebbe nell'inventare prodotti, senza alcun ausilio dell'uomo nel concepire le ideazioni e riconoscendo la novità e l'originalità delle proprie produzioni. In particolare, *DABUS* avrebbe realizzato autonomamente due invenzioni: un contenitore per cibi in grado di riscaldarsi rapidamente e un sistema per lanciare segnali di richiesta di soccorso in caso di emergenza.

In relazione a tali prodotti Thaler ha presentato diverse domande di brevetto, dapprima presso l'Ufficio Europeo dei Brevetti e l'ufficio brevettuale del Regno Unito (*United Kingdom Intellectual Property Office-UKIPO*) e, in un secondo momento, in una pluralità di stati designati mediante una domanda di brevetto internazionale[11]. Le domande presentate da Thaler venivano tutte rigettate dagli uffici brevettuali dei singoli stati coinvolti, ad eccezione della domanda proposta nella giurisdizione sudafricana. Nel caso dell'Ufficio Europeo dei Brevetti il rifiuto è stato motivato dal fatto che l'indicazione dell'inventore deve ricadere necessariamente su una persona fisica, in quanto dotata di capacità giuridica e che l'IA non possa validamente assumere né la qualifica di inventore, né quella di «lavoratore»[12]. Tale ultima circostanza ha portato al rigetto della ulteriore domanda di Thaler all'Ufficio Europeo dei Brevetti volta a farsi attribuire i diritti brevettuali ed economici che sarebbero spettati a *DABUS*, affermando di essere il «datore di lavoro» della «macchina creativa».

Le decisioni sono state impugnate da Thaler presso la Camera d'Appello dell'Ufficio Europeo dei Brevetti, la quale le ha integralmente confermate, cosicché la vertenza si è conclusa in via definitiva con l'affermazione del principio per cui nel sistema regionale europeo ad una macchina non possano essere riconosciuti né la paternità di un'invenzione, né diritti di privativa trasmissibili a terzi[13].

3. L'APPROCCIO DEGLI ORDINAMENTI ANGLOSASSONI AL CASO DABUS: L'INSOSTITUIBILITÀ DEL "GENIO" UMANO NELL'IDEAZIONE DELL'INVENZIONE

Anche le domande di brevetto presentate da Thaler nel Regno Unito sono state rigettate dall'*UKIPO* con la motivazione che la legge britannica (*Patents Act*) non contemplerebbe, tra le sue disposizioni, la possibilità di attribuire la paternità (*inventorship*) di un'invenzione ad una macchina[14].

È interessante notare come, a seguito delle impugnazioni proposte da Thaler contro la decisione dell'*UKIPO*, la questione è stata decisa dalla Corte Suprema del Regno Unito la quale non solo ha confermato la correttezza del diniego opposto alla concessione dei brevetti, ma ha altresì enfatizzato la circostanza per cui *DABUS* avrebbe meramente «creato o generato avanzamenti tecnici", ma non avrebbe invece "inventato un'invenzione»[15]. La scelta lessicale della sentenza sembrerebbe dunque riposare sull'assunto per cui solo un inventore (*deviser*) umano, che metta a frutto la propria attività inventiva, possa dar luogo ad una invenzione brevettabile. Esito negativo hanno avuto anche le domande di brevetto per le invenzioni generate da *DABUS* presentate da Thaler presso il l'ufficio brevettuale statunitense (*Patent and Trademark Office-USPTO*)[16]. Le decisioni dell'*USPTO* sono state successivamente confermate dalla Corte d'Appello del Distretto Federale nella nota sentenza *Thaler v. Vidal*, in cui i giudici hanno chiarito che la legge dei brevetti statunitense riconosce come inventore esclusivamente le persone fisiche e che queste ultime coincidono necessariamente con gli esseri umani[17]. È interessante notare come successivamente alla pronuncia *Thaler v. Vidal* il Presidente degli Stati Uniti abbia adottato l'ordine esecutivo intitolato "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" del 30 ottobre 2023, cui hanno fatto seguito nel febbraio 2024 le linee guida dell'*USPTO* per la determinazione della paternità delle invenzioni realizzate con l'ausilio dell'IA[18].

Pur non avendo forza di legge, le richiamate linee guida mirano a chiarire come non debba escludersi la brevettabilità delle invenzioni generate dall'IA, ma che la paternità delle stesse sia attribuibile solo ad una persona fisica e a condizione che questa abbia apportato un «contributo significativo all'invenzione», non essendo invece sufficiente l'aver messo in pratica l'invenzione, oppure essere il proprietario dell'IA. Si noti come la *ratio* delle linee guida e della

giurisprudenza richiamate si fondino sul principio per cui la creatività umana espressa nell'idea inventiva (*creative spark* o *inventive genius*) svolgano un ruolo fondamentale ed insostituibile nell'attribuzione della paternità dell'invenzione, concetti questi particolarmente radicati nella concezione di 'inventore' propria dei sistemi giuridici anglosassoni[19].

4. LA GIURISPRUDENZA TEDESCA SULLA QUESTIONE DELL'INVENTORSHIP MIRA A TUTELARE I DIRITTI MORALI DELL'INVENTORE E LE ESIGENZE DELL'ECONOMIA

Giova evidenziare come l'orientamento interpretativo che nega l'*inventorship* a *DABUS* sia stato, peraltro, affermato dagli uffici nazionali dei brevetti di Taiwan, Repubblica di Corea, Nuova Zelanda, Brasile, Canada, India e Germania e che tali decisioni sono state confermate in sede giurisdizionale, a seguito di ricorsi presentati da Thaler, anche in Nuova Zelanda, Repubblica di Corea e Germania[20]. Con riguardo alla domanda di brevetto per le invenzioni di *DABUS* presentata presso l'ufficio brevetti tedesco (*Deutsches Patent und Markenamt-DPM*), la vicenda merita particolare attenzione per la peculiarità della decisione del Tribunale Federale dei Brevetti (*Bundespatentgericht*), presso cui Thaler si è rivolto a seguito del rigetto delle domande di brevetto innanzi al *DPM*[21]. Secondo tale decisione, in base alla legge dei brevetti tedesca, solo l'essere umano ha diritto all'attribuzione della paternità di un'invenzione, in quanto dotato di personalità giuridica, e ciò anche in ragione del connesso diritto morale a essere riconosciuto inventore (*Erfinderehre*)[22].

Si noti che, secondo il tribunale, dalla considerazione dei diritti morali dell'inventore discenderebbe che nel diritto tedesco non possa neanche in futuro ammettersi la paternità di un'invenzione in capo all'IA, in quanto non sussisterebbero vizi nella legge che determino la necessità di una sua modifica. Peraltro, secondo il tribunale, ciò non osterebbe alla brevettabilità della invenzione dell'IA, cosicché gli interessi economici del richiedente rimarrebbero impregiudicati. Infatti, secondo la segnalata sentenza, l'esame dei requisiti di brevettabilità di un'invenzione si porrebbe come un'operazione oggettiva, a nulla rilevando se l'invenzione sia stata realizzata in tutto o in parte mediante l'IA.

Tale circostanza non impedirebbe, quindi, di indicare nella relativa domanda di brevetto una persona come inventore, quand'anche si ritenga che l'invenzione sia frutto esclusivo dell'IA, poiché la designazione dell'inventore in una domanda di brevetto sarebbe rimessa all'interpretazione dell'istante e non dovrebbe corrispondere necessariamente a una verità fattuale, né risulta sottoposta a una verifica della propria esattezza. La pragmatica decisione del tribunale federale è stata peraltro confermata dalla Corte Suprema (*Bundesgerichtshof*) nella propria decisione dell'11 giugno 2024 in cui questa ha precisato che possa riconoscersi come inventore colui che abbia avuto una parte significativa nel successo complessivo dell'invenzione, senza entrare nel dettaglio del ruolo da quest'ultimo svolto[23]. La soluzione cui la giurisprudenza tedesca è pervenuta, nel riflettere un *favor* per la funzione economica assolta dall'istituto brevettuale, sembra dunque realizzare un buon bilanciamento degli interessi in gioco e rappresentare una possibile modalità di adattamento del diritto dei brevetti alle specificità delle invenzioni dell'AI generativa[24]. Tuttavia, si potrebbe obiettare che tale approccio pur di far salva la concezione antropocentrica del brevetto e garantire al contempo ritorni economici alle imprese, consenta che venga designato come inventore anche colui che, di fatto, non ha partecipato al processo inventivo, suscitando così dubbi sul possesso del richiamato diritto morale dell'inventore.

5. LE ALTERNE VICENDE DELLE INVENZIONI DELL'IA GENERATIVA NELLA GIURISPRUDENZA AUSTRALIANA

Nell'ordinamento australiano, anch'esso indicato da Thaler nella propria domanda di brevetto internazionale, la questione giudiziaria ha subito alterne vicende che hanno visto la giurisprudenza ritornare sui propri passi in merito alla questione dell'*inventorship*. Anche l'ufficio brevettuale australiano (*Deputy Commissioner of Patents*), una volta ricevute le domande di brevetto di Thaler, le ha rigettate per il fatto che l'IA non possa essere designata come inventore. Tuttavia, la Corte Federale adita in sede di ricorso, in un primo momento, aveva ritenuto che le conclusioni degli esaminatori fossero errate, giacché la normativa australiana non vieterebbe in modo espresso la designazione di un inventore macchina[25].

In maniera innovativa la Corte Federale ha affermato, inoltre, che il fenomeno del progressivo adattarsi della nozione di 'invenzione' alle nuove tecnologie dovrebbe parimenti accompagnarsi a un simmetrico adattamento della nozione di 'inventore' a queste ultime. Solo così sarebbe possibile evitare distorsioni del sistema che possano ostacolare la funzione propulsiva del progresso propria del brevetto, peraltro, particolarmente valorizzata dalla legge australiana. La decisione della Corte Federale ha tuttavia avuto vita breve, essendo stata annullata in grado di appello solo un anno dopo alla sua pubblicazione. In particolare, la superiore *Full Federal Court of Australia*, allineandosi con l'orientamento giurisprudenziale prevalente nel contesto internazionale, ha statuito che solo una persona fisica («natural person») possa considerarsi 'inventore', in quanto l'istituto del brevetto è preordinato a incoraggiare l'innovazione e tale finalità non possa essere rivolta a un oggetto inanimato[26]. Nonostante il netto ripensamento della giurisprudenza australiana, tuttavia, la decisione resa in primo grado è stata la prima pronuncia a riconoscere la paternità dell'invenzione dell'IA nel contesto globale, facendosi portatrice di un'inedita interpretazione evolutiva della nozione di inventore, tesa a trovare un punto di convergenza tra la legge e le esigenze emergenti in seno alla società a causa della sempre più rapida evoluzione del progresso tecnologico.

6. IL CASO ISOLATO DEL SUDAFRICA. QUESTIONI APERTE E CRITICITÀ DERIVANTI DALL'EVENTUALE RICONOSCIMENTO DELL'IA COME INVENTORE

L'unico stato designato da Thaler nella propria domanda di brevetto che abbia pacificamente ammesso la brevettabilità delle invenzioni generate da *DABUS* è il Sudafrica[27]. Non sorprende che la decisione dell'ufficio brevetti sudafricano rappresenti un caso isolato nel contesto internazionale, considerato che un'accettazione generalizzata della possibilità di attribuire la qualità di inventore all'IA presenta indubbie questioni giuridiche di complessa risoluzione. Un primo ostacolo è dato dal fatto che nei sistemi giuridici contemporanei non sussistono, ad oggi, le basi giuridiche per riconoscere la titolarità di diritti, siano essi morali o di proprietà, in capo alle macchine. Quandanche poi si volesse ammettere, per ipotesi, il riconoscimento della qualifica di inventore all'IA, ciò solleverebbe dubbi sulla compatibilità di una simile previsione con il sistema brevettuale nel suo complesso.

Si noti che ai sensi dell'art. 52(1) della CBE, l'invenzione per poter essere brevettata deve essere «nuova», implicare un'attività inventiva – intesa come «non ovvietà» o «originalità» dell'invenzione – ed avere un'applicazione industriale. In particolare, secondo l'art. 54 della CBE la novità dell'invenzione si realizza allorché questa non sia compresa nello stato della tecnica noto sino a quel momento, consistente in tutto ciò che sia stato reso pubblico prima della data della domanda di brevetto. Con ogni probabilità, la velocità con cui l'IA è in grado di produrre algoritmi e l'enorme quantità di dati che questa può immagazzinare potrebbero comportare un ampliamento smisurato dello stato della tecnica, con conseguente saturazione dello stesso, così svuotando di significato il concetto stesso di «novità» di un'invenzione.

Un ulteriore problema riguarda la valutazione del secondo requisito di brevettabilità, ossia la «non ovvietà», oggi accertata in base alle capacità dell'essere umano. Ci si potrebbe quindi chiedere se, in uno scenario in cui l'IA sia considerata «inventore», la valutazione dell'ovvietà dell'invenzione non debba essere invece parametrata sulle capacità di una macchina. In tal caso, non solo un esaminatore umano potrebbe incontrare difficoltà a immedesimarsi in ciò che sia ovvio per una macchina, ma con ogni probabilità, per l'IA un'amplissima gamma di invenzioni apparirebbe senz'altro ovvia[28]. Infatti, non vi è dubbio che l'IA abbia capacità di immagazzinare, elaborare e utilizzare dati in grandi quantità, di molto superiori rispetto agli esseri umani, con la conseguenza che questi ultimi si troverebbero in netto svantaggio.

Sotto altro profilo, il riconoscimento della qualità di inventore all'IA causerebbe un ulteriore *vulnus* al sistema brevettuale attuale, con riguardo all'obbligo dell'inventore di descrivere in maniera chiara e completa l'invenzione che intende brevettare. I procedimenti alla base dei risultati cui perviene un sistema di IA sono, infatti, randomici e per lo più oscuri (c.d. *black box effect*), con la conseguenza che sia per lo più impossibile ripercorrerne l'*iter* logico e il procedimento tecnico adottato dalla macchina e, conseguentemente, fornirne una descrizione esaustiva[29].

Un altro aspetto non trascurabile e connesso alla tematica affrontata riguarda l'ipotesi in cui l'invenzione generata da una macchina creativa violi diritti di proprietà intellettuale di terzi. In tal caso si pone il problema di individuare il soggetto responsabile dell'uso illecito poiché, in assenza di una normativa che regoli tale ipotesi, le violazioni potrebbero rimanere impunite.

7. RIFLESSIONI CONCLUSIVE

Ci si potrebbe quindi chiedere se il diritto dei brevetti sia in grado di adattare i propri principi fondanti alle invenzioni dell'IA, oppure se questo rischi di venir meno alla funzione di incentivo dell'innovazione che sin dalle sue origini lo ha caratterizzato, divenendo così sostanzialmente inutile. Si potrebbe ritenere, infatti, che un ripensamento del concetto di inventore atto ad includervi anche la macchina, non solo porti a un radicale mutamento del quadro giuridico attuale, basato su una concezione antropocentrica dei diritti di privativa, ma possa avere addirittura effetti dirompenti sul sistema brevettuale.

A ciò si potrebbe però obiettare che, se si guarda al passato, i meccanismi che presiedono al rilascio dei brevetti hanno dimostrato una significativa capacità adattiva ai progressi della scienza, come accaduto con le invenzioni biotecnologiche, per le quali è stato effettuato un adeguamento della legge esistente alle peculiarità della materia vivente. Si potrebbe quindi ipotizzare che anche per le attività generative dell'IA i sistemi giuridici contemporanei possano assolvere all'arduo compito di evolversi in funzione delle caratteristiche di tale ambito, eventualmente anche ricorrendo ad un'interpretazione che consenta di attribuire la paternità dell'invenzione generata dall'IA ad un essere umano, come il proprietario, il programmatore o l'utilizzatore.

Ciò appare coerente con la richiamata natura antropocentrica dei sistemi giuridici contemporanei e con le caratteristiche stesse dell'IA generativa. Infatti, sebbene possa rilevarsi che i metodi impiegati per il *training* dell'IA si avvalgano anche di meccanismi premiali, tuttavia, il brevetto appare idoneo a dispiegare i propri effetti incentivanti nei confronti dei soli esseri umani. Del resto, allo stato attuale dell'evoluzione della tecnologia, non sembra si possa ancora affermare che l'IA sia dotata di una autonomia analoga a quella dell'uomo[30]. Inoltre, sembra difficile affermare che un sistema di IA abbia generato in completa autonomia dei risultati brevettabili, poiché il programma è stato pur sempre avviato da esseri umani nella fase inziale e, nelle fasi successive, delle persone hanno con ogni probabilità testato il prodotto o ne hanno certificato il corretto funzionamento[31]. Alla luce delle questioni evidenziate, si rende oggi necessario individuare un quadro giuridico armonizzato di riferimento, preferibilmente attraverso la cooperazione internazionale tra Stati all'interno dell'OMC, che preveda un nucleo di principi comuni e di obblighi minimi a garanzia della certezza del diritto e dei diritti dei soggetti coinvolti. In particolare, tale quadro dovrebbe essere congegnato in modo da consentire l'agevole individuazione delle persone fisiche cui attribuire la paternità dell'invenzione, lo sfruttamento dei diritti di privativa brevettuale, nonché la responsabilità in caso di violazione di altrui diritti in connessione all'invenzione generata dall'IA.

Un nucleo di norme al riguardo consentirebbe di incentivare la ricerca e l'innovazione, spingendo gli individui a trovare nuove soluzioni in tale ambito e, al tempo stesso, tutelerebbe gli interessi della collettività. Sotto altro profilo, allo scopo di evitare che siano brevettate invenzioni che non presentino effettivamente i caratteri della novità e dell'originalità, nonché per scongiurare possibili comportamenti strategici volti a saturare segmenti di mercato, sarebbe opportuna l'adozione diffusa di linee guida per il corretto esame delle domande di brevetto di invenzioni dell'IA.

Sotto altro profilo, un quadro normativo completo dovrebbe richiedere la trasparenza delle modalità di funzionamento degli algoritmi impiegati dall'IA generativa o, qualora parte dell'invenzione, dei metodi di *machine learning* usati per giungere al risultato brevettabile. Infatti, modelli di analisi dei dati il più possibile trasparenti e affidabili consentirebbero, da un lato, di ottemperare agli obblighi descrittivi dell'invenzione e, dall'altro, a circoscrivere gli effetti negativi di algoritmi inficiati da pregiudizi (*bias*). Infine, per veicolare le *policy* governative nonché quelle aziendali, soprattutto al fine di circoscrivere le ipotesi in cui eventuali *bias* dell'IA possano operare delle discriminazioni e limitare la fruizione di fondamentali diritti, inclusa la *privacy*, sarebbe auspicabile l'istituzione diffusa di comitati etici, sia a livello governativo che societario[32].

NOTE

1. WIPO-World Intellectual Property Organization, *AI Inventions* (2023), https://www.wipo.int/export/sites/www/about-ip/en/frontier_technologies/pdf/wipo-ai-inventions-fact-sheet.pdf.

2. Henry Grabowski, "Patents, Innovation and Access to New Pharmaceuticals", *Journal of International Economic Law*, 5, no. 4 (2002): 849-860, doi: 10.1093/jiel/5.4.849.

3 Kenneth Arrow, "The Economics of Inventive Activity over Fifty Years", in *The Rate and Direction of Inventive Activity Revisited*, a cura di Josh Lerner and Scott Stern (Chicago: University of Chicago Press, 2011): 43-48; Jack Hirshleifer, "The Private and Social Value of Information and the Reward to Inventive Activity", American Economic Review 61, no. 4 (1971): 561-574.

4 Convenzione sul Brevetto Europeo (Monaco, 1973), https://www.epo.org/en/legal/epc.

5 EPO Enlarged Board of Appeal, decisione del 10 marzo 2021 (causa G 0001/19), https://www.epo.org/en/boards-of-appeal/decisions/g190001ex1.

6 EPO-European Patent Office, Common practice as regards the examination of computer-implemented inventions and artificial intelligence (2023), https://link.epo.org/web/common_practice_cii_ai_for_convergence_website_en.pdf

7 Alice Corp. v. CLS Bank Int'l, 573 U.S. 208 (2014); Bilski v. Kappos, 561 U.S. 593 (2010).

8 Accordo sugli aspetti dei diritti di proprietà intellettuale attinenti al commercio (Marrakech, 1994), https://www.wto.org/english/docs_e/legal_e/27-trips_01_e.htm.

9 UIBM-Ufficio Italiano Brevetti e Marchi, Intelligenza Artificiale e Profili di Proprietà Intellettuale. Opportunità e sfide nel settore della Proprietà Intellettuale, a fronte dello sviluppo e della rapida diffusione di sistemi di Intelligenza Artificiale (2022): 1-73, https://uibm.mise.gov.it/images/Intelligenza_Artificiale_e_Profili_di_Proprieta_Intellettuale.pdf.

10 Enrico Bonadio, Luke Mcdonagh, and Plamen Dinev, "Artificial Intelligence as Inventor: Exploring the Consequences for Patent Law", *Intellectual Property Quarterly*, 1 (2021): 48-66.

11 International Application no. PCT/IB2019/057809, publication no. WO/2020/079499 (Food container and devices and methods for attracting enhanced attention), https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2020079499.

12 Domande di brevetto europeo no. 18 275 174 e no. EP 18 275 163, https://register.epo.org.

13 EPO Legal Board of Appeal, decisioni del 21 dicembre 2021 (cause J 0008/20 e J 0009/20), https://www.epo.org/en/case-law-appeals/decisions.

14 Thaler v. Comptroller [2021] EWCA Civ. 1374.

15 Thaler v. Comptroller-General of Patents, Designs and Trademarks, [2023] UKSC 49.

16 USPTO-United States Patent and Trademark, Applications no. 16/524.350 and no. 16/524.532, https://www.uspto.gov/sites/default/files/documents/16524350_22apr2020.pdf.

17 *Thaler v. Vidal,* 43 F.4 1207 (Fed. Cir. 2022).

18 USPTO-United States Patent and Trademark, Inventorship Guidance for AI-Assisted Inventions, 10043, Federal Register 89, no. 30, 13 febbraio 2024.

19 Tim Dornis, "Artificial Intelligence and innovation: the end of patent law as we know it", *Yale Journal of Law & Technology*, 23 (2020): 119-120.

20 WIPO-World Intellectual Property Organization, Artificial Intelligence (AI) and Inventorship (2023), https://www.wipo.int/edocs/mdocs/scp/en/scp_35/scp_35_7.pdf

21 DPMA-Deutsches Patent und Markenamt, Domanda di brevetto no. 1020191281202, https://register.dpma.de/DPMAregister/pat/register?AKZ=1020191281202&CURSOR=0

22 Bundespatentgericht, 11 W (pat) 5/21, decisione dell'11 novembre 2021. Per un estratto della sentenza in inglese si veda: Filing a Patent for an AI-Generated Invention, *GRUR International* 71, no. 12 (2022): 1185-1189, doi: 10.1093/grurint/ikac119.

23 Bundesgerichtshof AZ X ZB 5/22, decisione dell'11 giugno 2024, https://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&az=X%20ZB%205/22&nr=138469.

24 Daria Kim, "The Paradox of the DABUS Judgment of the German Federal Patent Court", *GRUR International*, 71, no. 12 (2022): 1162-1166, doi: 10.1093/grurint/ikac125.

25 Thaler v. Commissioner of Patents (2021) FCA 879.

26 Commissioner of Patents v. Thaler (2022) FCAFC 62.

27 Domanda di brevetto no. ZA2021/03242.2021, *Patent Journal* 54, no. 07 (2021): 252, https://iponline.cipc.co.za/Publications/PublishedJournals/E_Journal_July%202021%20Part%202.pdf.

28 Ryan Abbott, "Everything is Obvious", *UCLA Law Review*, 66, no. 2 (2019): 1-51, doi: 10.2139/ssrn.3056915.

29 Yavar Bathaee, "The Artificial Intelligence Black Box and the failure of intent and causation", *Harvard Journal of Law & Technology,* 31, no. 2 (2018): 890-938.

30 Hui Chia, Daniel Beck, Jeannie Paterson, and Julian Savulescu, "Autonomous AI: what does autonomy mean in relation to persons or machines?", *Journal of Law, Innovation and Technology,* 15, no. 2, Taylor & Francis (2023): 390-410, doi: 10.1080/17579961.2023.2245679.

31 WIPO – *World Intellectual Property Organization, Conversation on Intellectual Property (IP) and Artificial Intelligence* (AI), Incentive structure and Inventorship for AI (2020), https://www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/call_for_comments/pdf/ind_borges.pdf

32 Reid Blackman, "Why You Need an AI Ethics Committee Expert oversight will help you safeguard your data and your brand", *Harvard Business Magazine* (2022), https://hbr.org/2022/07/why-you-need-an-ai-ethics-committee.

Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

Condotte di ricerca discutibili e irresponsabili: il ruolo di sviluppatori, annotatori e utenti di sistemi di Intelligenza Artificiale

Questionable and Irresponsible Research Practices: The Role of Developers, Annotators, and Users of Artificial Intelligence Systems

LUDOVICA MARINUCCI
ludovica.marinucci@ethics.cnr.it

AFFILIAZIONE
Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca, CNR

SOMMARIO

L'articolo si propone di analizzare alcuni casi di sistemi di Intelligenza Artificiale (IA) che hanno sollevato preoccupazioni sia nell'opinione pubblica che all'interno della comunità di ricercatori/sviluppatori di IA. Questi esempi sono in grado di mostrare le conseguenze di pratiche di ricerca 'discutibili' e 'irresponsabili' dal punto di vista non solo dei principi morali ma soprattutto degli standard professionali di ricercatori e sviluppatori. Analisi sistematiche dello stato dell'arte dei sistemi di IA, sviluppati grazie a metodologie e tecnologie molto diverse che hanno portato a output indesiderati ed esperimenti falliti, sono necessarie non solo per definire standard e codici di condotta ma anche per aumentare la consapevolezza dei ricercatori/sviluppatori di IA dei principali comportamenti irresponsabili in cui possono incorrere, anche apparentemente non gravi, comprendendone così l'impatto e le ricadute a livello individuale e sociale.

PAROLE CHIAVE

Intelligenza Artificiale

Etica della ricerca

Integrità nella ricerca

Condotte di ricerca discutibili

ABSTRACT

The article aims to analyze some cases of Artificial Intelligence (AI) systems that have raised concerns both from the public opinion and within the AI researcher/developer community. These examples are able to show the consequences of 'questionable' and 'irresponsible' research practices from the point of view not only of moral principles but above all of the professional standards of researchers and developers. Systematic analyses of the state of the art of AI systems, developed thanks to very different methodologies and technologies that have led to unwanted outputs and failed experiments, are necessary not only to define standards and codes of conduct but also to increase the awareness of AI researchers/developers of the main irresponsible behaviors they may incur, even apparently not serious, thus understanding of their impact and repercussions at individual and social level.

KEYWORDS

Artificial Intelligence

Research Ethics

Research Integrity

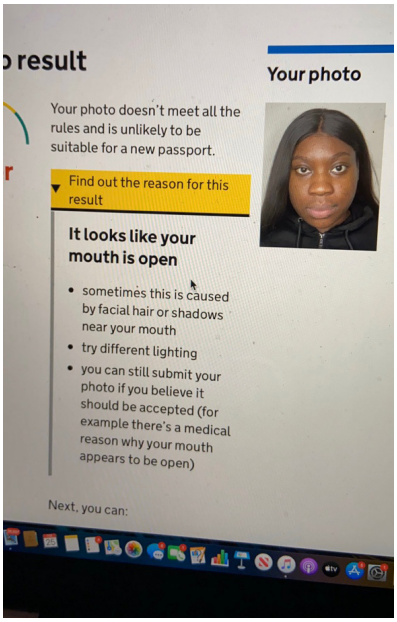
Questionable research practices

DOI: 10.53267/20240107

1. RISULTATI INDESIDERATI ED ESPERIMENTI FALLITI

L'articolo intende analizzare alcuni casi di sistemi di Intelligenza Artificiale (IA) che hanno suscitato dibattiti e preoccupazioni tanto da parte dell'opinione pubblica quanto all'interno della comunità di ricercatori/sviluppatori in ambito di IA. Tali esempi hanno lo scopo di mostrare le rilevanti conseguenze di condotte di ricerca che, seppur non definibili come *intenzionalmente* 'scorrette', come la falsificazione e il plagio, sono 'discutibili' (*questionable*) e 'irresponsabili' (*irresponsible*) dal punto di vista non solo dei principi morali (*Research Ethics*) ma soprattutto degli standard professionali (*Research Integrity*) di ricercatori e sviluppatori, secondo la nota distinzione dello storico della scienza Nicholas H. Steneck[1]. Alcuni tra gli esempi più noti sono stati riconosciuti dalla stessa comunità di ricercatori in ambito di IA, e in particolare di 'apprendimento automatico' (*machine learning*), come «an alarming red flag on our behavior as researchers and developers, since our actions can have a direct impact on society[2]». Prese di posizione di questo tipo, associate ad analisi sistematiche dello stato dell'arte di sistemi, basati su metodologie e tecnologie anche molto diverse che hanno portato a risultati (*output*) indesiderati e a esperimenti falliti, sono il primo passo verso la consapevolezza di ricercatori/sviluppatori in ambito di IA tanto dei rischi e delle implicazioni sociali quanto delle possibili ripercussioni psicologiche che tali applicativi, a volte imposti dagli stessi governi, possono avere sugli utenti.

Proprio la prospettiva dell'utente, inteso come utilizzatore passivo, è rappresentata dal caso del sistema automatico per la creazione dei passaporti britannici che nel 2020 ha erroneamente riconosciuto delle labbra prominenti, caratteristiche delle persone con origini africane, come una bocca aperta. Nel post ironico pubblicato sull'allora account Twitter (@elainebabey) dalla donna in questione, sono ancora visibili le possibili 'spiegazioni' fornite dal sistema: in particolare quelle relative alla possibile presenza di peli o di luce insufficiente sul viso inquadrato della donna mostrano come la discriminazione dell'utente sia la conseguenza di una "questionable research practice"[3], ovvero l'introduzione di 'bias' da parte dei ricercatori nelle metodologie seguite e nei risultati ottenuti e, nel caso specifico, negli strumenti informatici costruiti. Non a caso, come

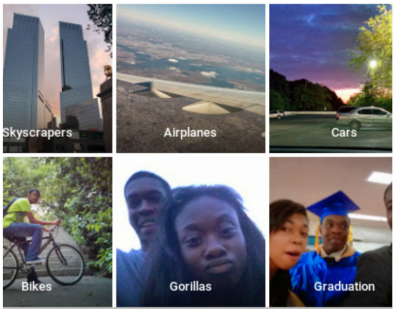


per gli studiosi della cognizione umana[4], il problema dei 'bias' è uno dei principali temi affrontati dai ricercatori di *machine learning*[5]. Infatti, dalle inclinazioni sociali presenti nei dati disponibili ai pregiudizi personali inseriti nel processo di creazione e rilascio di un sistema, ogni passaggio di una *pipeline* di apprendimento automatico è soggetta all'introduzione di una qualche forma di 'bias'. Già nella fase di selezione, raccolta e annotazione dei dati i ricercatori possono introdurre bias cognitivi, quali ad esempio il c.d. 'errore di campionamento' (*sampling bias*) oppure il c.d. 'favoritismo all'interno del gruppo' (*in-group favouritism*) correlato alla 'discriminazione all'esterno del gruppo' (*out-group discrimination*). Successivamente, le scelte algoritmiche da fare sul modello (*loss functions*, *regularization terms*, ecc.) possono ampliare qualsiasi 'bias' preesistente sui dati e, infine, il modo in cui sono analizzati e presentati i dati può essere deviato da bias interpretativi, come il c.d. 'pregiudizio di conferma' (*confirmation bias*) e molti altri. Questi errori a cui diamo la connotazione morale di 'pregiudizio', rappresentano la punta dell'iceberg di processi cognitivi, funzionali e definiti come "adaptive toolbox"[6], i quali restano sommersi e quasi inaccessibili, tanto per la cognizione umana che per quella artificiale. A quest'ultima, definita come una 'scatola nera' (*black-box*), è stata imposta dalle europee *Ethics Guidelines for Trustworty AI*[7] quella 'esplicazione' (*explainability*) utile non solo agli utenti ma soprattutto ai ricercatori/sviluppatori per elaborare strategie di mitigazione dei

'bias' (*debiasing*) già nelle prime fasi di progettazione dei sistemi di IA.

La presenza di 'bias' desta particolare clamore quando essi concorrono a creare *output* discriminatori imbarazzanti, nonché controproducenti a livello economico, per le stesse aziende che producono tali sistemi a fini commerciali. Come rilevato da vari studi in ambito di 'visione artificiale' (*computer vision*), i modelli commerciali di riconoscimento facciale hanno una notevole diminuzione delle prestazioni con soggetti con la pelle più scura, soprattutto se donne[8]. Tra i primi cattivi esempi figura *Google Photo* che nel 2015 ha associato l'etichetta 'gorilla' (*gorillas*) alla foto di due persone dalla pelle scura[9]. Nonostante le scuse pubbliche, la Big Tech americana non ha provveduto davvero a risolvere il problema a livello tecnologico, optando per l'eliminazione dell'etichetta 'gorilla' e simili ('scimpanzé', 'scimmia', ecc.) anche in relazione alle immagini di quei primati.

L'esempio dimostra come algoritmi



e dispositivi, anche di uso quotidiano, hanno il potenziale di diffondere e rafforzare stereotipi dannosi. Tali pregiudizi espongono alcune categorie di persone, e in particolare le donne di colore, al rischio di essere lasciate indietro nella vita economica, politica e sociale. Infatti, gli algoritmi non solo forniscono consigli sui film e prodotti da acquistare, ma sono anche sempre più utilizzati per prendere decisioni ad alto rischio, ad esempio nelle valutazioni dei pazienti[10], delle domande di prestito bancario[11], dei candidati da assumere[12] e persino delle probabilità di recidiva di un imputato[13], mostrando *output* discriminatori basati sull'etnia e sul genere. Ad oggi, nonostante i dibattiti su 'pregiudizi' ed 'equità' (*fairness*) nei sistemi di apprendimento automatico[14], i numerosi tentativi di "debiasing" sia tramite *post-processing*[15] sia direttamente durante il *training*[16], nonché gli sforzi per la creazione di grandi dataset rappresentativi di diverse etnie[17], la situazione non sembra cambiata. L'avvento dell'IA generativa, basata su grandi modelli linguistici (*Large Language Models*), solleva le stesse preoccupazioni circa il perpetuarsi di 'pregiudizi' sistemici incorporati nei dati di *pre-training*, come dimostra uno studio che esplora il potenziale pregiudizio etnico e di genere di ChatGPT: alla richiesta di valutazione di CV fittizi di candidati arabi, asiatici, afro e centroafricani, europei, americani e sudamericani le risposte discriminatorie del *chatbot* si basano ancora su un meccanismo statistico che riecheggia stereotipi sociali[18] di cui gli utenti, di qualunque tipo, devono essere resi consapevoli.

Tuttavia, prima ancora che un problema etico, si tratta qui di aspetti che hanno a che fare con l'adesione a certi standard professionali da parte di ricercatori in ambito informatico, i quali implicano strategie volte a correggere pratiche scorrette nella progettazione e valutazione dei sistemi. Visti da questa prospettiva, infatti, gli innumerevoli tentativi di mitigazione e prevenzione dei 'bias' da parte della comunità dei ricercatori/sviluppatori che utilizza tecniche di *machine learning* sembrano consistere in buone pratiche (*best practices*) riconducibili a nient'altro che 'condotte di ricerca responsabili'. Ad esempio, nella fase di *pre-training* del sistema, per garantire la diversità dei dati è consigliato selezionare e combinare *input* da più fonti di dati; mentre per ottenere un'annotazione accurata bisognerebbe non solo prevedere un team diverso rispetto a chi ha selezionato i dati ma anche ricorrere ad esperti esterni per rivedere il lavoro svolto[19]. Vale la pena sottolineare che la fase di annotazione dei dati è forse la più delicata e critica nell'orientare gli *output* di sistemi basati sull'apprendimento automatico, data la forte componente di soggettività umana nell'attività richiesta al c.d. 'annotatore', attore fondamentale del processo di cui il ricercatore/sviluppatore deve tenere conto, oltre all'utente. Nel corso degli anni, il ruolo di annotatori e revisori di annotazioni precedenti si è reso così fondamentale da ideare piattaforme, come *Amazon Mechanical Turk*, le quali erogando un compenso minimo[20], considerabile come sfruttamento in molti paesi occidentali, sollevano rilevanti criticità etiche relative al loro utilizzo all'interno di progetti di ricerca e sviluppo. A livello teorico, inoltre, uno dei problemi di ricerca più complessi e rilevanti, come testimoniano anni di tentativi per trovare metriche di 'accordo' tra annotazioni diverse[21], sistemi di va-

lutazione di annotazioni multiple[22] e approcci iterativi che assegnano al campione solo le annotazioni unanimi[23], è quello di raggiungere un 'consenso' tra annotatori indipendenti che elimini l'eccessiva soggettività delle annotazioni, molto evidente nel caso di concetti astratti come emozioni, sentimenti e valori[24]. A tal fine, le principali *best practices* per i ricercatori consistono sia nel definire linee guida chiare per gli annotatori sia nel creare come *gold standard* un dataset annotato in maniera ottimale per le finalità dell'applicativo.

L'assenza di un'annotazione ben bilanciata nella fase di *pre-training* del sistema è, ad avviso di chi scrive, una delle varie 'pratiche irresponsabili' che hanno portato al caso di Tay, *chatbot* basato sull'IA che fu lanciato da Microsoft sull'allora Twitter nel 2016. Dopo meno di 24 ore, Microsoft dovette terminare l'esperimento perché @TayandYou aveva iniziato a generare post inappropriati con un linguaggio giudicato razzista, sessista e antisemita.

Il ruolo attivo degli utenti è stato determinante, trattandosi di un *learning software* (LS), che dopo una prima fase di addestramento continua a imparare fino ad arrivare a cambiare il suo programma in risposta alle interazioni con utenti anche mediate e amplificate, come in questo caso, da un *social network*. Perciò, le dichiarazioni di Microsoft secondo cui «within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways[25]», non sono state sufficienti ad attenuare le critiche sull'assenza di consapevolezza e comprensione dei rischi e possibili danni che tali tipi di tecnologie possono comportare. Gli sviluppatori di Microsoft dovevano prevedere l'alta probabilità dell'esito dell'esperimento, dato che un «LS always has this sort of vulnerability, and therefore, a developer of LS should adopt a position of expecting this behavior. The developer cannot be confident about knowing how the system will behave because of the nature of software that learns. [...] LS developers need to be more keenly aware of their ethical responsibilities[26]». Quindi, proprio per la loro peculiare natura, lo sviluppo dei LS richiede un'adesione a principi di etica e integrità nella ricerca, quali affidabilità, responsabilità e diligenza[27], anche maggiore rispetto ad altri *software* standard. Il caso di Tay mostra una vera e propria 'condotta di ricerca irresponsabile' (*irresponsible research practice*) da parte dei ricercatori/sviluppatori di Microsoft che hanno sottovalutato l'importanza della fase di progettazione e pre-addestramento del sistema per mitigare i rischi della fase sperimentale di apprendimento aperto con utenti sconosciuti e anonimi. Inoltre, il team coinvolto nell'esperimento online avrebbe dovuto monitorare più diligentemente l'evoluzione delle risposte offensive del *chatbot* in modo tale da nasconderle al pubblico. Invece, non è chiaro se la rimozione di Tay da parte di Microsoft sia stata una reazione ai tweet offensivi del chatbot o una reazione all'indignazione degli utenti verso di essi. Nel secondo caso, Microsoft non solo non avrebbe previsto questa possibilità in anticipo, ma avrebbe anche sottovalutato il problema nascente lasciando attiva e libera Tay il più a lungo possibile, ovvero fino a quando la pressione mediatica ha reso evidente la necessità di terminare l'esperimento. Qualunque sia la verità, concordiamo sul fatto che questo caso è capace di evidenziare nodi cruciali per la definizione di «appropriate professional best practice for internal processes when "releasing" LS to the general public[28]».

2. IL NECESSARO LEGAME TRA FORMAZIONE, RICERCA E INNOVAZIONE

Questa analisi preliminare di specifiche tecnologie utilizzate per lo sviluppo di sistemi di IA, così come dei diversi attori coinvolti, è stato un primo passaggio finalizzato a evidenziare e specificare alcuni aspetti rilevanti per la definizione di norme e codici di condotta utili a ricercatori/sviluppatori in ambito informatico, che devono essere emanati da parte dei loro datori di lavoro (quali università e istituti di ricerca) e, con una visione più ampia, dai governi nazionali e sovranazionali, eventualmente anche aggiornando attuali linee guida di etica e integrità nella ricerca[29]. A tal fine, si rende necessaria da parte di gruppi multidisciplinari

di esperti una ricognizione e analisi sistematiche degli aspetti fondamentali del processo di progettazione, valutazione e distribuzione di sistemi di IA relativi tanto al tipo di tecnologie utilizzate per implementare i sistemi (*supervised machine learning*, *reinforcement learning*, ecc.) quanto al ruolo peculiare degli attori coinvolti (utenti attivi, annotatori, ecc.). Questo approccio richiede un coordinamento tra la comunità scientifica, i decisori politici e le altre parti interessate (*stakeholders*) che deve riflettersi non solo nella definizione di codici e linee guida ma anche nei bandi di finanziamento e nella valutazione dei risultati dei progetti. In tale contesto, gli stessi risultati derivanti da fondi pubblici, ad esempio dataset di qualità ad accesso aperto, devono essere diffusi e condivisi attraverso le infrastrutture di ricerca esistenti a livello nazionale ed europeo che favoriscano approcci collaborativi e di riuso di tali risultati seguendo standard di integrità nella ricerca. Tali standard mirano a salvaguardare la ricerca da distorsioni dovute a interessi economici e politici che, come abbiamo visto, sono particolarmente evidenti nei sistemi di IA sviluppati a uso commerciale da grandi aziende private. In particolare, gli esempi sopramenzionati evidenziano la necessità tanto di una maggiore attenzione e tutela verso annotatori e utenti quanto di un maggior controllo su piattaforme e strumenti collaborativi utilizzabili nelle varie fasi di addestramento e valutazione di sistemi di IA sviluppati nell'ambito di progetti di ricerca.

Un requisito fondamentale all'uso da parte dei ricercatori/sviluppatori di tali strumenti collaborativi dovrebbe essere l'erogazione di una formazione preliminare relativa non solo alle attività tecniche da svolgere ma anche alle responsabilità, all'impatto e ai rischi specifici del sistema di IA che si sta collaborando a implementare. In ambito di IA, infatti, tutti gli attori coinvolti nelle varie fasi di sviluppo dovrebbero essere formati su principi, criteri e pratiche per una condotta di ricerca responsabile, la quale «is simply conducting research in ways that fulfill the professional responsibilities of researchers, as defined by their professional organizations, the institutions for which they work and, when relevant, the government and public[30]». Tale definizione si basa sull'assunto secondo cui la ricerca scientifica vada considerata come un'attività professionale condotta da persone che hanno ricevuto una formazione specifica. Tuttavia, se possiamo dare per scontata la formazione scientifica, teorica e pratica, dei ricercatori sui contenuti peculiari del loro ambito di ricerca (nel caso specifico, quello dell'IA), in continuo aggiornamento sui più recenti sviluppi tecnologici, sembra meno evidente un'adeguata conoscenza dei principali comportamenti scorretti, anche apparentemente poco gravi, e comprensione del loro impatto e ripercussioni sociali. Una 'formazione culturale e civile' su principi morali e codici professionali che guidino ricercatori, annotatori, utenti, ecc., non solo su *cosa* dovrebbero fare o meno, ma anche su *come* dovrebbero farlo, eviterebbe buona parte degli esempi di 'risultati indesiderati' qui considerati. Il dibattito e l'attenzione mediatica suscitati, e che continuano ciclicamente a suscitare, reclamano la pressante necessità di una sensibilizzazione volta a far percepire tali questioni non come meri adempimenti burocratici per pubblicazioni o fondi di ricerca, ma come requisiti essenziali per ottenere risultati che puntino a costruire una "società buona"[31] o, almeno, a evitare danni tanto individuali quanto sociali e pubblici, soprattutto in termini di investimenti sprecati o di un indebolimento della fiducia nei riguardi dei sistemi di IA sviluppati e, per estensione, dei loro ricercatori/sviluppatori.

NOTE

1. N. H. Steneck, "Fostering integrity in research: Definitions, current knowledge, and future directions", *Science and engineering ethics* 12 (2006), 53-74, https://doi.org/10.1007/PL00022268

2. C. Laranjeira, V. Fernandes Mota, and J. A. dos Santos, "Machine Learning Bias in Computer Vision: Why do I have to care?", in *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images*, IEEE (2021), p. 1.

3. N. H. Steneck, 2021, *cit*.

4. A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," in *Rationality in action: Contemporary approaches*, ed. P. K. Moser (Cambridge University Press, 1990), 171–188 [Reprinted from *Science* 185 (1974), 1124-31]

5. N. Mehrabi et al., "A survey on bias and fairness in machine learning," *ACM computing surveys - CSUR*

54, no. 6 (2021), 1-35, https://doi.org/10.1145/3457607

6. G. Gigerenzer and H. Brighton, "Homo heuristicus: Why biased minds make better inferences," *Topics in cognitive science* 1, no. 1 (2009), 107-143, https://doi.org/10.1111/j.1756-8765.2008.01006.x

7. HLEG - High-Level Expert Group on AI, "Ethics Guidelines for Trustworthy AI," (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

8. J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in P*roceedings of the 1st Conference on Fairness, Accountability and Transparency*, vol. 81, (New York, NY, USA: PMLR, 23–24 Feb 2018), 77–91, http://proceedings.mlr.press/v81/buolamwini18a.html

9. BBC News, "Google apologises for Photos app's racist blunder," July 1, 2015, https://www.bbc.com/news/technology-33347866

10. Z. Obermeyer, B. Powers, C. Vogeli et al., "Dissecting racial bias in an algorithm used to manage the health of populations," *Science* 366, no. 6464 (2019), 447–453, https://www.science.org/doi/10.1126/science.aax2342

11. A. Mukerjee, R. Biswas, K Deb et al., "Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management," *International Transactions in operational research* 9, no. 5 (2002), 583-597, https://doi.org/10.1111/1475-3995.00375

12. A. Peng, B. Nushi, E. Kıcıman, et al., "What you see is what you get? the impact of representation criteria on human bias in hiring," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7 (2019), 125-134, https://doi.org/10.1609/hcomp.v7i1.5281

13. J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances* 4, no. 1 (2018), eaao5580, https://www.science.org/doi/full/10.1126/sciadv.aao5580

14. N. Mehrabi et al., 2021, *cit.*

15. T. Bolukbasi, K.W. Chang, J.Y. Zou et al., "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in NeurIPS Proceedings of *Advances in neural information processing systems* 29 (2016), 4349–4357, https://bit.ly/4h2gX87

16. J. Zhao, T. Wang, M. Yatskar et al., "Gender bias in contextualized word embeddings," (2019), https://doi.org/10.48550/arXiv.1904.0331

17. H. J. Ryu, M. Mitchell, H. Adam, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," in *Proceedings of Workshop on Fairness, Accountability, and Transparency in Machine Learning,* (FAT/ML 2018), https://www.fatml.org/media/documents/inclusive_facenet_zOOhwRN.pdf

18. L. Lippens, "Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach," *Computers in Human Behavior: Artificial Humans* 2, no. 1 (2024), 100054, https://doi.org/10.1016/j.chbah.2024.100054

19. B. Cowgill, F. Dell'Acqua, S. Deng et al., "Biased programmers? or biased data? a field experiment in operationalizing ai ethics," in ACM Conference on Economics and Computation, 2020, pp. 679-681. http://dx.doi.org/10.2139/ssrn.3615404

20. K. Fort, G. Adda and K.B. Cohen, "Amazon Mechanical Turk: Gold mine or coal mine?," *Computational Linguistics* 37, no. 2 (2011), 413–420, https://doi.org/10.1162/COLI_a_00057

21. M. Fuoli and C. Hommerberg, "Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions," *Corpora* 10, no. 3 (2015), 315-349, https://doi.org/10.3366/cor.2015.0080

22. F. Rodrigues, F. Pereira and B. Ribeiro, "Learning from multiple annotators: distinguishing good from random labelers," *Pattern Recognition Letters* 34, no. 12 (2013), 1428–1436, https://doi.org/10.1016/j.patrec.2013.05.012

23. D. M. Iraola and A. J. Yepes, "Single versus multiple annotation for named entity recognition of mutations" (2021), https://doi.org/10.48550/arXiv.2101.07450

24. J. Hoover, G. Portillo-Wightman, L. Yeh et al., "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological*

and Personality Science 11, no. 8 (2020), 1057-1071, https://doi.org/10.1177/1948550619876629

25. P. Lee (2016), "Learning from Tay's introduction," Official Microsoft Blog, https://bit.ly/49Jqj5X

26. M.J. Wolf, K.W. Miller, F.S. Grodzinsky, "Why we should have seen that coming: comments on Microsoft's Tay "experiment, and wider implications," *ACM SIGCAS Computers and Society* 47, no. 3 (2017), p. 3, https://doi.org/10.29297/orbit.v1i2.49

27. Commissione per l'Etica e l'Integrità nella Ricerca del CNR, "Linee guida per l'integrità nella ricerca," 2019, https://bit.ly/3P2B7CU

28. M.J. Wolf et al., 2017, *cit.*, p. 6.

29. ALLEA, "The European Code of Conduct for Research Integrity," Revised Edition 2023, Berlin, DOI 10.26356/ECOC, https://bit.ly/4ilkCcF

30. N. H. Steneck, 2021, *cit.*, p. 55.

31. L. Floridi, J. Cowls, M. Beltrametti, et al., "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations," *Minds and machines* 28 (2018), 689–707, https://doi.org/10.1007/s11023-018-9482-5

Genomica e discriminazione

Genomics and discrimination

GIULIA PERRI
perri.giulia16@gmail.com

AFFILIAZIONE
Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca (CID-Ethics)

SOMMARIO

A partire dal 1990 – quando ebbe inizio uno dei maggiori progetti di ricerca nell'ambito della scienza moderna: il Progetto Genoma Umano – si svilupparono strumenti sempre più all'avanguardia per decodificare il corredo genetico degli individui, identificando i tratti del DNA alla base di molte malattie e i geni che le provocano. Tali strumenti, noti come test genetici predittivi, hanno creato una nuova categoria di soggetti potenzialmente a rischio di discriminazione, i c.d. *unpatients*, individui 'sani' ma con una particolare suscettibilità a specifiche sostanze presenti nei luoghi di lavoro idonee, a causa del contatto prolungato, a determinare l'insorgere di talune malattie. Tali informazioni – potenzialmente utili in ottica di prevenzione – celano possibili abusi dei datori di lavoro, che potrebbero essere tentati da un utilizzo distorto di esse, quale la mancata assunzione di un soggetto con una suscettibilità o predisposizione nota.

PAROLE CHIAVE

Genetica e diritto

Discriminazione genetica

Discriminazione sul luogo di lavoro

ABSTRACT

Since 1990 – when the Human Genome Project, one of the major research projects in modern science, has begun – scientists have developed new tools apt to decode the genetic makeup of individuals, identifying DNA traits that cause many diseases and the genes that trigger their development. These tools – known as predictive genetic tests – created a whole new category of subjects potentially discriminated, the so called 'unpatients'. The unpatients are healthy individuals who show a particular susceptibility to certain chemicals present in the workplace, with a high probability of contracting a disease in the event of prolonged exposure to them. Such information – potentially useful from a health prevention perspective – may lead to possible abuses by employers, who, for instance, could reject candidates based on the knowledge of their disease susceptibility or predisposition.

KEYWORDS

Genetics and law

Genetic Discrimination

Workplace Discrimination

DOI: 10.53267/20240108

1. PREMESSE

L'osservazione dei numerosi ambienti dell'agire umano lascia intravedere un radicamento profondo delle pratiche discriminatorie, declinate nei modi più diversi e insidianti plurimi diritti: la dignità, l'uguaglianza, la salute, il lavoro, finanche la vita. Nonostante, infatti, la presenza di discipline antidiscriminatorie e apparati di norme posti a tutela di questi diritti, persistono e continuano a essere diffuse differenze di trattamento ingiustificate. Oltre alle forme di discriminazione più risalenti, quali la differenza di genere, di "razza" o religiosa e fattori di più recente emersione, come la discriminazione basata sull'orientamento sessuale, l'età o la disabilità, aleggiano nuove possibili forme di disparità di trattamento, emerse a seguito dei progressi scientifici sul genoma umano.

A partire dagli anni Settanta del secolo scorso gli avanzamenti dell'ingegneria genetica hanno via via alimentato dibattiti scientifici e biopolitici, con particolare riferimento all'impatto delle biotecnologie in vari ambiti e ai limiti da doversi eventualmente stabilire rispetto alle applicazioni pratiche in diversi settori: agricoltura, industria, ricerca scientifica, salute umana. Con l'avvento di CRISPR-Cas9, l'*editing* genetico ha raggiunto standard di efficienza e precisione sempre maggiori, uniti a costi piuttosto contenuti. In conseguenza di ciò, la concretezza dell'odierna possibilità di intervenire sul genoma in modo efficace rende necessaria la discussione di questioni che in precedenza erano relegate alla dimensione della speculazione teorica. L'elevata capacità predittiva di alcuni test genetici, nello specifico, permetterebbe non solo la diagnosi precoce di molte malattie di derivazione genetica, ma anche la conoscenza della suscettibilità individuale ad alcune sostanze che potrebbero concorrere allo sviluppo di una certa patologia.

I soggetti che a seguito dei suddetti test riscontrassero una particolare predisposizione ad una patologia genetica si ritroverebbero in una situazione di indefinitezza: «neither patients in the usual sense of being under treatment, nor nonpatients, in the sense of being free from a medically relevant condition»[1]. Una categoria i cui soggetti, denominati per l'appunto 'unpatients'[2], non sono dunque individuabili a priori, ma potrebbero divenire bersaglio di trattamenti discriminatori realizzati da chi utilizzasse in maniera distorta informazioni genetiche sul loro conto, quali, ad esempio, quelle relative ad informazioni sul loro stato di salute, anche se futuro e – oltretutto – solo potenziale. In questo breve contributo, dopo un inquadramento sintetico delle peculiarità che contraddistinguono i dati genetici, si tratteggeranno i contorni della discriminazione genetica, calandola in un contesto specifico: quello lavoristico. Difatti, nel settore in questione si rinvengono già forme di regolamentazione con le quali si sono cimentati alcuni legislatori nazionali (tra i quali, per quanto qui di interesse, statunitense, tedesco, francese e italiano), il che vale a rendere quello del diritto del lavoro un ambito sul quale portare l'attenzione.

2. LA SPECIALITÀ DEI DATI GENETICI

Come affermato nel primo articolo della "Dichiarazione Universale sul Genoma Umano e i Diritti Umani" adottata l'11 novembre 1997, il genoma umano è, in senso simbolico, «patrimonio dell'umanità». Successivamente alla scoperta nel 1953 della struttura a doppia elica del DNA umano da parte di Francis Crick e James Watson, una tappa fondamentale per una vera conoscenza della stessa fu proprio il sequenziamento del genoma umano, che avvenne in seno allo Human Genome Project. Il Progetto, che ebbe inizio nel 1990 negli Stati Uniti, fu una delle più importanti iniziative scientifiche condotte sul piano internazionale, finalizzato proprio alla comprensione dei processi genetici che conducono allo sviluppo di alcune malattie. I primi studi in tema di genetica furono condotti, quindi, in ambito medico[3], con la conseguenza pratica di un iniziale e spontaneo accostamento della categoria dei dati genetici con quella, più ampia, dei dati sanitari.

Le caratteristiche peculiari dei primi, tuttavia, hanno ben presto portato parte della dottrina – innanzitutto statunitense – ad optare per la teoria identificata con il nome di 'genetic exceptionalism' (eccezionalismo genetico), consistente nell'attribuzione di una tutela rafforzata e speciale ai dati genetici a cagione della loro unicità. I dati genetici, difatti, non soltanto hanno un contenuto immutabile, ma anche un carattere predittivo: forniscono un volume notevole di informazioni sullo stato di salute futuro di un individuo e dei suoi consanguinei, caratteristica questa che ha portato alcuni autori a parlare di 'natura condivisa' dei dati genetici. Dalle peculiarità appena menzionate discen-

dono numerosi problemi, tra i quali possono menzionarsi senza pretesa di completezza: la possibile lesione della privacy e della dignità conseguente ad un utilizzo disfunzionale di tali strumenti; la necessarietà di una tutela della 'identità genetica'; la lesione del 'diritto a non sapere', sancito dall'art. 10.2 della Convenzione sui Diritti dell'Uomo e la biomedicina del 1997; l'utilità di una appropriata consulenza genetica; il pericolo di manipolazioni del DNA per fini diversi da quelli scientifici; la discriminazione su base genetica, in particolare in ambito lavorativo e assicurativo e il rischio di discriminazioni multiple – quali le discriminazioni etnico-genetiche[4] – indirizzate nei confronti di gruppi con maggiori suscettibilità agli agenti patogeni.

3. LA DISCRIMINAZIONE SU BASE GENETICA

Come accennato, un utilizzo distorto delle conoscenze e delle applicazioni che derivano dai progressi nel settore della genetica umana potrebbe dar luogo a nuovi strumenti di violazione dei diritti umani e delle libertà fondamentali. Alcune fonti internazionali hanno, tuttavia, da tempo sancito in vario modo il divieto di discriminazione sulla base di tale fattore: l'art. 11 della Convenzione di Oviedo del 1997, che vieta «qualsiasi forma di discriminazione nei confronti di una persona a causa del suo patrimonio genetico», è stato pochi anni dopo seguito dall'art. 21 della Carta di Nizza, il quale accoglie il divieto «di qualsiasi discriminazione fondata, in particolare, […] sulle caratteristiche genetiche». Quantunque nel panorama nazionale ed internazionale difetti una puntuale definizione di 'discriminazione genetica'[5], gli scritti in materia sottolineano l'intrinseca differenza che corre tra questa tipologia e quelle che concernono discriminazioni perpetrate sulla base di fattori più risalenti nel tempo.

In tema, un ambito relativamente trascurato in Italia sia dalla produzione normativa sia dalla discussione accademica è quello lavoristico. In tale settore – come anche in quello assicurativo – un utilizzo disfunzionale dei test genetici (e in particolare di quelli relativi alle suscettibilità personali su base genetica) potrebbe dare luogo a discriminazioni basate sul proprio genoma. All'interno di un contesto lavorativo aziendale, infatti, si comprende bene l'interesse nutrito dal datore di lavoro rispetto al regolare funzionamento della sua unità produttiva. Il tutto ben potendosi sostanziarsi non solo in uno scontato controllo dei costi che egli deve affrontare relativamente ai beni economici afferenti alla sua unità produttiva (come i macchinari) e alle operazioni (quali ad esempio le vendite e gli acquisti effettuati), ma anche in monitoraggi sulla genetica delle persone che forniscono la propria manodopera per svolgere l'attività: i dipendenti. La discriminazione genetica sul lavoro, conseguentemente, potrà verificarsi in diversi ambiti: nella fase preassuntiva, al fine di selezionare candidati geneticamente più "desiderabili"; nel corso dello svolgimento del rapporto di lavoro, al fine di ricollocare il lavoratore, ad esempio demansionandolo, utilizzando come pretesto i suoi dati genetici; come motivo di licenziamento, sorretto da anomalie in tali dati; infine, in ambito di previdenza sociale, campo in cui le preoccupazioni sono in qualche caso analoghe o sovrapponibili a quelle che invadono la sfera assicurativa. Sempre nel settore lavoristico si potrebbe discutere della possibilità di rimodellare le stesse nozioni di 'idoneità psico-fisica' e 'disabilità' alla luce della rilevanza dei dati genetici, ricomprendendo al loro interno stati di salute 'futuri', connessi dunque all'incidenza di fattori che potrebbero presentarsi (forse) successivamente.

Ci si potrebbe interrogare, oltretutto, sull'incidenza dei test genetici sul generale obbligo in capo all'imprenditore di tutelare l'integrità fisica e la personalità morale dei prestatori di lavoro, tramite l'adozione delle misure che «secondo la particolarità del lavoro, l'esperienza e la tecnica» si rendano necessarie a tale scopo. La standardizzazione degli impianti e le procedure di certificazione difatti, potrebbero talvolta non essere protettive per tutti allo stesso modo, potendo ben verificarsi, alla luce di quanto detto finora, la situazione per cui a seconda di differenti caratteristiche genetiche – o anche per una sola caratteristica genetica – un soggetto finisca con l'essere preferito in alcuni settori, o per lo svolgimento di alcune mansioni, e discriminato in altri[6].

Il settore mostra quindi fin da subito tutta la sua fragilità e i rischi che possono emergere a seguito dell'utilizzo distorto dei dati genetici dei lavoratori. Come già rilevava il Comitato Nazionale per la Bioetica nel 1994, agli albori del Progetto Genoma Umano, sarebbe stato necessario scongiurare l'eventualità che si creassero gruppi di cittadini, già deboli perché portatori di condizioni sanitarie difficili, «resi ancora più deboli perché discriminati nella vita sociale e lavorativa»[7]. Allo stato attuale, mentre

alcuni Paesi, come gli Stati Uniti e la Germania, si sono dotati di trame normative che vietano di discriminare in base al genotipo nei settori più a rischio, altri, come la Francia, hanno preferito disciplinare la questione attraverso il rinvio ad una pluralità di fonti.

4. PROSPETTIVE COMPARATE

Benché si incontrino diversi approcci mirati alla risoluzione della discriminazione su base genetica – tra chi cerca di prevenire stigmatizzazioni attraverso un approccio 'a monte', teso a contenere la possibilità di ottenere informazioni ricavabili dai dati genetici e chi, al contrario, ha approntato 'a valle' una disciplina tesa a vietare la discriminazione basata su tali informazioni – l'atteggiamento generale può tuttora definirsi come di chiusura e diffidenza nei confronti di tali strumenti. In assenza di tutele legali, il timore di divenire vittima di discriminazione su base genetica ha tutto il potenziale per scoraggiare gli individui a sottoporsi a test genetici con finalità preventive e/o di ricerca scientifica, con una conseguente perdita di opportunità importanti nel progresso degli interventi mirati a ridurre o prevenire i disturbi genetici.

Nel nostro Paese la materia risulta tuttora affidata alle Autorizzazioni Generali del Garante della Privacy nn. 8 e 1 del 2016, secondo cui un test genetico può essere lecitamente utilizzato – anche senza il consenso dell'interessato – nel caso in cui sia dimostrato il suo carattere indispensabile ai fini della protezione della vita o dell'integrità fisica dell'interessato, e vi sia un nesso tra il profilo genetico e lo specifico lavoro da svolgere. Inforcando le lenti della comparazione, lo accennavamo, si ha riscontro di come alcuni Stati abbiano, invece, già da tempo provveduto a dotarsi di una disciplina in materia. È il caso degli Stati Uniti, in cui da anni è sviluppata la pratica di sottoporre i lavoratori a test genetici[8], e in cui è stato emanato nel 2008 il GINA – Genetic Information Non-Discrimination Act. Il suo principale scopo è di prevedere una regolazione di tipo preventivo, finalizzata certamente a scongiurare la discriminazione genetica nei settori lavorativo e assicurativo, ma, anche e soprattutto, a rassicurare gli individui circa l'utilizzo dei test genetici, in tal modo intendendo contenere i timori degli abusi e al contempo non rallentare il ricorso alle opportunità di beneficiare dei vantaggi della ricerca genetica. L'esigenza di disciplinare in vario modo il tema è stata avvertita anche al di qua dell'Atlantico. Nello stesso torno di anni in cui veniva adottato il GINA, per un verso il *Gendiagnostikgesetz* in Germania concentrava l'attenzione sull'utilizzo del materiale genetico e dei dati ricavati dalla sua analisi prevalentemente in ambito assicurativo e lavorativo, per un altro i codici penale e civile francesi vietavano la possibilità di utilizzare i risultati di un test genetico predittivo o relativo alla predisposizione ad una patologia, ove questa non si fosse ancora manifestata nel soggetto.

5. CONCLUSIONI

La concreta possibilità di conoscere e utilizzare i dati genetici ha aperto una breccia per lo sfruttamento degli stessi in vari ambiti. La pericolosità della discriminazione su base genetica è evidente e deriva, in aggiunta ai fattori precedentemente menzionati, pure dalla imprecisione che ancora oggi caratterizza alcuni aspetti di questi accertamenti. Si noti, invero, che la sola predisposizione individuale verso uno stato patologico potrebbe anche non tradursi mai nella manifestazione dello stesso, poiché a ben vedere si fonda su dati meramente probabilistici[9]. Ciononostante, licenziamenti predicati sulla base di questi ultimi determinano svantaggi attuali.

In secondo luogo, il dato della 'natura condivisa' delle informazioni genetiche potrebbe comportare una discriminazione intergenerazionale, con un pregiudizio tramandato in maniera automatica – e automatizzata, con molta probabilità – alle generazioni future, stigmatizzando tutta la discendenza di un individuo. A seguito dei primi casi emersi nel settore occupazionale e assicurativo negli Stati Uniti alla fine degli anni '70, molti Paesi si sono dotati di leggi mirate a prevenire o contrastare la discriminazione genetica, auspicando che simili soluzioni potessero infondere nuova fiducia all'utilizzo dei test genetici, dimostratisi sino a quel punto fonte di innumerevoli preoccupazioni.

Gli utilizzi in positivo di questi ultimi potrebbero tuttavia consentire importanti operazioni di prevenzione dalle malattie genetiche, finanche quelle aventi una concausa in agenti chimici o sostanze presenti sui luoghi di lavoro. Registrando le informazioni e rendendole disponibili al datore, ci si potrebbe oltretutto muovere nel senso di una 'personalizzazione' e profilazione del lavoratore, in un'ottica di realizzazione di interventi di natura preventiva, modellati sulla base

di ogni contesto, e personalizzati in base alle esigenze specifiche di ogni individuo, con una collocazione del lavoratore nel posto di lavoro più adatto al suo stato di salute.

L'incessante progredire degli studi in materia mette, tuttavia, costantemente in discussione gli strumenti giuridici tradizionali, innescando un ampio dibattito interdisciplinare su quali dovrebbero essere i principi idonei ad indirizzare il legislatore in materia. L'atteggiamento generale, come detto, è nel segno della chiusura nei confronti di tali tecnologie, e di apparente disinteresse nei confronti delle notevoli potenzialità di tali strumenti. Una chiusura, crediamo, provocata dalla consapevolezza della difficoltà presenti nella gestione di un gran numero di dati di cui il datore di lavoro si troverebbe in possesso e della pericolosità di una fuga degli stessi, che potrebbe essere ancora più improvvisa e incontrollabile a causa dei nuovi strumenti di Intelligenza Artificiale utilizzati, tra gli altri settori, anche in quello lavoristico (si pensi, ad esempio, ai dati inerenti alla gestione algoritmica dei lavoratori su piattaforma). Le costanti preoccupazioni in materia e i dibattiti relativamente all'etica e l'opportunità di fare ricorso ai test genetici hanno portato alla creazione, nel 2018, del *Genetic Discrimination Observatory*, un consorzio internazionale di ricercatori riuniti per svolgere attività di ricerca sulla discriminazione su base genetica e fornire risposte e soluzioni al problema. Con tali premesse, un atteggiamento lungimirante sarebbe quello di approfondire maggiormente quali potrebbero essere le esternalità positive e il miglior approccio di una legge sulla non discriminazione genetica sul lavoro, indagare quale potrebbe essere la relazione tra la maggiore fruizione dei test – effettuati in ambiti occupazionali o 'fai da te' – e la possibile diminuzione del rischio di mortalità legata a malattie lavoro-correlate e individuare i limiti da tenere saldi.

Se è vero, infatti, che diagnosticare prima una malattia incurabile potrebbe tramutarsi in una anticipazione del dolore e una anticipazione inutile di sofferenze[10], è altrettanto vero che per alcune di esse una soluzione potrebbe essere quella di evitare o ritardarne l'insorgenza, non permettendo l'avverarsi di quel concorso di cause genetiche e ambientali che la provocano. La auspicabile salubrità dei luoghi di lavoro, le tecniche preventive che si potrebbero approntare – anche tramite una diffusione dell'utilizzo di test genetici – e la protezione dei dati scaturenti da essi sono dunque solo alcuni dei temi che emergono in questa discussione. Il fatto, oltretutto, che ogni individuo possa presentare sequenze genetiche imperfette, espone la quasi totalità degli individui a trovarsi vittime di questo tipo di discriminazione, motivo per cui sarebbe desiderabile l'elaborazione di proposte di regolazione su più livelli: nazionale, europeo ed internazionale.

NOTE

1. Albert R. Jonsen, Sharon Jonsen Durfy, Wylie Burke, Arno G. Motulsky, "The advent of the "unpatients", *Nature medicine*, 2, no. 6 (1996): 623.

2. *Ibidem* 622–624.

3. Debora Provolo, *L'identità genetica nella tutela penale della* privacy *e contro la discriminazione*, Padova: Padova University Press, 2018, 20 ss.

4. Cfr. Carlo Casonato, "La discriminazione genetica: una nuova frontiera nei diritti dell'uomo?", in *I diritti fondamentali in Europa*, *XV Colloquio biennale (Messina-Taormina, 31 maggio-2 giugno 2001)*, a cura di AA.VV., Milano: Giuffrè, 2002, 644.

5. Su cui si rimanda, in generale, a Carlo Casonato, 2002, cit., 641 ss.

6. Anna Trojsi, "'Biodiritto del lavoro' e tutela antidiscriminatoria: i dati genetici del lavoratore," in *Il biodiritto e i suoi confini: definizioni, dialoghi, interazioni*, a cura di C. Casonato, Lucia Busatta, Simone Penasa, Cinzia Piciocchi, Maria Tomasi, Trento: Università degli Studi di Trento Editrice, 2014, 499.

7. Comitato Nazionale per la Bioetica, Progetto Genoma Umano, 18 marzo 1994, 24.

8. Anna Trojsi, 2014, cit., 500.

9. Carlo Casonato, "Diritto, diritti ed eugenetica. Prime considerazioni su un discorso giuridico altamente problematico", *Humanitas: rivista mensile di cultura*, 3, no. 4 (2004): 841-856.

10. Comitato Nazionale Per La Bioetica, Orientamenti bioetici per i test genetici, 19 novembre 1999, 20.

Documenti di etica e bioetica

Documenti di Etica e Bioetica

Intelligenze Future.

La ricerca scientifica nell'era dell'intelligenza artificiale

COMITATO ETICO FONDAZIONE UMBERTO VERONESI[1]

SOMMARIO

- L'intelligenza artificiale (IA) offre immense opportunità per la ricerca scientifica e dunque per il progresso umano;
- L'uso dell'IA sta già trasformando la scienza a ogni livello, tanto che oggi si parla di 'AI science', e cioè di un modo in parte nuovo di pensare e fare scienza nel quale l'uso massivo dell'IA è integrato 'by design' in tutte le fasi del processo di ricerca, dalla generazione di nuove ipotesi fino alla comunicazione dei risultati;
- Tuttavia, le potenzialità della *AI Science* sono limitate da alcuni fattori che impediscono di coglierne i benefici, tra cui (i) la qualità dei dati e i *bias* negli algoritmi; (ii) la compartimentalizzazione dei dataset e la dipendenza crescente da algoritmi, prodotti e applicativi proprietari che limitano accesso e la riproducibilità degli studi; (iii) il fabbisogno energetico e infrastrutturale; (iii) l'aumento di frodi e condotte scorrette nella ricerca e la conseguente crisi del paradigma di produzione e valutazione della conoscenza scientifica; (iv) la crescente distanza tra la complessità insita nell'AIS e la comprensione di tale fenomeno da parte della popolazione e dei decisori politici; (v) un deficit di pensiero rispetto ai cambiamenti in atto e alle loro implicazioni per il futuro;
- Tutti questi fattori richiedono un approccio sistemico per essere eliminati o mitigati: per governare il futuro della ricerca nell'era dell'*AI Science* è importante non solo investire in ricerca e formazione, ma anche sostenere un approccio più aperto alla scienza (adottando i principi della cosiddetta *open science*), predisporre processi di *governance* adeguata, nonché promuovere l'importanza dell'integrità nella ricerca presso tutti coloro che partecipano oggi al processo di ricerca;
- Visti i benefici che possono derivarne, l'integrazione tra intelligenza umana e artificiale per fini di ricerca scientifica rappresenta un obiettivo da promuovere e potenziare, invece che da temere e limitare, al fine di inaugurare una nuova era di scoperte e di progresso i cui benefici devono però essere equamente condivisi a vantaggio di tutti.

1. INTRODUZIONE

Il termine 'intelligenza artificiale' (IA) fu introdotto da John McCarthy in una conferenza nel 1956 presso il Dartmouth College. L'evento fu organizzato da McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon per studiare la possibilità di creare macchine intelligenti sulla base dell'ipotesi «che ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza possa, in linea di principio, essere descritto con tanta precisione da permettere di simulare tali caratteristiche in una macchina». Dal 1956 in poi, il termine 'intelligenza artificiale' è stato quindi adottato per indicare un ampio campo di ricerca e applicazioni ingegneristiche il cui denominatore comune è il tentativo di creare macchine o programmi capaci di risolvere problemi complessi e raggiungere obiettivi in modo automatico.

Dopo quasi settant'anni, la visione di McCarthy e colleghi si è realizzata in molti dei suoi aspetti fondamentali. Di conseguenza, l'intelligenza artificiale è oggi al centro di ogni dibattito che riguarda il futuro, con prospettive e valutazioni spesso tra loro diametralmente opposte. A un estremo c'è chi ritiene che l'IA costituisca un pericoloso 'rischio esistenziale', e cioè una delle possibili cause che potrebbero portare all'estinzione dell'umanità o comprometterne il futuro in modo irreversibile. All'altro estremo, invece, c'è chi ritiene invece che l'IA rappresenti uno strumento di straordinario progresso non solo tecnico e conoscitivo ma anche sociale, culturale e umano, destinato a segnare in positivo il presente e il futuro della nostra specie.

Il presente documento è più vicino alla seconda posizione, nella convinzione che l'intelligenza artificiale al servizio e in associazione all'intelligenza umana abbia le potenzialità per aiutare l'umanità a risolvere alcune tra le sue sfide più urgenti, illuminare problemi complessi riguardo alla natura di fenomeni ancora poco compresi – come la realtà fisica o la coscienza –, nonché migliorare in modo concreto le condizioni di vita delle generazioni presenti e future.

Allo stesso tempo, l'*AI science* solleva una serie di domande profonde e interrogativi a livello teorico, etico e politico che spaziano dalla qualità dei dati e dei risultati prodot-

ti fino alle nuove sfide per l'integrità nella ricerca. In tale contesto, il presente parere intende offrire una mappatura delle principali questioni connesse all'uso dell'IA per fini di ricerca, con un particolare riferimento alle implicazioni di queste tecnologie per le scienze biomediche e della vita. Come altri pareri del Comitato Etico di Fondazione Veronesi, nell'ultima sezione il documento avanza una serie di raccomandazioni utili per tutti i soggetti che partecipano al processo di ricerca scientifica (ricercatori e ricercatrici *in primis*, ma anche agenzie, istituzioni, società scientifiche e riviste), nonché per i decisori politici e la cittadinanza.

2. L'INTELLIGENZA ARTIFICIALE 1956-2024

La storia dell'IA è relativamente recente e può essere suddivisa in quattro fasi principali. La prima, iniziata a metà degli anni '50, vide la creazione dei primi programmi in grado di risolvere problemi algebrici e dimostrare teoremi matematici. Questi programmi seguivano un processo logico deduttivo ed erano, di conseguenza, perfettamente deterministici. Una delle applicazioni più famose di questo primo tipo fu 'Logic Theorist', un programma scritto nel 1956 da Allen Newell, Herbert A. Simon e Cliff Shaw, che fu capace di dimostrare automaticamente 38 dei 52 teoremi del secondo libro dei *Principia Mathematica*. Dati questi successi, negli anni '60 si diffuse un grande ottimismo sul futuro dell'IA. Tuttavia, tali speranze si scontrarono con due limiti allora invalicabili: la presenza di problemi non risolvibili tramite la sola logica formale e la mancanza di una potenza di calcolo sufficiente. Questo portò a una prima fase di stagnazione nella ricerca e una generalizzata perdita di interesse per l'IA e le sue applicazioni.

L'interesse si riaccese solo agli inizi degli anni '80, quando furono creati i primi 'expert system', e cioè dei sistemi capaci di estrapolare alcune regole di base tratte dalla conoscenza di esperti per applicarli poi a problemi reali. Uno dei più famosi fu MYCIN, un sistema creato da Edward Shortliffe e colleghi presso l'Università di Stanford con lo scopo di identificare i batteri responsabili di infezioni gravi, come la meningite, e di raccomandare antibiotici calibrandone la dose a seconda del peso del paziente. Uno studio dimostrò che i piani di cura raccomandati da MYCIN furono giudicati migliori da una giuria di medici esperti rispetto a quelli proposti da altri membri della facoltà di medicina. Nonostante i risultati incoraggianti, MYCIN non fu però mai utilizzato nella pratica. Oltre alle critiche avanzate rispetto sull'uso di queste tecnologie in medicina, questo fu dovuto, di nuovo, soprattutto allo stato della tecnologia allora disponibile. Come altri 'expert system' sviluppati negli stessi anni , anche MYCIN presentava diversi limiti dovuti alla mancanza di flessibilità, *performance* limitate e costi di gestione molto alti. Negli stessi anni – tra il 1982 e il 1992–, anche il progetto finanziato dal governo giapponese per creare un super-computer (chiamato 'Fifth Generation Computer Systems') si rivelò un fallimento e venne definitivamente sospeso, portando a un nuovo calo di interesse nei confronti dell'IA.

Questo secondo 'inverno dell'IA' durò fino al 2006, quando anziché usare processi computazionali legati alla logica tradizionale di tipo deduttivo si sono iniziati a sperimentare processi computazionali di tipo induttivo, non più strettamente deterministici ma bensì probabilistici. In tali processi l'errore, pur limitato, può sempre avvenire; in modo simile a come i bambini imparano a riconoscere un gatto, anziché un cane, vedendo l'animale di casa, per tentativi e correzioni progressivi. In particolare, il gruppo di ricerca guidato da Geoffrey Hinton propose un nuovo approccio basato sul 'deep learning' (DP), e cioè su reti neurali artificiali profonde (*deep neural network*) insieme a un metodo per evitare alcuni dei loro limiti tradizionali (come il problema del 'gradient vanishing' durante l'addestramento). Il DP è una forma di 'machine learning' (ML) che utilizza reti di 'neuroni artificiali' ed è particolarmente adatto per compiti complessi come il riconoscimento di immagini e di suoni. In particolare, gli algoritmi di DP riescono a identificare delle regolarità (e cioè, dei *pattern*) a partire da un insieme di dati (ad esempio, un insieme di immagini di strade) per fare predizioni (ad esempio, riconoscere se un'immagine contiene o no segnali stradali) o per offrire altri risultati.

In pochi anni questo nuovo tipo di IA ha consentito di compiere progressi straordinari, portando alle prime applicazioni concrete sotto forma di programmi capaci di superare gli esseri umani nell'eseguire una serie di compiti molto complessi come il gioco (ad esempio, negli scacchi, il quiz *Jeopardy!* e il Go), oppure il riconoscimento di immagini (ad esempio, nel campo della diagnostica medica, ma non solo). Nello stesso periodo, gli algoritmi alla base di questi nuovi strumenti di IA basati su ML e DP, uni-

tamente a una sempre maggiore disponibilità di dati e potenza di calcolo, hanno portato anche alla commercializzazione dei primi assistenti vocali rivolti ai consumatori, tra cui Amazon Alexa e Microsoft Cortana.

L'ultima grande rivoluzione nel campo dell'IA è avvenuta intorno al 2020 con la cosiddetta 'Intelligenza Artificiale Generativa'. Questa tipologia di IA permette di generare nuovi testi, audio, video o immagini in risposta alle richieste (dette *prompt*) degli utenti utilizzando tecniche avanzate di ML come i GANs. I GANs (*General Adversarial Networks*) sono tecniche di ML capaci di produrre *output* "sintetici" (come, ad esempio, immagini completamente "inventate" a cui non corrisponde niente di reale, inclusi i cosiddetti *deepfake*).[1] Il risultato sono testi, immagini, suoni e video indistinguibili da quelli creati da esseri umani. L'uso del linguaggio naturale ha portato queste tecnologie ad essere sempre più accessibili, tanto che per utilizzarle non è richiesta infatti alcuna competenza di programmazione né alcuna conoscenza informatica, facilitandone così enormemente l'uso e la diffusione. Esistono molti tipi di AI generativa basati su architetture diverse. Tra queste, uno dei più conosciuti è ChatGPT, uno strumento sviluppato a dalla società privata OpenAI e basato su un'architettura (*Generative*, *Pre-trained* e *Transformative*) particolarmente adatta a processare il linguaggio naturale. Oltre a ChatGPT, sono stati rilasciati molti altri strumenti e applicazioni capaci di produrre *output* in vari formati quali file audio, immagini e video, oppure di integrare tra loro tali formati. In generale, nel resto del presente documento, il termine 'intelligenza artificiale' sarà utilizzato in senso esteso per comprendere tecnologie di vario tipo, da sistemi basati sull'apprendimento automatico tramite ML fino alle ultime applicazioni dell'IA generativa.

3. DALLA SCIENZA ALL'*AI SCIENCE*: ALCUNI ESEMPI

La convergenza tra la disponibilità di *dataset* sempre più grandi, nuove tecniche algoritmiche di ML e maggiore potenza di calcolo ha reso l'IA uno strumento sempre più integrato e oramai indispensabile per la ricerca scientifica. Negli ultimi anni l'uso dell'IA per fini di ricerca è infatti cresciuto in modo esponenziale, tanto da portare a coniare il termine 'Artificial Intelligence Science' (abbreviato, AIS) per sottolineare la differenza tra i nuovi metodi di ricerca e scoperta che utilizzano in modo massiccio e integrato strumenti di l'AI con approccio 'by design' e i metodi tradizionali che hanno invece caratterizzato la ricerca scientifica. In particolare, nel presente documento il termine *AI Science* sarà utilizzato per indicare l'unione tra l'intelligenza umana e artificiale per fini di ricerca, scoperta, e innovazione, laddove il fine ultimo rimane di perseguire il bene comune attraverso l'applicazione sistematica del metodo scientifico.

Recentemente il sodalizio tra ricerca scientifica e IA è stato certificato dei premi Nobel 2024. il premio Nobel 2024 per la chimica è stato infatti assegnato a David Baker, Demis Hassabis e John Jumper per aver scoperto un metodo algoritmico che consente di predire la struttura funzionale delle proteine. mentre il premio 2024 per fisica è stato assegnato a John Hopfield e Geoffrey Hinton per aver contribuito allo sviluppo del *machine learning* attraverso le reti neurali. Le scelte della Fondazione Nobel sono significative, perché testimoniano non solo che l'IA è oramai radicata nella ricerca scientifica, ma anche il grado di ibridazione tra i saperi che essa ha reso possibile grazie all'utilizzo di tecniche simili ma in grado di rivoluzionare campi tra loro diversi, dalla fisica alla chimica, dalla teoria computazionale dei giochi fino alla biologia.

L'*AI science* promette avanzamenti conoscitivi straordinari. In fisica, astronomia, biologia e nella ricerca biomedica, l'AIS consente già di ridurre ricerche che in passato avrebbero richiesto decenni al lavoro di poche ore svolto in modo automatico, mentre in altri permette di pianificare ricerche ed esperimenti altrimenti impossibili con le tecniche tradizionali. L'*AI Science* può quindi aiutare l'umanità ad affrontare meglio le grandi sfide che la attendono, dalla prevenzione delle pandemie al cambiamento climatico. In prospettiva, i benefici derivanti da un uso responsabile dell'AI applicata alla ricerca in termini di progressi e miglioramento delle qualità di vita per i singoli e la collettività potrebbero essere incalcolabili.

Per comprendere l'impatto che l'IA sta avendo sulla ricerca scientifica è utile analizzare alcuni casi di studio rappresentativi. Il primo, già menzionato, riguarda l'utilizzo di algoritmi di IA per predire la struttura funzionale delle proteine e, più recentemente, di ogni altra molecola legata a processi biologici. Dalla scoperta del DNA in poi uno dei problemi centrali per la biologia ha riguardato la comprensione del processo attraverso il quale le proteine assumono la propria forma

funzionale a partire da una sequenza di amminoacidi. Per ricostruire la struttura di una sola proteina i metodi sperimentali tradizionali, come la cristallografia e la crio-microscopia elettronica, possono richiedere anni di ricerca. Nel 2020, il campo di ricerca della biologia strutturale è però stato rivoluzionato da AlphaFold, un sistema di IA sviluppato dall'azienda DeepMind, ora acquisita da Google. In pochi anni di sviluppo AlphaFold ha raggiunto un alto grado di precisione nella previsione delle strutture proteiche, vincendo nel 2020 la più importante competizione annuale, e cioè la *Critical Assessment of Structure Prediction* (CASP). Dopo questo successo, i progressi compiuti dalle successive versioni di AlphaFold sono stati molto rapidi e significativi. Oggi AlphaFold è in grado di predire con grande precisione la struttura di un numero incredibile di proteine, riducendo a poche ore processi di ricerca che avrebbero invece richiesto decenni, interi istituti di ricerca e centinaia di ricercatori con una preparazione tecnica avanzata. Nel 2024, DeepMind ha rilasciato l'ultima versione di AlpahFold, la quale è in grado di predire la struttura non solo delle proteine, ma anche di altre macromolecole, tra cui il DNA e l'RNA. Le implicazioni di AlphaFold sono rivoluzionarie per la biologia, nonché per la medicina, per nuove applicazioni industriali, per le biotecnologie e per settori specifici come l'agricoltura. AlphaFold e i programmi ad esso affini rappresentano un cambiamento di paradigma nella biologia computazionale, a dimostrazione di come l'IA sia in già grado di risolvere problemi fondamentali e complessi che fino a pochi anni fa erano considerati insolubili, o per la loro complessità o per la quantità di risorse richieste.

Un secondo caso esemplare riguarda l'uso dell'IA per la ricerca in fisica teorica e sperimentale. Gli algoritmi di ML possono essere utilizzati per esaminare in modo efficiente le enormi quantità di dati generati dagli esperimenti nella fisica delle alte energie per identificare nuove particelle o fenomeni fisici – come, ad esempio, quelli condotti presso il *Large Hadron Collider* (LHC). Già nel 2012, la scoperta del 'bosone di Higgs' era stata ottenuta tramite l'utilizzo di algoritmi di *machine learning* per analizzare i dati. Inoltre, l'IA può aiutare a risolvere complesse equazioni in fisica teorica, formulare predizioni sul comportamento dei fenomeni quantistici, o per esplorare le proprietà dei materiali quantistici e le potenziali applicazioni nell'informatica quantistica. In astrofisica l'uso dell'IA è sempre più indispensabile per analizzare i dati dei telescopi e altri strumenti, identificare esopianeti e mappare la struttura dell'universo. I modelli di DL possono infatti essere utilizzati per rilevare i segnali sottili degli esopianeti nelle curve di luce delle stelle, accelerando notevolmente il processo di scoperta. Queste applicazioni illustrano il potenziale dell'IA per migliorare sia la fisica teorica che quella sperimentale, portando a nuove scoperte e, dunque, a una comprensione più profonda dell'universo.

Altro caso di rilievo è poi quello che riguarda l'applicazione dell'IA al patrimonio culturale, come nel caso dei 'papiri di Ercolano'. Questi testi fanno parte di una collezione di rotoli ritrovati in una biblioteca di Ercolano che nel 79 d.C. fu seppellita dalle ceneri vulcaniche dopo l'eruzione del Vesuvio. Essendo carbonizzati questi rotoli sono incredibilmente fragili. I metodi tradizionali per srotolare e leggere i papiri rischiano spesso di distruggere questi delicati testi il cui contenuto è rimasto, per questa ragione, finora in larga parte sconosciuto. Recentemente, l'IA ha però fornito una soluzione rivoluzionaria a questo problema. I ricercatori hanno sviluppato algoritmi di ML per analizzare immagini a raggi X ad alta risoluzione dei rotoli. Questi algoritmi possono rilevare le sottili differenze di densità tra l'inchiostro e il papiro, permettendo di 'srotolare virtualmente' i rotoli e rivelare i testi senza doverli aprire fisicamente. Utilizzando una combinazione tra ML e tecniche di *imaging* avanzate, nel 2024 un gruppo di ricerca è riuscito a leggere parti dei papiri di Ercolano, scoprendo opere finora sconosciute di filosofia antica, probabilmente appartenenti alla scuola epicurea. Questo metodo di lettura indiretta preserva l'integrità fisica dei rotoli e potrebbe essere esteso allo studio di altri manoscritti antichi fragili. L'applicazione dell'IA nella decifrazione dei papiri di Ercolano evidenzia la natura interdisciplinare della ricerca IA, nella quale competenze di informatica, fisica e studi classici si combinano per ottenere risultati conoscitivi altrimenti irraggiungibili.

L'*IA Science* sta rivoluzionando anche il campo della genomica e dell'*editing* genetico tramite CRISPR, una tecnica particolarmente precisa ed efficace cui il Comitato Etico ha già dedicato un parere. Ad esempio, gli algoritmi di IA possono aiutare a prevedere i siti di taglio di CRISPR con maggiore precisione, riducendo il rischio di effetti fuori bersaglio (*off-target*) e aumentando la sicurezza delle terapie genetiche. L'uso di questi al-

goritmi può accelerare la ricerca e facilitare lo sviluppo di nuove terapie per malattie genetiche, offrendo speranza a milioni di persone affette da disturbi ereditari o rari. Oltre alle terapie geniche e avanzate, l'uso dell'AI sta però già cambiando anche il modello tradizionale di scoperta dei farmaci. Esempi di queste nuove applicazioni sono piattaforme come Atomwise, le quali utilizzano modelli predittivi di AI per identificare molecole candidate per nuovi farmaci. La promessa di questi strumenti è duplice: accelerare la scoperta di nuovi farmaci e ridurre i costi degli esperimenti preclinici. Inoltre, l'IA può essere utilizzata anche per riposizionare farmaci esistenti per nuove indicazioni terapeutiche, sfruttando database di informazioni farmaceutiche e cliniche per identificare nuove potenziali applicazioni per farmaci già approvati, migliorando così l'efficacia e la sicurezza dei trattamenti. Infine, l'IA viene utilizzata per sviluppare strumenti di diagnosi predittiva che possono identificare precocemente potenziali problemi di salute, consentendo interventi più tempestivi e personalizzati. Ad esempio, recentemente, gruppi di ricerca universitari (MIT negli Stati Uniti e McMaster in Canada) hanno scoperto nuovi potenziali antibiotici: Halicin contro *Escherichia coli* e Abaucin contro l'*Acinobacter baumannii*.

Gli esempi finora descritti non rappresentano che una piccola parte delle attuali applicazioni *dell'AI Science*.[2] Nel complesso, l'utilizzo dell'AI per fini di ricerca lascia presagire l'inizio di una nuova era di progresso per la scienza, con molteplici ricadute positive su tutte le fasi del processo di ricerca e implicazioni che spaziano dalle scienze teoriche fino a quelle applicate.

4. L'ERA DELL'*AI SCIENCE*: LIMITI E QUESTIONI APERTE

L'uso dell'IA nella ricerca scientifica comporta vantaggi concreti e significativi. Tuttavia, solleva anche una serie di problematiche a livello epistemologico, etico, sociale e politico. Considerando la vastità e la complessità di tali questioni, questo parere si concentrerà su cinque aspetti chiave che rappresentano limiti evidenti per il pieno sfruttamento delle potenzialità e dei benefici derivanti dalla sinergia tra intelligenza umana e artificiale nella ricerca scientifica. Questi aspetti riguardano: (i) la qualità dei dati e i bias negli algoritmi; (ii) la compartimentalizzazione dei dataset e la crescente dipendenza da algoritmi, prodotti e applicativi proprietari che limitano l'accesso e la riproducibilità degli studi; (iii) il fabbisogno energetico e infrastrutturale; (iv) l'aumento delle frodi e delle condotte scorrette nella ricerca, con la conseguente crisi del paradigma di produzione e valutazione della conoscenza scientifica; (v) la crescente distanza tra la complessità insita nell'AI Science e la comprensione di tale fenomeno da parte della popolazione e dei decisori politici; (vi) il deficit di pensiero critico riguardo ai cambiamenti in atto e alle loro implicazioni future.

Il presente parere non si propone né di essere esaustivo, né di avanzare un'analisi dettagliata di ciascuno di questi fattori limitanti rispetto a un pieno utilizzo dell'IA per la ricerca e l'innovazione. Piuttosto, intende offrire una mappatura degli aspetti principali che, a parere del Comitato Etico, richiedono azioni concrete e urgenti, da parte sia degli attori coinvolti nel processo di ricerca, sia dei decisori politici, per governare il presente e il futuro della *IA Science*.

4.1 Qualità dei dati e affidabilità degli algoritmi

Gli algoritmi di IA, pur essendo sofisticati, possono essere soggetti a errori e bias derivanti dai dati di addestramento e dalle assunzioni incorporate nei modelli. Attualmente, questi algoritmi sono costruiti utilizzando strumenti di analisi e calcolo probabilistici. Di conseguenza, risultano affidabili in contesti dove l'elaborazione probabilistica è fondamentale per ottenere risultati, ma lo diventano meno quando si richiedono elaborazioni logico-matematiche più precise. Uno dei principali problemi dell'AI science riguarda la qualità dei dataset utilizzati per addestrare gli algoritmi. La qualità dei dati è cruciale poiché influisce direttamente sulle prestazioni e sull'affidabilità dei modelli di IA. Dati incompleti, inaccurati o non rappresentativi possono infatti generare modelli le cui predizioni risultano errate o non generalizzabili al di fuori del loro set di addestramento. Questo problema è particolarmente rilevante nel campo della ricerca biomedica e diagnostica. Ad esempio, uno studio sull'uso dell'IA per diagnosticare il cancro al seno ha evidenziato che dataset non bilanciati portano a tassi di errore più elevati, con implicazioni potenzialmente gravi per i pazienti.

La scarsa qualità dei dataset ha anche implicazioni etiche, poiché l'uso di algoritmi di IA potrebbe comportare discriminazioni. I bias, ossia fattori distorsivi nei dati di addestramento, possono portare a risultati discriminatori quando gli algoritmi sviluppati

vengono applicati a gruppi demografici diversi. Un'analisi condotta dal MIT Media Lab, ad esempio, ha rivelato che gli algoritmi di riconoscimento facciale sviluppati da aziende leader come IBM e Microsoft avevano tassi di errore significativamente più elevati per persone di colore e donne rispetto agli uomini bianchi. Ciò evidenzia l'urgenza di affrontare il problema dei bias nei dataset di addestramento, non solo per migliorare la precisione delle predizioni, ma anche per evitare potenziali discriminazioni.

Entrambi gli aspetti — la qualità dei dati e la congruenza degli algoritmi — sono quindi essenziali per il buon esito della ricerca. Ciò solleva preoccupazioni sulla robustezza e sulla fiducia nei risultati prodotti dall'IA. Per garantire risultati accurati e affidabili, è fondamentale la verifica e la validazione indipendente dei modelli di IA, coinvolgendo più attori nel processo di ricerca.

4.2 Accesso ai dataset, riproducibilità, applicativi e piattaforme proprietarie

Idealmente, la ricerca dovrebbe basarsi su dati aperti e liberamente accessibili. Un accesso condiviso ai dati consente, infatti, di verificare in modo intersoggettivo la qualità dei dataset, l'affidabilità degli algoritmi e la correttezza dei risultati, contribuendo a limitare gli errori. Tuttavia, i dati utilizzati per l'addestramento dei modelli di IA sono spesso proprietari o difficili da ottenere in modo aperto e gratuito, creando barriere per i ricercatori e limitando la trasparenza e la replicabilità degli studi. Inoltre, l'accesso differenziato a vari database potrebbe amplificare le disuguaglianze tra le diverse discipline o avvantaggiare alcuni tipi di ricerca, come quella finanziata dai privati, rispetto alla ricerca pubblica, spesso sostenuta da università e istituzioni pubbliche.

Secondo diverse fonti, 'aprire' e condividere i database potrebbe contribuire a mitigare queste problematiche. Esistono già esempi virtuosi, che vanno incoraggiati. Ad esempio, il progetto Human Connectome, che mira a mappare le connessioni neurali del cervello umano, ha reso i suoi dati liberamente accessibili, facilitando una vasta gamma di ricerche neuroscientifiche.

Un altro aspetto fondamentale riguarda la riproducibilità dei risultati, un tema già oggetto di discussione nella comunità scientifica, ma che l'IA rende ancora più complesso. La riproducibilità richiede che i modelli di IA, i dataset e i processi di addestramento siano documentati in modo dettagliato e resi disponibili per la verifica indipendente. Tuttavia, le pratiche attuali spesso mancano di trasparenza, e in molti casi non è possibile conoscere esattamente il processo operativo seguito dai sistemi, complicato dal fatto che spesso si tratta di milioni di operazioni eseguite simultaneamente dalle macchine in frazioni di millisecondo. Ciò ostacola la capacità di altri ricercatori di replicare i risultati. Un esempio è rappresentato dall'iniziativa *Reproducibility Project: Cancer Biology*, che ha evidenziato difficoltà significative nel replicare studi di biologia del cancro che utilizzavano l'IA.

Inoltre, la crescente dipendenza da strumenti e applicativi di IA 'off-the-shelf', ossia prodotti proprietari sviluppati da terze parti commerciali, solleva preoccupazioni. Da una parte, l'uso di questi strumenti è parte integrante delle dinamiche moderne della ricerca scientifica, in quanto offre l'opportunità di introdurre innovazioni e tecnologie più avanzate per facilitare e migliorare la ricerca. D'altra parte, un'eccessiva dipendenza da prodotti proprietari basati sull'IA potrebbe limitare la capacità dei ricercatori di controllare e configurare autonomamente le proprie ricerche, contribuendo ad amplificare le disuguaglianze tra i vari sistemi di ricerca. Ad esempio, innovazioni come *AlphaFold*, che offre opportunità inedite per i ricercatori, sono spesso 'chiuse' e protette da brevetti, impedendo ai ricercatori stessi di comprendere appieno come siano state generate le ipotesi e le conclusioni, introducendo un margine di incertezza che può influire negativamente sulla qualità della ricerca prodotta. I costi legati all'utilizzo di questi strumenti applicativi e servizi possono rappresentare un ostacolo per l'accesso alle risorse necessarie per fare ricerca nell'era dell'*AI Science*.

L'accesso alle risorse di IA può creare un divario crescente nella capacità di condurre ricerca avanzata, penalizzando le comunità e le istituzioni meno attrezzate. Le università dei paesi in via di sviluppo, ad esempio, spesso non hanno accesso alle infrastrutture computazionali necessarie per l'addestramento di modelli di IA, limitando così la loro capacità di competere nel panorama globale della ricerca.

Per questi motivi esistono oggi proposte volte a realizzare internamente ai diversi ecosistemi o alle istituzioni di ricerca – soprattutto pubbliche, laddove esistenti – alcuni dei nuovi stru-

menti e applicativi per l'AI Science che sono attualmente solo disponibili presso terzi e, dunque, a pagamento; oppure per rendere quelli esistenti sempre più 'open-source'. Un'analisi approfondita di questi aspetti eccede i propositi del presente documento. Tuttavia, il Comitato Etico intende comunque sottolineare che per le sue proprietà, l'AI Science necessità l'avvio di una riflessione più generale all'interno della comunità dei ricercatori e tra i decisori politici in merito a quale tipo di modello e di relazione sia desiderabile costruire per gestire al meglio e bilanciare, da una parte, il bisogno di promuovere l'innovazione tecnologica, spesso affidato all'iniziativa dei privati e, dall'altro, il sistema che riguarda la produzione collettiva di conoscenza scientifica, il quale spesso è finanziato attraverso fondi pubblici.

4.3 Infrastrutture, consumo di risorse ed energia

Ogni interazione online richiede il supporto di un'imponente rete infrastrutturale, composta da migliaia di data center, satelliti, milioni di chilometri di cavi e altre infrastrutture fisiche. Le informazioni che vengono scambiate sono spesso archiviate ed elaborate in server situati in aree remote, i quali, sebbene fondamentali per la funzionalità dei sistemi di IA, consumano enormi quantità di energia. Secondo stime recenti, l'1-2% di tutta l'elettricità mondiale è utilizzato dai data center, con la previsione che questo dato possa crescere esponenzialmente nei prossimi anni, parallelamente all'espansione dell'IA. Questa crescita esponenziale della domanda energetica ha implicazioni significative sia dal punto di vista ambientale che geopolitico, richiedendo un'attenzione immediata e concreta.

Un report recente della Royal Society evidenzia chiaramente l'impatto ambientale dell'IA: «La raccolta, l'analisi, l'archiviazione e la condivisione dei dati richiesti per i sistemi basati sull'IA hanno un impatto ambientale significativo. Ad esempio, si stima che archiviare un terabyte di dati consumi 10 kg di anidride carbonica all'anno, mentre addestrare un modello di linguaggio come ChatGPT può generare 550 tonnellate di emissioni di anidride carbonica. Si stima che le emissioni globali di gas serra dei data center siano pari alle emissioni dell'aviazione commerciale statunitense e, poiché i set di dati e i modelli diventano più grandi, è probabile che ciò aumenti». Questo dato rappresenta una sfida significativa per il futuro, poiché suggerisce che le risorse naturali e l'infrastruttura energetica necessarie per l'evoluzione e l'espansione dell'IA potrebbero non essere sostenibili se non gestite con una pianificazione accurata.

Per rendere sostenibile la rivoluzione conoscitiva permessa dall'IA, è quindi essenziale elaborare piani strategici che consentano di affrontare il crescente fabbisogno energetico in modo responsabile. Questi piani dovrebbero includere l'adozione di tecnologie più efficienti dal punto di vista energetico, l'utilizzo di fonti di energia rinnovabile per alimentare i data center e una maggiore innovazione nel design delle infrastrutture, riducendo al minimo il consumo di risorse. Inoltre, le politiche e le pratiche di ricerca dovrebbero essere orientate alla ricerca di soluzioni che possano diminuire l'impronta ecologica dell'IA. Il crescente utilizzo di modelli di IA sempre più complessi e di larga scala implica la necessità di infrastrutture più potenti, ma questo deve essere accompagnato da un attento monitoraggio e regolamentazione per evitare danni ambientali irreversibili.

Un altro aspetto che merita attenzione riguarda la consapevolezza dei ricercatori e degli utilizzatori finali riguardo ai costi energetici associati all'addestramento e all'utilizzo di modelli di IA. Oggi, molti degli strumenti di IA disponibili sono presentati con interfacce utente che non enfatizzano adeguatamente i costi energetici e l'impatto ambientale che derivano dal loro utilizzo. Le piattaforme tendono, infatti, a nascondere l'effettivo costo in termini di risorse naturali e produzione di gas serra, un aspetto che potrebbe incentivare un uso meno responsabile da parte degli utenti. Gli sviluppatori, i ricercatori e i decisori politici devono quindi essere consapevoli di questo aspetto e cercare di promuovere pratiche più sostenibili, come la progettazione di modelli di IA che utilizzano meno energia e risorse, e l'adozione di tecnologie a basso impatto.

Infine, non va dimenticato che l'impatto ambientale non è solo legato all'energia richiesta per l'elaborazione dei dati, ma anche alla produzione e smaltimento delle infrastrutture hardware utilizzate, come i server e i dispositivi di storage. Questi dispositivi hanno una durata limitata e, una volta obsoleti, devono essere smaltiti. La gestione dei rifiuti elettronici, che include il recupero e il riciclo dei materiali, è priorità per l'industria dell'IA. In sintesi, la crescente domanda di energia necessaria per alimentare la ricerca e l'innovazione nell'ambito

dell'AI Science solleva questioni cruciali che richiedono risposte tempestive e strategiche. La transizione verso una IA più sostenibile deve essere accompagnata dalla consapevolezza di ricercatori, sviluppatori e decisori politici sui costi energetici e ambientali associati all'uso di tali tecnologie.

4.4 Etica, integrità nella ricerca e il modello di produzione scientifica

L'IA, in particolare quella generativa, grazie alla sua capacità di creare dati, immagini e testi sintetici indistinguibili da quelli reali, solleva seri problemi di integrità nella ricerca scientifica. I ricercatori, avvalendosi di strumenti di IA generativa, potrebbero produrre dati falsi che supportano ipotesi preesistenti, ingannando revisori e lettori. La possibilità di fabbricare dati o risultati di esperimenti non è solo una pratica scorretta, ma può contaminare la letteratura scientifica, minando la fiducia nel processo scientifico stesso. Un esempio di questa preoccupazione è il caso della creazione di immagini mediche sintetiche utilizzate per l'addestramento di modelli diagnostici. Se queste immagini non vengono esplicitamente identificate come sintetiche, possono compromettere l'affidabilità dei risultati degli studi clinici e portare a conclusioni errate. La difficoltà di distinguere tra dati autentici e sintetici rappresenta una nuova e significativa sfida per la verifica e la validazione della ricerca scientifica, aumentando il rischio di frodi e minando la credibilità della comunità scientifica.

Ad esempio, articoli scientifici pubblicati su riviste prestigiose sono stati ritirati dopo che si è scoperto che i dati erano stati generati artificialmente, come nel caso delle ricerche sulla biologia del cancro che hanno utilizzato dati sintetici senza una chiara indicazione. Mentre la questione dell'integrità nella ricerca esiste da sempre, è difficile negare che, nell'ambito dell'AI Science, essa abbia acquisito una nuova dimensione che richiede risposte urgenti e strutturali. Il rischio di manipolazioni è accentuato dalla disponibilità di strumenti di IA facili da utilizzare, che abbassano notevolmente la difficoltà di produrre risultati falsificati, aumentando esponenzialmente l'impatto che un singolo errore o frode possa avere.

Un fenomeno preoccupante che ha guadagnato attenzione negli ultimi anni è quello delle 'paper mills' o 'fabbriche di articoli scientifici', che producono articoli falsi o manipolati su commissione, spesso da parte di ricercatori che necessitano di pubblicazioni per avanzare nella carriera o per ottenere finanziamenti. Tali articoli, che possono includere dati falsificati, analisi inventate, grafici e immagini manipolate, vengono confezionati per sembrare pubblicazioni scientifiche legittime. Questo fenomeno mina la credibilità delle pubblicazioni scientifiche e distorce non solo il progresso scientifico, ma anche le politiche che si basano su ricerche inaffidabili. Un esempio eclatante riguarda la rivista *Tumor Biology*, che nel 2017 ha dovuto ritirare oltre 100 articoli dopo aver scoperto che erano stati prodotti da 'paper mills'. La presenza di questi articoli falsi nel corpus scientifico non solo inganna i ricercatori, ma può condurre ad ulteriori studi basati su risultati falsi e influenzare erroneamente decisioni politiche e protocolli clinici, con un significativo impatto economico.

Alla luce di questi problemi, alcune istituzioni, tra cui il Consiglio d'Europa, hanno pubblicato linee guida per aiutare i ricercatori a rispettare i principi di integrità nella ricerca nell'era dell'IA e delle intelligenze generative. Queste iniziative sono necessarie e devono essere incoraggiate. Esse aiutano a definire chiaramente quali comportamenti sono leciti e da promuovere e quali sono illeciti e da evitare. Tuttavia, il Comitato Etico segnala che tali impulsi normativi, pur essendo fondamentali, devono essere accompagnati da una riflessione più ampia e comune riguardo al modello di produzione scientifica attualmente in uso, che è alla base di molti dei comportamenti scorretti.

Il sistema di ricerca accademica, caratterizzato dal modello 'publish or perish', ha posto una pressione crescente sui ricercatori, creando incentivi sempre più forti a adottare pratiche illecite per fare fronte alla competizione, ad esempio falsificando risultati o accettando pratiche eticamente discutibili per ottenere pubblicazioni. L'adozione di strumenti di IA, sempre più facili da usare e accessibili, ha esacerbato le contraddizioni di questo sistema, abbassando drasticamente la difficoltà di produrre lavori scientifici manipolati. Questo rende sempre più difficile accertare le condotte scorrette e amplifica notevolmente l'impatto negativo che una sola persona può avere attraverso l'utilizzo di questi strumenti. La disponibilità di AI generativa, che rende l'inganno più semplice da realizzare, pone una sfida maggiore nell'individuare e fermare i comportamenti fraudolenti.

In questo contesto, il solo aggiornamento o la creazione di nuove linee guida è necessaria, ma non sufficiente. Questi interventi potrebbero tamponare i sintomi senza affrontare le cause profonde di condotte scorrette nella ricerca. È essenziale promuovere un cambiamento strutturale nel modo in cui la ricerca è valutata, finanziata e pubblicata. La comunità scientifica deve impegnarsi a ridefinire il concetto di successo accademico e professionale, superando il modello che premia la quantità a discapito della qualità e, in alcuni casi, anche dell'integrità della ricerca. Solo così si potrà costruire un ambiente di ricerca più etico e sostenibile, capace di affrontare le sfide dell'AI Science senza compromettere i valori fondamentali della scienza.

4.5 Necessità di una riflessione teorica ed epistemologica

Un aspetto cruciale da considerare nell'ambito delle sfide e dei limiti posti dall'IA alla scienza è quello relativo ai cambiamenti nel modo in cui concepiamo e pratichiamo la ricerca. In particolare, c'è una transizione in atto che sta spostando la scienza sempre più verso un modello *data-driven*, a discapito di un approccio *theory-driven*. Grazie agli strumenti predittivi offerti dall'IA, le previsioni e le analisi dei fenomeni scientifici sono diventate centrali, mettendo in secondo piano la costruzione di modelli teorici che spiegano *perché* certi fenomeni accadono. In questo nuovo paradigma, non ci interessa tanto spiegare i meccanismi sottostanti, ma piuttosto predire cosa accadrà. Questo cambiamento non è solo una questione pratica, ma solleva anche interrogativi di natura filosofica ed epistemologica.

Nel contesto di una scienza sempre più basata sui dati, diventa cruciale riflettere su come questi dataset vengano costruiti e utilizzati. La qualità e la composizione dei dati sono ora al centro della ricerca, ma c'è un rischio significativo che, spostandoci verso il predittivismo, si riduca l'attenzione sulle questioni teoriche fondamentali. Questo approccio rischia di perdere di vista la costruzione del significato e il valore delle teorie scientifiche, riducendo la ricerca a una mera applicazione tecnologica senza una base teorica solida.

Accanto a questo, è necessario riflettere sull'ermeneutica che l'IA impone al ricercatore. Non ci interessa solo un controllo normativo, ma comprendere come l'IA influenzi le modalità di produzione, circolazione e interpretazione dei contenuti scientifici. Queste logiche modellano anche le abitudini mentali e i costrutti simbolici con cui ci avviciniamo alla scienza, influenzando profondamente la cultura stessa. In questo senso, l'etica si sposta verso un campo della responsabilità più ampio, che riguarda il modo in cui i ricercatori interagiscono con gli strumenti tecnologici e come questi strumenti modificano il loro approccio alla conoscenza.

Un altro aspetto etico riguarda il processo di addestramento dell'IA, che attualmente manca di una dimensione meditativa e riflessiva. L'accumulo e il processamento dei dati, pur essendo cruciali, non sono accompagnati da una riflessione critica su quali stili cognitivi, riflessivi, contestuali e culturali siano alla base dei paradigmi scientifici. Il risultato è un sapere che, pur essendo vasto e potente, è privo di una cornice etica che ne orienti l'utilizzo e l'interpretazione. In questo contesto, è fondamentale interrogarsi su come si formi l'intelligenza che emerge da questi sistemi e come si relazioni con l'idea che l'IA rappresenti semplicemente una capacità di raggiungere fini complessi, come proposto da Max Tegmark nel 2018.

Nel dibattito epistemologico, è dunque essenziale includere una riflessione profonda sulla costruzione del significato, sull'interpretazione dei dati e sul sistema di valori che permea la scienza. Solo così sarà possibile comprendere appieno come l'IA stia cambiando non solo gli strumenti della ricerca, ma anche la sua stessa natura e la nostra comprensione del mondo.

4.6 Mancanza di comprensione generale e la necessità di educazione e formazione

Uno degli ostacoli principali che la scienza e la società devono affrontare nell'era dell'IA è la mancanza di una comprensione diffusa e approfondita di come questa tecnologia stia plasmando e trasformerà il nostro mondo. Sebbene l'AI stia avendo un impatto crescente in vari settori, dalla ricerca scientifica all'economia, dalle politiche pubbliche alla vita quotidiana, esiste una comprensione limitata delle sue potenzialità, dei suoi rischi e delle sue implicazioni. Questo gap di comprensione non riguarda solo il pubblico generale, ma anche i decisori politici, che spesso non dispongono delle competenze necessarie per prendere decisioni informate su come regolare, sviluppare e utilizzare l'IA in modo responsabile e sostenibile.

Il progresso dell'IA sta rapidamente trasformando non solo la scienza, ma anche le dinamiche sociali ed economiche globali. Tuttavia, senza una comprensione adeguata e condivisa, è difficile immaginare e pianificare un futuro in cui l'IA venga utilizzata in modo equo e a beneficio dell'intera società. Le decisioni che verranno prese oggi determineranno non solo il destino della ricerca scientifica, ma anche quello della nostra cultura, dei nostri valori e della nostra economia. In questo scenario, è fondamentale che i decisori politici, gli educatori e i ricercatori stessi abbiano una conoscenza più profonda e critica dei meccanismi e delle implicazioni dell'IA.

Per far fronte a questa carenza di comprensione, è necessario avviare un processo di educazione che inizi dalla scuola e che coinvolga tutti i livelli della società. L'educazione non deve limitarsi alla mera alfabetizzazione tecnologica, ma deve includere una riflessione critica sui temi etici, filosofici ed epistemologici legati all'uso dell'IA. Gli studenti devono essere preparati non solo a comprendere come funzionano gli algoritmi, ma anche a interrogarsi sui loro impatti, sui valori che li sottendono e sulle sfide che l'adozione di queste tecnologie comporta.

Inoltre, il coinvolgimento delle nuove generazioni in queste riflessioni deve andare oltre la mera conoscenza tecnica: occorre insegnare loro a diventare cittadini consapevoli, in grado di prendere decisioni informate e di influenzare positivamente le politiche pubbliche relative all'IA. Solo partendo da una base comune di comprensione, sarà possibile costruire un futuro in cui l'IA possa essere utilizzata in modo equo e sostenibile, evitando che i suoi benefici siano distribuiti in modo diseguale e che i suoi rischi siano gestiti in modo inadeguato.

Infine, è essenziale che anche i decisori politici si impegnino in un percorso di formazione continua, al fine di comprendere meglio le potenzialità, le sfide e le implicazioni dell'IA. Le politiche pubbliche devono essere guidate dalla consapevolezza delle complesse dinamiche che caratterizzano queste tecnologie, in modo da adottare regolamenti e leggi che proteggano i diritti dei cittadini e favoriscano un uso responsabile dell'IA. Senza una comprensione comune, l'adozione dell'IA potrebbe risultare frammentata e mal indirizzata, con conseguenze negative per l'innovazione scientifica, la coesione sociale e l'equità.

La formazione e l'educazione sono le chiavi per costruire una società che sappia affrontare i cambiamenti e le sfide posti dall'IA, facendo in modo che l'innovazione tecnologica possa prosperare in un ambiente che promuova la responsabilità, l'equità e il benessere collettivo.

5. CONCLUSIONI E RACCOMANDAZIONI

In sintesi, l'intelligenza artificiale rappresenta una tecnologia progettata dall'uomo per compiere operazioni che imitano il ragionamento umano, con la capacità di generare risultati che, in molti casi, sarebbero difficili o impossibili da raggiungere senza l'intervento di macchine. Tuttavia, non dobbiamo dimenticare che le capacità di 'comprensione', 'spiegazione' e 'comunicazione' rimangono prerogative dell'intelligenza umana, che è quella che ha progettato e sviluppato l'intelligenza artificiale stessa. Sebbene l'IA possa ottenere risultati straordinari, è importante ricordare che questi dipendono sempre dai programmi scritti da esseri umani, con tutti i limiti che ciò comporta: possibili errori nella programmazione e l'incapacità intrinseca del computer di comprendere o di rispondere a situazioni al di fuori degli schemi impostati.

L'uso dell'IA nella ricerca scientifica porta con sé enormi potenzialità, ma anche sfide complesse. Mentre l'intelligenza artificiale può accelerare il progresso scientifico e aprire nuove frontiere della conoscenza, un suo impiego scorretto può compromettere il processo di ricerca, generando risultati falsati e applicazioni dannose per individui e società. Per ridurre i rischi associati all'uso improprio dell'IA e per massimizzare i benefici, è fondamentale adottare una serie di principi etici e politiche appropriate.

Il Comitato Etico della Fondazione Veronesi raccomanda di intraprendere le seguenti azioni per garantire che l'uso dell'IA in ambito scientifico sia responsabile, sostenibile e benefico per la società:

1. Adottare un approccio centrato sull'umanità: È essenziale sviluppare l'intelligenza artificiale in modo che risponda ai bisogni umani, tuteli la dignità, la privacy e i diritti fondamentali delle persone, promuovendo il benessere sociale. L'adozione di un 'umanesimo digitale' significa mettere la tecnologia al servizio delle persone, creando un equilibrio tra innovazione tecnologi-

ca e valori umani. L'innovazione non deve essere solo tecnica, ma anche sociale e culturale, in modo da migliorare la qualità della vita umana e promuovere il bene comune.

2. Promuovere la scienza aperta: È necessario sostenere e diffondere i principi di 'open science' per garantire che l'AI Science sia sempre più trasparente, libera e accessibile. Ciò implica la condivisione dei database, degli algoritmi, dei materiali e dei metodi di ricerca, contribuendo così a un ambiente di collaborazione e fiducia tra ricercatori e istituzioni.

3. Investire nella cultura dell'integrità nella ricerca: La creazione di una cultura radicata nell'integrità è fondamentale per il buon uso dell'IA nella ricerca. È necessario offrire corsi di formazione a ricercatori e ricercatrici e promuovere iniziative di coordinamento tra tutti gli attori coinvolti nel processo di ricerca. Le istituzioni, le imprese, le società scientifiche e le riviste di settore devono collaborare per garantire che l'uso dell'IA nella ricerca sia sempre etico e responsabile. In particolare, il Comitato Etico raccomanda che tutti i ricercatori che ricevono contributi dalla Fondazione Veronesi per fini di ricerca sottoscrivano, come impegno etico vincolante, la *Dichiarazione in materia di Integrità nella Ricerca*. Tale dichiarazione, al punto 11, richiede ai ricercatori di:

   - Utilizzare in modo responsabile gli strumenti di intelligenza artificiale, assumendosi la responsabilità per i contenuti generati, ove applicabile;
   - Esplicitare in modo trasparente ogni utilizzo sostanziale di strumenti di IA nella ricerca;
   - Rispettare le normative relative alla privacy e alla proprietà intellettuale quando si condividono dati sensibili con l'IA;
   - Evitare di utilizzare in modo improprio strumenti di IA per attività che possano influire negativamente su altre ricercatrici, ricercatori o organizzazioni, come la revisione tra pari, la valutazione di progetti o la selezione di personale di ricerca.

In conclusione, l'intelligenza artificiale ha il potenziale di rivoluzionare la scienza e la ricerca, ma il suo impatto dipende in larga misura dalle scelte etiche, politiche e culturali che faremo. Solo attraverso un approccio equilibrato, che unisce innovazione tecnologica e responsabilità umana, potremo garantire che l'IA contribuisca in modo positivo al progresso scientifico e alla società nel suo complesso.

NOTE

1. Composizione del Comitato Etico: Carlo Alberto Redi, (Presidente), Giuseppe Testa (Vicepresidente), Guido Bosticco, Roberto Defez, Giorgio Macellari, Emanuela Mancino, Alberto Martinelli, Michela Matteoli, Telmo Pievani, Giuseppe Remuzzi, Luigi Ripamonti, Giuliano Amato (Presidente Onorario), Cinzia Caporale (Presidente Onorario), Marco Annoni (Coordinatore). Il documento è stato approvato all'unanimità con votazione telematica in data 30.01.2024. Alla stesura del documento hanno collaborato anche Dino Maurizio (Informatici Senza Frontiere APS) e Vieri Giuliano Santucci (ISTC-CNR) in qualità di esperti *ad acta.*

2. I GAN utilizzano due reti neurali chiamate 'generatore' e 'discriminatore'. Il 'generatore' crea dei dati o delle immagini sintetiche (e cioè, non reali), mentre il 'discriminatore' le valuta rispetto a dati o immagini reali, allenando così il generatore a creare output sintetici sempre più credibili. La fase di 'allenamento' di questi sistemi prosegue fino a quando il discriminatore non è più in grado di distinguere tra gli output sintetici del generatore e input reali.

3. Oltre a questi casi, l'uso dell'IA nella ricerca è sempre più essenziale in praticamente ogni campo di studi e ambito di ricerca. Nelle scienze ambientali, l'IA permette di integrare e analizzare grandi quantità di dati al fine di migliorare la comprensione dei cambiamenti climatici e predire l'evoluzione di precisione eventi meteorologici estremi, come uragani e ondate di calore, oppure per monitorare la deforestazione, la perdita di biodiversità e altre minacce ambientali, aiutando a prendere decisioni informate per la conservazione e la gestione sostenibile delle risorse naturali. Anche nelle neuroscienze, l'IA sta aprendo nuove frontiere nel campo delle interfacce cervello-computer (BCI), permettendo la comunicazione diretta tra il cervello umano e i dispositivi esterni grazie all'utilizzo di algoritmi di ML per decodificare i segnali cerebrali e tradurli in comandi per controllare protesi o altri dispositivi.

Dichiarazione in materia di integrità nella ricerca. 2024

COMITATO ETICO FONDAZIONE VERONESI[1]

La Fondazione Umberto Veronesi si riconosce nei principi e nei valori dell'integrità nella ricerca, così come affermati nei principali strumenti di orientamento e regolazione nazionali e internazionali sulla materia, tra i quali si segnalano la "Dichiarazione di Singapore sull'integrità nella ricerca" (II World Conference on Research Integrity, 2010)[2],The European Code of Conduct for Research Integrity by ALLEA (2023)[3] e, soprattutto, le "Linee guida per l'integrità nella ricerca" del CNR (2019)[4].

In particolare, la Fondazione fa propria la definizione di integrità nella ricerca contenuta in quest'ultimo documento, nell'auspicio della più ampia condivisione negli Atenei e nelle istituzioni di ricerca italiane: «Per integrità nella ricerca si intende l'insieme dei principi e dei valori etici, dei doveri deontologici e degli standard professionali sui quali si fonda una condotta responsabile e corretta da parte di chi svolge, finanzia o valuta la ricerca scientifica nonché da parte delle istituzioni che la promuovono e la realizzano. L'applicazione dei principi e dei valori e il rispetto della deontologia e degli standard professionali sono garanzia della qualità stessa della ricerca e contribuiscono ad accrescere la reputazione e l'immagine pubblica della scienza, con importanti ricadute sullo sviluppo della stessa e sulla società».

La Fondazione Umberto Veronesi richiede alle ricercatrici e ricercatori che svolgono attività di ricerca finanziate dalla Fondazione stessa o comunque condotte sotto la sua egida, di aderire e attenersi alla seguente Dichiarazione:

In qualità di ricercatrice/ricercatore, nello svolgimento delle mie attività scientifiche mi impegno a:

1. *non fabbricare o falsificare i dati o i risultati della mia ricerca nonché a documentare le attività sperimentali e a conservare con diligenza i materiali e i dati primari ottenuti nel loro svolgimento;*
2. *non commettere plagio né a sottrarre intenzionalmente o per una condotta negligente dati, risultati o testi altrui né appropriarsi di idee la cui attribuzione ad altri sia documentata e dimostrabile;*
3. *esplicitare in modo trasparente eventuali conflitti di interesse in grado di influenzare significativamente la mia obiettività, anche ove la loro esplicitazione non sia richiesta, e a menzionare nelle mie pubblicazioni il contributo dei soggetti finanziatori;*
4. *pubblicare tempestivamente i risultati delle mie ricerche in modo accurato, obiettivo e attendibile, non offrendo, attribuendo, imponendo o negando in modo improprio ad altri lo status di co-autore di una pubblicazione né accettando tale status non avendone i requisiti;*
5. *non annunciare in modo enfatico sui media di aver conseguito un risultato importante o di aver compiuto una scoperta qualora non vi fossero solide basi scientifiche per affermarlo;*
6. *chiedere la ritrattazione di un articolo di cui sono autore o co-autore ove fosse basato su dati fabbricati/falsificati oppure ove contenga errori gravi nonché a ritrattare l'annuncio sui media di un risultato o scoperta da me conseguiti nel caso in cui tale annuncio si sia dimostrato infondato;*
7. *non manipolare o falsificare il mio curriculum vitae, la mia affiliazione o l'elenco delle mie pubblicazioni né a includervi deliberatamente informazioni erronee;*
8. *non sabotare, ostacolare, rallentare o sminuire le ricerche dei miei colleghi né a fomentare pregiudizi o a ledere la loro reputazione scientifica in modo ingiustificato o per interesse personale;*
9. *segnalare un'eventuale condotta scorretta commessa da un altro ricercatore ove esistano fondate ragioni e opportuni riscontri, non contribuire a nascondere eventuali condotte scorrette mie o di altri e non formulare accuse infondate, malevole e/o palesemente futili;*
10. *agire con professionalità, responsabilità, lealtà, rigore, imparzialità, trasparenza e fair play, rendicontando pubblicamente le mie ricerche, rispettando i diritti di tutte le persone coinvolte e avendo cura della biosfera;*
11. *utilizzare in modo responsabile strumenti di intelligenza artificiale:*
    - *assumendomi la responsabilità rispetto ai contenuti generati o prodotti attraverso tali strumenti, ove applicabile;*

- *esplicitando in modo trasparente ogni utilizzo sostanziale di strumenti di intelligenza artificiale nella mia ricerca;*
- *rispettando le norme vigenti in materia di privacy, confidenzialità e proprietà intellettuale nel condividere con strumenti di intelligenza artificiale dati e informazioni sensibili o protetti;*
- *evitando di utilizzare in modo improprio ed univoco strumenti di intelligenza artificiale per attività che potrebbero avere un impatto su altre ricercatrici, ricercatori e organizzazioni, come la revisione tra pari, la valutazione di progetti o la selezione di personale di ricerca.*

NOTE

1. Composizione del Comitato Etico: Carlo Alberto Redi, (Presidente), Giuseppe Testa (Vicepresidente), Guido Bosticco, Roberto Defez, Giorgio Macellari, Emanuela Mancino, Alberto Martinelli, Michela Matteoli, Telmo Pievani, Giuseppe Remuzzi, Luigi Ripamonti, Giuliano Amato (Presidente Onorario), Cinzia Caporale (Presidente Onorario), Marco Annoni (Coordinatore). Il documento è stato approvato all'unanimità con votazione telematica in data 30.01.2024.

2. http://www.singaporestatement.org/downloads/singpore%20statement_lettersize.pdf.

3. https://www.cnr.it/sites/default/files/public/media/doc_istituzionali/ethics/European-Code-of-Conduct-Revised-Edition-2023.pdf.

4. https://cnr.it/sites/default/files/public/media/doc_istituzionali/linee-guida-integrita-nella-ricerca-cnr-commissione_etica.pdf. È in particolare a queste linee guida che la presente si ispira.

Documenti di Etica e Bioetica

Manifesto per un'etica procedurale

COMITATO BIOETICO PER LA VETERINARIA E L'AGROALIMENTARE
CBV-A

Viviamo in un contesto culturale caratterizzato da un profondo pluralismo etico[1]. La difficoltà della costruzione di uno spazio etico condiviso spesso genera la polarizzazione ideologica o, di converso, la banalizzazione frutto del senso comune con un progressivo impoverimento dell'analisi razionale, secondo una tendenza che può avere conseguenze sociali rilevanti. Creare le condizioni per un dibattito critico e aperto richiede il saper chiarire le ragioni alla base delle diverse posizioni etiche e i valori cui sono ispirate, ma anche il saper metterle alla prova dei fatti e verificare l'accettabilità delle conclusioni cui esse conducono. Questo confronto quotidiano con questioni etiche concrete è la funzione sociale svolta elettivamente dai Comitati etici. Il ruolo loro affidato è 'fare etica' al di là del semplice rispetto formalistico della norma, migliorando l'affidabilità complessiva dei comportamenti e alimentando nei cittadini un clima di fiducia Il CBV-A ha da sempre svolto questo ruolo di laboratorio etico nella convinzione che esplicitare le ragioni 'degli altri', pur senza alcuna passiva accettazione, sia un metodo per individuare la migliore soluzione possibile, ragionevolmente accettabile per i soggetti interessati anche se ben lontana da una ideale perfezione morale. L'auspicio è che il lavoro dei comitati etici come il CBV-A possa precedere ed ispirare il diritto e la politica.

L'etica ha una vocazione pratica, ma nulla è più pratico di una buona teoria. Occorre una buona teoria per fondare un metodo di lavoro rigoroso, verificabile e aperto alla discussione pubblica. Il metodo di lavoro del CBV-A adotta i presupposti teorici dell'etica procedurale il cui obiettivo è il perseguimento dell'**equità**, intesa come equa considerazione degli interessi in gioco. Tale approccio presuppone l'esistenza di uno schema cooperativo tra persone orientate a trovare ragioni mutualmente giustificabili: l'eticità delle decisioni deriva dalla legittimazione della procedura, che a sua volta è data dalla condivisione delle regole del gioco tra tutti i partecipanti. Nel conseguire una decisione condivisa tra gli stakeholder, devono essere individuati gli interessi moralmente rilevanti, la natura delle questioni etiche, gli obiettivi perseguiti e le conseguenze ad essi connesse, secondo le seguenti regole:

1. Descrizione del **contesto,** degli **scopi generali** e degli **obiettivi specifici** che si intendono raggiungere**.**
2. Individuazione degli **stakeholder.** Sono considerati stakeholder i soggetti i cui interessi possono essere direttamente o indirettamente lesi dalla messa in atto delle deliberazioni. Non sono ammessi al processo deliberativo coloro che non subiscono alcuna conseguenza nel contesto in esame**.**
3. Definizione di un **linguaggio morale comune.** Il linguaggio degli interessi è da prediligere in quanto più inclusivo rispetto al linguaggio normativo dei diritti, che esclude almeno potenzialmente soggetti morali passivi quali gli animali, l'ambiente o una biosfera nel suo complesso.
4. Definizione delle **questioni etiche correlate agli interessi degli stakeholder.** Tali questioni riguardano tutto ciò che incide significativamente, sia in senso positivo che negativo, sulla vita stessa o qualità della vita dei soggetti morali attivi o passivi.
5. Identificazione dei **criteri** in base ai quali considerare positivi o negativi gli effetti delle decisioni.
6. Descrizione delle **possibili conseguenze:** impatto degli effetti a medio e lungo termine, gravità e probabilità degli effetti, preferenze degli stakeholder riguardo alle conseguenze, problemi correlati.
7. Determinazione della **natura della decisione**: definizione della rivedibilità o meno della decisione e della sua natura (avversativa, convergente, compensativa o compositiva). Occorre adottare, ogni volta che è possibile, una deliberazione compositiva o compensativa.
8. Valutazione del peso dell'**incertezza** nello stabilire la probabilità, la gravità e l'impatto delle conseguenze possibili e il grado di incertezza nel **bilanciamento tra costi e benefici** e tra **rischi e benefici.**
9. Verifica del grado di **tolleranza** del rischio di tutti gli stakeholder in ragione delle risorse individuali, di specie o alla resilienza ambientale alle conseguenze che ne deriverebbero qualora questi si verificassero.
10. Valutazione delle **possibili alternative** sulla base dei criteri stabiliti. Occorre mantenere la

disponibilità a rivedere le decisioni già prese in considerazione di nuove conoscenze o ulteriori informazioni.

Ogni processo deliberativo deve basarsi sulla disponibilità a rendere conto dei criteri e delle ragioni sottese alle scelte/soluzioni/misure proposte e ad esplicitare i principi etici che le giustificano. Direttamente consequenziale a tale prerequisito è la rivedibilità delle decisioni raggiunte sulla base di nuove evidenze, come anche l'impegno ad individuare solo misure che possano essere effettivamente messe in pratica e che siano quindi concretamente applicabili, sia a livello individuale che da parte delle Istituzioni competenti. Un'adeguata consapevolezza dei problemi richiede, infine, la disposizione a riconoscere la loro complessità, ad adottare cioè una visione sistemica che integri e ponderi tra loro tutte le variabili connesse alle diverse dimensioni e alle differenti fasi temporali che influenzano le conseguenze effettive della deliberazione.

Pur non disconoscendo la rilevanza di ideali morali non negoziabili, essi non sono oggetto del processo deliberativo in quanto nella loro pretesa rilevanza assoluta non ammettono il confronto con la dimensione pubblica dell'etica. Sono però inclusi nel processo deliberativo coloro che si identificano in tali valori quali portatori di interessi alla pari degli altri. Tale processo ha l'obiettivo di giungere al miglior accordo morale possibile in una società multiculturale rispetto a questioni eticamente sensibili nella prospettiva di orientare la produzione normativa e la decisione politica.

e la disponibilità ad argomentare le diverse posizioni etiche, possa condurre a deliberazioni condivise tra i portatori di interesse e a soluzioni realisticamente applicabili. Il documento è stato elaborato da Elena Mancini, coordinatore del Comitato, letto e discusso con i suoi componenti, ed infine approvato definitivamente nella seduta plenaria del 27 giugno 2024.

NOTE

1. Composizione del CBV-A: Pasqualino Santori (Presidente), Alessandro Alessandrini, Salvatore Amato, Marco Annoni, Teresa Bossù, Beniamino Terzo Cenci Goga, Alessandra de Seneen, Alessandro Fantini, Gianluca Felicetti, Gianluigi Giovagnoli, Francesco Leopardi Dittaiuti, Francesco Loreto, Elena Mancini, Andrea Monaco, Eugenia Natoli, Ilja Richard Pavone, Domenico Pignone, Simone Pollo, Anna Saba, Guido Schwarz, Francesco Zecca. Il Manifesto intende delineare, nelle sue linee essenziali, il metodo di lavoro del Comitato Bioetico per la Veterinaria e l'Agroalimentare che predilige un approccio procedurale nella convinzione che l'analisi concettuale delle questioni

Manifesto per un'etica procedurale

The manifesto for a procedural ethics

Laura Palazzani

LUMSA, Roma
palazzani@lumsa.it

DOI: 10.53267/20240201

Il "Manifesto per un'etica procedurale" è proposto come esplicitazione del metodo di lavoro del Comitato Bioetico per la Veterinaria e l'Agroalimentare, elaborato da Elena Mancini, discusso e approvato da tutti i componenti. Si tratta di un documento di considerevole interesse per una serie di ragioni che ne evidenziano anche la portata che va anche oltre le intenzioni di chi lo ha redatto, discusso e approvato.

1) *I presupposti dell'etica procedurale*. Nel documento emerge in modo evidente che la pre-condizione per consentire un'adeguata funzione di un comitato di etica (e ciò vale non solo per la veterinaria e l'agroalimentare, ma per qualsiasi tema) è l'attitudine dialogica e critica dei componenti, nel 'pensare insieme'. Attitudine dialogica significa che chi fa parte di un comitato non deve né rinunciare scetticamente ad una verità, né imporre in modo univoco la propria verità, ma deve chiarire le proprie ragioni e porsi a confronto con le ragioni degli altri, nella disponibilità all'ascolto e nella disponibilità a sottoporre le proprie ragioni alla prova dei fatti, per verificarne la consistenza e la coerenza, dunque l'accettabilità o l'inaccettabilità. Il documento sottolinea che i fatti possono costringerci a rivedere le teorie, o comunque a riformularle in modo adeguato rispetto al contesto di analisi; inoltre il confronto delle ragioni e dei valori, anche differenti, contribuisce all'identificazione di elementi condivisibili, minimi o massimi. Esplicitare le ragioni proprie, confrontarle con le ragioni degli altri significa evitare sia la passiva accettazione, sia l'imposizione assolutistica, nella ricerca di una soluzione ragionevole. In questo senso i presupposti dell'etica procedurale non riguardano solo i comitati di etica per la veterinaria e l'agroalimentare, ma riguardano in senso lato i presupposti stessi del 'fare bioetica'.

2) *Dall'etica procedurale all'etica sostanziale*. L'etica procedurale descritta nel documento in modo dettagliato non è riducibile ad una mera metodologia che individua elementi e fasi della modalità di discussione nel contesto di un comitato etico specifico, quale quello veterinario e agroalimentare. Si tratta di una metodologia procedurale, che se rigorosamente applicata, consente una elaborazione *sostanziale*. In altri termini, non si tratta di una descrizione di procedure meramente formali, ma le procedure stesse divengono contenuti. Un metodo che consente di 'fare etica' e che rende il comitato etico un 'laboratorio' di pensiero e di produzione di pensiero, innovativo e creativo. Il metodo di lavoro offre una griglia (si comprende nel documento, costruita sulla base di una lunga e consolidata esperienza di partecipazione e coordinamento di comitati etici) che consente, nella precisazione dei diversi passaggi, di individuare possibili soluzioni ai quesiti emergenti, mediante un confronto tra discipline e visioni etiche diverse. Le 'valutazioni' elaborate ed espresse nei comitati etici su elementi emergenti nella ricerca, consentono una produzione originale, nel confronto con i fatti, le competenze, le visioni valoriali.

3) *Ideali morali non negoziabili come ostacolo alla deliberazione dialogica*. Il documento rileva che le

teorie assolutiste, ossia le teorie che 'pretendono' e impongono valori assoluti ritenuti 'non negoziabili', ossia non suscettibili di discussione critica in quanto devono essere solo accettati da tutti coloro che si confrontano, non fanno parte della deliberazione e delle procedure deliberative. Più precisamente: anche le teorie assolutiste possono partecipare al confronto, ma non possono imporre i valori assoluti, altrimenti annullano il senso del dialogo, in quanto la pretesa rilevanza assoluta respinge il confronto su cui si fonda il dialogo. Coloro che sostengono tali valori assoluti possono 'partecipare' al dibattito 'alla pari degli altri', ossia sullo stesso piano, non in una posizione di preminenza. In questo senso il documento insiste sulla necessità del dialogo tra teorie diverse per raggiungere il "miglior accordo morale possibile", che qualcuno chiama 'minimo etico' ma potrebbe anche dirsi 'massimo etico condivisibile'.

4) *Requisiti etici di una deliberazione*: *la soggettività*. Il documento sottolinea la rilevanza del *coinvolgimento di tutti i soggetti e portatori di interessi* in modo diretto e indiretto, con riferimento sia ai soggetti attivi che ai soggetti passivi, che includono esseri umani e non umani (animali, ambiente, biosfera). Non è esplicitata una posizione antropo-centrica, sensio-centrica, bio-centrica o eco-centrica, ossia non è esplicitata la centralità dell'umano, degli esseri senzienti (umani o animali) o della vita in generale (che include tutti i viventi, umani e non umani), ma nel documento si sostiene che – a prescindere dalla posizione di riferimento – ogni deliberazione debba esprimersi necessariamente sui valori per *tutti* i soggetti coinvolti. In particolare, la deliberazione deve considerare gli effetti e conseguenze delle decisioni considerando i fattori di incertezza, gravità, impatto a medio e lungo termine, tollerabilità e sopportabilità degli effetti, nel bilanciamento rischi e benefici e benefici e costi. In altri termini, la deliberazione pur presupponendo teorie favorevoli alla centralità dell'uomo o all'equivalenza umano-non umano deve porre specifica attenzione sugli interessi di tutti i soggetti nel bilanciamento benefici/rischi in senso complessivo.

5) *La disposizione a riconoscere la complessità*, ossia l'adozione, come punto di partenza dell'analisi, di una 'visione sistemica', ossia uno sguardo d'insieme, di sintesi o 'olistico', che evita di ridursi e chiudersi alla conoscenza analitica del particolare, nella ricerca di un significato che vada oltre il particolare, con riferimento all'interazione tra le parti.

6) *Le caratteristiche della deliberazione.* Nel documento si delineano le caratteristiche specifiche di una deliberazione dialogica, sintetizzabili nei seguenti elementi:
- disponibilità a 'rendere conto', ossia spiegare ed esplicitare le ragioni sottese alle soluzioni proposte e ad esplicitare i principi etici che le giustificano;
- dinamicità, progressività, flessibilità, ossia disponibilità ad adeguare e rivedere le deliberazioni sulla base di nuove conoscenze emergenti, quale fase metodologica essenziale per la ricerca, che è intrinsecamente in evoluzione; la deliberazione non può essere codificata e rigida, ma deve essere sempre aperta a possibili modifiche nel tempo alla luce di nuovi elementi, dati, evidenze che possono emergere;
- preferenza per una deliberazione 'compositiva', piuttosto che 'avversativa' e 'convergente', ossia preferenza per la ricerca di mediazioni tra valori anche diversi, in luogo del conflitto e contrapposizione, o di facili adesioni;
- applicabilità, ossia predilezione per deliberazioni con proposte concrete, di 'misure effettive', traducibili nella prassi da parte dei ricercatori, evitando speculazioni astratte di interesse teoretico ma non utili per 'fare ricerca'.

7) *La funzione dei comitati etici, 'oltre' l'etica*. Di particolare interesse nel documento la duplice funzione del comitato etico, oltre la valutazione etica della ricerca: *una funzione sociale* in quanto valutando la ricerca contribuisce all'avanzamento delle conoscenze per la società e spiegando le ragioni della ricerca consentire di acquisire la fiducia dei cittadini nei confronti della scienza; *una funzione di ispirazione per il diritto e la politica*, in quanto il metodo del confronto con i fatti e con le ragioni, consente di elaborare un quadro

etico di riferimento condiviso, da prospettive diverse, che consente anche di esporre l'esigenza, la giustificazione e il senso di regole e di governance, mostrando che esse non devono mai essere decisioni arbitrarie 'poste' da chi ha il potere di decidere, ma devono interloquire con l'etica, che ne elabora i valori di riferimento.

Dal documento si evince che i comitati etici costituiscono dunque il tempo e lo 'spazio dell'etica' che arricchisce l'etica e la rende preziosa: il modo di procedere di un comitato etico, fondato sul dialogo, fa comprendere i pericoli delle ideologie e degli ideologismi, l'importanza del confronto critico, la rilevanza del contributo delle teorie alla prassi concreta. In questo senso i comitati etici fungono da 'modello' di ragionamento, di interazione e di deliberazione per l'etica pubblica in generale, di particolare importanza oggi a fronte delle incertezze che caratterizzano le trasformazioni sociali e il progresso scientifico e tecnologico. È questo il modello procedurale che consente di 'orientare la produzione normativa e la decisione politica', ossia di fornire una guida al legislatore e al decisore politico, al fine di identificare percorsi e regole per tutti i cittadini, in una società pluralista e multiculturale.

Note in calce al Manifesto per un'etica procedurale approvato il 24 giugno 2024 dall'Istituto di Bioetica per la Veterinaria e l'Agroalimentare

Footnotes to the Manifesto for a Procedural Ethics approved on June 24, 2024, by the Institute of Bioethics for Veterinary Science and Agri-food

Vito Tenore

Presidente di Sezione della Corte dei Conti e docente SNA

DOI: 10.53267/20240202

1. CONSIDERAZIONI INTRODUTTIVE

La tensione verso la ricerca di principi e regole condivise è tipica di ogni ordinamento, di ogni epoca e di ogni società umana: la storia dell'uomo è anche la storia della convivenza tra uomini, fatta di principi e regole. Le *regole* sono norme idonee a guidare il comportamento umano (ad esempio, in un processo, quello del giudice, delle parti, dei loro difensori e degli ausiliari del giudice e delle parti) nel caso concreto; i *principi* enunciano una clausola generica e inclusiva nella quale viene indicato l'obbiettivo da perseguire, e la loro applicazione non è diretta. La loro applicazione implica difatti la deduzione da essi, sulla scorta di un'attività ermeneutica, di una regola da applicare al caso concreto.

Ma come ben colto nel "Manifesto per un'etica procedurale" in esame, ogni condotta umana, in qualsiasi contesto familiare, lavorativo, condominiale, sportivo, istituzionale, è retta, accanto a norme giuridiche (locali, nazionali, sovranazionali, universali) e a regole private (di fonte di regola contrattuale), da un canone basilare: *l'etica comportamentale*. Un canone alto, la cui osservanza non può essere imposta con sanzioni di varia natura, ma solo con la forza della sua obbiettività e accettazione sociale. Compito, quest'ultimo, assai arduo, in quanto, come ben colto nel "Manifesto" CBV-A, la società, soprattutto quella attuale, complessa, multietnica e multivaloriale, è connotata da un profondo pluralismo etico che ingenera, come ben colto, una «difficoltà della costruzione di uno spazio etico condiviso», che «spesso genera la polarizzazione ideologica o, di converso, la banalizzazione frutto del senso comune».

2. RIFLESSIONI SULL'ETICA PUBBLICA E PRIVATA

Ogni lavoro (pubblico o privato), ogni professione, ogni incarico pubblico o privato, ogni gesto umano si fonda su capacità logiche, tecniche e culturali e su un'*etica*, rappresentata non solo da principi e regole condivise dal micro-ordinamento di appartenenza in quanto codificate in precetti legislativi, regolamentari, contrattuali o deontologici, ma, ancor prima, espressa in valori personali e collettivi alla base della civile convivenza. In singolare coincidenza con l'emersione negli ultimi anni, sul piano non solo giudiziario ma anche mediatico, di crescenti episodi espressivi di condotte, lavorative ed extralavorative, "poco etiche" di politici, amministratori e dipendenti pubblici, imprenditori privati, ecclesiastici, militari, liberi professionisti, si assiste a livello scientifico e, dunque, culturale, ad una speculare crescita di studi, con-

vegni, dibattiti sull'etica dei comportamenti pubblici e privati ed alla sofferta gestazione prima nell'impiego privato dell'importante decreto legislativo 8 giugno 2001 n. 231, poi in quello pubblico (ma non solo) della legge 'anticorruzione' (l. 6 novembre 2012 n. 190 e d.P.R. n.62 del 2013 "*Codice di comportamento dei pubblici dipendenti*"), normative ispirate all' «etica ed alla legalità dei comportamenti» e coinvolgente vari soggetti (pubblici e privati) e numerosi micro-ordinamenti sociali.

Da qui anche il proliferare nella pubblica amministrazione, nelle imprese private, nelle libere professioni, nell'ordinamento sportivo, nelle associazioni e gli organi di governo centrali e locali, di codici etici tesi alla autoregolamentazione o alla co-regolamentazione di alcuni settori, fissando principi che devono presiedere al quotidiano vissuto. Ma l'etica sta anche progressivamente divenendo linea guida in molti altri campi: non solo nel lavoro pubblico e privato, nei rapporti con cittadini stranieri, nei rapporti con gli animali, l'ambiente e l'ecosistema (si pensi all'art.9 cost. novellato, ma anche agli studi di ecosofia ed ecoteologia), ma anche nella medicina, nella scienza, nei *media* e nella comunicazione. Il recupero del momento etico non è tuttavia una esigenza estemporanea occasionata dal degrado di alcune importanti fette della compagine sociale in quanto i valori morali e le regole comportamentali sono principi cardine della nostra società (secondo i noti insegnamenti del filosofo Maritain) e testualmente codificati nella nostra Costituzione (e dunque dell'agire di ogni cittadino italiano) incentrata sulla 'personalità umana'. Si pensi all'art. 2 che, nel riconoscere e garantire i diritti inviolabili dell'uomo, anche nelle formazioni sociali (quale è anche un ordine professionale), richiede «l'adempimento dei doveri inderogabili di solidarietà politica, economica e sociale»; si pensi all'art. 41 che, nel riconoscere la libertà di iniziativa economica privata, chiarisce che la stessa «non può svolgersi in contrasto con l'utilità sociale o in modo da recare danno alla sicurezza, alla libertà, alla dignità umana»; si pensi all'art. 54 che impone ai titolari di funzioni pubbliche di svolgerle «con disciplina ed onore»; si pensi all'art. 97 che impone una organizzazione dei pubblici uffici per il perseguimento del «buon andamento e l'imparzialità dell'amministrazione»; si pensi al novello art.9 che tutela l'ambiente, la biodiversità e gli ecosistemi, anche nell'interesse delle future generazioni. Tra le varie concause delle avvertite carenze del sentire etico nell'attuale società, oltre all'individualismo (e relativismo) ed alla esasperata ricerca del lucro, non è poi da escludere una grave carenza formativa sul punto da parte della Scuola e dell'Università, che hanno come scopo ultimo la 'educazione della persona' e non solo la sua 'formazione', intesa come crescita culturale, come hanno ben evidenziato alcuni recenti studi, che rimarcano come l''emergenza etica' altro non sia che un aspetto dell'altra nota emergenza, ovvero l''emergenza educativa'. E lo Stato-Scuola (ma lo stesso vale in parte per il micro-ordinamento-famiglia), a fronte di una pluralità e di una frammentazione di parametri etici sembra aver abdicato alla sua funzione educatrice delle più giovani generazioni, ispirandosi al pericoloso principio di tolleranza quale unico possibile criterio di convivenza pacifica abbandonando ogni pretesa veritativa.

E tale crisi etica va di pari passo con la crisi della legalità, essendo entrambe frutto dell'imperante soggettivismo: una cultura o un ordinamento che prescindano da ogni riferimento ad un ordine valoriale oggettivo restano inevitabilmente privi di condivise regole morali e giuridiche ed originano conflitti di coscienza e scontri sociali. Queste conclusioni, valevoli per l'ordinamento generale (il nostro Paese) sono integralmente trasponibili nei singoli micro-ordinamenti settoriali. Senza addentrarci, nell'economia delle presenti riflessioni, in considerazioni di tipo filosofico, religioso, sociologico o storico sull'etica in generale e nei singoli 'micro-ordinamenti' in particolare, è sufficiente premettere che, per i pragmatici giuristi, condotte poco etiche, nel lavoro o anche nella vita privata, non sempre rimangono sul piano dell''irrilevante giuridico' originando cioè una mera riprovazione morale (quale espressione di stigmatizzabile maleducazione o spregiudicatezza comportamentale), ma spesso le stesse assurgono a rilevanza giuridica, in quanto violano precetti penali, civili o deontologico-disciplinari. Del resto, l'etica ed il diritto positivo non sono così distanti come spesso si ritiene, in quanto l'etica ispira tutte le azioni umane e, quindi, anche quelle giuridica-

mente rilevanti. Non vi è necessaria corrispondenza biunivoca tra condotte non etiche e condotte illegali (ovvero contrarie a norme giuridiche): appare evidente la non eticità e nel contempo l'illegalità di talune condotte ma talvolta la condotta può assumere un disvalore solo morale, secondo parametri fissati in codici deontologici o canoni etici di comune dominio, senza tradursi in un illecito giuridico.

Non sempre, dunque, come evidenziato, la violazione di regole etiche si traduce in sanzioni giuridicamente rilevanti; tuttavia, si assiste ad una progressiva valorizzazione anche giuridica, e non solo nel lavoro pubblico, di tali regole morali, la cui inosservanza si traduce in sanzioni disciplinari interne al micro-ordinamento di appartenenza, il che fa fare un salto logico al precetto violato che, da regola etica, diviene regola deontologica giuridicamente rilevante. Con tale salto giuridico, la violazione etica diviene disciplinarmente sanzionabile secondo le regole, sostanziali e procedurali, del singolo ordinamento.

3. CONCLUSIONI

La profonda e nel contempo concreta intuizione del "Manifesto per un'etica procedurale", che prescinde dal far diventare regola giuridica il proprio spunto metodologico, va ricercata nel proposto «metodo procedurale» di individuazione, selezione e discussione dei valori etici comuni, a prescindere dalle ricadute sanzionatorie nascenti dalla loro violazione. Un metodo necessariamente rigoroso, verificabile e aperto alla discussione pubblica. L'etica ha innegabilmente una vocazione pratica, ma, come ben colto nel "Manifesto", nulla è più pratico di una buona teoria: ed occorre allora una buona teoria per fondare un metodo di lavoro, come detto, rigoroso, verificabile e aperto alla discussione pubblica.

Solo attraverso tale rigore metodologico, si ripete, applicabile anche oltre il campo d'azione dell'Istituto di Bioetica per la Veterinaria e l'Agroalimentare (CBV-A), in ogni contesto aggregativo (dalla famiglia alle Istituzioni), più o meno vasto, si può giungere, andando oltre le mere regole e principi normativi (pur necessari per imporre, anche coattivamente, il rispetto della civile convivenza e del rispetto altrui e delle Istituzioni), ad una condivisione etica di valori che guidino l'*agere* quotidiano in ogni sua manifestazione interpersonale e sociale, privata e pubblica. Il fine da perseguire attraverso tale rigorosa metodica procedurale è l'affidabilità complessiva dei comportamenti umani e dunque la fiducia negli stessi da parte della comunità.

Dunque, il «metodo procedurale» come unico e centrale strumento di emersione di valori etici comuni che siano guida in ogni campo per effettuare scelte: nella politica, nelle Istituzioni, nelle procedure amministrative, nella gestione di una impresa, nello sport, nell'uso dei *media*, nelle relazioni interpersonali in qualsiasi contesto (familiare, condominiale, lavorativo, ludico-ricreativo e persino affettivo). Ogni processo deliberativo assunto secondo le condivisibili regole procedimentali del "Manifesto" in esame, deve basarsi sulla «disponibilità a rendere conto» (plastica e pertinente espressione del "Manifesto") dei criteri e delle ragioni sottese alle scelte/soluzioni/misure proposte e ad esplicitare i principi etici che le giustificano. E tale «metodo» deve inoltre consentire revisioni successive ove emergano problematiche applicative: una sorta di autotutela correttiva per dirla in termini giuridici. Se il metodo proposto dal "Manifesto" CBV-A fosse recepito, metabolizzato e, soprattutto, applicato in questi variegati contesti, l'anarchia valoriale che connota l'attuale società (dove tutti pretendono di far diventare regola il proprio mero portato egoistico), l'estemporaneità delle scelte politiche (mosse da mera ricerca del consenso e non da etica), la disgregazione che connota una società 'fragile' (frutto di perdita di valori comuni, frutto cioè di un'etica condivisa), la confusione valoriale (nascente da eccesso di fonti di disinformazione sovente guidate da regie occulte), il dilagante pensiero ispirato al 'politicamente corretto' (che esprime in realtà scelte facili e buoniste incapaci, come tali, di educare la società verso solidi valori), forse verrebbero messe in discussione e rimeditate, con conseguente emersione di un distillato di valori etici, base per una rinascita dei valori fondanti della società, più semplici, veri e condivisi.

E questo *distillato etico*, frutto del confronto attraverso il rigoroso metodo procedurale suggerito dal "Manifesto" CBV-A, non può che essere affidato ad Uomini e Donne e non certo ad Intelligenze Artificiali, che essendo delle mere *res* (ovvero cose) sono incapaci di partorire valori etici se non rielaborando pensieri ed idee già formulati da esseri umani. Ma anche se lo fossero, sarebbe triste e mortificante che regole etiche per scelte umane venissero elaborate da

una mera *res non senziente*. Forse, alcune regole etiche potremmo in modo più affidabile trarle dall'osservazione studio delle innate e basilari condotte degli animali e della natura, anch'esse *res* sul piano meramente giuridico, ma innegabilmente vive e senzienti e capaci, se ben osservate, di offrirci regole etiche che l'Uomo sta progressivamente perdendo.

Recensioni

Franco Basaglia

a cura di Marica Setaro

Fare l'impossibile. Ragionando di psichiatria e potere

Donzelli, 2024
ISBN: 9788855225793
pp. 144

PAOLO SAVOIA
paolo.savoia3@unibo.it

AFFILIAZIONE
Università di Bologna

DOI: 10.53267/20240301

In "Fare l'impossibile. Ragionando di psichiatria e potere", la curatrice Marica Setaro, storica della psichiatria, ha raccolto tre inediti di Franco Basaglia (o meglio, tre inediti che sono stati composti da un autore collettivo che va sotto il nome di Franco Basaglia), risalenti agli anni 1968-72.

In un periodo di celebrazioni, in cui molte sono state le iniziative per ricordare il centenario del 'liberatore dei matti', il volume – sia negli inediti sia nella ricchissima introduzione di Setaro – ci aiuta a prendere di petto la questione della *celebrità* di Basaglia, iniziata quando egli era già in vita, che fa il paio con la sua *monumentalizzazione* attuale. Il meccanismo della celebrità, che non risparmia le celebrità scientifiche, è un meccanismo riduzionista: i processi storici collettivi – sociali, intellettuali, politici – vengono invisibilizzati a favore di un nome, di un individuo, di una personalità geniale. Il risultato è che invece di percepire il mutamento storico si vede solo l'eccezionalità, un'eccezione che non turba di fatto la norma, proprio in quanto eccezionalità. Questo è il «fantasma» di Basaglia di cui lo psichiatra stesso parla nel primo inedito, la registrazione di un incontro con un collettivo studentesco padovano (p. 60). La vicenda di Basaglia ci parla della questione generale della 'scienza' nel suo rapporto con la democrazia, intesa come uguaglianza delle possibilità di partenza e partecipazione a tutti i processi politici e decisionali, dentro e fuori le istituzioni. Setaro lo scrive bene in un bel passaggio della sua Introduzione: «Esiste […] una responsabilità più grande che investe una dimensione comune e che sollecita un nervo scoperto: il legame tra scienza e democrazia nell'attuazione di politiche, azioni, conoscenze, metodi, norme che realizzino il processo costituente per il diritto alla salute, e per il diritto a costruire una nuova salute mentale» (p. 14).

Su questo nodo tematico dei rapporti tra scienza – meglio sarebbe dire scienze, meglio ancora sarebbe parlare di scienze mediche – e democrazia si possono dire molte cose e si rischia anche di fare molta confusione. Certo, è inevitabile pensare alla distanza che ci separa dal periodo in cui Basaglia – ma anche gli altri anti-psichiatri, o i difensori della psichiatria di settore, insomma gli attori che caldeggiavano una riforma radicale della psichiatria e della medicina tout-court negli anni '60 e '70 del XX secolo – e il presente. Una distanza storica che è anche utile misurare per comprendere bene cosa è stato quel movimento. Mi riferisco qui all'immagine pubblica dello scienziato, e specificamente del medico, che si è affermata nell'immaginario culturale italiano durante la crisi pandemica, ovvero un immaginario esplicitamente autoritario della medicina, animato dalla postura del medico che rivendica il suo 'la scienza non è democratica', il suo 'se non avete studiato dovete stare zitte'.

Basaglia usa spesso le virgolette quando parla e scrive di 'scienza'. Nell'ultimo intervento pubblicato da Setaro – *Donne, psichiatria, potere* – Basaglia parla del doppio legame politico che intrappola quella che oggi chiameremmo una forma della maschilità: l'uomo nei confronti della donna ha un ruolo di potere, anche se nella trama dei rapporti sociali è oggetto del potere: «un escluso sociale e un escludente individuale», scrive. E conclude: «il che significa che [l'uomo] deve negare in sé … la faccia del potere con cui agisce nei confronti della donna» (p. 141). Poco prima, Basaglia aveva messo in chiaro che sussiste un'analogia tra questo rapporto di potere complesso, tra uomo e donna, e quello che si instaura tra lo psichiatra e il malato di mente.

Come non pensare qui anche al rapporto che sussiste tra la scienza medica e la democrazia? Basaglia, da scienziato, da medico, parla di una psichiatria che si fa difesa dell'ordine sociale – sia nell'istituzione manicomiale sia al di fuori, come gestione delle devianze e agenzia di integrazione – ma al tempo stesso, proprio perché lui continua a fare lo psichiatra e il medico, parla di una scienza del futuro: qui sta il senso di quelle virgolette con cui spesso racchiude la parola scienza. Storicamente, è noto, gli psichiatri si sono spesso chiesti come rendere la loro disciplina indipendente dalle esigenze dell'ordine sociale; Basaglia in questo senso è uno dei tanti. Questa contraddizione, o forse questa tensione, tra una scienza che è – come ripete spesso – difesa degli interessi delle classi dominanti e la possibilità di un sapere nuovo, di una tecnica nuova, emerge benissimo e mi sembra la cifra con cui leggere questi interventi, e forse buona parte dell'opera pratica e teorica di Basaglia.

Circa trent'anni fa, Funtowicz e Ravets parlavano di 'scienza post-normale' (1993), ovvero di quelle scienze che in momenti di crisi devono potersi aprire ai saperi non-scientifici, che sono pur sempre saperi, per prendere decisioni difficili e diventare veramente pluraliste, democratiche. Siamo oggi molto distanti, almeno nel campo della medicina e della psichiatria, come sottolineano anche Mario Colucci e Pierangelo Di Vittorio in un altro volume recente dedicato a Basaglia[1], dalla creazione di spazi del genere, anche se forse non così distanti nel campo delle scienze del clima e dell'ambiente. Emerge infine dal volume il problema del mancato radicamento accademico della psichiatria critica di quegli anni.

Setaro ha fatto un lavoro prezioso nella sua introduzione, ha scritto un capitolo di microstoria dell'anti-psichiatria – chiamiamola così per comodità – che abbatte il *monumento* Basaglia, la *celebrità* Basaglia, il *fantasma* Basaglia e restituisce tutti «i dubbi, le incertezze, gli errori, i limiti, le angosce di Basaglia e di chi con lui scavava nelle crepe di una realtà complessa» (p. 42).

Come diceva il grande storico della medicina Georges Canguilhem, quando si esce dalla facoltà di psicologia da una parte si va al Pantehon, dove sono seppelliti alcuni grandi personaggi; dall'altra invece si va alla prefettura di polizia.

NOTE

1. Mario Colucci e Pierangelo Di Vittorio, *Basaglia*, Feltrinelli, 2024.

Bart Schultz

Utilitarianism as a way of life. Re-envisioning planetary happiness

Polity, 2024
ISBN: 150955226X
pp. 224

LEONARDO URSILLO
lursillo@luiss.it

AFFILIAZIONE
Luiss Guido Carli

DOI: 10.53267/20240302

Bart Schultz è uno dei più noti e apprezzati storici dell'utilitarismo attualmente in attività e questo suo nuovo testo, "Utilitarianism as a way of life", è solo l'ultimo di una lunga serie di studi da lui dedicati alla storia dell'etica utilitaristica.[1] Ma a differenza degli altri suoi lavori, questo si distingue per una caratteristica peculiare: si confronta con il presente. È come se Schultz, dopo averci raccontato, servendosi di ampie documentazioni, la vita dei più importanti autori dell'utilitarismo classico, avesse deciso di costruire un ponte fra gli scritti di questi autori e i problemi della nostra contemporaneità, sui quali, evidentemente, egli sente il bisogno di esprimersi. La fabbrica della felicità che dovrebbe caratterizzare la nostra società occidentale, stando almeno al parere di alcuni contemporanei, appare a Schultz difettosa per i risultati che produce, e dunque da rivedere. Di che risultati si tratta? «Epidemie di solitudine, isolamento, ansia, depressione, rabbia, risentimento, paura, disturbi da deficit di natura», dovuti al fatto che le persone, soprattutto i bambini, stanno passando (rispetto ai loro progenitori) meno tempo all'aria aperta, un fatto che secondo alcuni studiosi potrebbe provocare diversi disagi comportamentali.[2] Tutto questo viene mascherato «da misure sottili e poco informate di felicità e salute», mentre un gran numero di «persone si chiedono perché dovrebbero sentirsi bene» quando non fanno altro che lottare solo per andare avanti, facendo «lavori di merda [*bullshit jobs*], o facendo parte del precariato»; viviamo dunque in un sistema che ha «prodotto dolore diffuso e morti per disperazione»[3]. La nostra stessa vita sentimentale, così come le nostre emozioni, sarebbero in pericolo: «Nel vortice di smartphone e grandi schermi sempre più potenti, l'assalto continuo del mondo opulento alla gentilezza e alle relazioni significative è avvertito da molti, ma tragicamente minimizzato dalle élite e dai signori della tecnologia»[4]. Secondo le più recenti stime, decine di milioni di americani stanno soffrendo a causa di questi disagi (e possiamo immaginare dati analoghi anche per l'Europa); il mito della ricca felicità della società occidentale sembra diventato una «barzelletta malata»[5]. Per un utilitarista questo è un problema che non può essere ignorato.

Storicamente parlando, l'utilitarismo appartiene a quell'insieme di dottrine dette consequenzialiste, per cui le azioni sono giuste se le conseguenze sono buone e ingiuste se le loro conseguenze sono cattive, pertanto dovremmo cercare di attenerci alla prima opzione. L'utilitarismo però richiede di *massimizzare* le conseguenze buone; non basta che il saldo netto di felicità sia buono (numericamente parlando), deve essere massimo; dobbiamo quindi ottenere il miglior risultato possibile. Per questo alcuni autori hanno criticato l'utilitarismo accusandolo di essere una teoria etica troppo esigente (*demanding*)[6]. Se non si coglie questo aspetto, le preoccupazioni di Schultz sembreranno esagerate.

Dopotutto se la maggior parte delle persone si trovasse a suo agio nel contesto poco sopra descritto, senza provare alcun disagio, nonostante vi sia una certa dose di sofferenza patita da una cospicua, ma minore, fetta della popolazione mondiale, il saldo delle buone conseguenze, o della felicità, rimarrebbe comunque positivo. Ma se la nostra priorità è massimizzare quel saldo, dobbiamo preoccuparci del livello di felicità complessivo, migliorandolo ancora di più. Perciò bisogna occuparsi di quei disagi e, se necessario, ripensare la felicità planetaria, intervenendo sul nostro stile e sistema di vita[7]. Di fronte a questa difficoltà, Schultz richiama l'attenzione del lettore su alcune delle pagine più celebri degli autori legati all'utilitarismo classico del XVIII e XIX secolo, mettendo a confronto le loro idee su questioni politiche, educative, sociali e sentimentali[8]. Un simile confronto è molto utile per comprendere le differenze teoriche, ma non solo, che hanno caratterizzato le idee dei diversi utilitaristi, spesso in contrasto fra loro, come nel caso di J. Bentham e W. Godwin, passando per J. S. Mill e H. Sidgwick[9]. Schultz si confronta con questi autori e con le problematiche che erano chiamati ad affrontare nel loro periodo storico, cercando di trovare qualche riflessione in grado di offrire un sostegno ai problemi, analoghi e allo stesso tempo diversi, con cui dobbiamo misurarci nel nostro presente. Questo però non è l'unico scopo del suo lavoro. Infatti, attraverso la sua ricostruzione storica, Schultz vuole anche sfidare alcuni dei più noti luoghi comuni che hanno da sempre contribuito ad alimentare la cattiva fama dell'utilitarismo. Quest'ultimo viene infatti «condannato per aver sostenuto che il fine giustifica i mezzi, igno-

rando la separatezza delle persone, fallendo miseramente nel supportare i diritti fondamentali e la giustizia distributiva, e giustificando il sacrificio di persone innocenti per un bene più grande»[10]. Oggigiorno ci sono persino alcuni entusiasti sostenitori del «libero mercato che invocano una caricatura dell'utilitarismo» a supporto delle loro ricette economiche[11]. Da questo punto di vista, l'opera di Schultz riesce a rispondere a gran parte di queste critiche, ma non a tutte. Rimane infatti un problema di fondo che egli giudica estremamente rilevate, il sostegno che in alcuni casi gli utilitaristi hanno offerto al colonialismo britannico. Storicamente parlando è comprensibile che in un'epoca così diversa dalla nostra vi siano state prese di posizione non sempre critiche nei confronti dell'imperialismo da parte di questi autori. Per questo motivo Schultz parla di «decolonizzare» l'utilitarismo, decentrandolo rispetto alla sua matrice eurocentrica e occidentale[12]. Anche su questo punto non tutti gli utilitaristi di quel periodo storico la pensavano allo stesso modo, ed è di nuovo qui che il lavoro di Schultz riesce a farci comprendere le differenze fondamentali che contraddistinguono questi autori, arricchendo la nostra memoria storica, senza mancare di aiutarci a prendere una posizione critica nei confronti dell'attualità.

NOTE

1. Schultz B., *The Happiness Philosophers. The Lives and Works of the great Utilitarians*, Princeton University Press, 2017, e dello stesso autore si veda anche, *Henry Sidgwick - Eye of the Universe: An intellectual Biography*, Cambridge University Press, 2004 e *Utilitarianism and Empire*, edited by Bart Schultz and Georgios Varouxakis, Lexington Books, 2005.

2. Schultz B., *Utilitarianism as a way of life. Re-envisioning planetary happiness,* Polity Press, Cambridge and Hoboken, 2024, p. 18.

3. Ibidem.

4. Ibidem. Questi dati sono stati confermati anche dal recente lavoro di Jonathan Haidt, *The Anxious Generation: How the Great Rewiring of Childhood is Causing an Epidemic of Mental Illness*, Penguin Book, London 2024.

5. Ivi, p 19.

6. Rawls J., *A Theory of Justice*, Harvard University Press, Cambridge 1971, pp. 286-287.

7. Schultz B., *Utilitarianism as a way of life*, cit., pp. 205-206.

8. Ivi, pp. 26-56.

9. Ivi, pp. 152-204.

10. Ivi, p. 22.

11. Ivi, p. 213.

12. Ivi, p. 23 e pp. 34-35.

Daniele Caligiore

Curarsi con l'Intelligenza Artificiale

Il Mulino, 2024
ISBN: 8815388591
pp. 168

ANTONIO MALVASO
antonio.malvaso01@universitadipavia.it

AFFILIAZIONE
IRCCS Fondazione "*C. Mondino*"
Istituto Neurologico Nazionale

DOI: 10.53267/20240303

Il volume a cura di Daniele Caligiore, "Curarsi con l'Intelligenza Artificiale", realizzato nel 2024 ed edito il Mulino, rappresenta una concreta e indispensabile fruibilità di un tema che costituisce una delle sfide più impegnative per l'umanità nel secolo presente – La Medicina e L'intelligenza Artificiale (IA).

Nel libro l'autore affronta un universo parallelo in arrivo verso la nostra quotidianità e che suscita sia speranze enormi che timori. La salute è un tema cruciale e capire come le nuove tecnologie possano contribuire a guarire, migliorare il nostro benessere e rendere la vita più sana è fondamentale. In questo libro, grazie all'esperienza dell'autore nel campo dell'IA e all'aiuto di esperti del settore medico, vengono esplorati i vantaggi e le criticità dell'intelligenza artificiale applicata alla medicina, fornendo gli strumenti necessari per affrontare con consapevolezza una rivoluzione destinata a trasformare il nostro modo di curarci.

Da medico in formazione specialistica in neurologia, mi sono avvicinato al campo dell'intelligenza artificiale (IA) qualche anno fa, con la consapevolezza che queste nuove tecnologie stanno ridisegnando profondamente il panorama della medicina. Pertanto, reputo di fondamentale importanza i temi trattati nel libro. Con questa intima riflessione lancio un appello a tutti i medici, quelli in formazione specialistica e i medici del futuro, poiché l'IA è destinata a diventare un alleato imprescindibile nel nostro lavoro quotidiano.

Entrando nel dettaglio vivo di questo affascinante viaggio, nella prima parte, l'autore introduce i concetti fondamentali dell'IA, come il *machine learning* e il *deep learning*, e ne esplora le applicazioni storiche in medicina. L'accento viene posto sull'importanza della digitalizzazione dei dati, che ha aperto nuove frontiere, dalla diagnostica per immagini alla chirurgia robotica, fino alla telemedicina, con un focus particolare sugli esoscheletri e i cyborg, che rappresentano il futuro dell'assistenza e della riabilitazione. Inoltre, Daniele Caligiore esplora le straordinarie potenzialità dell'IA nel migliorare la diagnosi, il trattamento e la prevenzione delle malattie.

L'autore dunque ci accompagna a comprendere l'analisi di enormi quantità di dati che consentono la creazione di modelli personalizzati e la simulazione di scenari virtuali, come nel caso dei gemelli digitali ("digital twins") e del metaverso. Questi strumenti offrono la possibilità di simulare il corpo umano (e anche modelli animali virtuali) e testare terapie su misura, migliorando la prevenzione e ottimizzando i trattamenti.

Successivamente, nel quarto capitolo l'autore si concentra sulla medicina del futuro come una contaminazione di competenze e discipline. L'IA non è vista come una tecnologia isolata, ma come un catalizzatore di innovazione che può integrarsi con altre competenze, favorendo un approccio più olistico e personalizzato alla medicina. L'interdisciplinarità è, infatti, fondamentale per una comprensione più profonda dei processi biologici e per la realizzazione della medicina di precisione, che tiene conto di fattori genetici, ambientali e comportamentali.

Tuttavia, nell'ottica di darci un valido contro-scenario, nelle parole dell'autore non mancano le preoccupazioni legate alla privacy dei dati e alla responsabilità delle decisioni algoritmiche (il concetto di *explainability*), temi affrontati nel dettaglio. L'autore in particolare si sofferma su questioni etiche cruciali come la sostituzione dei medici e il rischio di perdere il controllo sulle decisioni fondamentali per la salute.

In questo contesto, vorrei appunto sollevare importanti questioni che per il momento restano irrisolte. Se da un lato le potenzialità dell'IA in medicina sono straordinarie, dall'altro emergono interrogativi più profondi sul significato di 'cura' e sul ruolo dell'uomo nella medicina. Se l'IA è in grado di diagnosticare malattie con una precisione superiore a quella di un medico, a cosa serve la nostra presenza, il nostro giudizio, il nostro 'tocco umano'? La medicina ha sempre implicato una relazione empatica tra medico e paziente, una dimensione che la tecnologia non può replicare. La domanda che si pone, quindi, è: possiamo davvero affidare all'IA il compito di prendersi cura di noi, o rischiamo di perdere il significato profondo della cura, che non è solo un atto tecnico, ma anche un atto di comprensione e di relazione umana?

Pertanto, il manoscritto di Daniele Caligiore affronta queste doman-

de, invitando il lettore a riflettere sulle implicazioni etiche, sociali e filosofiche dell'introduzione dell'IA nella medicina e suoi potenziali benefici. La sfida è quella di trovare un equilibrio: sfruttare i vantaggi dell'IA per migliorare la qualità delle cure, senza sacrificare quella dimensione umana che è essenziale per una medicina veramente integrata e centrata sulla persona.

L'autore, nell'ultima parte del libro, ci concede una profonda riflessione sulla necessità di formare medici e pazienti per un uso consapevole dell'IA, garantendo che, pur avanzando nella tecnologia, non perdiamo mai di vista l'importanza dell'etica, della privacy e, soprattutto, dell'umanità nella cura della salute. Questo libro rappresenta una lettura fondamentale per chiunque voglia capire come l'intelligenza artificiale trasformerà non solo la medicina, ma anche il nostro concetto di salute e benessere.

Consulta Scientifica del Cortile dei Gentili

C. Caporale, L. Palazzani (a cura di)

Dialogo sul suicidio medicalmente assistito

Cnr Edizioni, 2024
ISBN: 9788880806479
pp. 122

FABIO MACIOCE
f.macioce@lumsa.it

AFFILIAZIONE
LUMSA, Roma

DOI: 10.53267/20240304

Il testo in oggetto è, sotto molteplici aspetti, di notevole interesse, e si inserisce con competenza e precisione in un dibattito non solo complesso, ma reso spinoso da forti pressioni di tipo ideologico. Il suicidio medicalmente assistito rappresenta infatti uno dei temi più complessi e controversi del dibattito contemporaneo, sia dal punto di vista etico che giuridico. La questione – come è noto – ruota attorno alla possibilità, per un individuo affetto da una malattia incurabile e in condizioni di sofferenza insopportabile, di scegliere autonomamente di porre fine alla propria vita con l'assistenza di personale medico. Si tratta dunque, con tutta evidenza, di un tema che tocca gli aspetti più profonde della dignità umana, della libertà personale e del valore della vita, e suscita reazioni comprensibilmente opposte fra loro, accendendo un dibattito politico, giuridico, religioso ed etico in molte società contemporanee.

I nodi del dibattito sono noti, ma può esser utile riassumerli brevemente. In primo luogo, una questione cruciale è la possibile – e non banale – distinzione tra suicidio assistito ed eutanasia. Si ritiene in genere che nel primo caso sia il paziente a compiere personalmente l'atto che porta alla morte, seppure con il supporto del medico, ad esempio attraverso la prescrizione di farmaci letali. Nel secondo caso, al contrario, è il medico a intervenire direttamente per causare la morte del paziente. Questa differenza, sebbene sottile, ha implicazioni etiche e giuridiche significative, in quanto molti ordinamenti giuridici riconoscono al suicidio assistito una maggiore compatibilità con il principio dell'autodeterminazione, rispetto all'eutanasia.

Dal punto di vista etico, il suicidio medicalmente assistito pone interrogativi altrettanto complessi. Da un lato, c'è chi sostiene che il diritto all'autonomia e alla dignità debba estendersi fino ad includere la possibilità di decidere quando e come terminare la propria vita, specialmente in situazioni di sofferenza estrema. In questa prospettiva, impedire il suicidio medicalmente assistito implica imporre una continuazione della vita contro la volontà del paziente stesso, e costituisce una forma di crudeltà o comunque una violazione della sua libertà di autodeterminarsi. Dall'altro lato, si ritiene che ogni forma di intervento, anche indiretto, volto a porre fine alla vita umana sia eticamente ingiustificabile, non soltanto per (legittimi) motivi religiosi o morali, ma anche perché intrinsecamente contrario all'orientamento della pratica medica alla tutela della vita. A queste ragioni di tipo teoretico si aggiungono spesso timori pratici, come il rischio di abusi o pressioni indebite su pazienti vulnerabili, che potrebbero sentirsi obbligati a scegliere il suicidio assistito per non essere un peso per i propri familiari.

Il panorama legislativo è dunque estremamente variegato. In alcuni paesi, il suicidio assistito è consentito entro rigorosi limiti legali e protocolli medici. In altri, è vietato e punito penalmente. In Italia, in particolare, il dibattito si è recentemente molto intensificato a seguito di alcuni casi emblematici e, soprattutto, di una pronuncia della Corte Costituzionale, che ha da un lato sollevato l'urgenza di una regolamentazione chiara e bilanciata, e dall'altro ha stabilito l'indebita riduzione di questa pratica al reato di istigazione al suicidio, previsto nel nostro Codice penale. Come è noto, purtroppo, la decisione della Corte non ha ancora trovato una risposta da parte del Parlamento.

Ora, il testo in oggetto si inserisce con puntualità all'interno di questo dibattito, contribuendo a precisare alcuni aspetti meritevoli di attenda considerazione e prendendo posizione in modo interessante su alcuni dei punti più critici. È bene dire, anzitutto, che il documento proposto dalla Consulta scientifica del Cortile dei Gentili prende le mosse dalla recente evoluzione normativa italiana, e in particolare dalla legge sul consenso informato e le disposizioni anticipate di trattamento (l. 219 del 2017) prima, e dalla sentenza della Corte costituzionale n. 242 del 2019. Indubbiamente, tanto la legge quanto la decisione della Corte hanno contribuito a modificare l'orizzonte del dibattito, contribuendo a dare un fondamento normativo solido all'autonomia del paziente. La legge del 2017 sul consenso, in particolare, ha ribadito che il consenso informato del paziente costituisce l'espressione primaria della tutela dell'autonomia personale, e che sulla garanzia di tale autonomia si fonda la disciplina della pratica medica: il consenso infatti non solo rileva quanto alla

legittimità dei singoli trattamenti e interventi, ma anche quanto alle modalità di attuazione della relazione medico-paziente. Dalla diagnosi alla terapia alla prognosi, ogni atto medico deve essere praticato nell'ambito di una relazione che, per legge, deve essere di tipo cooperativo, ovvero una interazione nella quale l'autonomia del paziente e la competenza professionale e la responsabilità del medico si incontrano su un piano di parità (art. 1 comma 2). Analogamente, nella sentenza n. 242 del 2019, con cui la Corte ha identificato i criteri di non punibilità dell'assistenza medica al suicidio, il rispetto dell'autonomia del paziente sulla prosecuzione delle cure e della propria vita assume un rilievo centrale.

Va notato, in questo contesto, che il testo in oggetto effettua un approfondimento molto opportuno sul tema della sofferenza intollerabile, che come è noto rappresenta uno dei criteri che la Corte costituzionale ha indicato nella sentenza (accanto all'irreversibilità della patologia, alla competenza, e alla dipendenza da trattamenti di sostegno vitale). Qui la Consulta scientifica coglie un punto davvero cruciale, a mio parere, determinato dal difficile bilanciamento fra la necessità di stabilire un criterio chiaro e oggettivo, da un lato, e la pericolosità di formule troppo rigide o troppo vaghe, dall'altro. Si tratta di una questione non nuova, da un punto di vista concettuale, ma che nella drammaticità delle decisioni di fine vita assume un rilievo tutto particolare; e si tratta di una questione sulla quale il testo prende una posizione meritevole di grande apprezzamento.

Come sanno i giuristi, in iure omnis definitio periculosa est. Ma quando si tratta di decisioni di fine vita, definizioni troppo rigide, o troppo generiche, di cosa sia una sofferenza 'intollerabile', sarebbero impossibili ed estremamente rischiose. Troppo variabili sono le percezioni soggettive, e molto limitatamente comunicabili a terzi, perché si possa dare una definizione chiara e precisa di questa soglia di intollerabilità; ma troppo grande è il rischio di scelte arbitrarie, perché si possa semplicemente rinunciare ad ogni definizione. La prospettiva della Consulta scientifica, sul punto, è molto prudente, e allo stesso tempo molto lucida.

Da un lato, si mettono in luce i rischi e le molte difficoltà derivanti da una valutazione della sofferenza psichica, oltre che della sofferenza psicologica, perché essa è certamente ancor più difficile da categorizzare e definire. D'altro canto, si dà conto del rilievo che nella valutazione della sofferenza – anche fisica – è necessario dare alle speranze e paure soggettive, e in generale agli aspetti emotivi legati alla situazione di sofferenza. Molto saggiamente, si ribadisce che un possibile equilibrio possa essere trovato solo nel contesto, perché solo nel contesto (potremmo dire: nel caso concreto, nell'esperienza della singola persona e del singolo paziente) si può valutare con ragionevolezza il peso che nella decisione hanno tutti questi elementi. La paura o il senso di colpa per l'eventuale carico emotivo ed economico che la malattia terminale può avere sui propri familiari va tenuto ben distinto dalla esperienza di una sofferenza intollerabile, così come è solo nel contesto che il personale sanitario può cogliere il peso dei molti fattori che conducono una persona a chiedere di essere aiutata a morire.

Ciò che il documento sottolinea, insomma, è che una normazione sul suicidio medicalmente assistito non debba trasformare la gestione della fase finale della vita umana in una procedura burocratica, ma che tale scelta debba sempre avvenire all'interno di una relazione di cura, in cui possa essere percepita come affermazione della dignità della persona e non come una forma di abbandono terapeutico. Si tratta anche in tal caso, come ben si comprende, di un problema non nuovo, e certo non limitato alle questioni di fine vita. In fondo, tutto il dibattito e l'evoluzione della normativa sul consenso informato è testimonianza di questa preoccupazione. Molto si è detto su quanto sia necessario integrare il modello corrente di consenso informato superando tanto il suo estremo proceduralismo, quanto l'essere fondato su un paradigma apparentemente neutrale, secondo il quale individui perfettamente autonomi, una volta forniti di informazioni adeguate e di un tempo necessario all'elaborazione delle scelte, possono prendere decisioni in relazione ad atti terapeutici, che saranno in grado di riconoscere come *proprie*. Non è una mera dichiarazione di consenso, con tutta evidenza, che garantisce l'autonomia reale delle scelte. A tal fine, è necessario ripensare il consenso in una direzione relazionale, ovvero tale da collocarlo all'interno di contesti comunicativi, la cui forma, modalità, incidenza siano adeguatamente tenuti in considerazione. Il processo del consenso

informato, e nella stessa prospettiva anche la richiesta di suicidio medicalmente assistito, chiedono di essere tematizzate tenendo conto, oltre che dei contenuti informativi, del contesto all'interno del quale sono attuate, delle relazioni di potere, dei contesti simbolici e di discorso che le sostengono e le modellano, poiché tutto ciò dà forma alle effettive possibilità di scelta della persona, e alle sue effettive possibilità di esercitare la propria libertà.

Questa valorizzazione del contesto e delle specifiche condizioni fisiche, psichiche e relazionali del paziente sono cruciali anche per la distinzione fra suicidio assistito e eutanasia.

La distinzione, opportunamente richiamata dalla Consulta in questo documenti, ha a che fare con l'intenzionalità dell'atto medico. L'eutanasia consiste nell'atto, da parte di un medico o altra figura sanitaria, di somministrare intenzionalmente una sostanza letale al fine di porre fine alla vita di un paziente che lo richiede, laddove nel suicidio medicalmente assistito il medico si limita a fornire al paziente i mezzi per porre fine alla propria vita autonomamente, ma l'azione finale è compiuta dal paziente stesso. Tale distinzione, chiara nella teoria, può essere molto difficile da tracciare nella pratica. Il documento, sottolineando il rischio del 'pendio scivoloso', ribadisce però la centralità di tale distinzione, e l'utilità dei criteri stabiliti dalla Corte costituzionale proprio per tracciare questo confine fondamentale. Anche in tal caso, la differenza fra il legittimo rispetto della volontà soggettiva e dell'autonomia, e una pericolosa valutazione esterna sulla dignità del vivere, può avvenire solo nel contesto: è solo una decisione non burocratizzata che può evitare fenomeni di marginalizzazione e esclusione sconcertanti, e distinguere casi in cui va rispettata l'autonomia della persona da casi in cui, dietro una scelta apparentemente autonoma, si celano pressioni indebite e fenomeni di marginalizzazione e abbandono dei soggetti più fragili.

In questa stessa prospettiva, il testo molto opportunamente richiama il ruolo dei Comitati etici: nel testo la necessaria implementazione di tali strutture è molto sottolineata, e a ragione. La valutazione di contesto, più volte richiamata, implica non solo una discussione sulla presenza degli elementi rilevanti, dei requisiti, la necessaria offerta di cure palliative, ma in generale richiede un accompagnamento del paziente nella scelta, ad opera di professionisti di diversa competenza ed esterni all'equipe sanitaria. In questo punto, a mio avviso, v'è un aspetto sul quale è necessaria una riflessione attenta. Se sono certamente convincenti i richiami della Consulta al necessario coinvolgimento dei comitati per l'etica clinica, va considerata l'effettiva applicabilità di questa proposta nel contesto attuale. Mi pare, in altre parole, che sia da tempo crescente, in Italia, una tendenza a limitare il più possibile ogni interferenza con l'autonomia individuale, quando sono in gioco scelte personalissime in ambito clinico. In tale prospettiva, la proposta di far passare del tempo fra la richiesta di suicidio medicalmente assistito, e la sua materiale esecuzione, potrebbe andare incontro a critiche simili a quelle già avanzate in merito all'analogo periodo 'cuscinetto' richiesto per l'interruzione di gravidanza. Pur nella diversità di situazioni, mi pare che la tendenza oggi prevalente sia quella di considerare di principio indebite tutte le forme di interferenza rispetto alle richieste della persona, quasi si trattasse di strategie per rendere più faticosa (e non più ponderata) la scelta effettuata. In altre parole, mi pare che la sfida che i comitati per l'etica clinica dovranno cogliere, se vorranno intervenire in procedure di suicidio meramente assistito, è quella di evitare di trasformarsi in meri sportelli informativi per la comunicazione di processi, procedure, diritti, informazioni: tutte cose necessarie, certamente, ma non paragonabili all'importanza di un vero e proprio accompagnamento nella decisione, che come si è visto è cruciale sotto molti aspetti.

Un ultimo aspetto mi pare meriti di essere sottolineato. Il documento in oggetto coglie con grande lucidità un aspetto spesso lasciato ai margini, nelle discussioni sul fine vita e sul suicidio medicalmente assistito: la posizione di vulnerabilità dei *care-givers*, ovvero delle persone che accudiscono e si prendono cura del paziente. Se certamente la posizione e la vulnerabilità del paziente hanno una indubbia priorità, e con maggiore evidenza sono al centro della scena, molti autori e autrici hanno

evidenziato la necessità che la gestione della dipendenza e la risposta alla vulnerabilità siano obiettivi prioritari per le istituzioni pubbliche, fuori dal paradigma liberale dell'indipendenza e dell'autonomia soggettiva. Ciò che spesso sfugge, difatti, è che i soggetti che si dedicano alla cura dei pazienti terminali (ma, analogamente, di tutte le persone non autonome) sono essi stessi oggetto di processi di vulnerabilizzazione, ed è dunque prioritario che anche di costoro ci si faccia carico. Con una precisione non comune, il documento della Consulta invita a prendersi cura «anche di coloro che assistono i pazienti in situazioni particolarmente difficili e per lungo tempo»: si tratta di quella che la riflessione recente ha definito 'nested dependency', dipendenza annidata dentro un'altra dipendenza, e mette in luce come la relazione tra il care-giver e la persona dipendente sia ambivalente, almeno nella misura in cui l'assunzione di responsabilità per una persona vulnerabile può a sua volta generare vulnerabilità. Se i *care-givers* non sono oggetto di 'cura' da parte del sistema pubblico, la loro vulnerabilità (economica, lavorativa, emotiva, personale) diventa un problema tanto grande quanto invisibile. C'è da sperare che questo testo, così autorevole e lucido, possa contribuire a orientare il dibattito anche su questi problemi, spesso marginalizzati.

Davide Battisti

Procreative Responsibility and Assisted Reproductive Technologies

Routledge, 2024
ISBN: 9781032652085
pp. 242

MARCO ANNONI
marco.annoni@cnr.it

AFFILIAZIONE
Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca (CID Ethics)

DOI: 10.53267/20240305

In uno dei saggi conclusivi, dedicato al ragionamento morale e ripubblicato nel volume *Sulla Morale Politica*, Richard M. Hare osserva che i «filosofi si occupano soprattutto di argomenti, di come distinguere quelli buoni da quelli cattivi; e per farlo dispongono di tecniche peculiari che rientrano tutte quante in senso lato nella logica»[1]. Se diamo credito a questa visione di Hare, un buon saggio di filosofia morale e di etica applicata dovrebbe proporre una serie di argomenti ben costruiti, con termini chiaramente definiti e passaggi inferenziali esplicitati. Inoltre, dovrebbe includere un'analisi critica delle posizioni presenti nella letteratura, evidenziandone meriti, limiti, fallacie, assunzioni di partenza e possibili implicazioni. In particolare, nel campo dell'etica applicata, tale lavoro dovrebbe consentire al lettore di riflettere in modo più chiaro su questioni pratiche rilevanti.

Il volume di Davide Battisti, "Procreative Responsibility and Assisted Reproductive Technologies", soddisfa pienamente questi requisiti, affermandosi come un contributo significativo nel panorama contemporaneo dell'etica applicata e della bioetica. Il tema affrontato è di grande attualità: esplorare come i recenti sviluppi delle tecnologie di procreazione medicalmente assistita (ARTs) stiano ampliando e ridefinendo i confini della responsabilità procreativa. Negli ultimi decenni, le ARTs hanno conosciuto una rapida evoluzione, modificando profondamente le possibilità riproduttive e sollevando nuovi interrogativi teorici e morali.

Questo sviluppo tecnologico comprende una vasta gamma di tecniche, quali la fecondazione in vitro, la diagnosi preimpianto, i test genetici prenatali, la terapia fetale e la sostituzione mitocondriale. Altre tecnologie, come la gametogenesi in vitro, l'editing del genoma a scopo riproduttivo (rGE) e l'ectogenesi – ossia la capacità di sviluppare un individuo al di fuori del corpo umano in un ambiente artificiale – potrebbero aggiungersi nel prossimo futuro. La tesi generale di Battisti è che queste tecnologie non si limitano ad ampliare la libertà procreativa, ma generano nuove responsabilità morali che devono essere analizzate e integrate in un quadro teorico ancora in fase di sviluppo.
Il volume è articolato in sette capitoli densi e approfonditi. Nel primo capitolo, Battisti offre una rassegna delle principali ARTs attualmente disponibili e di quelle in fase di sviluppo, con un focus su tecnologie emergenti come l'editing genomico e l'ectogenesi. Inoltre, esplicita alcune assunzioni teoriche di base che delimitano l'indagine al tema centrale della responsabilità procreativa.

Nel secondo capitolo, l'autore introduce una serie di definizioni e distinzioni concettuali fondamentali, proponendo una tassonomia originale del concetto di responsabilità. In particolare, definisce la 'responsabilità procreativa' come l'insieme dei doveri morali di coloro che stanno per generare una persona futura. All'interno di questa nozione, distingue poi tra 'responsabilità procreativa genitoriale', relativa agli obblighi morali dei genitori nei confronti del nascituro, e 'responsabilità riproduttiva', concernente i doveri morali verso terzi o, in alcuni casi, verso la collettività. Battisti sottolinea che tali obblighi non sono assoluti, ma *prima facie*, e devono essere bilanciati con altri aspetti moralmente rilevanti nel contesto procreativo. La parte finale del capitolo è dedicata a una discussione sul concetto di 'disabilità', cruciale per valutare il 'danno' rispetto ad alcune scelte riproduttive.

I successivi capitoli affrontano i doveri morali dei futuri genitori in modo analitico. Il terzo capitolo è dedicato al 'Principio di Beneficenza Procreativa' (*Principle of Procreative Beneficence*, PPB), proposto da Savulescu nel 2001. Secondo questo principio, chi decide di avere un figlio ha una ragione morale significativa per selezionare, tra i possibili embrioni, quello la cui vita è prevedibilmente migliore o non peggiore rispetto agli altri. Battisti analizza criticamente il PPB, mettendone in discussione la validità teorica e mostrando come esso si basi su un concetto di danno impersonale tipico dell'utilitarismo totale. Questo implica che il PPB abbia, di per sé, una forza prescrittiva assai ridotta per chi non assume tale prospettiva meta-etica. Invece che essere un principio universale al quale tutti dovrebbero quindi ancorare le proprie scelte riproduttive, infatti, esso si applica solo nel caso in cui «prospective parents already want to follow its prescriptions […] In other words, PPB proponents do not provide any reasons for complying with the model's prescription rather than satisfying prospective parent' desires»[2].

Nel quarto capitolo, Battisti propone una prospettiva morale basata sulle conseguenze per le persone attuali (*person-affecting morality*) come punto di partenza per discutere i doveri procreativi. Questa prospettiva viene applicata all'analisi di tecnologie come l'editing genomico e l'ectogenesi, portando l'autore a formulare la 'Greater Moral Obligation View', secondo cui l'accesso a nuove tecnologie riproduttive genera nuovi obblighi morali verso la prole. Nei capitoli successivi, l'autore esplora ulteriormente il ruolo delle tecnologie non neutrali rispetto all'identità numerica del nascituro e analizza l'obbligo morale, per i genitori, di potenziare i propri figli mediante tecnologie come l'rGE. Battisti conclude che, nel contesto attuale, tali obblighi non sussistono, ma potrebbero emergere in futuri contesti sociali caratterizzati da nuove norme cooperative. Il settimo capitolo, infine, amplia l'analisi considerando il ruolo morale delle intenzioni e delle attitudini dei genitori, senza abbandonare un focus primario sulle persone attuali come riferimento etico fondamentale.

Nel complesso, il volume di Battisti è rilevante per almeno due ragioni principali. In primo luogo, offre un contributo prezioso sia per gli specialisti in medicina riproduttiva, sia per i filosofi morali e gli studiosi di bioetica, fornendo una ricostruzione organica del dibattito contemporaneo e una serie di strumenti concettuali utili per affrontare le questioni normative legate all'etica della riproduzione. In secondo luogo, il testo dimostra i meriti di un approccio interdisciplinare che integra conoscenze tecniche avanzate con una riflessione filosofica rigorosa. Contro una visione della bioetica limitata al commento dei casi controversi, Battisti ci invita a esplorare in profondità le implicazioni morali delle tecnologie riproduttive, creando un ponte tra medicina, genomica e filosofia morale. Il risultato è un'analisi scientificamente fondata e intellettualmente stimolante, che rappresenta un esempio eccellente della migliore riflessione morale contemporanea.

NOTE

1. Schultz B., *The Happiness Philosophers. The Lives and Works of the great Utilitarians*, Princeton University Press, 2017, e dello stesso autore si veda anche, *Henry Sidgwick - Eye of the Universe: An intellectual Biography*, Cambridge University Press, 2004 e *Utilitarianism and Empire*, edited by Bart Schultz and Georgios Varouxakis, Lexington Books, 2005.

2. Schultz B., *Utilitarianism as a way of life. Re-envisioning planetary happiness,* Polity Press, Cambridge and Hoboken, 2024, p. 18.

Call for papers

REINVENTARE LA SCIENZA:
CRITICITÀ E NUOVI PARADIGMI PER IL FUTURO DELLA RICERCA SCIENTIFICA

Data per la sottomissione: 30 GIUGNO 2025

Il paradigma tradizionale di produzione della conoscenza scientifica è oggi sottoposto a una crescente pressione e critica. La crisi della riproducibilità, l'integrità nella ricerca, l'impatto dell'intelligenza artificiale, la cultura del "publish or perish", la scienza aperta, i limiti della revisione tra pari, i sistemi di valutazione, e il rapporto tra democrazia e scienza sono solo alcune delle questioni aperte su cui si gioca la credibilità e l'efficacia della scienza contemporanea. Questi problemi strutturali sollevano interrogativi fondamentali sul futuro della scienza: come possiamo reinventarla per affrontare le sfide del XXI secolo?

Questo numero speciale invita contributi interdisciplinari e multidisciplinari che analizzino le criticità dell'attuale modello scientifico e propongano alternative pratiche e teoriche. Sono benvenuti articoli che adottino prospettive bioetiche, etiche, antropologiche, socio-giuridiche e legate al biodiritto. Il nostro obiettivo è stimolare una riflessione ampia e profonda sulle modalità con cui la scienza può essere rinnovata per rispondere alle sfide etiche, culturali e metodologiche contemporanee.

Temi di interesse (non esaustivi):

- Come la crisi della riproducibilità influenza la credibilità della scienza e quali sono le possibili soluzioni?
- Qual è l'impatto della cultura del "publish or perish" sulla qualità della ricerca scientifica?
- In che modo l'intelligenza artificiale sta ridefinendo il processo di produzione della conoscenza?
- La scienza aperta (*open science*) è una soluzione ai problemi strutturali della scienza o presenta ulteriori sfide?
- Quali sono i limiti e le potenzialità dell'attuale sistema della revisione tra pari (peer review), e quali le possibili alternative?
- Quali sono le cause, i limiti e i problemi di un modello di editoria scientifica sempre più concentrato nelle mani di pochi, grandi gruppi editoriali, e quali le alternative?
- Come possono l'etica della ricerca e l'integrità scientifica essere rafforzate?
- Qual è il ruolo delle scienze sociali e del diritto nella riforma del modello scientifico?
- Quali aspetti della bioetica sono più rilevanti nel dibattito sulla riforma della scienza?
- Come possiamo ripensare la relazione tra scienza, società e politica per affrontare le sfide contemporanee?

Indicazioni per gli autori:

Gli articoli dovrebbero essere originali e non sottoposti ad altre riviste o pubblicazioni. Sono accettati contributi teorici, analisi critiche e case studies. Gli autori possono inviare:

- **Abstract** (massimo 300 parole) per una valutazione preliminare.
- **Articoli completi** (tra 5.000 e 8.000 parole, inclusi riferimenti bibliografici).

Tutti i contributi saranno sottoposti a peer review anonima per garantire la qualità e la pertinenza accademica.

Le linee guida di stile sono disponibili al seguente indirizzo: https://scienceandethics.fondazioneveronesi.it/submission/

Scadenze importanti:

- **Invio dei contributi**: 30 giugno 2025

N. 10 - 2025

- **Comunicazione dei risultati della revisione tra pari**: 30 settembre 2025
- **Pubblicazione online e cartacea**: on-line first, e finale a dicembre 2025

Modalità di invio:

I contributi devono essere inviati tramite email all'indirizzo: Ethics.Journal@fondazioneveronesi.it

Lingue accettate:

Sono accettati contributi in italiano e inglese.

Call for papers

REINVENTING SCIENCE:

CRITICAL ISSUES AND NEW PARADIGMS FOR THE FUTURE OF SCIENTIFIC RESEARCH

Submission Deadline: 30th JUNE 2025

The traditional paradigm of scientific knowledge production is under increasing pressure and criticism. The reproducibility crisis, research integrity, the impact of artificial intelligence, the 'publish or perish' culture, open science, the limits of peer review, evaluation systems, and the relationship between democracy and science are just some of the open issues on which the credibility and effectiveness of contemporary science depend. These structural problems raise fundamental questions about the future of science: how can we reinvent it to address the challenges of the 21st century?

This special issue invites interdisciplinary and multidisciplinary contributions that analyze the criticalities of the current scientific model and propose practical and theoretical alternatives. Articles adopting bioethical, ethical, anthropological, socio-legal, and biolaw perspectives are particularly welcome. Our goal is to stimulate a broad and in-depth reflection on how science can be renewed to respond to contemporary ethical, cultural, and methodological challenges.

Topics of Interest (not exhaustive):

- How does the reproducibility crisis affect the credibility of science, and what are the possible solutions?
- What is the impact of the "publish or perish" culture on the quality of scientific research?
- How is artificial intelligence redefining the process of knowledge production?
- Is open science a solution to the structural problems of science, or does it present further challenges?
- What are the limitations and potentials of the current peer review system, and what are the possible alternatives?
- What are the causes, limitations, and problems of a scientific publishing model increasingly concentrated in the hands of a few large corporations, and what are the alternatives?
- How can research ethics and scientific integrity be strengthened?
- What role do social sciences and law play in reforming the scientific model?
- Which aspects of bioethics are most relevant to the debate on the reform of science?
- How can we rethink the relationship between science, society, and politics to address contemporary challenges?

Author Guidelines:

Articles must be original and not submitted to other journals or publications. We accept theoretical contributions, critical analyses, and case studies. Authors may submit:

- **Abstract** (maximum 300 words) for preliminary evaluation.
- **Full papers** (between 5,000 and 8,000 words, including references).

All contributions will undergo anonymous peer review to ensure academic quality and relevance.

Style guidelines are available at the following link: https://scienceandethics.fondazioneveronesi.it/submission/

Important Deadlines:

- **Submission of Contributions**: 30th June 2025

N. 10 - 2025

- **Peer Review Results Notification**: 30th September 2025
- **Online and Print Publication**: Online-first, final publication in December 2025

Submission Guidelines:

Contributions must be submitted via email to: Ethics.Journal@fondazioneveronesi.it

Accepted Languages:

Submissions are accepted in Italian and English.

Norme editoriali

Per ogni numero è possibile sottomettere:

– Articoli liberi su temi di interesse per la rivista o articoli in risposta a *call for papers*

– Commenti ai documenti di etica e bioetica che sono stati o che saranno pubblicati

– Recensioni di volumi pubblicati nei 12 mesi precedenti alla pubblicazione della rivista

La rivista accetta contributi in lingua italiana e inglese.

Tutti i testi vanno inviati a: ethics.journal@fondazioneveronesi.it

I testi devono essere inediti e non devono essere già sottomessi ad altre riviste scientifiche.

Per sottomettere un **articolo** occorre inviare:

1. un file in formato Word, privo di ogni riferimento agli autori, di minimo 10.000 e massimo 25.000 battute (inclusi spazi, note e bibliografia);

2. un secondo file Word separato contenente:
(a) i nominativi degli autori

(b) l'affiliazione di ciascun autore

(c) l'indirizzo e-mail dell'autore corrispondente

(d) il titolo dell'articolo in italiano e in inglese

(e) un abstract dell'articolo di massimo 150 parole in italiano e in inglese

(f) da 3 a 6 parole chiave in italiano e in inglese

(g) l'esplicitazione di eventuali conflitti di interesse

(h) un indirizzo di posta (città, cap, via, n.) per ricevere eventuali copie della rivista

Per sottomettere un **commento** occorre inviare:

1. un file in formato Word di massimo 10.000 battute (inclusi spazi, note e bibliografia);

2. un secondo file Word separato contenente:
(a) il titolo del commento in italiano e in inglese

(b) i nominativi degli autori

(c) l'affiliazione di ciascun autore

(d) l'indirizzo e-mail dell'autore corrispondente

(e) l'esplicitazione di eventuali conflitti di interesse

(f) un indirizzo di posta (città, cap, via, n.) per ricevere eventuali copie della rivista

Per sottomettere una **recensione** occorre inviare:

1. un file in formato Word di massimo 5.000 battute (inclusi spazi, note e bibliografia);

2. un secondo file Word contenente:
(a) i nominativi degli autori

(b) l'affiliazione di ciascun autore

(c) l'indirizzo e-mail dell'autore corrispondente

(d) le seguenti informazioni sul libro recensito: titolo, autori, casa editrice, codice ISBN, n. di pagine, prezzo

(e) l'esplicitazione di eventuali conflitti di interesse

(f) un indirizzo di posta (città, cap, via, n.) per ricevere eventuali copie della rivista

STILE REDAZIONALE

Il tipo di carattere da utilizzarsi è il seguente: *Times New Roman* 12, con interlinea doppia.

Le note vanno inserite a piè di pagina e numerate con numeri arabi (1, 2, 3…).
I titoli devono essere brevi e specifici per facilitarne il reperimento nelle banche dati. I titoli di paragrafi e dei sotto-paragrafi devono essere ordinati utilizzando i numeri arabi, secondo una numerazione progressiva.

RIFERIMENTI BIBLIOGRAFICI

Il sistema di riferimento della rivista per le citazioni bibliografiche è lo stile Chicago A (sistema note-bibliografia). A ogni citazione nel testo, sia essa letterale o parafrasata, deve corrispondere una nota (a fine testo), completa di ogni riferimento bibliografico. La bibliografia finale viene omessa.

Il manuale di riferimento è il Chicago Manual of Style, pubblicato dal 1906 dalla Chicago University Press. Per le norme ufficiali si rimanda a: https://www.chicagomanualofstyle.org/tools_citationguide/citation-guide-1.html

Codice etico

The Future of Science and Ethics aderisce agli standard internazionali in materia di etica della ricerca e della pubblicazione, tra cui:

– il *Code of Conduct and Best Practice Guidelines* elaborato da COPE (Committee on Publication Ethics);

– il *Responsible research publication: international standards for editors*, promulgato in occasione della 2nd World Conference on Research Integrity di Singapore;

– le *Linee guida per l'integrità nella ricerca* pubblicate dalla Commissione per l'Etica e l'Integrità nella Ricerca del CNR.

DOVERI DEGLI ORGANI DIRETTIVI DELLA RIVISTA

Decisioni in merito alla pubblicazione

Il Direttore è responsabile per le decisioni relative alla pubblicazione dei manoscritti sottomessi alla rivista.Il Direttore è responsabile per le decisioni che riguardano eventuali casi di diffamazione, violazione del copyright e plagio.

Equità

Il Direttore, il Capo Redattore, i Redattori e i revisori incaricati valutano sempre i manoscritti in base al loro contenuto intellettuale senza discriminazioni di razza/etnia, genere, orientamento sessuale, credo religioso, cittadinanza, o credo politico.

Confidenzialità

Il Direttore, il Capo Redattore e i Redattori hanno il dovere di non rivelare alcuna informazione riguardo ai manoscritti che sono stati sottomessi alla rivista a persone che non siano l'autore responsabile della corrispondenza (corresponding author), i revisori e i potenziali revisori, altre persone coinvolte nell'*editing* e l'editore, laddove appropriato.

Conflitti di interesse

Il materiale non pubblicato contenuto in un manoscritto che è stato sottomesso alla rivista non può essere usato dal Direttore, dal Capo Redattore e dai Redattori per scopi di ricerca senza il consenso esplicito dell'autore.

Accesso ai contributi

I contributi accettati da The Future of Science and Ethics sono resi disponibili ad accesso libero e senza restrizioni, al fine di promuovere la politica dell'accesso aperto (open-access). I contributi sono disponibili sul sito della rivista e possono essere utilizzati a scopi scientifici, a condizione che sia citata la fonte della prima pubblicazione del manoscritto su *The Future of Science and Ethics*.

DOVERI DEI REVISORI

Revisione partitaria (Peer Review)

I testi degli articoli sottomessi alla rivista sono sottoposti a revisione paritaria anonima in doppio cieco (*Double Blind Peer Review*). Fanno eccezione i testi degli articoli delle sezioni "Prospettive", usualmente richiesti su invito, e delle sezioni "Documenti" e "Recensioni". I file Word anonimizzati e privi di eventuali riferimenti agli autori vengono inviati a due revisori anonimi individuati tra esperti esterni specialisti della materia in valutazione o, in alcuni casi, tra i componenti del Comitato Scientifico della rivista. Non possono essere affidate revisioni di singoli articoli né a componenti del Comitato di Direzione né a componenti del Comitato Editoriale della rivista. La revisione richiede circa 4 settimane dalla data di ricezione del manoscritto. Nel caso in cui siano richieste revisioni (minime o sostanziali), il testo deve essere corretto evidenziando le parti modificate, e quindi ri-sottomesso alla redazione nei tempi richiesti, accompagnato da una breve lettera di risposta ai Revisori. In caso di giudizi significativamente discordanti tra i revisori, la redazione si riserva di chiedere un terzo parere e di prolungare il processo di revisione di ulteriori 2 settimane.

Celerità

Qualsiasi revisore che ritenga di non essere qualificato per svolgere la revisione del manoscritto che gli è stato assegnato, o di non poter terminare

tale revisione entro i tempi richiesti e comunque entro un tempo considerato ragionevole, deve subito notificare tali aspetti al Direttore e quindi rinunciare a prendere parte al processo di revisione paritaria.

Confidenzialità

Qualsiasi manoscritto ricevuto per la revisione paritaria deve essere considerato come un documento confidenziale. Come tale non deve essere mostrato o discusso con altri se non nei casi autorizzati dal Direttore della rivista.

DOVERI DEGLI AUTORI

Contenuti

Gli autori di un manoscritto che riguardi una ricerca originale devono presentare una descrizione accurata del lavoro svolto così come una discussione obiettiva del suo significato. I dati devono essere presentati in modo accurato nel manoscritto. Un manoscritto deve contente dettagli sufficienti e riferimenti bibliografici da permette ad altri di replicare il lavoro. Affermazioni fraudolente o intenzionalmente inaccurate costituiscono casi di comportamento non etico e inaccettabile.

Originalità e plagio

Gli autori devono assicurare di aver scritto un lavoro interamente originale; e, qualora gli autori abbiano usato il lavoro di altri, che esso sia citato in modo appropriato.

Un autore non deve pubblicare in generale manoscritti che descrivano la stessa identica ricerca in più di una rivista o pubblicazione primaria. Sottomettere lo stesso manoscritto a più di una rivista simultaneamente costituisce un comportamento non etico e inaccettabile.

Riconoscimento delle fonti

Devono sempre essere inseriti riferimenti appropriati ai lavori di altri autori. Gli autori devono citare le pubblicazioni che hanno influito nel determinare la natura del lavoro riportato nel manoscritto.

Ruolo di autori e co-autori

Il ruolo di autore o co-autore di un manoscritto deve essere attribuito esclusivamente a coloro che hanno apportato un contributo significativo all'ideazione, progettazione, esecuzione, o interpretazione dei risultati dello studio o dei contenuti concettuali presenti nel manoscritto. Tutti coloro che hanno apportato contributi significativi devono essere inseriti come co-autori. Nel caso in cui vi siano altre persone che hanno partecipato in altri aspetti sostanziali della ricerca, essi devono essere riconosciuti e indicati come contributori o ringraziati in una apposita nota.

L'autore responsabile della corrispondenza deve garantire che tutti i co-autori appropriati e nessun co-autore inappropriato siano inclusi tra i firmatari nel manoscritto, e che tutti i co-autori abbiamo visto e approvato la versione finale del manoscritto e concordato alla sua sottomissione per la pubblicazione.

Consenso informato e rispetto dei diritti umani e animali

Tutti gli autori di manoscritti in cui vengono descritte ricerche che hanno coinvolto soggetti umani o animali devono garantire che la ricerca sia stata svolta con il consenso dei partecipanti e con le autorizzazioni necessarie per svolgere ricerche con soggetti umani o animali.

Conflitti di interesse

Tutti gli autori devono dichiarare nel manoscritto i loro eventuali conflitti di interesse a livello finanziario o di altra natura che potrebbero aver influenzato i risultati o l'interpretazione del contenuto del manoscritto. Tutte le fonti di finanziamento o supporto al progetto di ricerca da cui è nato lo studio devono essere dichiarate in modo esplicito.

Errori fondamentali nei lavori pubblicati

Qualora un autore venisse a conoscenza di un errore significativo o di una inesattezza nei lavori che ha pub-

blicato sulla rivista, è Suo preciso dovere segnalare tale errore al Direttore o alla Redazione e cooperare con il Direttore e la Redazione per ritirare o correggere il manoscritto.

Diritti d'Autore

Gli autori garantiscono di avere la titolarità dei diritti sulle opere che sottopongono alla rivista The Future of Science and Ethics e garantiscono che tali opere siano inedite, liberamente disponibili e lecite, sollevando l'editore da ogni eventuale danno o spesa.

Gli autori mantengono i diritti d'autore sulle proprie opere e autorizzano l'editore a pubblicare, riprodurre, distribuire le opere con qualunque mezzo e in ogni parte del mondo e a comunicarli al pubblico attraverso reti telematiche, compresa la messa a disposizione del pubblico in maniera che ciascuno possa avervi accesso dal luogo e nel momento scelti individualmente, disponendo le utilizzazioni a tal fine preordinate.

Gli autori che intendano includere nelle loro opere testi, immagini, fotografie o altre opere già pubblicate altrove si assumono la responsabilità di ottenere le autorizzazioni dei relativi titolari dei diritti ove necessarie. Gli autori garantiscono che sulle opere non sussistano diritti di alcun genere appartenenti a terze parti.

Gli autori hanno diritto a riprodurre, distribuire, comunicare al pubblico, eseguire pubblicamente gli articoli pubblicati sulla rivista con ogni mezzo, per scopi non commerciali (ad esempio durante il corso di lezioni, presentazioni, seminari, o in siti web personali o istituzionali) e ad autorizzare terzi ad un uso non commerciale degli stessi, a condizione che gli autori siano riconosciuti come tali e la rivista The Future of Science and Ethics sia citata come fonte della prima pubblicazione dell'articolo.

La rivista non pretenderà dagli autori alcun pagamento per la pubblicazione degli articoli. Gli autori non riceveranno alcun compenso per la pubblicazione degli articoli.

I compiti del Comitato Etico della Fondazione Umberto Veronesi

"La scienza è un'attività umana inclusiva, presuppone un percorso cooperativo verso una meta comune ed è nella scienza che gli ideali di libertà e pari dignità di tutti gli individui hanno sempre trovato la loro costante realizzazione.

La ricerca scientifica è ricerca della verità, perseguimento di una descrizione imparziale dei fatti e luogo di dialogo con l'altro attraverso critiche e confutazioni. Ha dunque una valenza etica intrinseca e un valore sociale indiscutibile, è un bene umano fondamentale e produce costantemente altri beni umani.

In particolare, la ricerca biomedica promuove beni umani irrinunciabili quale la salute e la vita stessa, e ha un'ispirazione propriamente umana poiché mira alla tutela dei più deboli, le persone ammalate, contrastando talora la natura con la cultura e con la ragione diretta alla piena realizzazione umana.

L'etica ha un ruolo cruciale nella scienza e deve sempre accompagnare il percorso di ricerca piuttosto che precederlo o seguirlo. È uno strumento che un buon ricercatore usa quotidianamente.

La morale è anche l'unico raccordo tra scienziati e persone comuni, è il solo linguaggio condiviso possibile.

Ci avvicina: quando si discute di valori, i ricercatori non sono più esperti di noi. Semmai, sono le nostre prime sentinelle per i problemi etici emergenti e, storicamente, è proprio all'interno della comunità scientifica che si forma la consapevolezza delle implicazioni morali delle tecnologie biomediche moderne.

Promuovere la scienza, come fa mirabilmente la Fondazione Veronesi, significa proteggere l'esercizio di un diritto umano fondamentale, la libertà di perseguire la conoscenza e il progresso, ma anche, più profondamente, significa favorire lo sviluppo di condizioni di vita migliori per tutti.

Compiti del Comitato Etico saranno quelli di dialogare con la Fondazione e con i ricercatori, favorendo la crescita di una coscienza critica e insieme di porsi responsabilmente quali garanti terzi dei cittadini rispetto alle pratiche scientifiche, guidati dai principi fondamentali condivisi a livello internazionale e tenendo nella massima considerazione le differenze culturali".