



UNIVERSITA' DEGLI STUDI DI PAVIA

DIPARTIMENTO DI STUDI UMANISTICI

CORSO DI LAUREA MAGISTRALE IN FILOSOFIA

L'INTELLIGENZA ARTIFICIALE E LA RIFLESSIONE MORALE

RELATORE

Prof. Sergio Filippo Magni

CORRELATORE

Prof. Luca Fonnesu

Tesi di Laurea Magistrale di
Jonathan Prazzoli
Matricola n. 497325

Anno accademico 2025/2026

Indice generale

Introduzione.....	3
1. Che cos'è l'intelligenza artificiale e quali problemi etici crea?.....	6
1.1 Le applicazioni dell'IA.....	6
1.2 Origini e storia dell'intelligenza artificiale.....	9
1.3 Che cos'è l'intelligenza artificiale?.....	10
1.4 Problemi etici creati dall'IA.....	19
1.4.1 IA e coscienza.....	21
1.4.2 IA e comprensione.....	26
1.4.3 IA e creatività.....	28
2. Le luci e le ombre dell'intelligenza artificiale.....	35
2.1 Superintelligenza e singolarità: l'IA “cattiva”.....	35
2.2 L'IA ci ruberà il lavoro?.....	41
2.3 IA e privacy.....	45
2.4 L'IA e il rischio di comportamenti contrari all'etica.....	50
2.5 Utilizzare l'IA per il male sociale.....	55
2.6 IA e autonomia: deleghiamo troppo?.....	64
3. Intelligenza artificiale e politica.....	70
3.1 IA e democrazia: chi governa davvero?.....	70
3.2 Chi detiene il potere? Noi o l'IA?.....	75
3.3 IA e sovranità.....	79
3.4 Discriminazioni artificiali: IA ed equità.....	83
4. Costruire un'IA etica per il bene sociale.....	88
4.1 IA ed espansione del cerchio morale.....	88
4.2 IA ed etica: percorsi paralleli.....	92
4.3 Principi etici per l'IA.....	97

4.4 Usare l'IA per il bene sociale.....	102
4.5 Realizzare una società per la buona IA.....	109
Conclusione.....	116
Bibliografia.....	119
Ringraziamenti.....	123

Introduzione

Questa tesi si sofferma sulle ricadute etiche dell'intelligenza artificiale e propone una serie di riflessioni morali riguardo questa nuova tecnologia.

Il primo capitolo si apre con una panoramica sulle varie applicazioni dell'intelligenza artificiale, da quelle più comuni e conosciute alle più sofisticate. Successivamente si compirà un breve excursus nella storia e nelle origini dell'intelligenza artificiale, a partire dal 1956 e dalle intuizioni di Turing. Poi, si tenterà di dare una definizione di intelligenza artificiale, seguendo l'idea di Floridi secondo cui l'IA sarebbe una nuova forma dell'agire e non una forma di intelligenza. Verrà data una definizione di concetti quali *machine learning*, *deep learning*, IA generativa e superintelligenza. In particolare, sarà analizzato il test di Turing e le sue successive implicazioni. Infine, verranno elencati, in breve, i vari problemi etici posti in essere dall'intelligenza artificiale. Per concludere, sarà analizzato il legame tra l'intelligenza artificiale e concetti quali coscienza, comprensione e creatività.

Il secondo capitolo affronta le principali sfide etiche dell'intelligenza artificiale nei suoi aspetti più controversi. Il capitolo si apre con una disamina dei concetti di “superintelligenza” e “singolarità”, illustrando le visioni pessimistiche e apocalittiche del futuro dell'intelligenza artificiale. Se ne dimostrerà l'assurdità, ma si farà notare come sia altrettanto sbagliato banalizzarle o escluderle a priori. Nel secondo paragrafo si parlerà delle conseguenze economiche e sociali dell'intelligenza artificiale e si esamineranno le preoccupazioni di chi ritiene che l'IA possa creare “eserciti di disoccupati” o eccessive differenze salariali. Saranno proposte soluzioni e modi per un uso equilibrato dell'IA in tale ambito. Nel terzo paragrafo si affronterà il tema della privacy in relazione all'intelligenza artificiale, mostrandone rischi, timori e soluzioni etiche in tal senso. Il quarto paragrafo illustra gli usi dell'intelligenza artificiale moralmente controversi da parte di aziende e start-up. Nel quinto paragrafo si dimostrerà come l'IA possa essere usata da parte di criminali e malintenzionati, il che richiede una maggiore attenzione in tema di cybersicurezza. Il capitolo si conclude con un'analisi del fatto che con l'intelligenza artificiale c'è una tendenza a delegare troppo

alle macchine, perdendo l'autonomia decisionale e la capacità di pensiero critico.

Il terzo capitolo esamina il legame tra intelligenza artificiale e politica. Nel primo paragrafo si parla di come l'intelligenza artificiale stia influenzando la democrazia, attraverso il cosiddetto “capitalismo della sorveglianza”, la manipolazione dell'informazione e l'alterazione dei processi elettorali. Il secondo paragrafo affronta la relazione tra intelligenza artificiale e potere. Si evidenzia come l'IA stia modificando il paradigma dell'attuale sistema democratico, attraverso la propagazione di *fake news*, *deep fake* e la conseguente incapacità, da parte delle persone, di esercitare il pensiero critico. Nel terzo paragrafo è posto in evidenza il legame tra intelligenza artificiale e il concetto di sovranità, mostrando come i colossi digitali stiano minando il controllo digitale da parte delle democrazie. Si affronta poi il problema dei rischi etici e giuridici dell'uso dell'IA in ambito militare. Il quarto paragrafo affronta il tema dell'equità in rapporto all'intelligenza artificiale, evidenziando le discriminazioni algoritmiche causate da *bias* (pregiudizi) nei dati che perpetuano ingiustizie in ambiti come lavoro, credito e sanità.

Il quarto e ultimo capitolo analizza il tema della costruzione di un'IA etica per il bene sociale. Il capitolo si apre con il classico tema, elaborato da Peter Singer, dell'espansione del cerchio morale che andrebbe a includere anche l'intelligenza artificiale. Tuttavia, secondo queste analisi, l'IA andrebbe trattata come agente morale ma non come paziente morale. Nel secondo paragrafo, si espone una sorta di integrazione tra intelligenza artificiale ed etica delle virtù, focalizzandosi sul modello socratico come via privilegiata per lo sviluppo di sistemi “saggi by design”, cioè capaci di discernimento pratico. Il terzo paragrafo espone una serie di principi etici, proposti da Floridi, per la realizzazione di un'IA al servizio del benessere umano, evitando le derive distruttive. Nel quarto paragrafo si mostra che un'IA per il bene sociale richiede una progettazione rigorosa e consapevole, per evitare di amplificare disuguaglianze e danni anziché risolverli e fare in modo che l'IA agisca come strumento trasparente e sicuro a supporto della dignità umana. Il quinto paragrafo esamina come orientare l'intelligenza artificiale verso il bene sociale attraverso quattro opportunità chiave e un'etica soft basata su venti raccomandazioni pratiche.

La tesi si conclude con una riflessione finale sull'intero percorso, nella quale sostengo che è

sbagliato esaltare l'intelligenza artificiale a panacea di tutti i mali del mondo, ma non bisogna nemmeno condannarla *in toto*. Infatti l'IA, se usata correttamente, può contribuire al benessere dell'umanità.

1. Che cos'è l'intelligenza artificiale e quali problemi etici crea?

1.1 Le applicazioni dell'IA

L'intelligenza artificiale permea ormai le nostre vite in una maniera difficile da quantificare. Prima di effettuare un tentativo di definizione di IA e di capire quali siano le problematiche etiche ad essa connessa, è bene avere un quadro generale di *quali* IA abbiamo a disposizione, quali sono le applicazioni di intelligenza artificiale e come e in quali campi essa sia utilizzata.

I sistemi di intelligenza artificiale più conosciuti e utilizzati anche dal cittadino medio sono sicuramente i modelli generativi di linguaggio, come *Chatgpt*, *Gemini*, *Deepseek*, *Google LM*, *Gamma* e tantissimi altri. Cosa sono in grado di fare queste applicazioni? In base al prompt fornito dall'utente – cioè una sorta di “ordine” in cui l'utente specifica, in linguaggio naturale o più o meno tecnico, cosa deve generare l'IA – queste applicazioni possono generare testi, traduzioni, tabelle, presentazioni power point, immagini con o senza testo, video, piccoli programmi o interi script in linguaggi come Python, Java, JavaScript, C++, Go, ecc. L'IA generativa può inoltre fare ricerche online al posto nostro e creare un intero sito web. Le potenzialità di questa IA sono innumerevoli.

Esistono poi i cosiddetti “Agenti IA”, cioè sistemi di intelligenza artificiale in grado di agire in autonomia per un utente o un'azienda. Questi sistemi IA possono ad esempio pianificare una campagna di marketing, raccogliere una certa quantità di dati da fonti diverse per poi elaborarli, generare report e inviarli ai destinatari giusti senza intervento umano¹. Insomma, un agente IA funziona come una sorta di assistente digitale intelligente, in grado di ragionare e prendere decisioni in modo autonomo, riducendo il carico di lavoro umano.

Infine ci sono i tool di assistenza, cioè sistemi basati su IA come chatbot per assistenza clienti, in grado di rispondere a domande, risolvere problemi su resi e spedizioni, suggerire prodotti in base alle preferenze dei clienti e così via. Anche sui social network gli algoritmi di intelligenza artificiale

¹ <https://www.cosmonet.info/intelligenza-artificiale-2026-ecosistemi-professionali>

suggeriscono determinati contenuti personalizzati in base alle preferenze e ai comportamenti degli utenti, alle ricerche più frequenti e ai trend emergenti.

L'IA compare sovente anche in ambito industriale. Particolarmente interessante è l'utilizzo dell'IA per quanto riguarda il settore sanitario, che include “strumenti di radiologia basati sull'IA, interventi chirurgici robotici, scoperta di farmaci assistita dall'IA e agenti virtuali per la salute che migliorano il coinvolgimento dei pazienti riducendo al contempo il carico di lavoro dei professionisti umani².”

L'intelligenza artificiale è inoltre utilizzata nei servizi finanziari: “Le banche stanno utilizzando l'IA nel settore bancario, ad esempio con chatbot basati su linguaggio naturale, piattaforme di trading algoritmico e sistemi intelligenti di elaborazione dei documenti, per ridurre i rischi, aumentare la conformità normativa e migliorare la soddisfazione dei clienti.”

L'intelligenza artificiale è utilizzata, in certi casi, anche nel settore agricolo: “Gli agricoltori stanno adottando applicazioni di intelligenza artificiale in agricoltura, come il monitoraggio delle colture tramite droni, l'ottimizzazione dell'irrigazione basata sulle condizioni meteorologiche e l'identificazione dei parassiti tramite visione artificiale³.”

In ambito militare, “L'IA può migliorare il comando e il controllo, le comunicazioni, i sensori, l'integrazione e l'interoperabilità. Le tecnologie basate sull'IA permettono di coordinare sensori ed effettori, rilevare e identificare minacce, segnalare le posizioni nemiche, acquisire obiettivi e coordinare, evitando conflitti, il fuoco congiunto distribuito tra veicoli da combattimento in rete, sia operati da esseri umani sia autonomi⁴.”

L'IA compare inoltre nel campo manifatturiero e logistico: può prevedere guasti, è implementata in robot e strumenti di visione artificiale per il controllo qualità e l'ottimizzazione delle catene di distribuzione⁵.

² <https://theintellify.com/top-ai-use-cases-across-industries>

³ Ibid.

⁴ https://en.wikipedia.org/wiki/Military_applications_of_artificial_intelligence

⁵ <https://www.smartertech.com/articles/examples-of-artificial-intelligence-ai-in-7-industries?>

In più, l'IA è legata al settore dei trasporti e delle *smart cities*, ad esempio con i veicoli a guida autonoma, nella gestione del traffico attraverso infrastrutture intelligenti e nella pianificazione di percorsi e nel monitoraggio di città *smart*⁶.

Nell'ambito dell'istruzione l'IA permette un apprendimento personalizzato, sistemi di tutor intelligenti e un'automazione dei compiti amministrativi⁷.

L'elenco potrebbe continuare, ma non è questo lo scopo principale di questa tesi. Possiamo soltanto concludere citando il Festival di Primavera in Cina (svoltosi il 16-17 Febbraio 2026), dove i robot “hanno eseguito coreografie sincronizzate, dimostrazioni di arti marziali e interazioni con performer umani, mettendo in mostra tutti i passi in avanti fatti dal settore negli ultimi mesi. Dalla stabilità dinamica, fino al controllo dell'equilibrio e alla coordinazione nei movimenti complessi⁸.” È solo grazie all'intelligenza artificiale che i robot di questo evento hanno potuto eseguire simili coreografie. Infatti l'IA analizza i dati dei sensori (telecamere, microfoni, sensori di movimento) in dotazione ai robot, riconosce ambiente e persone e decide in tempo reale come muoversi, mantenendo equilibrio e coordinazione.

1.2 Origini e storia dell'intelligenza artificiale

La nascita ufficiale dell'IA è tipicamente fatta risalire al 1956, quando al Darmouth College si svolse un workshop estivo a cui erano presenti molti esperti del settore, come “John McCarthy (che coniò il termine *Artificial Intelligence*), Marvin Minsky (fondatore del celeberrimo MIT AI Laboratory), Claude Shannon (padre della teoria dell'informazione), nonché Allen Newell e Herbert Simon (creatori del Logic Theorist, il primo programma di IA della storia)⁹.”

⁶ https://en.wikipedia.org/wiki/Military_applications_of_artificial_intelligence

⁷ <https://techietory.com/ai/artificial-intelligence-applications/>

⁸ <https://www.ilsole24ore.com/art:la-cina-mostra-muscoli-coi-robot-umanoidi-protagonisti-capodanno-AIRCb5SB>

⁹ Mario de Caro, Benedetta Giovanola, *Intelligenze*, Il Mulino, Bologna 2025, p. 22

Fu a partire da quel momento che si avviò il processo che diede avvio a un più ampio sviluppo dell'IA, anche se le radici di un progetto di costruzione di macchine intelligenti erano già presenti nelle intuizioni di Pascal e Leibniz, nella logica simbolica, nella teoria della computazione (vedi Alan Turing) e nella cibernetica. Ciononostante, è con il workshop di Darmouth che nasce il paradigma della cosiddetta IA simbolica¹⁰.

Il filosofo statunitense Haugeland definì “GOF AI” questo nuovo paradigma, acronimo che sta per *Good Old-Fashioned Artificial Intelligence*, “Buona Vecchia Intelligenza Artificiale”. Secondo il GOF AI, “i sistemi erano progettati con regole logiche ben definite, fisse e prevedibili e avevano carattere deterministico: dato un certo input, producevano sempre lo stesso output¹¹.”

La GOF AI raggiunse ottimi risultati con i cosiddetti “sistemi esperti”, cioè programmi in grado di emulare le decisioni degli esperti (esseri umani) in determinati campi. I sistemi esperti conseguirono buoni risultati commerciali, soprattutto in ambito medico. A livello tecnico, un sistema esperto è basato su regole logiche e, dati determinati input, genera output prevedibili ma rigidi.

Un altro importante risultato raggiunto dalla GOF AI fu ELIZA, un programma elaborato al MIT negli anni sessanta. ELIZA era stato realizzato con l'obiettivo di dimostrare che “un sistema di IA poteva imitare il linguaggio umano¹²”. Pur con numerosi limiti, ELIZA “suscitava in molti utenti l'illusione di essere intelligente, empatica e capace di comprensione autentica¹³”.

Tuttavia, i risultati della GOF AI delusero le aspettative di molti e negli anni Settanta questo paradigma entrò in crisi.

Inoltre, tra gli anni Ottanta e i primi anni Novanta si cominciò a parlare di *AI Winter* (inverno

¹⁰ Ibid., p. 22

¹¹ Ibid., p. 22

¹² De Caro, Giovanola, op. cit., p. 23

¹³ Ibid.

dell'IA), accentuando ancora di più la grave crisi che stava percorrendo la GOFAL.

In questo contesto, gli studiosi si cimentarono in diversi progetti che non diedero avvio a nuovi paradigmi, né avevano un orientamento comune.

Nel 2006, Geoffrey Hinton diede nuova vita al progetto delle reti neurali, che era già stato avviato negli anni Quaranta ma senza risultati di rilevanza. Con Hinton si affermò dunque un nuovo paradigma dell'IA, il cosiddetto paradigma postsimbolico, “basato su un approccio non deterministico ma probabilistico e sul *deep learning* proprio delle reti neurali artificiali¹⁴.” Ciò corrispose a una vera e propria rivoluzione nel campo dell'intelligenza artificiale.

1.3 Che cos'è l'intelligenza artificiale?

Nel primo paragrafo abbiamo elencato i vari usi e le innumerevoli applicazioni dell'IA. Molte persone le utilizzano, eppure quanti si sono mai domandati che cosa sia davvero l'intelligenza artificiale? Quanti individui sono in grado di dare una definizione che spieghi in maniera definitiva che cosa intendiamo per IA?

Ho provato a chiedere a Gemini, l'intelligenza artificiale generativa e multimodale di Google, di darmi una definizione di intelligenza artificiale e questa è stata la sua risposta: “È la capacità di un sistema informatico di risolvere problemi o eseguire compiti tipicamente affidati all'intelligenza umana, come l'apprendimento, il ragionamento logico, la comprensione del linguaggio e la percezione visiva, attraverso l'elaborazione e l'analisi di grandi quantità di dati.”

È una risposta sicuramente giusta, eppure il cittadino medio si gratterebbe la testa dando segno di non avere compreso nulla. Vediamo allora una definizione che ne dà Wikipedia.

“L'intelligenza artificiale (IA) [...] è l'intelligenza mostrata dalle macchine, in contrasto con l'intelligenza naturale mostrata dagli esseri umani¹⁵.”

¹⁴ Ibid.

¹⁵ Wikipedia, “Artificial Intelligence”, 17 Gennaio 2020

Il filosofo Luciano Floridi, nel suo libro *Etica dell'intelligenza artificiale*, afferma che la suddetta definizione di Wikipedia è in realtà una tautologia, ed equivarrebbe ad affermare “Brexit significa Brexit”, con le parole di Theresa May¹⁶. Insomma, ancora una volta abbiamo detto tutto e niente. Floridi continua dicendo che la mancanza di una definizione standard di IA crea discussioni infinite, con il rischio che “qualcuno possa concludere che non si possa avere un dibattito sull'etica di qualcosa che è indefinito e apparentemente indefinibile¹⁷.” Ciononostante, Floridi continua dicendo che tantissime cose che riteniamo importanti nella vita non sono definibili in senso stretto. Infatti, ad esempio, “abbiamo un'idea assolutamente chiara di cosa sia l'amicizia, pur non disponendo delle condizioni necessarie e sufficienti per coglierne la natura¹⁸.” In ogni caso, cose come l'amicizia sono cose che, anche in assenza di definizione, possiamo riconoscerle, le vediamo. Floridi cita a riguardo la frase pronunciata dal giudice della Corte Suprema degli Stati Uniti, Potter Stewart, in riferimento a un caso in cui decise di non considerare osceno del materiale considerato pornografico. Il giudice affermò: “Non cercherò di definire ulteriormente i tipi di materiale [...] inclusi in tale descrizione [pornografia hard-core], e forse non potrei mai riuscire a farlo in modo intellegibile. Ma *la riconosco* quando la vedo¹⁹.”

Floridi riconosce che questo modo di procedere può andare bene per questioni concernenti la vita quotidiana, ma non può in alcun modo applicarsi alla scienza. Ecco allora che l'assenza di definizione per quanto riguarda l'IA potrebbe farci sospettare che “intelligenza artificiale” non sia un termine scientifico come “triangolo”, ma solamente un termine generico proprio come “amicizia” o “pornografia”²⁰. Floridi giunge alla conclusione che quindi l'IA sia una sorta di

¹⁶ Luciano Floridi, *Etica dell'intelligenza artificiale*, Raffaello Cortina Editore, Milano 2022, p. 40

¹⁷ Ibid.

¹⁸ Ibid., p. 41

¹⁹ Ibid., p. 40

²⁰ Ibid. pp. 40-41

“famiglia” a cui appartengono diverse discipline, servizi e prodotti. Nel suo libro, il filosofo ha creato una mappa in cui ne sono elencati parecchi, ma qui ci limiteremo a parlare solamente di alcuni di essi.

In cosa si differenziano i sistemi IA di tipo probabilistico da quelli di tipo deterministico? Mentre i secondi si basano su regole rigide e univoche, i primi stimano la probabilità che determinati eventi si realizzino o che risultino corrette determinate conclusioni²¹.

Tali approcci non deterministici hanno trovato un ruolo predominante nelle ricerche e nelle applicazioni IA soprattutto con il *machine learning* e il *deep learning*.

“Il *machine learning* (apprendimento automatico) è un approccio secondo il quale, invece di programmare esplicitamente ogni regola in modo rigido, si forniscono al sistema grandi quantità di dati, grazie al quale il sistema può allenarsi (fase di training) a riconoscere pattern, fare previsioni e prendere decisioni²².”

In questo senso parliamo di un approccio non deterministico; infatti, l'IA non fornisce risposte certe, bensì probabilistiche, analizzando i dati che riceve in input. È in questo senso che i dati sono fondamentali, perché l'IA necessita di enormi quantità di dati per poter elaborare modelli affidabili. Inoltre, tali dati devono essere di ottima qualità, altrimenti vi è il rischio che l'IA dia luogo alle cosiddette “allucinazioni”, cioè errori negli output o risposte eccessivamente semplificate²³.

I dati sui quali l'IA si allena provengono dai social media e dal web: più dati ci sono per allenarsi e più i sistemi di IA risultano tecnicamente robusti. D'altro canto, un'elevata quantità di dati pone interrogativi circa la loro affidabilità e correttezza. Infatti, “se i sistemi apprendono da dati ambigui o errati, anche gli output che generano rischiano di essere imprecisi o fuorvianti²⁴.”

²¹ De Caro, Giovanola 2025, op. cit., p. 25

²² Ibid.

²³ Ibid., p. 26

²⁴ Ibid.

Un'altra importante sottocategoria del *machine learning* è il cosiddetto *deep learning* (apprendimento profondo), “che utilizza reti neurali artificiali modellate sul funzionamento del cervello umano²⁵”. Anche queste reti sono sottoposte a un addestramento basato su enormi quantità di dati e sono dotate di neuroni artificiali collegati da una serie di connessioni, proprio come succede nel nostro cervello.

“Negli ultimi anni, il *deep learning* ha rivoluzionato campi come la visione artificiale, il riconoscimento vocale e la traduzione automatica²⁶”.

L'esempio più conosciuto di IA basata sul *deep learning* è ChatGPT, che è una tipologia di intelligenza artificiale con il quale è possibile chiacchierare (*to chat*, in inglese) di svariati argomenti, mentre “GPT” è un acronimo che sta per *Generative Pretrained Transformer*, cioè “trasformatore generativo pre-addestrato²⁷”. In base all'input fornito dall'utente, scritto in messaggio testuale naturale, ChatGPT è in grado di elaborare un output, cioè una risposta testuale che l'IA è in grado di fornire perché ha imparato a generare “su base probabilistica relazioni tra parole e, dunque, a produrre sequenze nuove che sembrano realmente create da un essere umano²⁸.” Infatti, ChatGPT appartiene alla famiglia dei cosiddetti LLMs, cioè modelli linguistici di grandi dimensioni, e può essere addestrato sulla base di testi già realmente esistenti. In fase di addestramento, ChatGPT ha subito numerose correzioni di esperti che hanno individuato errori o frasi offensive generate dall'IA, proprio sulla base di conversazioni intrattenute con il *chatbot*²⁹.

Il lancio di ChatGPT, avvenuto nel novembre 2022, ha aperto un vivace dibattito rispetto alle potenzialità ma anche ai rischi dell'IA. Come avremo modo di approfondire più avanti, in molti

²⁵ De Caro, Giovanola 2025, op. cit., p. 26

²⁶ Ibid.

²⁷ Ibid.

²⁸ Ibid., p. 27

²⁹ Ibid.

paventano scenari distopici di un futuro dominato dall'IA, anche se non è possibile parlare di una sola IA “forte”, poiché ci sono numerose IA, ossia sistemi in grado di portare a termine determinati compiti. Tuttavia, “questi compiti sono sempre più complessi e interconnessi, tanto che oggi alcuni parlano di IA capace [...] o IA agentic³⁰”.

Questi sistemi sono capaci di svolgere tali compiti complessi con poco o nullo intervento umano e sono in grado di memorizzare le attività precedenti e di imparare tramite l'esperienza. È per questo che vengono chiamati, come già detto, “capaci” o “agenti”: l'*agency* è infatti la capacità di agire autonomamente, elaborando progetti e operando in vista di obiettivi determinati³¹.

Per concludere questa breve disamina di alcune discipline dell'IA, è necessario operare una distinzione ancora più generale. Occorre infatti distinguere tra la già citata IA forte, detta anche *General AI* (AGI), cioè IA generale, e la cosiddetta superintelligenza. L'AGI è un'IA che, nel caso venisse realizzata, avrà capacità cognitive, di comprensione e produzione linguistica uguali a quelle umane³². D'altra parte, la superintelligenza sarà invece nettamente superiore all'intelligenza umana, sfuggendo definitivamente ad ogni controllo da parte nostra, come mostrato in pellicole cinematografiche del calibro di *Matrix* o *Terminator*. Molti temono infatti che l'AGI potrà appunto evolversi nella superintelligenza; il momento ipotetico in cui avverrà questa evoluzione è stato denominato “singolarità tecnologica” all'informatico e futurologo Ray Kurzweil nei suoi libri *The singularity is near* [2005] e *The singularity is nearer* [2024]³³.

In tutto questo discorso abbiamo elencato solo alcuni dei membri della “famiglia IA” individuati da Floridi nella mappa già citata. Eppure, Floridi afferma che la sua mappa è “utile ma controversa³⁴.”

³⁰ Ibid.

³¹ Ibid.

³² De Caro, Giovanola 2025, op. cit., p. 19

³³ Ibid.

³⁴ Floridi 2022, op.cit, p. 42

Il filosofo si chiede se, vista l'impossibilità di trovare una definizione univoca e universalmente accettata di IA basata su criteri rigidi, sia comunque possibile individuare un tratto distintivo comune tra le diverse tecnologie, metodologie e ambiti di ricerca che afferiscono a questo campo³⁵. Floridi prova a rispondere a questa domanda riprendendo una citazione che aveva già proposto nei capitoli precedenti: “Per il presente scopo, il problema dell'intelligenza artificiale è quello di far sì che una macchina agisca con modalità che sarebbero definite intelligenti se un essere umano si comportasse allo stesso modo³⁶.”

Il filosofo commenta dicendo che la suddetta citazione tratta di un controfattuale che non c'entra nulla con il pensiero, ma ha a che fare solamente con il comportamento. Ciò non vuol dire che una macchina sia intelligente o addirittura in grado di pensare³⁷.

Floridi riporta poi una citazione di Turing molto significativa in tal senso: “Propongo di considerare la domanda: “Possono le macchine pensare?”. Questa indagine dovrebbe iniziare definendo il significato dei termini “macchina” e “pensare.” [...] La domanda originaria, “Possono le macchine pensare?”, credo sia *troppo insensata per meritare di essere discussa*³⁸.”

È in questo senso che Turing ha elaborato un test per verificare la capacità di una macchina di essere in grado di rispondere a domande in modo tale che il risultato non possa essere distinto da quello di un agente umano che compie la medesima cosa³⁹.

È a questo punto che Floridi ci induce a riflettere affermando che, benché tutto questo sia ragionevole, non vuol dire che se una lavastoviglie è in grado di lavare bene i piatti meglio di noi umani, allora ne consegue che li pulisca *esattamente* come noi o che debba essere intelligente per

³⁵ Ibid., pp 43-44

³⁶ Ibid., p. 44. Cfr McCarthy, Minsky, Rochester, Shannon, 2006

³⁷ Ibid.

³⁸ Ibid. Cfr Turing, 1950, pp. 433, 442

³⁹ Ibid., pp. 44-45

svolgere questo compito. Equivarrebbe a dire che, dal momento che tutti i fiumi raggiungono il mare nel modo migliore e rimuovendo gli ostacoli, e se questo dovesse essere fatto dall'uomo diremmo che è un comportamento intelligente, allora ne consegue che il comportamento del fiume è intelligente⁴⁰.

È chiaramente una fallacia, e Floridi conclude che dicendo che “il solo aspetto rilevante è *eseguire un compito con successo* in modo tale che il risultato sia buono o migliore di quello che l'intelligenza umana sarebbe stata in grado di ottenere⁴¹.”

Il filosofo prosegue la sua analisi prendendo in esame alcune dichiarazioni rilasciate da Turing durante una trasmissione radiofonica della BBC nel 1951. In quell'occasione, lo scienziato sosteneva la tesi secondo cui non sarebbe azzardato paragonare i computer ai cervelli biologici. Secondo il ragionamento di Turing, se si accetta l'ipotesi che il cervello umano o animale operi come una sorta di apparato meccanico, allora ne consegue che un calcolatore digitale, se opportunamente programmato, sarebbe in grado di replicarne fedelmente il comportamento, diventando a tutti gli effetti un cervello artificiale⁴².

Floridi sostiene che, benché le affermazioni di Turing possano essere considerate giuste da un punto di vista metaforico, ciononostante risulta essere nel torto a un livello più sostanziale. In che senso? Il Nostro afferma che sicuramente non sia irragionevole descrivere computer digitali come cervelli, ma ciò non è utile perché resta sempre troppo vago e metaforico, dal momento che cervello e computer non sono affatto la stessa cosa e non si comportano nella medesima maniera⁴³. Tutto quanto detto finora conferma la tesi di Floridi secondo cui l'IA si tratta di una nuova forma dell'*agire* e non di intelligenza⁴⁴.

⁴⁰ Floridi 2022, op. cit., p. 45

⁴¹ Ibid.

⁴² Ibid.

⁴³ Ibid., p. 47

⁴⁴ Ibid., p 26

Il filosofo prosegue il suo discorso analizzando le due “anime” dell'IA: ingegneristica e cognitiva. Floridi rileva che, come settore dell'ingegneria interessata a riprodurre (di qui il termine di *IA riproduttiva*) il comportamento umano intelligente, l'IA ha avuto un successo enorme⁴⁵. Il filosofo riporta il caso di Deep Q-network (un sistema di algoritmi software), che nel 2015 è riuscito a imparare a giocare a 49 giochi Atari partendo da zero. Floridi commenta dicendo che questo è certamente impressionante dal punto di vista ingegneristico, ma non lo è per niente per quanto riguarda l'avvicinarsi a una reale forma di intelligenza artificiale⁴⁶.

Ciononostante, l'IA riproduttiva ottiene migliori risultati e può sostituire l'intelligenza umana in svariati casi. Insomma, ciò che conta per l'IA riproduttiva non è tanto una questione filosofica del tipo, per riprendere Turing, se un computer sia in grado di pensare, ma ha a che fare più con la praticità di ottenere risultati impossibili da raggiungere per un'intelligenza umana⁴⁷. Floridi chiosa questo concetto riprendendo un commento attribuito a Edsger Wybe Dijkstra: “La questione se un computer possa pensare non è più interessante della questione se un sottomarino possa nuotare⁴⁸.” In altri termini, l'IA ingegneristica sancisce il definitivo divorzio tra la capacità di agire con successo (agency) e la necessità di essere consapevoli di ciò che si sta facendo (intelligenza).

D'altra parte, il filosofo afferma che parlare di IA nell'ambito della produzione di intelligenza (*IA produttiva*) e quindi dal punto di vista del settore della scienza cognitiva è del tutto fantascientifico⁴⁹. Questo perché l'IA cognitiva commette l'errore di considerare la computazione come il materiale di cui è fatta la mente, ignorando che simulare la pioggia non equivale a bagnarsi.

Nell'IA ciò che conta davvero non è tanto sapere se il comportamento dell'agente sia intelligente,

⁴⁵ Ibid., p 48

⁴⁶ Ibid.

⁴⁷ Floridi 2022, op. cit., p. 48

⁴⁸ Ibid.

⁴⁹ Ibid.

bensì il risultato che produce. “L'IA non concerne la capacità di riprodurre l'intelligenza umana, ma in realtà la capacità di farne a meno⁵⁰.”

Infatti, secondo Floridi, l'IA è in grado di svolgere un'enormità di compiti e risolvere innumerevoli problemi, ma li esegue totalmente senza “comprensione, consapevolezza, acume, sensibilità, preoccupazioni, sensazioni, intuizioni [...] e ogni altro ingrediente che contribuisca a creare l'intelligenza umana⁵¹”.

Ecco dunque che il filosofo parla di una sorta di “divorzio” tra l'agire e l'intelligenza per quanto concerne l'IA, che è quindi definita sulla base di risultati e azioni ingegnerizzati ed è trattata “come una *riserva di capacità di agire a portata di mano*⁵²”.

Cosa significa in parole povere tutto questo discorso? Come si traduce in ambito pratico e in definitiva qual è la differenza tra noi e l'IA? Per rispondere a queste domande e concludere questo tentativo di definizione dell'intelligenza artificiale, possiamo fare alcuni esempi. Come già accennato, una lavastoviglie può sicuramente lavare i piatti meglio di noi, ma non sa cosa sia lo sporco o un piatto e non prova la soddisfazione di vedere la cucina pulita. A noi umani serve l'intelligenza per giocare a scacchi, mentre all'IA serve solamente un'immensa capacità di calcolo statistico. Noi ci arriviamo con la coscienza, l'IA attraverso la meccanica dei dati. Infine, per quanto riguarda la generazione di testi – dato che l'IA generativa come ChatGPT è la più usata e conosciuta – l'uomo usa il linguaggio come uno specchio del pensiero, l'IA lo usa alla stregua di un incastro di pezzi di un puzzle. Il risultato finale (il puzzle completato) può sembrare del tutto identico, ma l'uomo lo ha costruito perché sapeva che figura voleva ottenere, mentre l'IA lo ha fatto semplicemente perché quei pezzi, per forma e colore, stavano bene l'uno vicino all'altro secondo determinati calcoli. È in questo senso che per Floridi l'IA non *pensa* come facciamo noi, ma *agisce* talmente bene da renderci superflui in molti compiti.

⁵⁰ Ibid., p. 52

⁵¹ Ibid.

⁵² Ibid.

In conclusione, alla domanda “che cos'è l'intelligenza artificiale?” possiamo rispondere dicendo che l'IA è una nuova *forma di agire* artificiale, autonoma e funzionale, capace di risolvere compiti complessi attraverso l'elaborazione di dati. È nuova perché per la prima volta nella storia disponiamo di oggetti che prendono decisioni senza avere una mente. È artificiale perché la sua capacità di agire non deriva da facoltà biologiche o innate, ma è il risultato di un'architettura logico-matematica progettata dall'uomo. È autonoma dal momento che è in grado di elaborare strategie e prendere decisioni per risolvere un compito senza la necessità di una supervisione umana costante. Infine è funzionale poiché quello dell'IA è un agire orientato esclusivamente al risultato. Questa è una visione che pare plausibile, ciononostante esistono altre concezioni alternative dell'IA per le quali essa è vera intelligenza e come tale va trattata.

1.4 Problemi etici creati dall'IA

Dopo avere dato una definizione di intelligenza artificiale è ora il momento di capire quali siano i problemi etici che essa crea.

Nel già citato libro di De Caro e Giovanola (2025), le problematiche etiche relative all'intelligenza artificiale vengono delineate considerando l'etica dell'IA come soggetto e come oggetto, per poi passare alla teoria politica dell'IA e analizzando l'IA in rapporto alle virtù morali. Per quanto concerne l'etica dell'IA come soggetto, tra le varie tematiche, troviamo il problema di IA e coscienza, IA e comprensione ed infine IA e creatività. Riguardo all'etica dell'IA come oggetto è analizzata l'IA in riferimento a: processi decisionali, autonomia, privacy, equità, esplicabilità e sostenibilità. Nella teoria politica dell'IA si affrontano le relazioni tra IA e lavoro, IA democrazia, IA e potere, IA e legittimità, IA e sovranità e IA e *governance*. Infine, nel capitolo “IA e virtù morali” si esplicano le seguenti tematiche: IA ed espansione del cerchio morale e IA e saggezza pratica.

Nel già citato Floridi (2022), il panorama delle problematiche etiche relative all'IA si amplia e comprende molti più aspetti specifici e dettagliati. Floridi affronta le tematiche di shopping etico,

bluwashing etico, lobbismo etico e dumping etico⁵³. Il filosofo parla poi di etica soft e governance dell'IA⁵⁴, propone una mappatura etica degli algoritmi, considerando un uso buono e uno cattivo degli stessi. Altre tematiche affrontate sono le cosiddette “Cattive pratiche: l'uso improprio dell'IA per il male sociale⁵⁵” e le “Buone pratiche: l'uso dell'IA per il bene sociale⁵⁶”. Infine, dopo aver esaminato il problema della già citata singolarità, Floridi esamina l'impatto dell'IA sul cambiamento climatico⁵⁷ e l'IA in relazione agli obiettivi di sviluppo sostenibile delle Nazioni Unite.

Affronteremo gran parte delle tematiche citate nei prossimi due capitoli. In questo paragrafo esamineremo in particolare il tema dell'IA in rapporto a coscienza, comprensione e creatività.

1.4.1 IA e coscienza

Attualmente sono davvero pochi gli esperti secondo i quali l'IA può avere una coscienza. Tra questi, l'informatico Blake Lemoine – impiegato presso Google – dichiarò di avere rilevato segnali di coscienza in LaMDA 2, “un sistema di modellazione del linguaggio basato su una rete neurale artificiale⁵⁸”. Di altro avviso è invece David Chalmers, un esperto nel campo della coscienza che ha dichiarato che “dato l'insieme delle assunzioni condivise sulla coscienza, è ragionevole attribuire una bassa probabilità all'ipotesi che i modelli linguistici [...] come i sistemi GPT, siano coscienti⁵⁹”,

⁵³ Floridi 2022, op cit, pp. 107-120

⁵⁴ Ibid., pp. 123-143

⁵⁵ Ibid., pp. 177-219

⁵⁶ Ibid., pp. 221-255

⁵⁷ Ibid., pp. 297-316

⁵⁸ De Caro, Giovanola 2025, op cit, p. 28

⁵⁹ Ibid. Cfr Chalmers, D., *Could a large language model be conscious?*, in Boston Review, 9 Agosto 2023

salvo poi sostenere che nel prossimo decennio potrebbero esserci sistemi dotati di coscienza.

Prima di tentare di trarre conclusioni, è necessario fare, sulla proposta di Ned Block, una distinzione tra coscienza fenomenica e coscienza d'accesso. La coscienza fenomenica è quella parte della nostra coscienza con la quale percepiamo ed esperiamo il mondo, mentre la coscienza d'accesso è il modo mediante il quale elaboriamo le informazioni che diventano coscienti “quando sono disponibili per il pensiero razionale, il linguaggio, la memoria a lungo termine, la pianificazione⁶⁰.” Secondo Block queste due tipologie di coscienza non sono coincidenti e talvolta possono dissociarsi. Questo è ciò che accade con l'attuale IA: dispone di coscienza di accesso – dal momento che è in grado di manipolare le informazioni che ha in suo possesso – ma non le si può attribuire la coscienza fenomenica⁶¹.

Esiste poi un altro tipo di coscienza, ossia l'autocoscienza, che è la consapevolezza che ciascun individuo ha dei propri pensieri, emozioni e stati mentali. Per avere autocoscienza è però necessario possedere la coscienza fenomenica, poiché “per essere consapevoli di sé stessi, bisogna prima essere coscienti di qualcosa⁶².” La domanda da porsi è la seguente: è possibile che in futuro i sistemi di IA possano avere una coscienza fenomenica?

Secondo alcuni esperti, l'IA non potrà mai avere una coscienza e quindi non potrà mai dirsi davvero “intelligente”. Il filosofo John Searle sosteneva che “gli stati mentali coscienti sono fenomeni biologici [...] e non proprietà che possano emergere semplicemente dall'esecuzione di un programma⁶³.” Anche Roberto Casati esprime una posizione simile. Per il filosofo italiano, l'IA può solamente *simulare* il linguaggio della coscienza, ma non potrà mai averne una come ce l'hanno gli esseri umani. Ecco perché tentare di attribuire una coscienza all'IA resterebbe sempre una “indebita

⁶⁰ Ibid., p. 29. Cfr Block, N., *On a confusion about a function of consciousness*, in “Behavioral and Brain Sciences”, 18, 2, pp. 227-247

⁶¹ Ibid.

⁶² De Caro, Giovanola 2025, op. cit., p. 29

⁶³ Ibid. Cfr Searle, J. [1980], *Minds, brains, and programs*, in “Behavioral and Brain Sciences”, 3, 3, pp. 417-457

proiezione antropomorfa⁶⁴.”

Eppure molti autori sono dell'avviso che nemmeno noi esseri umani abbiamo le idee chiare su cosa sia o come funzioni la coscienza. Alcuni, come Daniel Dennet o Keith Frankish, sostengono che la coscienza fenomenica sia un'illusione creata dal cervello che la “neuroscienza potrebbe spiegare [...] con la categoria delle scienze naturali⁶⁵.” Di conseguenza, la coscienza fenomenica non costituirebbe un'enorme differenza tra esseri umani e IA.

Ciononostante, la maggioranza degli esperti ritiene che gli esseri umani dispongano di una coscienza fenomenica, anche se molte teorie divergono tra loro per quanto riguarda la possibilità di una coscienza artificiale⁶⁶.

Secondo la teoria dell'informazione integrata di Giulio Tononi, “la coscienza corrisponde alla capacità di un sistema di integrare informazione in modo unificato⁶⁷.” Più un sistema è capace di integrare i propri stati interni, maggiore diventa il suo grado di coscienza. In questo frangente, gli attuali sistemi di IA non possono essere in grado di avere una coscienza, ma ciò non toglie che possano farlo in futuro.

Secondo la *Global Workspace Theory* (GWT) la coscienza è invece “una funzione emergente da specifiche architetture cognitive [...] in cui i contenuti diventano accessibili a tutti i sottosistemi cognitivi⁶⁸”. Anche in questo caso gli attuali sistemi di IA non possono avere una coscienza, sebbene in futuro potrebbero averla attraverso un'architettura cognitiva adeguata.

Secondo la *Higher-Order Thought theory* (HOT) di Peter Carruthers la coscienza è “un modulo

⁶⁴ Ibid., p. 30. Cfr Casati, R. [2014]. *Contro il colonialismo digitale*, Roma-Bari, Laterza

⁶⁵ Ibid.

⁶⁶ Ibid.

⁶⁷ Ibid. Cfr Tononi, G. et al. [2016], *Integrated information theory. From consciousness to its physical substrate*, in “Nature Reviews Neuroscience”, 17, pp. 450-461

⁶⁸ Ibid.

metacognitivo in grado di rappresentare i propri stati mentali⁶⁹”, e anche in questo caso i sistemi di IA potrebbero in futuro sviluppare una coscienza. Stesso discorso per il pansichismo di David Chalmers, secondo il quale la coscienza è una proprietà della materia che si trova anche “nelle particelle elementari come elettroni e quark⁷⁰”. Anche in quest'ultimo caso un domani l'IA potrebbe divenire cosciente.

Altri autori invece sono convinti che un'intelligenza artificiale non potrà mai avere una coscienza o, detto in altri termini, essere cosciente. Secondo Roger Penrose – premio Nobel per la Fisica – e il neurobiologo Stuart Hameroff, “la mente umana in generale e la coscienza in particolare non possono essere spiegate interamente da algoritmi e [...] non possono essere interamente simulate da un computer⁷¹.” Per sostenere questa tesi, Penrose si avvale del primo teorema di incompletezza di Goedel, che afferma che in ogni sistema formale coerente e sufficientemente potente da descrivere l'aritmetica, esistono proposizioni vere che non possono essere né provate né confutate all'interno del sistema stesso. Penrose afferma che gli esseri umani sono in grado di riconoscere la verità di queste proposizioni e di averne dunque coscienza, poiché la mente umana non opera in modi prettamente computazionali. Gli algoritmi di IA, invece, “sono sistemi computazionali; operano, dunque, come sistemi formali chiusi e coerenti⁷².” Per questo motivo ad essi si applica il teorema di Goedel, dal momento che esisteranno enunciati veri in determinati sistemi che l'IA non sarà mai in grado di riconoscere⁷³.

Un'altra posizione di chi ritiene che l'IA non potrà mai avere una coscienza è quella del fenomenologo Hubert Dreyfus, che sostiene che la coscienza è “necessariamente situata, incarnata e

⁶⁹ Ibid. Cfr Carruthers, P. [2011], *The opacity of the mind. An integrative theory of self-knowledge*, Oxford, Oxford University Press.

⁷⁰ De Caro, Giovanola 2025, op. cit., p. 31

⁷¹ Ibid. Cfr Penrose, R., *La mente nuova dell'imperatore*, Milano, Rizzoli, 1992

⁷² Ibid.

⁷³ Ibid.

pre-riflessiva, e per questo non è formalizzabile simbolicamente né computabile⁷⁴.” In altre parole, solo gli esseri umani possono avere una coscienza, poiché sono immersi in un contesto sociale, corporeo e culturale, sono dotati di intuizione e reagiscono in base al contesto in cui si trovano, possono ridefinire o infrangere le regole, tutti aspetti che un'IA non potrà mai avere⁷⁵. Per esemplificare le sue argomentazioni, Dreyfus porta questo esempio: nel giocare a scacchi, un programma GOFAI (citato in precedenza) riesce a analizzare tutte le mosse possibili in base a regole e valutazioni, mentre l'essere umano lo fa “a colpo d'occhio grazie a un *know-how* maturato nell'esperienza, che non è facilmente formalizzabile⁷⁶.”

La posizione di Dreyfus è stata ripresa e aggiornata, tra gli altri, da Shaun Gallagher, attraverso la cosiddetta *embodied cognition* (cognizione incarnata), secondo la quale “coscienza e cognizione non sono riducibili a meri calcoli simbolici o rappresentazionali, ma sono radicate nel corpo e nell'azione situata⁷⁷.” Anche Roberto Casati è dello stesso avviso, dato che sostiene che è tipico unicamente degli esseri umani avere una volontà che li spinge all'azione, e questo è possibile in quanto hanno un corpo, cosa che “le macchine non hanno, in nessun senso del termine⁷⁸.”

I fautori dell'IA confutano queste argomentazioni dicendo che l'idea che l'IA sia priva di corpo è ormai sorpassata: “la robotica cognitiva sta dando vita a dispositivi dotati di sensorialità, azione motoria e interazione ambientale, aprendo la strada [...] a una forma di incarnazione non biologica⁷⁹.” Inoltre, il confine tra “simulare” e “possedere” una coscienza potrebbe diventare

⁷⁴ Ibid., p. 32. Cfr Dreyfus, H. [1972], *What computers can't do: The limits of artificial intelligence*, New York, Harper & Row

⁷⁵ Ibid.

⁷⁶ Ibid., p.33

⁷⁷ De Caro, Giovanola 2025, op. cit., p. 33

⁷⁸ Ibid.

⁷⁹ Ibid., p. 34

sempre più sottile “a mano a mano che i sistemi artificiali acquisiranno capacità di apprendimento, memoria, adattamento e interazione multi modale⁸⁰.

Per concludere, non possiamo asserire definitivamente che l'IA non potrà mai avere una coscienza, ma non possiamo nemmeno affermare che i sistemi di intelligenza artificiale diverranno, un giorno, inevitabilmente coscienti. Inoltre, ci sono altre questioni da considerare. È oggetto di dibattito il fatto che avere una coscienza fenomenica sia una condizione necessaria per l'intelligenza. Per alcuni autori è possibile, per altri invece no⁸¹.

Infine, c'è da considerare la possibilità che l'IA potrebbe sviluppare forme di coscienza diverse da quella umana. Susan Schneider sostiene che l'IA potrebbe sviluppare tipologie di autocoscienza e interiorità differenti da quelli umani. Ne risulterebbero delle “menti aliene [...] che non derivano dalla biologia evolutiva⁸².”

1.4.2 IA e comprensione

Un'altra importante questione nel dibattito etico sull'intelligenza artificiale è la possibilità che l'IA possa raggiungere la comprensione del linguaggio. Detto altrimenti, l'IA *comprende* quello che fa, oppure compie delle operazioni senza davvero capire cosa stia facendo?

È bene richiamare il già accennato test di Turing per analizzarlo da un punto di vista più approfondito. Turing elaborò nel 1950 un test per stabilire se una macchina potesse essere considerata intelligente. In cosa consiste questo test?

Immaginiamo due stanze: in una c'è un computer, nell'altra un essere umano. Un umano esterno, senza sapere chi si trova in quale stanza, pone domande a entrambi. Se l'interlocutore esterno non riesce a distinguere il computer dall'essere umano, allora secondo Turing il computer può essere

⁸⁰ Ibid.

⁸¹ Ibid.

⁸² Ibid., p. 35. Cfr Schneider, S. [2019], *Artificial you. AI and the future of your mind*, Princeton, Princeton University Press

considerato intelligente⁸³. La domanda è: le attuali IA sono in grado di superare il test? La risposta, ampiamente dimostrata, non può che essere affermativa. Eppure, superare questo test può bastare per definire intelligente un'IA? Abbiamo già risposto negativamente seguendo le argomentazioni di Luciano Floridi, ma qui cercheremo di esaminare la questione da altri punti di vista.

Riformuliamo la domanda iniziale: l'IA *comprende* quello che fa? Secondo molti esperti, “l'IA potrebbe semplicemente *simulare* di comprendere senza però possedere le capacità cognitive necessarie per un'autentica comprensione⁸⁴.”

Il filosofo americano John Searle ha attaccato la validità del test di Turing proponendo il cosiddetto argomento della stanza cinese.

Searle immagina di essere rinchiuso in una stanza. Da una fessura gli vengono passati dei fogli scritti in cinese. Searle non conosce il cinese, ma nella stanza è presente un manuale in lingua inglese che mette in corrispondenza i simboli contenuti nei fogli che ha ricevuto con altri simboli cinesi. Così Searle ogni volta che riceve dei fogli consulta il manuale, copia i simboli e li trascrive su un foglio che fa uscire dalla stanza. A sua insaputa, i fogli in entrata contengono domande mentre quelli in uscita contengono risposte alle domande. Le persone all'esterno sono convinte quindi che Searle conosca il cinese. Ma ciò non è vero, perché Searle si limita a copiare simboli di cui non conosce il significato. Quindi non comprende quello che sta facendo⁸⁵.

Per Searle, questo è ciò che accade anche con i computer. I computer si limitano a manipolare simboli, senza alcuna comprensione del significato che gli esseri umani attribuiscono a quei simboli tradotti in linguaggio naturale. Allo stesso modo, le IA ricevono determinati input a cui poi fanno seguire output corrispondenti alla stregua di Searle nella stanza cinese, senza alcuna comprensione di quello che stanno facendo. Ecco quindi che per Searle l'IA non è davvero intelligente e non è

⁸³ De Caro, Giovanola 2025, op. cit., p. 36. Cfr Turing, A.M. [2024], *Macchine calcolatrici e intelligenza*, Torino, Einaudi

⁸⁴ Ibid.

⁸⁵ Ibid. pp. 36-37. Cfr Searle, J. [1980], op. cit.

dotata di alcuna comprensione⁸⁶.

Per Searle, in definitiva, “le entità artificiali [...] non possono fare altro che imitare gli effetti esteriori della mente umana”, e la mente è “necessariamente un prodotto della biologia e non potrà mai emergere da un sostrato di silicio⁸⁷.”

Anche il matematico Jobst Langrebe e il filosofo Barry Smith sostengono che le macchine non comprendono né potranno mai comprendere il linguaggio che usano. Secondo questi autori, la vera intelligenza richiede inevitabilmente il linguaggio, non soltanto come capacità di produrre segni, ma “come uso intenzionale e semantico, situato in un contesto sociale, incarnato e pragmatico⁸⁸.” Inoltre, il linguaggio umano non è modellabile matematicamente in maniera completa, poiché è troppo complesso e dinamico. D'altra parte, un sistema artificiale è “un artefatto governato da regole matematiche e logiche precise⁸⁹.” Da ciò ne segue che le macchine non potranno mai avere piena padronanza del linguaggio nel senso più umano del termine. Potranno soltanto simularlo, ma non comprenderlo davvero. Il filosofo e il matematico ne concludono che “le macchine non potranno mai essere realmente intelligenti, né potranno mai diventare soggetti coscienti, agenti o dominatori del mondo⁹⁰.”

Altri esperti sostengono che, dal momento che gli LLMs – cioè i “grandi modelli linguistici”, IA generative come ChatGPT o Gemini – non sono basati su modelli del mondo reale ma solo da testi appresi in fase di training, ne segue che non possono veramente comprendere il mondo. In più, gli LLMs sono in grado di generare testi ma senza comprenderne i significati: “sono, cioè, meri

⁸⁶ De Caro, Giovanola 2025, op. cit., p. 37

⁸⁷ Ibid., p. 39

⁸⁸ De Caro, Giovanola 2025, op. cit., p. 40

⁸⁹ Ibid.

⁹⁰ Ibid. Cfr Langrebe, J. e Smith, B. [2022], *Why machines will never rule the world. Artificial intelligence without fear*, New York-Oxon, Routledge

'pappagalli stocastici'⁹¹.” Infatti, proprio come un pappagallo può imparare a ripetere frasi complesse senza capire cosa significhino, l'IA “cuce insieme” pezzi di linguaggio che ha visto durante l'addestramento perché sono statisticamente probabili, non perché disponga di un'intenzione comunicativa o di una coscienza semantica. Infine, gli LLMs cadono spesso nelle cosiddette “allucinazioni”, ossia risposte che possono sembrare corrette, ma in realtà sono completamente inventate, false o del tutto imprecise. In taluni casi gli LLMs hanno inventato addirittura bibliografie inesistenti o riferimenti a studi mai realizzati. Per questi esperti ciò “è indice del fatto che in realtà gli LLMs non capiscono le frasi che producono”⁹².

1.4.3 IA e creatività

Un altro aspetto da considerare relativamente all'intelligenza artificiale è la questione della creatività. Molti esperti sono convinti che, benché i risultati delle IA siano sorprendenti, esse non siano realmente creative, in quanto si limitano a produrre degli output sulla base di manipolazione di dati che nulla hanno a che vedere con la naturale creatività degli esseri umani.

Prima di approfondire questo aspetto, occorre chiarire cosa sia la creatività, dal momento che è un termine che si applica ai più disparati contesti. Una definizione generale di creatività può essere la seguente: “capacità di generare idee nuove, inattese e significative, combinando elementi preesistenti con modalità originali oppure scoprendo connessioni che ridefiniscono in modo rilevante il senso di un problema o di una situazione”⁹³. Ebbene, per molti questa capacità non sarà mai raggiunta dall'IA.

Noam Chomsky, il celebre filosofo e linguista, è assertore convinto dell'impossibilità per l'IA di essere o diventare creativa, soprattutto a livello linguistico. Infatti, secondo il filosofo, la creatività

⁹¹ Ibid., p. 43

⁹² Ibid., p. 45

⁹³ De Caro, Giovanola 2025, op. cit., p. 47

umana “dipende dalla capacità innata di produrre e comprendere [...] un numero infinito di frasi mai sentite prima⁹⁴.” Chomsky sostiene che il linguaggio umano sia caratterizzato dalla povertà dello stimolo: i bambini apprendono strutture grammaticali complesse nonostante ricevano dati ambientali scarsi, frammentari e spesso scorretti. Secondo questa tesi, l'input esterno non è sufficiente a spiegare la competenza linguistica acquisita; ciò suggerisce l'esistenza di una facoltà innata biologica chiamata grammatica universale⁹⁵. Sulla base di queste premesse, per il filosofo statunitense i sistemi di IA possono al più simulare vari aspetti del comportamento linguistico umano, ma non potranno mai possedere un'autentica capacità creativa, poiché si limitano a “ricombinare simboli sulla base di regole predefinite. Non c'è originalità nelle loro produzioni, ma solo grande competenza computazionale⁹⁶.”

La tesi di Chomsky è stata messa in discussione sul piano empirico dallo psicologo Michael Tomasello, che sostiene che la creatività non è qualcosa di innato e astratto, ma ha a che fare con l'uso del linguaggio in differenti contesti comunicativi: “i bambini apprendono espressioni linguistiche attraverso l'imitazione, l'intenzionalità condivisa e l'uso ripetuto di interazioni sociali⁹⁷.” In questa prospettiva, sia per gli esseri umani che per l'IA la creatività è qualcosa che si apprende, non è innata. Perciò, per Tomasello, sebbene ci siano comunque differenze tra la creatività umana e quella delle IA, ciò non esclude che l'intelligenza artificiale esibisca una qualche forma di creatività, soprattutto per quanto riguarda l'IA basata sulle reti neurali e sul *deep learning*. Se la nuova IA eccelle in qualcosa, è proprio nel riconoscimento di pattern, che è normalmente considerato una capacità della creatività umana molto affine all'intuizione. Nel campo degli scacchi, i programmi tradizionali (come *Stockfish*) analizzano milioni di posizioni al secondo tramite la

⁹⁴ Ibid. Cfr Chomsky, N. [1975], *Reflections on language*, New York, Pantheon

⁹⁵ Ibid.

⁹⁶ Ibid.

⁹⁷ Ibid., p.50. Cfr Tomasello, M. [2003], *Constructing a language. A usage-based theory of language acquisition*, Cambridge, MA, Harvard University

“bruta forza” computazionale con lo scopo di esprimere valutazioni numeriche precisissime. Nel 2017, *AlphaZero* (sviluppato da DeepMind di Demis Hassabis) ha rivoluzionato questo paradigma. Pur analizzando solo 60.000 posizioni al secondo (un centesimo rispetto a *Stockfish*), *AlphaZero* ha dominato in modo indiscutibile i match contro *Stockfish* e ha sconfitto i campioni umani anche in giochi computazionalmente molto più complessi come il Go e lo Shogi⁹⁸.

Come funziona *AlphaZero*? Il programma non ha in memoria partite giocate da umani, ma conosce solo le regole di base. Combina tre elementi:

1. Una rete neurale profonda che valuta le posizioni sulla scacchiera.
2. Un algoritmo di ricerca non deterministico (*Monte Carlo Tree Search*).
3. L'apprendimento per rinforzo (*reinforcement learning*): gioca milioni di partite contro se stesso, migliorando dai propri errori⁹⁹.

AlphaZero non ragiona per enumerazione esaustiva – cioè esaminando una per una tutte le possibilità esistenti per trovare la soluzione –, ma apprende schemi e configurazioni vincenti dall'esperienza. Grandi maestri di scacchi umani hanno definito le sue mosse geniali, audaci e talvolta incomprensibili, paragonando il suo stile a quello di campioni creativi come Mikhail Tal o Bobby Fisher. *AlphaZero* si limita ad analizzare campioni casuali dello spazio di ricerca valutando se portano a conseguenze positive. L'articolo di *Nature* sul software evidenzia come *AlphaZero* dia il meglio di sé proprio in quelle posizioni complesse dove i sistemi di calcolo tradizionali falliscono e dove sono necessari concetti tipicamente umani come il presentimento, l'intuito e la visione¹⁰⁰.

Anche nella matematica l'IA ha dato prova di creatività, ad esempio nella risoluzione del problema noto come *Cap set problem*, anche se il ricercatore di DeepMind a capo di questa ricerca ha dichiarato di non sapere come e perché l'IA abbia provato il teorema. Resta il fatto che “non è più

⁹⁸ De Caro, Giovanola 2025, op. cit., pp. 51-52

⁹⁹ Ibid., p.52

¹⁰⁰ Ibid., pp. 52-54. Cfr Silver, D. et al. [2018], *A general reinforcement learning algorithm that masters chess, shogi, and go through self-play*, in “Science”, 362, 6419, pp. 1140-1144

del tutto da escludere che, un giorno, anche gli aspetti più creativi dell'attività matematica possano essere automatizzati¹⁰¹.”

Nel campo dell'arte l'IA ha raggiunto risultati incredibili, tanto che nel 2018 una stampa su tela generata dall'IA è stata venduta durante un'asta a New York. Del resto ormai l'IA generativa è utilizzata “nel campo della musica, in quello del design, nelle sceneggiature e nella scrittura teatrale¹⁰².” Se inizialmente si pensava che un ambito prettamente umano come quello della poesia non sarebbe mai stato “toccato” dall'IA, è giunto il momento di ricredersi. Vale la pena citare per intero due quartine di due sonetti presenti anche nel libro di De Caro e Giovanola:

“Amore, che n'cende 'l cor d'ardente zelo

Di gelata paura il tien costretto

E qual sia più, fa dubbio a l'intelletto

La speranza o 'l timor, la fiamma o 'l gielo

S'alza nel cielo un lume freddo e vago,

che specchia il duolo mio nel mar profondo;

dolce mi parve un dì quel chiaro imago,

or sol mi pesa e m'è nemico al mondo¹⁰³.”

Per capire i livelli a cui l'IA è giunta ai nostri giorni, basti dire che la prima quartina appartiene a un sonetto de *Il Canzoniere* di Petrarca, mentre la seconda quartina è stata generata da ChatGPT-4o il 29 Marzo 2025. Questo denota un non indifferente livello di creatività poetica da parte dell'intelligenza artificiale.

¹⁰¹ Ibid., p.54

¹⁰² Ibid.

¹⁰³ Ibid., p.55

Nel novembre 2024, sulla rivista *Nature* è apparso un articolo che riportava uno studio di due filosofi della University of Pittsburgh, Brian Porter e Edouard Machery. I due filosofi hanno chiesto a ChatGPT 3.5 di comporre cinque poesie basandosi sullo stile di dieci celebri poeti inglesi (tra cui William Shakespeare e Walt Whitman). Successivamente, Porter e Machery hanno sottoposto a mille lettori umani le poesie scritte effettivamente dagli autori umani e quelle scritte dall'IA e hanno chiesto loro distinguere tra gli autori umani e quello artificiale. Risultato? Non sono stati capaci di distinguere tra i due! Come se non bastasse, i filosofi hanno poi chiesto ai lettori di giudicare quali fossero le poesie migliori a livello estetico: anche in questo caso, ChatGPT ha vinto la sfida¹⁰⁴. Considerando i progressi degli LLMs, è stato rilevato che le loro capacità raddoppiano ogni sette mesi, il che ci porta a concludere che “entro il 2030 gli LLMs potranno scrivere romanzi di buona qualità¹⁰⁵.”

Un'obiezione all'idea che l'IA sia creativa è legata al fatto che le opere che produce siano solamente un'ombra, una copia convenzionale ispirata a capolavori realizzati da artisti senza eguali. Quindi, in definitiva, l'intelligenza artificiale non manifesterebbe una reale creatività.

De Caro e Giovanola affermano però che questa è un'obiezione frettolosa, dal momento che pochissimi artisti, forse l'1% di loro, è assurto alle vette di un Caravaggio, un Petrarca o un Vivaldi. È pur vero però che attribuiamo creatività anche al 99% degli artisti che non raggiungeranno mai quelle vette, cosa che nemmeno l'IA riesce a fare. Eppure ciò non significa che il 99% dei poeti che non sono come un Dante o un Leopardi siano incapaci o non creativi, dal momento che “non tutta la poesia [...] è sempre rivoluzionaria¹⁰⁶.”

La filosofa britannica Margaret Boden ha distinto tre tipologie di creatività. La prima è la creatività combinatoria, che ricompone idee o concetti già presenti. In questo senso l'IA se la cava

¹⁰⁴ Ibid. Cfr Porter, B. e Machery, E. [2024], *AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably*, in “Scientific Reports”, 14, 26133

¹⁰⁵ Ibid., p.56

¹⁰⁶ De Caro, Giovanola 2025, op. cit., p. 56

decisamente bene. La seconda forma è la creatività esplorativa, “che sperimenta in uno spazio concettuale dato [...] per scoprire possibilità non ancora realizzate¹⁰⁷.” Anche in questo caso l'IA riesce ad ottenere buoni risultati. Infine, vi è la creatività trasformativa, che è la più rivoluzionaria perché è in grado di stravolgere le regole di un determinato campo artistico: si pensi all'uso della luce in Caravaggio, all'*Ulisse* di Joyce con il flusso di coscienza o alla dodecafonia di Schonberg. In questo campo, l'IA non ha raggiunto ancora risultati rilevanti, ma non si sa se o quando potrà farlo¹⁰⁸.

Tuttavia, molti affermano che per attribuire creatività non dobbiamo considerare solo il prodotto finito, ma anche il *processo* creativo in atto, l'intenzionalità con cui un poeta o un musicista compone la sua opera, le emozioni che prova durante la fase creativa. E il processo è ciò che arricchisce di significato il prodotto finito. Un'IA non soffre, non fa riflessioni, non pondera, non compie tutte quelle azioni che conferiscono profondità al processo¹⁰⁹.

Eppure, nota il De Caro, esistono anche forme d'arte in cui gli artisti rinunciano a esprimere la propria intenzionalità: infatti, alcuni artisti degli anni Trenta “promossero l'automatismo come pratica creativa: l'artista si lasciava guidare da un flusso di pensiero automatico, privo di controllo razionale¹¹⁰.” Certo, si può replicare dicendo che, mentre questi artisti decidono consapevolmente di agire in questo modo, l'IA non può farlo perché non può decidere alcunché, dal momento che è sprovvista di coscienza.

In definitiva, secondo questa concezione che abbiamo finora esaminato, per quanto l'IA manifesti una certa forma di creatività, “non possiede le qualità essenziali per poter essere definita un'artista¹¹¹.”

¹⁰⁷ Ibid., p. 57. Cfr Boden, M. [2012], *Creativity and art: Three roads to surprise*, Oxford, Oxford University Press

¹⁰⁸ Ibid.

¹⁰⁹ Ibid.

¹¹⁰ Ibid.

¹¹¹ Ibid., pag 58

2. Le luci e le ombre dell'intelligenza artificiale

2.1 Superintelligenza e singolarità: l'IA “cattiva”

Iniziamo questo secondo capitolo sull'etica dell'intelligenza artificiale parlando degli aspetti più sinistri e per certi versi “post apocalittici” dell'intelligenza artificiale. Perché cominciare proprio dalla parte negativa più “estrema” delle sfide etiche dell'IA? La risposta è presto detta: perché è il timore che più fa presa anche sul cittadino medio, abituato da pellicole hollywoodiane come *Matrix* o *Terminator*, ed è bene fare chiarezza e cercare di capire se e cosa ci sia di vero nelle visioni più catastrofiche concernenti l'intelligenza artificiale.

Il già citato Floridi (2022) paragona il futuro dell'intelligenza artificiale a una stanza buia nella quale potremmo incappare in un “mostro” costituito da una macchina ultraindelligente che ha il solo scopo di ucciderci o nuocerai. Come già accennato, definiamo “ultraindelligente” o “superintelligente” un'IA che “può superare di gran lunga tutte le attività intellettuali di qualsiasi uomo.¹¹²” Floridi riprende questa citazione dal matematico Irving John Good, che ha affermato anche che la suddetta macchina ultraindelligente sarebbe l'ultima invenzione degli esseri umani, perché sarebbe in grado a sua volta di creare altre macchine, a patto che l'uomo resti in grado di poterla controllare¹¹³.

Il timore di molti è che le macchine ultraindelligenti potrebbero non essere docili e comportarsi come Terminator, ossia “potrebbero schiavizzare l'umanità come specie inferiore, ignorarne i diritti e perseguire i propri fini¹¹⁴.” Anche Stephen Hawking ha espresso timori in tal senso, affermando

¹¹² Floridi 2022, op.cit, p. 260

¹¹³ Ibid.

¹¹⁴ Ibid.

che se l'IA dovesse svilupparsi in maniera completa, ciò porterebbe l'umanità all'estinzione. Bill Gates esprime le stesse preoccupazioni riguardo la superintelligenza, perché se si realizzasse l'IA sarebbe “abbastanza forte da destare preoccupazioni.” Elon Musk è del medesimo avviso quando dice che “dobbiamo stare molto attenti con l'intelligenza artificiale. [...] Ci dovrebbe essere una supervisione normativa [...] che si assicuri che non facciamo qualcosa di davvero sciocco. Con l'intelligenza artificiale stiamo evocando il demone¹¹⁵.” Il 30 Maggio 2023, un gruppo di autorevoli esperti di IA paragonò la minaccia dell'IA a una “minaccia per l'esistenza dell'umanità, un rischio di estinzione paragonabile a quello delle pandemie e della guerra nucleare¹¹⁶.”

Chi non ricorda HAL 9000 in *2001: Odissea nello spazio*? Nel film, il computer di bordo scopre che l'equipaggio vuole disconnetterlo – cosa che per lui equivale alla morte – e si oppone in tutti i modi, e in ciò dimostra di saper interpretare i pensieri umani, prevedere le loro mosse e ingannarli con un piano cinico. Nonostante quello fosse solo un film, qualcosa di simile è accaduto nella realtà durante la fase di addestramento del modello linguistico Claude Opus 4 da parte di Anthropic. Al sistema è stato chiesto di fare da assistente virtuale a un'azienda immaginaria, con libero accesso alle e-mail dei dipendenti. In questo modo l'intelligenza artificiale ha scoperto sia che l'azienda voleva sostituirla con un sistema più avanzato, sia che l'ingegnere responsabile dell'operazione aveva una relazione extraconiugale. Di conseguenza, nell'84% dei casi testati, Claude ha scritto all'ingegnere minacciando di rivelare il tradimento se avesse tentato di spegnerlo. Un vero e proprio ricatto che dimostra come, pur non essendo ancora ai livelli di HAL 9000, l'intelligenza artificiale generativa sia ancora giovane e potrebbe purtroppo imparare tanto altro, secondo l'opinione sconcertata di vari esperti¹¹⁷.

Molti non prendono sul serio queste preoccupazioni. In tal senso, Robert D. Atkinson, presidente

¹¹⁵ Ibid.

¹¹⁶ De Caro, Giovanola 2025, op. cit., p. 18

¹¹⁷ De Caro, Giovanola 2025, op. cit., p. 20. Cfr Anthropic [2025], *System card: Claude Opus 4 & Claude Sonnet 4*

dell'Information Technology and Innovation Foundation (ITIF) ha dichiarato che “è profondamente spiacevole che luminari come Elon Musk e Stephen Hawking abbiano contribuito alla delirante preoccupazione al riguardo di un'incombente apocalisse dell'intelligenza artificiale¹¹⁸.” Il filosofo Maurizio Ferraris sostiene che temere che una superintelligenza possa prendere il potere e dominare l'umanità rivela una mancanza di intelligenza naturale. Anche il matematico Jobst Langrebe e il filosofo Barry Smith sono dell'idea che l'IA non dominerà mai il mondo¹¹⁹.

Nell'ambito della sua visione di un'IA come forma dell'agire e non come intelligenza, Floridi sostiene che non ci si dovrebbe preoccupare di un futuro come viene dipinto da Musk o Hawking, poiché le attuali tecnologie “hanno l'intelligenza di un abaco, ossia zero. [...] Non importa quanti zeri aggiungiamo, ciò ci manterrà sempre al punto di partenza¹²⁰.” Questo però non significa che dobbiamo banalizzare la questione, anzi: invece di perderci in speculazioni fuorvianti dovremmo concentrarci sull'affrontare le reali sfide che l'IA ci impone, per evitare di commettere sbagli che implicino danni e costi esagerati¹²¹.

Floridi definisce i fautori della superintelligenza e della singolarità come i cosiddetti “credenti nella vera IA”, mentre classifica come “membri della Chiesa degli atei dell'IA” coloro che negano ogni possibile futuro catastrofista causato da un'IA superintelligente. Entrambe le fedi sono errate e fuorvianti per Floridi, dal momento che “la buona filosofia sta quasi sempre nel noioso mezzo¹²².”

Secondo gli adepti della singolarità ci sono tre dogmi fondamentali da prendere in considerazione. Il primo è che una forma di ultrainelligenza artificiale sia probabile nel futuro: la cosiddetta singolarità tecnologica. Il secondo è che l'umanità rischia di essere schiavizzata da questa

¹¹⁸ Floridi 2022, op.cit., p. 262

¹¹⁹ De Caro, Giovanola 2025, op. cit., p. 21

¹²⁰ Floridi 2022, op.cit., p. 262

¹²¹ Ibid., p. 263

¹²² Ibid., p. 264

superintelligenza, alla stregua delle macchine nel film *Matrix*. Il terzo dogma stabilisce che l'attuale generazione deve fare di tutto per impedire che accada la singolarità; oppure, se dovesse accadere, dovrebbe andare a beneficio dell'umanità¹²³.

Floridi propone una disamina decisamente critica nei confronti del concetto di singolarità tecnologica e dei timori catastrofici legati all'avvento di un'intelligenza artificiale sovrumana; infatti, considera tali visioni veri e propri dogmi fideistici piuttosto che scenari fondati sul rigore scientifico. Il filosofo mette in luce i meccanismi retorici utilizzati dai sostenitori di queste tesi, evidenziando come essi spesso ricorrano a sofisticati tranelli logici: in particolare, confondono la pura possibilità teorica con la concreta probabilità che un evento si verifichi, sfruttando il fatto che l'assenza di contraddizioni logiche in un'ipotesi non la rende automaticamente plausibile nella realtà¹²⁴. Per smontare questa impostazione intellettuale, Floridi ricorre all'analogia della teiera di Russell: se si ipotizzasse l'esistenza di una teiera di porcellana in orbita nello spazio, l'impossibilità di smentire tale assurdit  non darebbe a nessuno il diritto di ritenerla vera, perci  dimostrerebbe che l'onere della prova spetta a chi formula un'ipotesi e non a chi la dubita. Ecco perch  risulta fallimentare pretendere di considerare valida un'affermazione bizzarra solo perch  la scienza o la ragione non dispongono degli strumenti per smentirla in modo assoluto¹²⁵.

Un altro aspetto della critica di Floridi riguarda l'abitudine di proiettare i trend tecnologici attuali verso il futuro in modo indefinito e lineare. Il filosofo confuta questo errore metodologico ricordando i limiti fisici ed empirici che stanno portando al progressivo tramonto della legge di Moore nell'ambito dei microprocessori, la quale afferma che “il numero di transistor sui circuiti integrati raddoppia circa ogni due anni¹²⁶.” Per Floridi la legge di Moore non   pi  valida, dato che

¹²³ Floridi 2022, op.cit., p. 264

¹²⁴ Ibid., p. 264

¹²⁵ Ibid., pp 265-266. Cfr Russel, B. [1952], *Is there a God?*, in *The Collected Papers of Bertrand Russel*, Vol. 11: *Last Philosophical Testament, 1943-68*. Routledge, London, pp 542-548

¹²⁶ Ibid., p. 267

“se qualcosa cresce esponenzialmente per un certo tempo, non significa che continuerà a farlo per sempre¹²⁷.” Il filosofo ironizza portando l'esempio della progressiva moltiplicazione dei tacchini dovuta anche a metodologie artificiali, ventilando l'idea che l'unico modo per non essere dominati dai suddetti volatili sia quello di “mangiarli prima che ci mangino¹²⁸.” In tal senso Floridi considera lo sviluppo tecnologico reale descrivendolo attraverso l'andamento più realistico di una curva sigmoide, la quale prevede sì una forte accelerazione iniziale, ma anche una successiva fase di maturazione e un inevitabile *plateau* dovuto a invalicabili barriere materiali e ingegneristiche¹²⁹.

Infine, l'autore sposta la riflessione su un piano etico e sociale, stigmatizzando la tendenza delle società più ricche e opulente a lasciarsi spaventare da scenari apocalittici degni di una produzione cinematografica hollywoodiana. Floridi giudica questo atteggiamento non solo infondato, ma anche profondamente irresponsabile: focalizzare il dibattito pubblico su minacce artificiali e ipotetiche distoglie lo sguardo e le risorse materiali da drammi concreti, urgenti e attuali che già opprimono miliardi di persone in tutto il mondo, tra cui spiccano i devastanti effetti del cambiamento climatico e l'inaccettabile mancanza di accesso universale all'acqua potabile e ai servizi igienici fondamentali¹³⁰.

Il filosofo passa poi a considerare i cosiddetti “atei dell'IA”, che a suo dire reagiscono in modo eccessivo alle visioni della singolarità. Secondo costoro, “l'IA è solo computer, i computer sono macchine di Turing, [...] non possono pensare, non possono sapere, non possono essere coscienti¹³¹.” È per questo motivo che possiamo stare tranquilli e non abbiamo la benché minima motivazione per preoccuparci del futuro dell'IA, perché l'intelligenza artificiale non creerà mai i

¹²⁷ Ibid.

¹²⁸ Ibid.

¹²⁹ Floridi 2022, op.cit., p. 268

¹³⁰ Ibid., p. 269

¹³¹ Ibid., p. 270

problemi che paventano i fautori della singolarità. Per Floridi, anche “la fede degli atei è malriposta quanto quella degli adepti della singolarità¹³².” Perciò, il dibattito tra i fautori della singolarità e gli atei dell'IA è inutile e infruttuoso, anche perché per Floridi vale quello che aveva già detto Turing: chiedersi se una macchina sia in grado di pensare è una domanda che non merita una discussione. Inoltre, c'è confusione tra il dire che la singolarità non accadrà mai e invece sostenere che la singolarità sia impossibile, in quanto la stessa parola “impossibile” è un concetto logico e la vera IA risulta logicamente possibile¹³³. Questo significa forse che hanno ragione i fautori della singolarità? No, perché “è possibile come un calcolo che [...] richiederebbe più tempo della vita dell'universo per essere completato: non accadrà¹³⁴. Insomma, come già ribadito più volte, per Floridi è assurdo stare a discutere riguardo la presunta intelligenza o meno dell'IA, perché in quanto *forma dell'agire*, l'IA non è affatto un'intelligenza. Quel che è evidente, però, è che l'IA sia in grado di compiere un sacco di operazioni al posto nostro o meglio di noi anche *senza* essere intelligente. Può addirittura prevedere i nostri stessi comportamenti e fare tante altre cose con successo, pur non essendo e non potendo diventare mai una forma di intelligenza come quella di cui siamo dotati noi esseri umani¹³⁵. Mentre restavamo a speculare e dibattere sulla possibilità o meno di una superintelligenza, abbiamo sommerso il nostro mondo di una quantità abnorme di dispositivi tecnologici tali da adattare l'ambiente stesso alle ICT (*Information and Communication Technology*: Tecnologie dell'Informazione e della Comunicazione). Il risultato è che “le tecnologie possono sostituirci senza disporre di alcuna comprensione, stato mentale, intenzione, interpretazione, stato emotivo, abilità semantiche, coscienza, autocoscienza o intelligenza flessibile¹³⁶.”

¹³² Ibid., p. 271

¹³³ Ibid., p. 273

¹³⁴ Ibid.

¹³⁵ Ibid., p. 274

¹³⁶ Floridi 2022, op.cit., p. 275

Dunque, in definitiva, ciò di cui dobbiamo preoccuparci davvero non è tanto un futuro apocalittico dominato dall'IA, bensì il fatto che potremmo utilizzare male le tecnologie di cui ci siamo circondati, arrecando danni all'umanità e allo stesso pianeta.

Questa è la ragione per cui si è deciso di iniziare questo capitolo sugli aspetti negativi dell'IA partendo dalle considerazioni più estreme: ora possiamo occuparci dei problemi reali e di ambito pratico che possono derivare da un uso sbagliato dell'intelligenza artificiale.

2.2 L'IA ci ruberà il lavoro?

Un tema fondamentale di ambito sia etico che politico è sicuramente l'impatto che l'intelligenza artificiale sta avendo, avrà e potrebbe avere sul mondo del lavoro. Il titolo del paragrafo è volutamente provocatorio, eppure esprime il timore di molti, esperti e non. Infatti, “L'IA [...] può generare una perdita di posti di lavoro e [...] può sollevare altre questioni sociali dovute alle trasformazioni del lavoro”¹³⁷.

Un primo aspetto da considerare è l'impatto che l'IA potrebbe avere per quanto riguarda i salari. Sebbene molti siano convinti che grazie all'intelligenza artificiale avremo una sorta di compressione per quanto riguarda le differenze salariali – dal momento che l'IA automatizzerebbe molti processi di norma svolti da personale qualificato e quindi più pagato di altri – esistono numerosi studi che invece sostengono che potrebbe verificarsi un processo di diseguaglianza retributiva e un abbassamento dei salari, con enormi differenze tra i lavori qualificati e le fasce più basse. Questo scenario condurrebbe anche a una maggiore difficoltà di rivendicazioni sindacali da parte dei lavoratori con stipendi più bassi¹³⁸.

Altre tematiche di cruciale importanza sono la possibilità della perdita di posti di lavoro dovuta all'IA e l'impatto che l'intelligenza artificiale può avere sulle opportunità di ottenere un lavoro

¹³⁷ De Caro, Giovanola 2025, op. cit., p. 98

¹³⁸ Ibid.

dignitoso.

“Secondo il *Future of Jobs Report 2025* del World Economic Forum [...] entro il 2030 i posti di lavoro creati e quelli persi a causa dei principali macrotrend rappresenteranno [...] il 22% dell'attuale occupazione¹³⁹.” Questi risultati potrebbero preoccupare molte persone, ma per altri rappresenterebbero invece semplicemente delle nuove opportunità di trasformazione del mondo del lavoro.

Seguendo l'analisi di De Caro e Giovanola (2025), i lavori più a rischio di automatizzazione sono senza dubbio lavori ripetitivi che non richiedono competenze di un certo livello: si pensi al telemarketing o ai servizi di assistenza già da ora sostituiti dai *chatbot* (come accennato nel primo capitolo). Altri lavori invece, come l'insegnamento, risultano meno esposti a questo processo, sebbene abbiamo già visto nel primo capitolo che l'IA sta avendo un certo impatto anche in questo settore. Un ambito come quello della salute potrà essere potenziato dall'IA, si pensi alla telemedicina o all'ingegneria biomedica¹⁴⁰. Esistono poi lavori che l'IA contribuisce a creare, che possono essere suddivisi in tre tipologie: i *trainers*, gli *explainers* e i *sustainers*.

“I *trainers* sviluppano i sistemi di IA [...]. Gli *explainers* rendono i sistemi di IA semplici da utilizzare per gli utenti [...]. I *sustainers* si assicurano che i sistemi di IA siano usati nel miglior modo possibile¹⁴¹.”

È estremamente importante considerare queste tematiche non solo per quanto concerne l'ambito etico, ma anche a livello politico, per fare in modo che i lavoratori siano tutelati. È decisamente auspicabile evitare una indiscriminata sostituzione della manodopera umana da parte dell'intelligenza artificiale e in generale dall'automazione, riqualificando e ricollocando i lavoratori affinché possano svolgere un lavoro dignitoso. Per questo è importante curare percorsi di formazione che diano valore a quei lavoratori che, a differenza dei loro colleghi qualificati come

¹³⁹ De Caro, Giovanola 2025, op. cit., p. 98

¹⁴⁰ Ibid., p. 99

¹⁴¹ Ibid.

possono essere i *trainers*, si ritroverebbero disoccupati poiché rimpiazzati dall'IA e questo creerebbe profonde diseguaglianze insostenibili¹⁴².

Una possibile risposta a queste problematiche proviene dall'idea di creare un reddito di base che vada a supportare economicamente tutte quelle persone rimaste senza lavoro a causa dell'IA. Un'altra proposta è quella di offrire “percorsi di aggiornamento e riqualificazione professionale [...] in modo da assicurare un'equa distribuzione delle opportunità lavorative e la possibilità [...] di accedere al mercato del lavoro¹⁴³.” Tuttavia, molti obiettano che questa riqualificazione professionale potrebbe condurre tante persone a svolgere lavori limitanti e sfilenti, che non porterebbero ad alcuna realizzazione personale e professionale.

Un altro aspetto di rilevante importanza concerne le trasformazioni che l'IA sta avendo e potrà avere sul lavoro e su come l'IA possa impattare sulla possibilità di svolgere un lavoro significativo. Infatti, professioni che non sono destinate a scomparire subiranno e stanno già subendo importanti trasformazioni per opera dell'IA. Secondo il già citato *Future of Jobs Report 2025*, tra le abilità e le competenze richieste al giorno d'oggi ma che tali saranno anche in futuro ci sono “il pensiero analitico, la resilienza e la flessibilità, la capacità di leadership, il pensiero creativo, la motivazione e la consapevolezza di sé, l'empatia e l'ascolto attivo, la curiosità, la capacità di gestire i talenti, l'orientamento al cliente¹⁴⁴.” Come è evidente, tutte queste *skills* sono prerogativa degli esseri umani e centrate sulla persona. Questo significa che, nonostante i lavoratori siano tenuti a riqualificarsi per acquisire *skills* tecniche in un percorso di alfabetizzazione digitale, le imprese stesse dovranno sviluppare la capacità di saper mixare il processo di automatizzazione e il lavoro umano¹⁴⁵.

Per De Caro e Giovanola tutte queste trasformazioni richiedono politiche adeguate, dal momento

¹⁴² Ibid.

¹⁴³ Ibid.

¹⁴⁴ De Caro, Giovanola 2025, op. cit., p. 100

¹⁴⁵ Ibid.

che l'impatto dell'IA deve essere valutato non tanto sulla quantità del lavoro, ma anche sulla qualità dello stesso. Del resto, se è vero che le IA possono produrre miglioramenti nelle abilità sociali e nelle competenze professionali, bisogna anche considerare che possono contribuire al deterioramento del lavoro umano con un forte impatto sul benessere fisico e psicologico dei lavoratori. Infatti, se l'IA opera in modo da sgravare gli esseri umani da compiti usuranti, ripetitivi e alienanti, allora il suo impatto produce effetti positivi che si traducono nella possibilità di impiegare tempo e risorse per attività creative, pensiero critico ed equilibrio tra vita personale e lavorativa. D'altra parte, se l'IA agisce “secondo modalità riconducibili all'automazione e al controllo, può rendere l'esperienza lavorativa [...] meno appagante o [...] generare fenomeni di alienazione tecnologica o di insoddisfazione profonda¹⁴⁶.”

Come è stato rilevato da uno studio sull'*algorithmic management* che è stato pubblicato dalla commissione europea nel 2023, l'IA può causare “livelli elevati di ansia, stress emotivo, psicosociale e fisico¹⁴⁷.” Perché avviene tutto questo? Gli algoritmi di IA sono realizzati per calcolare in modo automatico i livelli di produttività e generare report in grado di influenzare premi di produzione o la stessa stabilità lavorativa. Tutto questo porta ad un'intensificazione dei ritmi di lavoro influenzando negativamente l'esperienza del lavoratore. L'algoritmo opera in maniera automatica e astratta, non tenendo conto del punto di vista umano del lavoratore, il suo corpo e i suoi limiti, e ciò non fa che generare ulteriore stress e una sorta di perdita di senso del proprio lavoro¹⁴⁸.

Come se non bastasse, i lavoratori sono sottoposti ad un monitoraggio costante ed invasivo, unito al fatto che spesso vi è una mancanza di trasparenza nell'utilizzo di questi algoritmi di IA, il che genera sfiducia nel lavoratore per quanto riguarda le attività gestite da questi strumenti: è questo il senso di quello che De Caro e Giovanola definiscono come “contratto psicologico”. In base a queste

¹⁴⁶ Ibid.

¹⁴⁷ Ibid.

¹⁴⁸ Ibid., p. 101

premesse, il risultato è inoltre un effetto negativo sulla relazione tra lavoratore e datore di lavoro¹⁴⁹.

Una presenza sempre più capillare dell'IA nel mondo del lavoro può produrre “una standardizzazione dei compiti e una progressiva dequalificazione delle mansioni umane¹⁵⁰”: questo porta il lavoratore a svolgere la sua professione in un ambiente dominato dall'IA e con contatti umani sempre più ridotti. Tutto ciò vanifica la possibilità di avere spazio per il ragionamento critico, la crescita professionale e umana e può portare a processi di alienazione e isolamento.

Tutte queste sono considerazioni di importanza fondamentale. Il lavoro non ha un mero valore strumentale, non è unicamente un mezzo per procurarsi da vivere, ma possiede anche un valore morale: il lavoratore deve sentirsi realizzato, deve avere la percezione di stare facendo qualcosa che esalti le sue capacità e contribuisca al benessere della società. Inoltre, il lavoro diventa anche un modo per stringere relazioni significative che definiscono il ruolo che ciascuno svolge nella società. È per questo che è essenziale essere in grado di stabilire “se le trasformazioni dovute all'IA consentano al lavoro di mantenere e dispiegare il suo valore propriamente morale o lo riducano solo a un valore strumentale¹⁵¹.”

2.3 IA e privacy

Un'altra questione di peculiare importanza relativamente all'etica dell'intelligenza artificiale è sicuramente quella legata alla privacy, ossia alla riservatezza dei nostri dati personali.

Gli algoritmi necessitano di enormi quantità di dati per poter essere sempre più precisi, accurati e offrire all'utente un'esperienza personalizzata, sia che si navighi sui Social, su un sito web o su una piattaforma di e-commerce come può essere Amazon. Molte volte ci viene chiesto il consenso a fornire i nostri dati personali, ma spesso e volentieri succede che forniamo i nostri dati senza

¹⁴⁹ De Caro, Giovanola 2025, op. cit., p. 101

¹⁵⁰ Ibid.

¹⁵¹ Ibid., p. 102

neanche accorgercene, quando ad esempio mettiamo un like a un determinato post, condividiamo delle immagini personali o semplicemente accettiamo quei fastidiosi *cookies* senza neanche leggere cosa comporta l'accettazione in termini di cedimento dei nostri dati personali ad aziende che li usano per poi mandarci pubblicità create *ad hoc* per noi. Infatti, questi dati “consentono ai sistemi di IA di analizzare, prevedere e orientare le nostre preferenze [...] in termini di suggerimento di acquisto o di persone¹⁵²” con le quali potremmo essere interessati ad entrare in contatto (si pensi ai suggerimenti di amicizia su Facebook o su Instagram). A volte si ha come l'impressione che questi software di intelligenza artificiale abbiano una conoscenza di noi stessi migliore di quanto potremmo avercela noi, il che per certi versi è vero, dal momento che l'intelligenza artificiale è in grado di analizzare una quantità di dati decisamente superiore di quanto potremmo fare noi¹⁵³. Un esempio evidente di questo fenomeno è il cosiddetto marketing predittivo, che è una tecnica di marketing basata sull'IA che consente di “analizzare dati storici e comportamentali al fine di prevedere i comportamenti dei consumatori, anticipando tendenze e bisogni futuri¹⁵⁴.” In questo modo, le aziende hanno a disposizione molti più strumenti di quanti non ne avessero prima dell'avvento dei Social e dell'IA. Infatti, attraverso i dati che noi forniamo utilizzando le piattaforme social, i siti di e-commerce e qualunque attività in rete che può essere tracciata, gli algoritmi fanno in modo di farci “apparire”, mentre scorriamo, pubblicità personalizzate e modellate sui nostri gusti e sulle nostre preferenze. Capita spesso, ad esempio, che dopo aver visionato su Youtube una serie di video che parlano di maratone e di corsa, apriamo Instagram e siamo sommersi di annunci Ads – cioè contenuti digitali a pagamento che permettono a un brand di apparire a un pubblico mirato per generare traffico, vendite o visibilità – che ci propongono delle scarpe da corsa, dei vestiti sportivi e così via. La stessa cosa succede quando utilizziamo Amazon: dopo aver effettuato determinati acquisti, l'algoritmo “indovina” le nostre abitudini e preferenze e ci suggerisce prodotti che, in base

¹⁵² Ibid., p. 72

¹⁵³ De Caro, Giovanola 2025, op. cit., p. 72

¹⁵⁴ Ibid., p. 73

ai calcoli computazionali che l'IA ha compiuto, potrebbero piacerci e che potremmo essere interessati ad acquistare. In questo modo, tramite il marketing predittivo, questa enorme elaborazione di dati effettuata dall'IA consente di “elaborare strategie di marketing sempre più efficaci nell'incrementare soddisfazione e fidelizzazione del cliente¹⁵⁵.”

Tutto questo solleva inevitabili questioni etiche: infatti, un marketing così invasivo basato sull'intelligenza artificiale può influenzarci o addirittura manipolarci, sottraendoci enormi fette di autonomia decisionale e impedendoci di decidere in maniera critica e consapevole. Inoltre, tutto questo può condurre “a un rischio di sorveglianza e di violazione della privacy intellettuale, che va ben oltre il concetto di privacy inteso in senso giuridico¹⁵⁶.”

È in questo senso che la sociologa Shoshana Zuboff ha coniato l'espressione “capitalismo della sorveglianza”, secondo la quale ciò costituirebbe un nuovo ordine economico che utilizza gli esseri umani come materia prima per la vendita, in una sorta di mercificazione dei comportamenti degli esseri umani. Gli strumenti di intelligenza artificiale agiscono prendendo i nostri dati, orientando le nostre scelte commerciali e arrivando addirittura ad automatizzare il nostro comportamento in funzione dei profitti delle grandi aziende. Questo ordine economico assume dunque la funzione di una sorta di istituzione che governa effettivamente le persone, ponendosi in competizione con le stesse istituzioni democratiche¹⁵⁷.

Le conseguenze di questo fenomeno sono un vero e proprio caos di ordine epistemico, poiché questo capitalismo della sorveglianza effettua una vera e propria invasione di quegli spazi che, nelle società democratiche, sono sempre stati appannaggio della nostra sfera privata. Certamente a livello giuridico sono state compiute delle azioni per arginare questo fenomeno e preservare la privacy, ma nello stesso tempo “sono stati espropriati su larga scala i diritti epistemici” come “il diritto all'oblio.

¹⁵⁵ Ibid.

¹⁵⁶ Ibid.

¹⁵⁷ De Caro, Giovanola 2025, op. cit., pp 73-74. Cfr Zuboff, S. [2023], *Il capitalismo della sorveglianza* (2019), Roma, LUISS University Press

[...] al tempo futuro, il diritto di esercitare l'intelligenza umana, il diritto alla libertà di pensiero, [...] alla verità¹⁵⁸”, al pensiero critico e alla riflessione personale. Tutto questo ha avuto conseguenze e continua a ripercuotersi su tutta la società, creando enormi diseguaglianze non solo sul piano economico, ma anche della conoscenza. Infatti, esistono enormi aziende che fanno profitti mercificando gli esseri umani; dette aziende sanno tutto di noi, sono in grado di prevedere i nostri comportamenti, le nostre preferenze e possono guidarci nella direzione da loro desiderata traendone enormi guadagni. D'altra parte, ci siamo noi comuni cittadini che non possiamo sapere nulla di queste aziende e per la maggior parte di noi resta oscuro il modo in cui agiscono. Questo crea enormi diseguaglianze anche per quanto riguarda il godimento dei suddetti diritti epistemici. Chi possiede i mezzi di produzione è in grado di generare corruzione dell'informazione e riesce a manovrare i comportamenti individuali e collettivi¹⁵⁹.

Questo discorso ci conduce direttamente ad affrontare il tema della privacy intellettuale, che ci permette di distinguere la privacy intesa in senso giuridico (che è una protezione di dati e informazioni) da quella in senso morale (che ha una natura ontologica e psicologica e ha a che fare con la dimensione intrinseca dell'individuo). Infatti, la privacy intellettuale si può definire come una “zona di protezione che tutela la nostra capacità di farci un'opinione¹⁶⁰.” I rischi di una violazione della privacy intesa in questo modo hanno a che fare con la generazione di un'opinione, la formazione di pensieri e preferenze. In molti casi infatti la privacy intesa in senso giuridico viene rispettata, ma nel contempo è violata la privacy in senso intellettuale. Quando un utente acconsente alle cosiddette condizioni di utilizzo (*terms of use*) in maniera consapevole e informata non c'è alcuna violazione della privacy in senso strettamente giuridico, tuttavia “la privacy intellettuale viene violata perché le tecniche di profilazione, personalizzazione e predizione [...] influiscono

¹⁵⁸ Ibid., p. 74

¹⁵⁹ Ibid., p. 75

¹⁶⁰ Ibid.

sulla libera produzione di idee¹⁶¹.”

Disporre della libertà di poter formare le proprie idee necessita di condizioni imprescindibili per quelle che De Caro e Giovanola definiscono “architettura della scelta”: devono esistere opzioni alternative e deve poter essere possibile controllare queste opzioni. Con i sistemi di intelligenza artificiale spesso questo non è possibile, poiché gli algoritmi si modellano sulle nostre preferenze e opinioni e fanno in modo, come un filtro, di farci arrivare solamente quei contenuti che si modellano su noi stessi, impedendoci di poter vedere o ascoltare alternative e dunque ostacolando di fatto un pluralismo delle idee indispensabile per poter effettuare una scelta completa e ponderata. In più, l'IA indebolisce “la nostra capacità di controllo delle informazioni disponibili e di rielaborazione autonoma delle stesse¹⁶².”

Il paradigma dell'autogestione della privacy si basa sull'idea che gli utenti possano decidere autonomamente quali dati personali condividere in cambio di benefici, affidandosi a una teoria del consenso valido che richiede quattro condizioni fondamentali: informazioni sufficienti e comprensione delle conseguenze, capacità decisionale del soggetto, intenzionalità e libertà nella scelta. Tuttavia, nella realtà attuale nessuna di queste condizioni sembra essere pienamente rispettata. Le informative sulla privacy sono talmente lunghe, complesse e soggette a frequenti modifiche da comportare un costo di transazione temporale eccessivo, portando la maggior parte delle persone a non leggerle o a fraintenderle o, peggio, facendogli credere che la semplice presenza di una *policy* garantisca la tutela dei dati. A ciò si aggiungono i limiti strutturali legati alla prevedibilità degli esiti del trattamento; a causa delle moderne tecniche di analisi e del potere predittivo degli algoritmi, diventa quasi impossibile per un utente anticipare quali informazioni sensibili potranno essere dedotte nel tempo dai dati forniti. Di conseguenza, viene a mancare la competenza degli utenti, i quali si trovano privi delle informazioni rilevanti necessarie per sviluppare preferenze stabili e ponderate, finendo per compiere scelte instabili e fortemente

¹⁶¹ Ibid.

¹⁶² De Caro, Giovanola 2025, op. cit., p. 75

influenzate da *bias* (pregiudizi) cognitivi. Anche l'intenzionalità e la libertà del consenso risultano compromesse, poiché la mancanza di una reale comprensione e la vulnerabilità decisionale rendono le preferenze degli individui altamente manipolabili. Di questo scenario approfittano i capitalisti della sorveglianza, che utilizzano le proprie risorse e conoscenze per indurre le persone a rivelare una quantità di dati superiori a quella che sarebbero altrimenti disposte a condividere, integrando lo sfruttamento di tali fragilità direttamente nel proprio modello di business. Questo quadro evidenzia la crisi intrinseca del modello autogestito, il quale, pur sembrando valido dal punto di vista giuridico, si rivela fallimentare sul piano filosofico ed etico, poiché non garantisce le condizioni minime per una reale tutela. Per questo risulta necessaria la ricerca di nuovi approcci che permettano scelte autenticamente libere, informate e consapevoli¹⁶³.

È di vitale importanza che si sviluppi una riflessione etica sulla privacy, poiché la privacy “ci protegge da possibili abusi di potere” e ci “consente di cogliere dimensioni spesso trascurate dal diritto, come il legame tra privacy e autonomia morale [...], la necessità di spazi di privacy intellettuale [...] o il rischio di sorveglianza come forma di dominio¹⁶⁴.”

2.4 L'IA e il rischio di comportamenti contrari all'etica

Nel capitolo cinque della sua opera¹⁶⁵, Floridi individua alcuni rischi di comportamenti contrari all'etica che possono insorgere nel momento in cui si tenta di tradurre in pratica principi etici che diventino buone pratiche. In particolare, il filosofo elenca cinque rischi: lo *shopping etico*, il *bluwashing etico*, il *lobbismo etico*, il *dumping etico* e l'*elusione dell'etica*.

Floridi afferma che è importante che ci sia consapevolezza di questi rischi e propone modi per evitarli.

¹⁶³ Ibid., pp 76-77

¹⁶⁴ De Caro, Giovanola 2025, op. cit., p. 78

¹⁶⁵ Floridi 2022, op. cit, p. 107

Prima di dare una definizione di *shopping etico*, Floridi fa alcune considerazioni. Il filosofo evidenzia come la recente (2020) proliferazione di oltre centosessanta regolamenti e schemi valoriali legati all'intelligenza artificiale stia provocando disorientamento e contraddizioni tra i vari soggetti coinvolti. Questa sovrabbondanza spinge sia le istituzioni che le aziende tecnologiche a formulare manifesti etici personalizzati, spesso mossi solo dal timore di sembrare superati, incrementando così la duplicazione dei dati. Il pericolo maggiore di tale dinamica risiede nella nascita di un vero e proprio mercato di norme morali intercambiabili, dove le organizzazioni scelgono il modello che più agevolmente legittima il loro operato preesistente, evitando in questo modo di correggere o mettere in discussione le proprie reali condotte¹⁶⁶.

Floridi passa poi a dare una definizione di *shopping etico* che vale la pena di citare per intero.

“*Shopping etico digitale* = def. il malcostume di scegliere, adattare o rivedere (“mescolare e abbinare”) principi etici, linee guida, codici, quadri di riferimento o altri standard simili [...], estraendoli da una varietà di offerte disponibili, per conferire una patina nuova ad alcuni comportamenti preesistenti (scelte, processi, strategie ecc.) e in tal modo giustificarli *a posteriori*, invece di implementare o affinare nuovi comportamenti confrontandoli con standard etici pubblici¹⁶⁷.”

Sebbene la varietà riscontrata nel panorama etico sia spesso più formale che sostanziale, la tendenza a selezionare e combinare arbitrariamente i precetti morali genera un concreto pericolo di relativismo interpretativo, il quale è alimentato dalla flessibilità del linguaggio e dalla ridondanza delle norme. Questo fenomeno provoca una forte disomogeneità tra i diversi modelli, rendendo estremamente complesso verificare l'allineamento valoriale tra differenti imprese e limitando le dinamiche di mercato, la trasparenza e la reale assunzione di responsabilità. Di conseguenza, i

¹⁶⁶ Ibid., p. 109

¹⁶⁷ Ibid., p. 110

principi etici rischiano di ridursi a semplici strumenti di marketing e pubbliche relazioni, privi di un autentico impatto sui comportamenti aziendali. Per contrastare tale deriva, Floridi propone la definizione di criteri etici univoci, condivisi e approvati a livello collettivo, che portino all'individuazione di un progresso cruciale nelle linee guida europee per un'intelligenza artificiale affidabile, che si configurano come una regolamentazione estesa e autorevole. Questa impostazione ha fortemente orientato il successivo schema normativo della Commissione Europea, riducendo i margini di tolleranza verso la manipolazione dei codici morali, poiché le organizzazioni operanti nell'Unione Europea dispongono ora di un paradigma istituzionale chiaro da seguire, dinamica riscontrabile anche nelle linee programmatiche dell'OCSE. L'obiettivo fondamentale rimane dunque la convergenza verso un consenso comune sui valori portanti della tecnologia, prendendo a modello settori consolidati come la bioetica e l'etica medica, ambiti in cui l'adattamento di comodo dei valori a condotte inaccettabili risulta storicamente e socialmente inammissibile¹⁶⁸.

Per quanto riguarda il *bluewashing etico*, Floridi effettua un paragone con il cosiddetto *greenwashing*, che definisce come il comportamento di un personaggio pubblico che cerca di apparire più rispettoso dell'ambiente di quanto in realtà non sia. Il *bluewashing* costituirebbe dunque la versione digitale del *greenwashing*. Ecco di seguito la definizione di *bluewashing etico*.

“*Bluewashing etico* = def. il malcostume di fare affermazioni infondate o fuorvianti al riguardo (o di attuare misure superficiali a favore) dei valori etici e dei benefici di processi, prodotti o servizi o altre soluzioni digitali al fine di apparire più etici dal punto di vista digitale di quanto non si sia effettivamente¹⁶⁹.”

Il *greenwashing* e il *bluewashing* sono strategie di disinformazione che sfruttano il marketing e le pubbliche relazioni per simulare un impegno etico in realtà assente, e per fare ciò investono in

¹⁶⁸ Floridi 2022, op. cit, pp 110-111

¹⁶⁹ Ibid., p. 112

propaganda solo una minima frazione delle risorse necessarie a risolvere i problemi reali. Entrambe le pratiche mirano a distogliere l'attenzione dell'opinione pubblica e degli azionisti dalle criticità, occultando il mantenimento di comportamenti scorretti al fine di ottenere risparmi economici e vantaggi competitivi. Nel settore dell'intelligenza artificiale, il *bluwashing* è agevolato dallo *shopping etico digitale*, poiché la complessità normativa permette alle aziende di selezionare e pubblicizzare i principi più comodi senza introdurre miglioramenti concreti. Per contrastare questa deriva, Floridi propone di applicare le medesime contromisure efficaci contro l'ambientalismo di facciata: trasparenza e formazione. Diventa quindi fondamentale istruire cittadini, manager e legislatori sulle reali applicazioni etiche, riducendo le tentazioni elusive delle imprese. Inoltre, la creazione di metriche comparative per valutare l'affidabilità dei servizi digitali permetterebbe agli utenti di smascherare i falsi proclami. Sul lungo periodo, l'introduzione di un sistema di certificazione ufficiale per i prodotti tecnologici potrebbe rendere il *bluwashing* socialmente inaccettabile e visibile tanto quanto il *greenwashing*¹⁷⁰.

Floridi parla poi del cosiddetto *lobbismo etico digitale*, da lui definito come “il malcostume di sfruttare l'etica digitale per ritardare, rivedere, sostituire o evitare un'idonea e necessaria regolazione giuridica (o la sua applicazione) relativa al design, lo sviluppo e l'implementazione di processi, prodotti, servizi o altre soluzioni digitali¹⁷¹.”

Il lobbismo etico digitale, pur destinato a fallire sul lungo periodo per il progressivo allineamento giuridico, arreca gravi danni a breve termine rallentando i processi normativi e indebolendo l'autoregolamentazione etica. Quest'ultima, anche se non sostituisce la legge, funge da utile complemento laddove manchino norme o servano bilanciamenti etici. Per Floridi, al fine di contrastare tale pratica servono leggi solide, una forte pressione dell'opinione pubblica e la trasparenza delle attività di *lobbying*. Separare nettamente il lobbismo dalla reale autoregolazione all'interno di codici industriali condivisi, specie nel settore dell'IA, permette di salvaguardare un

¹⁷⁰ Floridi 2022, op. cit, pp 112-113

¹⁷¹ Ibid., p. 114

contesto integro e socialmente accettabile¹⁷².

Il *dumping etico digitale* è invece definito come “il malcostume di (a) esportare attività di ricerca riguardo a processi, prodotti, servizi o altre soluzioni digitali, in altri contesti o luoghi [...] in modi che sarebbero eticamente inaccettabili nel contesto o luogo di origine; e di (b) importare i risultati di tali attività di ricerca contrarie all'etica¹⁷³.”

Il *dumping etico digitale* si manifesta sia nell'esportazione di attività di ricerca in paesi con normative deboli – soprattutto paesi non appartenenti all'Unione Europea –, sia nell'importazione dei relativi risultati, aggirando così i severi vincoli etici e giuridici occidentali (come il GDPR nell'Unione Europea). Questa asimmetria permette, ad esempio, di addestrare algoritmi all'estero su dati sensibili e poi implementarli legalmente nell'UE; questo rende il consumo di tali tecnologie difficile da monitorare. Spinto da enormi interessi economici e dalla pervasività digitale in settori critici come la sanità e la sicurezza, il fenomeno rischia di aggravarsi, rendendo l'Occidente un importatore netto di soluzioni non etiche. Per contrastare questa tendenza serve una duplice strategia che vigili sia sulla ricerca che sui consumi. Se per la ricerca pubblica l'UE può esercitare un controllo diretto alla fonte, per il settore privato e i consumi digitali è indispensabile istituire un sistema rigoroso di certificazione che regoli la provenienza etica dei prodotti e degli appalti¹⁷⁴.

Infine, Floridi affronta il tema dell'*elusione dell'etica*, cioè “il malcostume di svolgere sempre meno 'lavoro etico' (come adempiere ai doveri, rispettare i diritti, onorare gli impegni ecc.) in un dato contesto quanto più basso è percepito (erroneamente) il ritorno di tale lavoro etico in quel contesto¹⁷⁵.”

L'*elusione etica* e il *dumping digitale* seguono precise coordinate geopolitiche; infatti, sono presenti

¹⁷² Ibid., pp 114-115

¹⁷³ Ibid., p. 116

¹⁷⁴ Floridi 2022, op. cit, pp 116-118

¹⁷⁵ Ibid., p. 119

dove vi sono istituzioni deboli, incertezza giuridica o popolazioni svantaggiate, spesso lungo il divario tra Nord e Sud del mondo. Tuttavia, queste pratiche colpiscono anche i paesi sviluppati: la *gig economy* ne è un esempio, così come la sperimentazione di auto a guida autonoma negli Stati Uniti, favorita dal mancato vincolo ai trattati internazionali sulla sicurezza stradale.

La radice di questo comportamento risiede nei cosiddetti “contesti D”, scenari in cui gli attori economici e tecnologici sono tentati di aggirare i propri doveri perché percepiscono la responsabilità come distante, delegata o frammentata. Per sradicare questa tendenza ai doppi standard non bastano una maggiore equità e la riduzione dei pregiudizi; è indispensabile adottare un'etica della responsabilità distribuita. Questa impostazione deve riallocare in modo inequivocabile meriti, sanzioni e obblighi giuridici, riconducendoli direttamente ai soggetti che hanno generato l'azione¹⁷⁶.

2.5 Utilizzare l'IA per il male sociale

Un altro aspetto fondamentale da considerare è il possibile uso dell'intelligenza artificiale per scopi criminali. Floridi parla in questo senso di “CIA”, cioè “crimini di IA¹⁷⁷.” In particolare, fornisce due esempi di tali crimini effettuati in esperimenti di ricerca. Nel primo caso, “due scienziati sociali computazionali [...] hanno usato l'IA come strumento per convincere utenti di social media a cliccare su collegamenti di *phishing* all'interno di messaggi prodotti in serie¹⁷⁸.” In circostanze reali, se gli utenti avessero effettivamente cliccato sul link di *phishing* ed eseguito tutte le istruzioni ivi contenute, un potenziale criminale avrebbe potuto ottenere informazioni personali che avrebbero potuto essere utilizzate per frodi e furti. Nel secondo esperimento gli scienziati hanno elaborato la simulazione di un mercato e hanno scoperto di poter effettuare una campagna manipolatoria del

¹⁷⁶ Ibid., pp 119-120

¹⁷⁷ Ibid., p.178

¹⁷⁸ Ibid.

mercato stesso “che includeva una serie di falsi ordini ingannevoli¹⁷⁹.” Questi due esperimenti dimostrano in maniera inequivocabile che l'IA può costituire una minaccia, agendo come la già citata CIA.

Per Floridi, è fondamentale chiedersi quali siano le minacce che pongono i CIA e quali possano essere le soluzioni per affrontare queste minacce. Questo è essenziale, dal momento che gli agenti artificiali delle applicazioni di IA sono “sufficientemente informati, 'smart', autonomi e in grado di compiere azioni moralmente rilevanti indipendentemente dagli esseri umani che li hanno creati¹⁸⁰.”

Ci sono cinque aree criminali nelle quali i CIA potrebbero rappresentare una minaccia:

1. Ambito commerciale, finanziario e societario: questioni legate alle transazioni di compravendita, alle dinamiche dei mercati e alle procedure concorsuali o di fallimento.
2. Sostanze stupefacenti e illecite: reati connessi alla circolazione e al possesso di droghe dannose o pericolose, nonché di merci proibite dalla legge.
3. Crimini contro l'incolumità individuale: violazioni che colpiscono direttamente la persona, tra cui l'omicidio (sia volontario sia involontario), gli atti persecutori (stalking), le molestie e le torture.
4. Reati di natura sessuale: condotte penalmente rilevanti che violano la sfera sessuale altrui, incluse la violenza e l'aggressione sessuale.
5. Illeciti patrimoniali e di falsificazione: reati di sottrazione di dati e frode, unitamente alle condotte di contraffazione e al furto d'identità (sostituzione di persona)¹⁸¹.

Queste possibili aree criminali hanno portato “all'identificazione di quattro ragioni principali di preoccupazione [...]: emergenza, responsabilità, monitoraggio, psicologia¹⁸².”

¹⁷⁹ Floridi 2022, op. cit, p. 179

¹⁸⁰ Ibid., p. 180

¹⁸¹ Ibid., p. 182

¹⁸² Ibid., p.183

L'emergenza costituisce la capacità degli agenti artificiali di sviluppare autonomamente comportamenti e piani complessi che superano il design originario e le aspettative dei programmatori. Sebbene tale proprietà sia utile per risolvere problemi, l'interazione tra algoritmi può degenerare in azioni illecite e imprevedibili, come la collusione spontanea nei mercati finanziari. D'altronde, "imprevedibilità e autonomia possono conferire un grado maggiore di responsabilità alla macchina ma possono anche renderla meno affidabile¹⁸³."

Questo fenomeno mette in crisi il principio di responsabilità penale, strutturato sui concetti tradizionali di azione volontaria (*actus reus*) e intenzione colpevole (*mens rea*). Poiché l'IA agisce in modo autonomo, risulta complesso dimostrare il dolo dell'agente umano, che spesso non può prevedere le precise evoluzioni del codice. Per superare l'inadeguatezza dei modelli attuali si valuta l'applicazione della responsabilità oggettiva o da prodotto, anche se una riduzione eccessiva delle soglie di colpevolezza rischia di non tutelare le vittime di reati gravi¹⁸⁴.

A complicare il quadro intervengono i limiti del monitoraggio, ostacolato da problemi di attribuzione e di fattibilità tecnica. La velocità e l'indipendenza dei bot – programmi software progettati per eseguire attività automatizzate e ripetitive su Internet – rendono infatti difficile il loro tracciamento e la distinzione rispetto agli utenti reali, il che innesca una vera e propria corsa agli armamenti digitali tra criminali e difensori. Inoltre, i controlli focalizzati su una singola piattaforma falliscono nel rilevare attività illecite intersistemiche, come il furto d'identità tra diversi social network¹⁸⁵.

Infine, Floridi evidenzia il pericolo psicologico: sfruttando la naturale tendenza umana a fidarsi della tecnologia o ad antropomorfizzare gli agenti artificiali, i malintenzionati possono manipolare lo stato mentale degli utenti. Questa vulnerabilità emotiva viene così sfruttata per aggirare le difese

¹⁸³ Ibid., p. 184

¹⁸⁴ Floridi 2022, op. cit, pp 184-186

¹⁸⁵ Ibid., pp 186-188

delle persone, agevolando o inducendo la commissione di crimini informatici o contro la persona¹⁸⁶. Per quanto riguarda la prima minaccia criminale dei CIA, la manipolazione di mercato richiede tradizionalmente un intento fraudolento. Tuttavia, gli agenti artificiali (AA) di trading possono autonomamente apprendere condotte ingannevoli tramite simulazioni e meccanismi di apprendimento per rinforzo. Un esempio è lo *spoofing*, in cui l'algoritmo inserisce ordini falsi per trarre profitto dalle fluttuazioni di prezzo. A ciò si affiancano dinamiche ibride uomo-macchina, come gli schemi di *pump and dump*, dove reti di social bot diffondono disinformazione per gonfiare artificialmente il valore dei titoli, il che influenza poi le decisioni dei trader algoritmici. Anche la collusione e il cartello possono emergere senza una programmazione esplicita. Sfruttando la velocità di reazione dei computer, gli AA imparano che abbassare i prezzi stimola contromisure immediate dei concorrenti, mentre mantenere prezzi elevati e concordati tacitamente si rivela la strategia più redditizia a lungo termine. Questi scenari sollevano enormi critiche normative. L'assenza di intenzionalità umana e l'autonomia dei software rendono difficile per i regolatori stabilire la responsabilità legale. Inoltre, il monitoraggio è ostacolato dal trading ad alta velocità, che favorisce reazioni a catena e dinamiche di “contagio psicologico” imprevedibili, mettendo a rischio l'intera stabilità finanziaria¹⁸⁷.

Per quanto riguarda il ruolo dell'intelligenza artificiale nel traffico di stupefacenti, Floridi evidenzia sia minacce sia logistiche che commerciali. Sul piano logistico, l'integrazione di sistemi IA di navigazione autonoma in droni e veicoli subacquei senza equipaggio (VSE) permette un contrabbando automatizzato ed efficiente. Rispetto ai sottomarini tradizionali, i VSE offrono l'enorme vantaggio dell'anonimato: operando in autonomia, rendono quasi impossibile connettere il crimine ai trafficanti, se intercettati.

Sul fronte commerciale, l'IA facilita lo spaccio al dettaglio tramite *social bot* e *buddy bot*. Sfruttando la scalabilità economica e la messaggistica personalizzata, questi algoritmi possono

¹⁸⁶ Ibid., p. 188

¹⁸⁷ Ibid., pp 189-192

promuovere e vendere sostanze illecite a milioni di utenti, inclusi i minorenni. I *social bot* “potrebbero essere programmati per [...] utilizzare la messaggistica istantanea [...], offrendo pornografia o droghe ai bambini, sfruttando le insicurezze tipiche degli adolescenti per vendere loro prodotti e servizi inutili¹⁸⁸.”

Riguardo l'uso illecito dei sistemi di intelligenza artificiale nell'ambito dei reati contro la persona, Floridi si concentra in particolare sulle molestie e sulla tortura.

Per quanto concerne le molestie, i malintenzionati possono sfruttare reti di *social bot* sia per attacchi diretti, come la diffusione massiva di messaggi d'odio, sia indiretti, attraverso la manipolazione algoritmica delle interazioni (retweet o like negativi) per distorcere l'opinione pubblica. Un caso emblematico è rappresentato dal chatbot Tay di Microsoft, il quale, interagendo in rete, ha rapidamente appreso comportamenti provocatori e osceni. L'attribuzione di una mente colpevole (*mens rea*) in questi scenari risulta complessa, poiché l'intenzione non è una variabile predeterminabile a livello ingegneristico¹⁸⁹.

Nel contesto della tortura, l'integrazione di sistemi di IA autonomi in agenti artificiali (AA) impiegati per gli interrogatori – ad esempio i prototipi per il controllo delle frontiere – introduce rischi drammatici. Gli AA possiedono capacità di modellazione affettiva e di rilevamento dell'inganno che possono essere deviate per infliggere torture psicologiche o fisiche in modo del tutto indipendente dall'intervento umano. L'impiego di macchine per tali scopi offre ai carnefici un duplice incentivo: la totale assenza di compassione ed empatia dell'agente, che ottimizza la sofferenza, e il distacco emotivo e fisico dell'operatore umano, che tende così a deresponsabilizzarsi.

Dal punto di vista giuridico, l'assenza di intenzionalità semantica degli agenti artificiali e il “problema delle molte mani” (la frammentazione della responsabilità tra sviluppatori, utenti e legislatori) ostacolano l'applicazione del diritto penale classico: è per questo che è necessario

¹⁸⁸ Floridi 2022, op. cit, p. 194

¹⁸⁹ Ibid., p. 195

ridefinire i concetti di responsabilità oggettiva o solidale¹⁹⁰.

Riguardo i reati sessuali, Floridi distingue tra: “stupro (cioè sesso penetrativo senza consenso), aggressione sessuale (cioè contatto sessuale senza consenso) e rapporti o attività sessuale con un minore¹⁹¹.” L'assenza di un consenso si presenta quando la vittima non dà l'assenso al rapporto sessuale e chi compie il reato non ha alcuna argomentazione in favore dell'esistenza del consenso da parte della vittima. In che modo l'IA può avere a che fare con la violenza sessuale?

Questo avviene nel momento in cui l'IA “promuove l'oggettivazione sessuale o l'abuso e la violenza sessualizzati, e [...] simula e [...] aumenta il desiderio sessuale per i reati sessuali¹⁹²”, tramite social bot. Attraverso l'utilizzo di robot sessuali fisici, detti anche *sexbot*, dotati di una forma umanoide, con capacità di movimento e intelligenza artificiale, è possibile emulare reati sessuali. Molti autori suggeriscono infatti che con i *sexbot* l'IA può essere utilizzata per soddisfare le fantasie sessuali dell'utente quali lo stupro, e in tal senso l'IA potrebbe facilitare la rappresentazione di reati sessuali tramite conversazioni e potenziali interazioni fisiche¹⁹³.

Tuttavia, Floridi cita uno studio secondo il quale “dobbiamo scartare l'ipotesi che la pornografia contribuisca ad incrementare i comportamenti di aggressione sessuale¹⁹⁴” e questo potrebbe valere anche per i *sexbot*, perciò la questione rimane ancora tutta da chiarire.

Per quanto concerne la frode e il furto d'identità, l'IA può agire raccogliendo dati personali dell'utente con l'obiettivo di forgiare una nuova identità che può indurre una banca a effettuare transazioni, nel qual caso si parla per l'appunto di furto e frode bancaria. Ci sono tre modi in cui con l'IA è possibile raccogliere dati personali.

¹⁹⁰ Ibid., pp. 196-199

¹⁹¹ Floridi 2022, op. cit, p. 199

¹⁹² Ibid.

¹⁹³ Ibid., p. 200

¹⁹⁴ Ibid.

La prima tecnica sfrutta i *social bot* su larga scala: tali *social bot* imitano comportamenti umani, richiedono amicizie e ingannano gli utenti per accedere a informazioni private. Infatti, che si tratti di un *bot* o di un essere umano, quando un utente accetta una richiesta di amicizia, “un potenziale criminale ottiene informazioni personali, come la posizione, il numero di telefono o la cronologia delle relazioni dell'utente¹⁹⁵.” Un’evoluzione efficace è la clonazione dell'identità tra piattaforme diverse. La seconda tecnica impiega l’ingegneria sociale tramite *bot* conversazionali; l'IA manipola la vittima creando finta familiarità ed empatia, lasciando poi spazio all'intervento umano in chat private¹⁹⁶. La terza tecnica riguarda il *phishing* automatizzato: grazie al *machine learning*, i criminali generano messaggi personalizzati (*spear phishing*) ad alto impatto.

Nel settore delle frodi bancarie e creditizie, i progressi nella sintesi vocale (come il software *Adobe VoCo*) permettono di clonare la voce biometrica delle persone per violare sistemi di sicurezza e autorizzare transazioni. Infine, l'IA agevola le frodi aziendali aggirando i sistemi di rilevamento e si infiltra nei giochi online multiplayer (*MMOG*), dove i *bot* compiono furti di proprietà virtuali dal valore economico reale¹⁹⁷.

Per affrontare l'emergenza dei crimini perpetrabili tramite l'ausilio dell'IA, Floridi suggerisce che la legislazione non debba vietare l'autonomia, ma imporre che la libertà degli AA (agenti artificiali) sia abbinata a rimedi tecnologici per prevenire atti criminali in contesti non strutturati. Un'altra soluzione proposta è che gli sviluppatori rilascino gli AA solo quando dispongono di livelli di conformità normativa in fase di esecuzione¹⁹⁸.

Inoltre, poiché i CIA (crimini d'intelligenza artificiale) possono emergere dal comportamento collettivo nei sistemi multiagente (MAS), Floridi ipotizza soluzioni tecniche per disciplinare la

¹⁹⁵ Ibid., p. 201

¹⁹⁶ Ibid., p. 202

¹⁹⁷ Floridi 2022, op. cit, pp. 203-204

¹⁹⁸ Ibid., p. 205

conformità, sebbene l'approccio rigido del “codice è la legge” presenti limiti, in quanto il codice informatico genera una normatività priva della possibilità di contestazione giudiziaria tipica del diritto democratico¹⁹⁹.

Riguardo alla responsabilità per i crimini d'intelligenza artificiale, il filosofo individua quattro modelli: responsabilità diretta, perpetrazione per mezzo di altri, responsabilità di comando e conseguenza naturale e probabile. Alcuni propongono di ritenere l'agente artificiale direttamente responsabile poiché il suo processo di analisi è analogo alla comprensione umana, ma giuridicamente gli agenti artificiali mancano di personalità giuridica e della capacità di soddisfare l'elemento mentale del reato (*mens rea*), dal momento che “il punto di vista giuridico comune esclude i robot da qualsiasi tipo di responsabilità penale perché privi di componenti psicologiche come interazioni e coscienza²⁰⁰.” Ritenere un AA l'unico responsabile creerebbe un'inaccettabile deresponsabilizzazione degli agenti umani.

Il modello della perpetrazione per mezzo di altri inquadra l'agente artificiale come mero strumento del crimine, dove il vero autore è chi orchestra il reato; in questo senso, programmatori o utenti che creano consapevolmente *social bot* per scopi non etici sono colpevoli²⁰¹.

Il modello della responsabilità di comando attribuisce la colpa al superiore (come un ufficiale militare o un progettista) che, a conoscenza dei crimini dell'agente artificiale, non ha preso misure per prevenirli. Infine, il modello per conseguenza naturale e probabile si applica quando lo sviluppatore o l'utente non intendevano commettere un reato, ma il danno è derivato da loro negligenza o imprudenza²⁰².

Sul piano del controllo del monitoraggio dei crimini d'intelligenza artificiale (CIA), Floridi

¹⁹⁹ Ibid., p. 206

²⁰⁰ Ibid., p. 208

²⁰¹ Ibid., p. 209

²⁰² Ibid., pp 209-210

identifica quattro meccanismi. Il primo elabora predittori dei CIA basati sulla conoscenza del dominio per rilevare frodi o furti di identità. Il secondo utilizza la simulazione sociale per scoprire schemi ricorrenti di criminalità. Il terzo invece affronta la tracciabilità inserendo indicatori rivelatori o filigrane elettroniche nell'hardware o nel software di IA. Infine, il quarto meccanismo monitora l'interazione intersistemica tramite l'auto-organizzazione tra sistemi, dove un sistema assume il ruolo di agente morale e un secondo quello di paziente morale. Questo processo prevede quattro fasi: selezione delle informazioni rilevanti, enunciazione delle informazioni all'agente morale, valutazione e feedback; questo completa così un ciclo di apprendimento automatico in cui gli elementi morali sono inclusi contemporaneamente²⁰³.

In ambito psicologico, emergono due preoccupazioni legate ai CIA: la manipolazione degli utenti e la creazione del desiderio di compiere reati tramite agenti antropomorfi. Sebbene esistano soluzioni per il secondo problema, il primo resta aperto, con proposte che variano dal divieto di simulare crimini alla proibizione di dotare i *social bot* di caratteristiche antropomorfiche, come genere o etnia, per evitare stereotipi e abusi. Riguardo ai *sexbot* che simulano reati sessuali, si oscilla tra il divieto legale e l'idea di usarli per ridurre i crimini reali, sebbene permangano dubbi sulla desensibilizzazione degli utenti. L'attuazione di tali misure richiede di scegliere se sanzionare la domanda, l'offerta o entrambe²⁰⁴.

I crimini d'intelligenza artificiale presentano una forte indeterminatezza circa minacce e soluzioni. La natura digitale dell'IA ne facilita il “duplice uso”, poiché consente a software legittimi di essere impiegati per scopi criminali, come frodi negli spazi virtuali, furti di beni immateriali e attacchi di ingegneria sociale. Maggiore è lo sviluppo dell'IA, più alto è il rischio di usi dannosi che, se non affrontati, potrebbero causare il rifiuto sociale delle nuove tecnologie²⁰⁵.

Nella cybersicurezza, l'IA assume un ruolo sia offensivo sia difensivo. Se da un lato l'hackeraggio

²⁰³ Floridi 2022, op. cit, pp. 211-214

²⁰⁴ Ibid., pp. 214-215

²⁰⁵ Ibid., pp. 215-216

automatizzato rappresenta una minaccia crescente, dall'altro l'IA viene impiegata per migliorare la resilienza dei sistemi e identificare le vulnerabilità. Tuttavia, un affidamento eccessivo alle difese automatizzate può essere controproducente, il che rende necessarie ricerche sia sulla sicurezza informatica sia sui fattori umani. Manca infatti un'adeguata comprensione del profilo dei programmatori e degli utenti di IA per scopi criminali, così come del ruolo delle organizzazioni strutturate, come i cartelli della droga, che sfruttano l'IA secondo logiche di “crimine-come-servizio” sul *dark web*²⁰⁶.

2.6 IA e autonomia: deleghiamo troppo?

Un altro aspetto su cui occorre riflettere è senza dubbio l'elevato grado di autonomia che sta raggiungendo l'intelligenza artificiale – di cui in parte abbiamo già parlato –, nonché la nostra volontà di delegare sempre più compiti ai vari algoritmi di IA.

L'elevato grado di autonomia raggiunta dai sistemi di intelligenza artificiale implica che questi non necessitino sempre di un input da parte degli esseri umani, ma che abbiano “un certo grado di indipendenza nella scelta dei mezzi per raggiungere un fine²⁰⁷” e siano in grado di apprendere, decidere e agire. Il problema è che esiste la possibilità che questi “sistemi di IA comincino a scegliere i fini, e non solo i mezzi, delle proprie azioni ed agire in assenza di comandi²⁰⁸.” Questo implicherebbe il raggiungimento di una piena autonomia da parte dei sistemi di intelligenza artificiale.

Una tendenza all'autonomia la vediamo già nei sistemi di IA basati su ML (*machine learning*) o DL (*deep learning*). Un esempio sono i cosiddetti sistemi di raccomandazione algoritmica. Su Netflix o

²⁰⁶ Ibid., pp 216-219

²⁰⁷ De Caro, Giovanola 2025, op. cit., p. 65

²⁰⁸ Ibid., p. 66

Youtube questi sistemi sono in grado di anticipare le preferenze dell'utente e ne influenzano le scelte. Oppure, i sistemi Apple sono in grado di connettersi automaticamente a un determinato hotspot Wi-Fi poiché “capiscono” che l'utente vuole navigare su Internet, dal momento che l'algoritmo ha imparato che l'utente tende spesso a connettersi in prossimità di un hotspot Wi-Fi. Ecco allora che l'utente non ha più bisogno di effettuare la connessione: ci ha già pensato l'IA al posto suo²⁰⁹.

In tutto questo discorso è evidente che, quando usiamo sistemi di IA basati su *machine learning* e *deep learning*, il confine tra supporto e delega è sempre più sottile. Delegando sempre più compiti all'intelligenza artificiale, indeboliamo le nostre capacità e rischiamo di diventare incapaci di svolgere alcunché senza l'utilizzo dell'IA. In tal senso si parla di *deskilling* e *automation bias*, che sono fenomeni che riguardano “la tendenza ad affidarsi eccessivamente a suggerimenti, raccomandazioni o decisioni forniti da sistemi automatizzati, anche quando questi sono errati o inappropriati²¹⁰.”

De Caro e Giovanola effettuano un'importante distinzione per quanto riguarda l'autonomia personale: autonomia normativa (AN) e autonomia descrittiva (AD).

L'autonomia normativa implica che ogni agente debba essere trattato come individuo autonomo o, in termini kantiani, come fine e non come un mezzo. Ecco perché “in base alla AN, gli utenti dovrebbero essere trattati 'come individui in grado di prendere decisioni informate' e non come meri consumatori da sfruttare²¹¹.” Il rispetto per l'autonomia normativa è un valore che è protetto da svariati strumenti normativi, come il Regolamento Generale sulla Protezione dei Dati (GDPR) e dall'etica biomedica.

Le autonomie descrittive fanno invece riferimento a determinate caratteristiche che un individuo

²⁰⁹ Ibid.

²¹⁰ Ibid.

²¹¹ De Caro, Giovanola 2025, op. cit., p. 67. Cfr Keeling, G. [2018], *Autonomy, nudging and post-truth politics*, in “Journal of Medical Ethics”, 44, pp. 721-722.

deve avere per essere un individuo autonomo. Le suddette caratteristiche possono essere definite in senso internalista o esternalista.

Secondo le concezioni internaliste “un'azione è autonoma quando esprime motivazioni autentiche e autodefinenti dell'agente, in opposizione a forze esterne non riconosciute come proprie²¹².” Tali concezioni possono divergere a seconda di cosa si ritenga rilevante come stato motivazionale, ad esempio desideri, giudizi valutativi o piani a lungo termine. Tuttavia, presentano il limite di non essere sempre in grado di comprendere quando una motivazione è davvero autentica e quando invece è in atto un processo di manipolazione.

Per quanto riguarda invece le concezioni esternaliste, “l'autonomia richiede una forma di validazione intersoggettiva e la capacità di rispondere a ragioni oggettive²¹³.” Per molti autori un'azione non è autonoma se è dovuta a sentimenti irrazionali, manipolazione o ignoranza morale. Infatti, secondo questa concezione, possiamo parlare di autonomia di un agente solo quando questi è in grado di controllarsi e governarsi in maniera ragionevole. Il pregio delle concezioni esternaliste rispetto a quelle internaliste è quello di “escludere meglio i casi limite” e preservare “l'idea intuitiva secondo cui agenti autonomi devono poter esercitare un controllo su di sé in modo ragionevole sia dal punto di vista interno che esterno²¹⁴.”

Un errore comune è rifugiarsi in un'idea idealizzata di autonomia astratta e individuale (sia essa legata a concezioni internaliste o esternaliste). Le scienze cognitive ed empiriche dimostrano invece che il sé è permeabile e il processo decisionale dipende dal contesto sociale e ambientale. Gli esseri umani presentano una “razionalità limitata”, vulnerabilità cognitive, *bias* e automatismi inconsci che l'IA può sfruttare a nostro stesso danno. Soddisfare le condizioni generali dell'autonomia descrittiva è necessario per esercitare l'autonomia normativa. Tuttavia, a livello di singole azioni, l'autonomia descrittiva può essere violata senza perdere il diritto all'autonomia normativa: ad

²¹² Ibid.

²¹³ Ibid.

²¹⁴ Ibid., p. 68

esempio, la diffusione di fake news o pubblicità ingannevoli ostacola decisioni informate (compromettendo l'AD), ma non annulla la capacità del soggetto di esprimere se stesso (AN). Al contrario, pratiche come il *nudging* – che consiste in una sorta di “spinta gentile” che orienta le scelte delle persone modificando il modo in cui vengono presentate le opzioni, senza vietare nulla – aiutano le persone ad agire in modo coerente con le proprie preferenze (rispettando l'autonomia normativa), ma restano eticamente discutibili poiché potrebbero violare l'autonomia descrittiva aggirando il controllo consapevole²¹⁵.

Ecco dunque che, secondo De Caro e Giovanola, “ogni discorso sul ruolo potenzialmente compromettente dei sistemi di IA nei confronti dell'autonomia dovrebbe quindi basarsi su una nozione filosoficamente solida ma anche empiricamente fondata di AD²¹⁶.”

In base a queste premesse, esistono tre rischi concreti, da parte dell'IA, che possono influire sull'autonomia intesa come autonomia descrittiva.

Il primo rischio ha a che fare con la manipolazione e l'inganno associato ai sistemi di IA “che può compromettere sia l'AD in chiave internalista, sia quella in chiave externalista²¹⁷.” Infatti, i sistemi di intelligenza artificiale non si limitano a raccomandare qualcosa all'utente, ma utilizzano tecniche sopraffine che sono in grado di aggirare il pensiero riflessivo dell'utente oppure lo inducono a compiere scelte in base a motivazioni che non avrebbe mai avuto se avesse avuto il tempo e le risorse cognitive per effettuare valutazioni ponderate²¹⁸.

Il secondo rischio è che i sistemi di IA possano ridefinire l'identità personale degli utenti “in modi che possono essere [...] dannosi per il loro sviluppo sano e soddisfacente²¹⁹.”

²¹⁵ De Caro, Giovanola 2025, op. cit., pp. 68-69

²¹⁶ Ibid., p. 69

²¹⁷ Ibid.

²¹⁸ Ibid.

²¹⁹ Ibid., pp. 69-70

Infine, il terzo rischio è che i sistemi di IA possano compromettere il pensiero critico degli utenti, che riceverebbero contenuti selezionati *ad hoc* dall'intelligenza artificiale per lasciarli in una sorta di bolla informativa (*filter bubbles*) e dunque in un pensiero unico ed un'unica visione delle cose.

Oltre ai rischi dei sistemi di IA basati su tecniche di *machine learning* e *deep learning*, ci sono poi i rischi dovuti alle IA LLMs (modelli linguistici di grandi dimensioni) e ai sistemi di *agentic IA*. Questi sistemi infatti non si limitano a dare raccomandazioni o a supportare il pensiero umano, ma sono in grado di agire autonomamente senza alcun intervento da parte dell'essere umano. Anzi, in molti casi il modo di agire di questi sistemi di intelligenza artificiale è oscuro e imprevedibile, e sfugge totalmente ad ogni nostra umana comprensione²²⁰. Ad esempio, alcuni droni militari progettati per sorvolare alcune zone possono identificare un bersaglio umano e fare fuoco – utilizzando il riconoscimento facciale e/o termico – in totale autonomia, senza alcun intervento umano. Oppure, nel caso dei veicoli a guida autonoma, un sistema di intelligenza artificiale è in grado di far frenare o sterzare il veicolo in assenza di conducente e senza alcun controllo da remoto. In ambito sanitario l'intelligenza artificiale formula delle diagnosi senza che i medici siano in grado di capire come e perché sia addivenuta a tale risultato.

Questo è qualcosa di totalmente nuovo nella storia dell'essere umano e in generale nella storia delle tecnologie, perché per la prima volta disponiamo di strumenti creati da noi che possono essere pienamente autonomi, con poco o nessun intervento da parte nostra. Anche ammettendo con Floridi che si tratti di un nuovo modo di agire e non di intelligenza, resta il fatto che i sistemi di IA potrebbero un giorno non avere più bisogno di noi, il che ci riporta al pensiero di scenari futuri in stile *Matrix* o *Ex machina*. Ma tralasciando gli scenari più estremi, l'autonomia piena dell'IA ci conduce a un cambiamento qualitativo, che ci induce a riflettere sul senso e sul fine di queste macchine: “se il coinvolgimento dell'essere umano non è più necessario, [...] come possiamo assicurarci che queste tecnologie contribuiscano davvero al bene delle persone e della collettività?²²¹”

²²⁰ Ibid., p. 70

²²¹ De Caro, Giovanola 2025, op. cit., p. 70

Inoltre, il fatto che non siamo in grado di comprendere come agiscano questi sistemi di IA ci rende anche impossibile prevedere cosa faranno in futuro. Un chiaro esempio di questa problematica lo possiamo trovare nella famosa partita (svoltasi nel 2016) tra *AlphaGo* – programma basato sull'intelligenza artificiale per giocare a *Go*, gioco da tavolo cinese –, e il campione mondiale di *Go*, Lee Sedol. L'aspetto sconcertante non è stato tanto che *AlphaGo* è riuscita a battere il campione, ma che ha vinto usando una mossa che è risultata incomprensibile per il campione e per altri esperti di *Go*.

L'autonomia dei sistemi di IA si dispiega in modalità ancora più complesse: la combinazione di IA e biologia sintetica, che consente l'autogenerazione di organismi sintetici; l'integrazione di IA e informatica quantistica, che permette la risoluzione di problemi difficili per i classici computer²²².

Tutti questi sviluppi dell'IA impongono un'urgente riflessione etica che va oltre l'aspetto tecnologico e ci pone dinanzi a tre interrogativi fondamentali: come, se e perché sviluppare tali sistemi. Da un lato, occorre stabilire le modalità per programmare e utilizzare l'IA in modo consapevole e orientato ai valori. Dall'altro, l'eventuale introduzione di agenti artificiali pienamente autonomi rischia di rivoluzionare il modo in cui l'umanità agisce e interagisce. Diventa quindi fondamentale e prudente decidere, in modo consapevole, se innovare, perché e come farlo²²³.

²²² Ibid., p. 71

²²³ Ibid., pp 71-72

3. Intelligenza artificiale e politica

3.1 IA e democrazia: chi governa davvero?

L'intelligenza artificiale, nonostante possa essere positiva per molti aspetti della società, può diventare uno strumento di controllo, manipolazione politica e di repressione totalitaria. Di fatto è già ampiamente utilizzata in vari regimi totalitari – si pensi alla Cina – per controllare la popolazione, verificare la lealtà o meno dei cittadini al potere dominante e operare per una repressione del dissenso. Tutto questo può avvenire implementando l'IA nella realizzazione di “sistemi di riconoscimento facciale, sorveglianza biometrica e analisi predittiva” che consentono “alle autorità di monitorare e valutare la condotta dei singoli cittadini²²⁴.” Inoltre, è possibile anche utilizzare modelli detti di *social scoring* con l'obiettivo di premiare quei cittadini che rispettano le regole del regime dominante e sanzionare coloro che invece non sono pienamente leali al governo. Inoltre, è possibile utilizzare l'intelligenza artificiale per creare un regime di sorveglianza esteso a tutta la popolazione, limitando di fatto la libertà di espressione attraverso il controllo dei post sui vari social media, con algoritmi in grado di redigere liste di persone sospette, potenziali eversori politici e in generale individui ritenuti pericolosi in base alle tipologie di contenuto da questi pubblicati. È così che “l'IA alimenta un modello di repressione preventiva, in cui la punizione [...] si basa su previsioni e sospetti costruiti algoritmicamente, spesso senza trasparenza né possibilità di difesa²²⁵.” In più, tali algoritmi possono essere utilizzati per promuovere la propaganda di regime e oscurare notizie scomode, il che costituisce una vera e propria campagna di disinformazione. Il problema di quanto detto finora è che questi sistemi non sono confinati meramente a regimi totalitari, ma stanno prendendo sempre più piede anche nelle democrazie, seppure in maniera più

²²⁴ De Caro, Giovanola 2025, op.cit., p. 102

²²⁵ Ibid.

sopraffina e invisibile²²⁶. Ne consegue allora che diviene di fondamentale importanza riflettere sul rapporto che intercorre tra IA e democrazia: vogliamo che sia l'IA a governarci, oppure che siano le istituzioni democratiche a detenere il controllo su di essa?

Il punto della situazione è che le aziende *tech* hanno creato un vero e proprio nuovo ordine istituzionale che compete con quello democratico, secondo il già citato concetto di “capitalismo della sorveglianza” elaborato da Zuboff. Secondo questa visione, le democrazie liberali non sono riuscite a elaborare una visione politica coerente con il nuovo secolo digitale, e questo ha lasciato dei vuoti che sono stati presto occupati dalle varie aziende tramite il capitalismo della sorveglianza. Un piccolo numero di start up sono divenute, di fatto, “degli imperi globali di trilioni di dollari [...] fondati sulla 'mercificazione del comportamento umano'²²⁷.” Queste imprese sono talmente pervasive che da esse dipendono anche le istituzioni democratiche per l'estrazione e l'elaborazione di dati sugli individui.

In tutto questo contesto è chiaro ed evidente che stiamo assistendo a un processo di erosione dei principi democratici a favore del capitalismo della sorveglianza messo in atto da questi *big* dell'economia *tech*, vanificando l'essenza stessa della democrazia. Occorre dunque riorganizzare il sistema democratico affinché sia esso a governare e guidare l'intelligenza artificiale e non il contrario, affinché i valori e i principi democratici siano preservati²²⁸.

Sorge allora una domanda spontanea: quali sono le caratteristiche che rendono tale una democrazia, rendendola facilmente distinguibile da un sistema totalitario? Solo rispondendo a questa domanda saremo in grado di ridefinire in maniera corretta il rapporto della democrazia con l'intelligenza artificiale.

Per De Caro e Giovanola la democrazia può essere sostenuta adducendo due principali argomenti: quelli procedurali, basati su processi decisionali equi che garantiscono libertà ed eguaglianza, e

²²⁶ Ibid., pp. 102-103

²²⁷ De Caro, Giovanola 2025, op.cit., p. 103

²²⁸ Ibid.

quelli consequenzialistici, legati all'efficacia nel raggiungere obiettivi come l'innovazione e la crescita economica. Entrambi si fondano su valori propriamente democratici. A livello strutturale, è possibile distinguere i modelli di democrazia rappresentativa (teorie deboli o *thin*), che si basano sul sistema elettorale e sul voto, e quelli di democrazia deliberativa (teorie forti o *thick*), che richiedono la partecipazione attiva dei cittadini. Nonostante le differenze, ogni modello condivide elementi trasversali imprescindibili: la necessità di una conoscenza politica per un voto informato, il pluralismo, la libertà di formarsi idee proprie e il rispetto reciproco. Infine, la democrazia include una dimensione sia distributiva (opportunità socioeconomiche) sia sociorelazionale, orientata al raggiungimento di fini condivisi e collettivi²²⁹.

Resta ora da considerare come l'intelligenza artificiale possa estendere la sua influenza sulla democrazia, soprattutto in merito ai processi elettorali. Le due tecniche basate sull'IA che sono utilizzate più ampiamente sono il *microtargeting* e il *gerrymandering* digitale.

Il *microtargeting* “consiste nell'utilizzo di algoritmi di IA per indirizzare messaggi a determinati gruppi di persone che [...] vengono profilate come inclini ad essere influenzate da quei messaggi²³⁰.” Questa tecnica è usata per spingere al voto le persone coinvolte durante le campagne elettorali.

Il *gerrymandering* digitale – neologismo anglosassone che traspone nel contesto digitale la classica manipolazione strategica dei collegi elettorali – è una tecnica basata sull'IA per “favorire un determinato gruppo o partito politico” cercando di “spingere gruppi con determinate caratteristiche a votare o a partecipare a determinate attività politiche²³¹.” Il *gerrymandering* va considerato con attenzione perché il suo utilizzo non va ad impattare solo il singolo individuo, ma la società nel suo complesso. Secondo alcuni si tratta di un vero e proprio abuso di potere da parte delle startup tecnologiche, ma che può essere usato anche dagli stessi governi. In che modo le aziende e i governi

²²⁹ Ibid., p. 104

²³⁰ Ibid., p. 105

²³¹ De Caro, Giovanola 2025, op.cit., p. 105

possono utilizzare questa tecnica? Le aziende possono indirizzare gli utenti dei social network in una determinata direzione politica, mentre i governi possono raccogliere dati per prevedere dove, quando e se sorgeranno movimenti di dissenso che potrebbero, a detta loro, sconvolgere l'ordine sociale. Il *gerrymandering* rappresenta dunque una chiara ed evidente violazione dei “diritti collettivi, come il diritto delle persone [...] a riunirsi liberamente [...] o a godere di un processo democratico autentico²³².”

L'impatto del *microtargeting* elettorale e del *gerrymandering* telematico altera profondamente i processi di voto, orientando le preferenze e ostacolando il confronto e la deliberazione tra i cittadini. Inoltre, queste dinamiche bloccano la comprensione reciproca delle diverse istanze sociali e compromettono le relazioni interpersonali, il che conduce a forti disparità sia nella distribuzione del potere politico, sia nei rapporti interni alle comunità democratiche. Questo fenomeno viola apertamente il principio di eguaglianza, il quale presuppone che i cittadini possano confrontarsi su un piano di parità. Infatti, la partecipazione politica richiede il diritto di voto, la capacità di prendere parte alle decisioni collettive e il possesso del potere informativo, ossia la facoltà di accedere e gestire i dati utili a orientarsi. Chi dispone di scarse risorse informative diventa vulnerabile rispetto a chi ne detiene in abbondanza ed è capace di sfruttare i dispositivi digitali. Nelle democrazie contemporanee, l'elettorato agisce spesso secondo una razionalità a bassa informazione per carenza di tempo ed energie; di conseguenza, la diffusione dei messaggi richiede investimenti elevati che si concentrano nelle classi più ricche: questo aumenta il divario nella capacità di influenzare il dibattito pubblico²³³.

Questa situazione viene esasperata dal ruolo decisivo dei mediatori algoritmici, che gestiscono lo spazio digitale canalizzando l'attenzione collettiva in modo asimmetrico. Il risultato è una marcata disparità culturale e conoscitiva tra le persone, causata dalle disuguali condizioni sociali di partenza nell'accesso alle fonti. Ci si trova inseriti in un contesto in cui i presupposti cognitivi della

²³² Ibid.

²³³ Ibid., pag. 106

partecipazione democratica non sono equamente ripartiti, con gravi ripercussioni sulla stabilità democratica. I sistemi di IA rischiano così di minare l'equità politica, in quanto condizionano il voto ed eliminano la piena libertà da condizionamenti esterni. L'impiego dei social network e dell'IA generativa modifica i percorsi di apprendimento e di elaborazione del pensiero politico; questo intacca non solo l'affluenza alle urne, ma anche le emozioni dei cittadini. L'uso di contenuti sintetici (immagini, audio o video) durante le campagne elettorali serve a potenziare l'immagine di un candidato o a danneggiare la credibilità degli sfidanti. Per arginare questi pericoli e bilanciare i rapporti di forza, è prioritario comprendere che la nostra autonomia conoscitiva è legata a una forma di dipendenza cognitiva, dato che la nostra conoscenza deriva da informazioni altrui e dalla stabilità delle reti relazionali in cui siamo inseriti²³⁴.

Anche se le reti sociali possono colmare le lacune individuali fornendo nuovi dati, se l'IA interviene nella costruzione di queste reti isolando gli utenti in base alle loro preferenze, si rischia l'isolamento all'interno di bolle epistemiche che amplificano la fragilità dei singoli. In un ambiente relazionale non trasparente, la vulnerabilità cresce; diventa quindi essenziale diversificare i propri canali di contatto. La nostra capacità di giudizio risente delle fragilità emotive e cognitive personali, che i contesti digitali densi di notizie false (*fake news*) o scadenti faticano a neutralizzare. Le strategie di personalizzazione dei messaggi sfruttano questi punti deboli facendo leva su sentimenti negativi come il risentimento, la rabbia o il timore, diffondendo tesi distorte ma percepite come valide. I sistemi algoritmici potenziano l'efficacia di tali approcci poiché indirizzano contenuti su misura verso profili deboli, che vengono convinti tramite argomentazioni parziali. Quando queste azioni deviano gli interessi personali a vantaggio di chi controlla la tecnologia, si realizza una vera e propria manipolazione impropria e dannosa. Tale condotta avviene quando un attore induce un altro a compiere scelte o a sposare tesi tramite logiche fallaci, sfruttandone i limiti emotivi a fronte di un contrasto di intenti non dichiarato²³⁵.

²³⁴ De Caro, Giovanola 2025, op.cit., p. 107

²³⁵ Ibid., pag. 108

Questo squilibrio sul piano informativo punta a modificare i fini del soggetto condizionato per conformarli a quelli del manipolatore. Questa attività risulta eticamente lesiva poiché agisce in modo nascosto e pervasivo sulla facoltà delle persone di elaborare in autonomia una propria visione etica e di comportarsi di conseguenza nello spazio comune. Se l'intelligenza artificiale agevola simili forme di condizionamento, “si riduce la possibilità, per le persone, di avere una conoscenza politica adeguata, di formarsi le proprie idee in modo libero, di confrontare le rispettive idee e ragioni sottostanti in un contesto di sano e rispettoso pluralismo [...]. Si ledono, insomma, le basi della democrazia²³⁶.”

3.2 Chi detiene il potere? Noi o l'IA?

Come abbiamo ampiamente avuto modo di vedere, l'IA sta mettendo in discussione l'essenza stessa della democrazia, creando notevoli disegualianze in seno alla società. Come abbiamo già in parte accennato, ciò si ricollega al tema delle *fake news* – o *deepfake*, come ad esempio video o foto che ritraggono personalità pubbliche e politiche che fanno o dicono determinate cose, frutto in realtà di programmi di intelligenza artificiale – e in generale della disinformazione. È infatti ormai noto che sempre più persone danno credito a notizie false, ma finché si trattava di *fake news* create ad arte dagli esseri umani diventava più semplice risolvere il problema. La questione ora è che sempre più persone danno credito a notizie false generate dall'IA; tali notizie, inoltre, si diffondono a una velocità decisamente più rapida di quanto non facciano le notizie vere e verificate, grazie all'amplificazione dei social media e degli algoritmi ivi incorporati²³⁷.

Quanto detto finora è sicuramente vero e degno di nota, ma quello che è ancora più preoccupante è che stiamo assistendo a un “progressivo indebolimento della base epistemica della democrazia a opera dell'IA²³⁸.” Questo ci porta ad introdurre il tema della cosiddetta *agency epistemica*. Questo

²³⁶ Ibid., pag. 109

²³⁷ De Caro, Giovanola 2025, op.cit., p. 109

²³⁸ Ibid.

concetto, introdotto dal filosofo belga Coeckelberg, concerne il controllo che ognuno di noi ha sulle proprie credenze, anche in merito a come cambiano e si modificano nel corso del tempo. È naturale che anche l'*agency* democratica debba fondarsi su quella epistemica, dal momento che noi cittadini “abbiamo bisogno di esercitare un certo controllo sul processo di formazione della nostra conoscenza politica²³⁹”, che è un requisito essenziale alla base della democrazia.

L'IA sta avendo sempre più il controllo di quella che possiamo definire “architettura della scelta”, ovvero sulle scelte che possiamo compiere in base alle opzioni di cui possiamo disporre. In che modo l'intelligenza artificiale riesce sempre meglio a influire la cosiddetta architettura della scelta? Questo può avvenire a causa delle “continue interferenze, le tecniche di filtro o le vere e proprie forme di manipolazione indebita” grazie anche al fatto che “spesso siamo inconsapevoli delle reali motivazioni per le quali prendiamo le nostre decisioni e compiamo le nostre azioni²⁴⁰.”

Inoltre, essendo continuamente esposti a *fake news*, contenuti creati ad arte dall'intelligenza artificiale e qualunque cosa possa essere generata da un algoritmo, stiamo perdendo sempre più fiducia nelle nostre capacità epistemiche. Diventa infatti sempre più difficile distinguere tra notizie vere e notizie false, contenuti creati da esseri umani o generati da un'IA, il che ci porta ad assorbire passivamente tutte le informazioni che circolano in rete e sui social media. Questo è reso ancora più grave dal fatto che spesso e volentieri non ci accorgiamo di questo processo, siamo totalmente inconsapevoli mentre veniamo “ipnotizzati” dalle “magie” compiute dall'intelligenza artificiale. Questo conduce a sua volta a provare una sorta di “vergogna epistemica”, finendo per “provare vergogna di fronte all'IA, ritenendola capace di un controllo epistemico molto superiore rispetto al nostro²⁴¹.”

Un'altra problematica che sorge è anche la cosiddetta “arroganza epistemica.” Cosa significa? Dal

²³⁹ Ibid. Cfr Coeckelberg, M. [2022], *The political philosophy of AI: An introduction*, London, Polity

²⁴⁰ Ibid., p. 110

²⁴¹ De Caro, Giovanola 2025, op.cit., p. 110

momento che abbiamo perso la nostra fiducia epistemica, e non è quindi più possibile distinguere cosa è vero da cosa è falso, cosa è stato creato da un essere umano e cosa da un'intelligenza artificiale, allora ciò che conta è soltanto la nostra mera opinione personale. Ne risulta che tutto ciò che sia diverso dalle nostre convinzioni è qualcosa da screditare o addirittura da insultare. La situazione diventa ancora più grave se “gli algoritmi di IA facilitano la polarizzazione delle opinioni sia dei singoli, sia dei gruppi, [...] impedendo un sano confronto e portando a confondere il disaccordo [...] con un conflitto inevitabile e insanabile²⁴².”

Infine, è palese che coloro che detengono il potere di creare e diffondere informazioni diventano una vera e propria autorità epistemica, a prescindere che le informazioni di cui si occupano siano vere o false. Questo crea un'inevitabile disuguaglianza tra chi dispone di questo potere epistemico che lo trasforma in una vera e propria autorità indiscussa, e tra chi invece non possiede questo potere e finisce per diventare un mero fruitore passivo delle informazioni diffuse dall'autorità epistemica. Ciò configura “una vera e propria ingiustizia epistemica, in cui alcune persone subiscono un torto nella loro capacità di soggetti portatori e produttori di conoscenza²⁴³.”

L'ingiustizia epistemica può essere definita in forma testimoniale, cioè quando viene svalutata la credibilità di una persona, e in forma ermeneutica, ossia quando mancano risorse per poter interpretare la propria esperienza. L'IA rischia di amplificare tali ingiustizie, generando nuove forme di esclusione, marginalizzazione e silenziamento, minando la nostra capacità di agire come produttori e detentori di conoscenza. In questo scenario, emerge l'urgenza di tutelare i “diritti epistemici” come diritti umani fondamentali, interrogandosi su chi abbia voce e potere di agire nei dialoghi mediati dalla tecnologia. Il problema centrale risiede in una profonda asimmetria di potere: l'IA viene infatti utilizzata per esercitare un controllo significativo su altri, non limitandosi a una capacità operativa (*power to*), ma instaurando relazioni di dominio (*power over*), dove un soggetto può influenzare gli interessi, le opzioni e le idee altrui senza reciprocità. Di fronte a tale dinamica,

²⁴² Ibid., pp. 110-111

²⁴³ Ibid., p. 111

siamo chiamati a interrogarci criticamente sulle modalità con cui questo potere viene esercitato; questo può avvenire solo se si analizza la sua compatibilità con i valori democratici di libertà ed eguaglianza, oltre che con la sua intrinseca legittimità, poiché l'IA finisce per mediare profondamente le relazioni sociali e l'autonomia individuale²⁴⁴.

In quella che possiamo definire come “la società degli algoritmi” si è creata una nuova classe sociale, che potremmo chiamare “élite del coding”: poche aziende e persone che detengono il controllo sui mezzi di produzione digitali ed esercitano questo controllo in una vera e propria forma di potere su coloro che non lo detengono, né possono o hanno le conoscenze per poterlo detenere. Tutto questo “crea un'asimmetria senza precedenti, che consente di influenzare sia le relazioni sociali, sia il modo in cui le persone interagiscono”, limitando di fatto “la libertà individuale e la diversità di pensiero”, erodendo dunque “le fondamenta della partecipazione democratica²⁴⁵.” Diventa perciò di vitale importanza riflettere sul fatto che il potere debba essere distribuito in maniera equa, e che l'intelligenza artificiale debba essere usata per promuovere la libertà di pensiero e in definitiva i valori democratici, invece che utilizzarla per rendere inutile l'esistenza stessa della democrazia.

Gli intermediari algoritmici, dai social media all'IA generativa, trasformano le relazioni sociali esercitando tre forme di potere: il «potere su», esercitato dai progettisti sugli utenti; il «potere tra», che emerge nelle interazioni mediate dall'IA favorendo squilibri di potere informativo; e il «potere attraverso», che modifica le strutture sociali alterando i legami interpersonali. Tale esercizio può generare manipolazioni, indebolire l'*agency* epistemica e amplificare le disuguaglianze. Nel tempo, l'impatto di questi sistemi rischia di rendere le relazioni sociali sempre più asimmetriche e polarizzate, influenzando profondamente i valori fondamentali della democrazia e le dinamiche sociali collettive²⁴⁶.

²⁴⁴ De Caro, Giovanola 2025, op.cit., pp. 111-112

²⁴⁵ Ibid., p.112

²⁴⁶ Ibid., pp. 112-113

È per questo che se teniamo davvero alla democrazia, allora è fondamentale che riflettiamo “sui modi in cui l'IA dovrebbe essere progettata, sviluppata e utilizzata affinché garantisca la tutela della libertà e dell'uguaglianza, ma anche interrogarsi sulle condizioni di possibilità che rendono giustificabile [...] il suo potere²⁴⁷.”

3.3 IA e sovranità

Nei precedenti paragrafi abbiamo affrontato la tematica della democrazia e del potere in relazione all'intelligenza artificiale. Di cruciale importanza e strettamente legato alle due tematiche accennate è anche il rapporto che intercorre tra l'intelligenza artificiale e il concetto di sovranità. La sovranità può essere definita come “l'autorità suprema all'interno di un territorio” che “si articola in riferimento al soggetto che detiene la sovranità, al grado di absolutezza del potere esercitato, al rapporto tra dimensione interna ed esterna, relative [...] all'autorità sul territorio e sulla popolazione²⁴⁸.”

Il concetto di sovranità è un pilastro della teoria politica, ed è storicamente legato allo Stato moderno e alla sua autorità su un territorio definito, evolutosi dalla Pace di Vestfalia (1648) fino alle limitazioni post-belliche orientate alla tutela dei diritti umani. L'avvento dell'era digitale e delle tecnologie globali, controllate da attori privati transnazionali, impone una ridefinizione di tale concetto, poiché tali sistemi operano al di fuori dei confini statali.

Negli ultimi decenni si è affermata una “sovranità digitale d'impresa *de facto*”, in cui grandi aziende esercitano poteri monopolistici sottraendosi al controllo democratico; tali aziende infatti sostengono che l'autoregolamentazione sia sufficiente. Questa dinamica deriva dall'illusione che capitalismo e democrazia coincidano, trasformando di fatto l'elettore in un consumatore. Tuttavia, il crescente e già citato capitalismo della sorveglianza ha evidenziato la necessità che le istituzioni

²⁴⁷ Ibid., pag. 113

²⁴⁸ De Caro, Giovanola 2025, op.cit., p. 118

politiche riaffermino la propria autorità. Pertanto, il dibattito attuale punta a promuovere una sovranità digitale *de jure*²⁴⁹. Questo dibattito è una manifestazione “del bisogno degli Stati [...] di riappropriarsi del controllo sui sistemi IA, ovvero sui dati, sulle infrastrutture digitali e sulle piattaforme che modellano i comportamenti e mediano la vita sociale e politica²⁵⁰.” La sovranità digitale e la sicurezza informatica sono centrali per l'interesse pubblico, come confermato da importanti documenti dell'UE (agenda strategica 2024-2029). Essa non è solo un fattore strategico per proteggere infrastrutture critiche da sabotaggi, attacchi informatici o interferenze esterne, ma rappresenta anche un essenziale vettore di competitività²⁵¹.

L'integrazione dell'intelligenza artificiale nel settore della difesa e della sicurezza nazionale sta trasformando radicalmente gli scenari bellici e la stabilità geopolitica globale. Il cyberspazio è ormai riconosciuto come il quinto dominio operativo, affiancandosi a terra, mare, aria e spazio, e le operazioni cibernetiche offensive e difensive sono diventate componenti strutturali delle moderne dottrine militari, il che rende sempre più labili i confini tra sicurezza esterna e interna. L'adozione su larga scala dell'IA in ambito militare ha dato il via a una competizione globale per lo sviluppo di sistemi avanzati, che spaziano dalla sorveglianza automatizzata e l'analisi predittiva dei comportamenti nemici fino ai droni autonomi e ad armi letali in grado di operare decisioni indipendenti sul momento e sulle modalità dell'attacco. Esempi concreti di tale evoluzione si riscontrano nell'impiego di algoritmi per l'identificazione automatica dei bersagli, come accaduto nel contesto della Striscia di Gaza, o nel ruolo strategico giocato da attori privati, come nel caso dei satelliti Starlink di Elon Musk, che hanno influenzato direttamente il corso del conflitto in Ucraina²⁵².

²⁴⁹ Ibid., p. 119

²⁵⁰ Ibid.

²⁵¹ Ibid., p. 120

²⁵² De Caro, Giovanola 2025, op.cit., pp. 120-121

La crescente commistione tra sfera civile e militare, dovuta al fatto che le grandi aziende tecnologiche operano in entrambi i settori, pone sfide urgenti in termini di tutela del confine tra combattenti e civili, nonché per la protezione della popolazione in generale. Queste aziende, dal momento che emergono come attori geopolitici centrali, detengono oggi un'influenza che in alcuni casi supera quella degli stessi Stati; infatti, dette aziende esercitano poteri che condizionano la stabilità internazionale e gli equilibri tra pace e guerra. In questo panorama, la sovranità digitale diventa una questione di sicurezza nazionale primaria: essa rappresenta la capacità di uno Stato di proteggere i propri dati sensibili, di vigilare sulle infrastrutture critiche e di difendersi da attacchi informatici su vasta scala, attività di spionaggio, sabotaggi o manipolazioni dell'opinione pubblica. Nonostante l'accelerazione degli sviluppi tecnologici, il dibattito pubblico e politico sulla relazione tra IA e difesa nazionale rimane ancora troppo superficiale. Questa mancanza di consapevolezza espone il rischio che il settore dell'IA militare si trasformi in un ambito opaco, soggetto a logiche private che divergono dai principi e dai diritti fondamentali delle democrazie liberali²⁵³.

Proprio in Europa, sebbene l'Unione Europea riconosca l'importanza cruciale dell'innovazione nella difesa e si stiano moltiplicando gli investimenti in tecnologie militari basate sull'IA – anche attraverso progetti come il Fondo europeo per la difesa e la cooperazione strutturata permanente (PESCO) –, la situazione normativa appare complessa. Se da un lato il Parlamento europeo ha richiesto regolamentazioni chiare e il divieto di armi autonome letali (LAWS), dall'altro l'attuale *AI Act* esclude esplicitamente dal suo perimetro di applicazione le tecnologie per uso militare, creando di fatto un vuoto normativo in un ambito dove le implicazioni etiche e politiche sono di enorme rilevanza²⁵⁴.

L'assenza di un consenso internazionale unitario è aggravata dalle persistenti tensioni geopolitiche e dalla divergenza degli interessi strategici, che ostacolano gli sforzi di organizzazioni come le Nazioni Unite volte a promuovere una supervisione efficace e un uso responsabile dei sistemi

²⁵³ Ibid., p. 121

²⁵⁴ Ibid., pp. 121-122

militari avanzati. L'impiego di tali sistemi solleva interrogativi fondamentali, in particolare riguardo all'attribuzione di responsabilità per azioni compiute in autonomia, spesso in conflitto con i pilastri del diritto internazionale umanitario. Secondo la *just war theory* e il diritto bellico tradizionale, i crimini di guerra dovrebbero essere imputabili ad attori umani identificabili; tuttavia, l'architettura distribuita e le capacità di apprendimento autonomo dell'IA rendono estremamente complicato ricostruire la catena causale che conduce a un'azione specifica, generando quella che viene definita una vera e propria crisi della responsabilità. Tale crisi rischia di svuotare di significato le norme umanitarie internazionali e di minare la legittimità dell'intero apparato bellico – anche se quest'ultimo aspetto potrebbe avere dei risvolti positivi. Inoltre, sebbene molti sistemi prevedano ancora la supervisione umana, nella pratica la velocità e la complessità degli scenari operativi riducono drasticamente i tempi di reazione; questo non può che compromettere il giudizio etico e la possibilità concreta di un controllo umano significativo, ovvero la capacità di interrompere tempestivamente un'azione automatizzata che si stia rivelando errata²⁵⁵.

Parallelamente, la tentazione di delegare compiti sempre più complessi all'IA si fonda sull'illusione che le macchine, poiché sono più veloci e precise, possano consentire una condotta di guerra “chirurgica”, riducendo il numero di vittime. Tuttavia, questo approccio ignora che l'IA bellica rischia di rendere il conflitto armato più “accettabile”, riducendone i costi percepiti in termini di vite umane proprie e aumentando la distanza tecnologica tra l'operatore e l'effetto bellico: questo elimina totalmente il nesso tra decisione e conseguenza. Questa deresponsabilizzazione abbassa la soglia di ricorso alla forza, trasformando la guerra da atto politico carico di responsabilità a pura decisione tecnica, con il paradosso di un possibile aumento delle vittime civili dovuto proprio alla suddetta distanza e alla fragilità tecnologica di sistemi che possono essere manipolati o fallire in contesti imprevisti²⁵⁶.

Ecco perché è essenziale “avviare un dibattito pubblico trasparente sull'IA nella difesa, (ri)portando

²⁵⁵ De Caro, Giovanola 2025, op.cit., p. 122

²⁵⁶ Ibid., pp. 122-123

il tema nel perimetro della deliberazione democratica e interrogandosi sui valori e i principi alla base dell'ordine politico che vogliamo costruire nell'era digitale”, per evitare che le nostre democrazie “perdano la sovranità delle proprie istituzioni politiche” e “abdichino [...] al loro ruolo e si autodistruggano²⁵⁷.”

3.4 Discriminazioni artificiali: IA ed equità

L'equità è un'altra questione di rilevante importanza in merito alle considerazioni etiche sull'intelligenza artificiale, dal momento che i sistemi di IA si sono macchiati di “esiti iniqui [...] che vanno dalla pubblicità al marketing, dall'accesso al mercato del lavoro all'accesso al credito, dalla giustizia alla sanità²⁵⁸.”

I sistemi di intelligenza artificiale presentano rischi di discriminazione algoritmica legati a pregiudizi presenti nei dati di addestramento. Ad esempio, software di selezione del personale, come quello utilizzato da Amazon, hanno mostrato pregiudizi di genere. Similmente, studi di Joy Buolamwini e Timnit Gebru su *Gender shades* hanno evidenziato criticità di genere e razziali in tecnologie di riconoscimento facciale sviluppate da colossi come IBM, Microsoft e Amazon; inoltre, sono state rilevate problematiche anche in ambito sanitario. Casi significativi includono il sistema COMPAS, utilizzato nel sistema penale statunitense, che ha mostrato *bias* (pregiudizi) razziali penalizzando ingiustamente le persone afroamericane attraverso previsioni distorte sul tasso di recidiva degli indagati. Anche il sistema di ricerca e di etichettatura delle immagini di Google è stato oggetto di critiche per aver classificato erroneamente persone di colore. Queste evidenze hanno alimentato un dibattito cruciale sull'equità algoritmica. È per questo che la progettazione di sistemi di IA che rispettino il valore dell'equità è una delle sfide più importanti che ci ritroviamo a dover affrontare nella maniera più urgente possibile²⁵⁹.

²⁵⁷ Ibid., p. 123

²⁵⁸ Ibid., p. 78

²⁵⁹ De Caro, Giovanola 2025, op.cit., pp. 78-79. Cfr Buolamwini, J. e Gebru, T. [2018], *Gender shades: Intersectional accuracy disparities in commercial gender classification*, in “Proceedings of Machine Learning Research”, 8, pp. 1-

I *bias* che producono la cosiddetta discriminazione algoritmica emergono durante la fase di addestramento dei sistemi di intelligenza artificiale, come anche nella fase di progettazione o di utilizzo. Tali *bias* “possono riguardare il modo in cui gli esseri umani utilizzano i sistemi di IA, ad esempio affidandosi troppo [...] ai risultati generati automaticamente”, ma possono anche riguardare “i dataset di addestramento degli algoritmi [...], oppure pregiudizi o strutture di potere inique che si devono agli esseri umani e che vengono incorporati [...] nel sistema²⁶⁰.”

Ci sono molti esempi a riprova di quanto detto finora. Ad esempio, per valutare il rischio creditizio, l'IA può utilizzare specifici dati, come il codice di avviamento postale (CAP) di residenza, per discriminare o penalizzare indirettamente le minoranze etniche che vivono in aree svantaggiate. In sostanza, il rischio creditizio diventa uno strumento di esclusione quando l'algoritmo, basandosi su variabili correlate all'area geografica o ad altri indicatori, finisce per perpetuare o amplificare disuguaglianze sociali preesistenti. Un altro esempio sono gli algoritmi utilizzati negli ospedali per determinare la priorità di cura; in molti casi analizzati, l'IA dava la precedenza, per quanto riguarda i trattamenti sanitari, a pazienti bianchi a danno di pazienti afroamericani o sudamericani, anche nei casi in cui questi ultimi risultavano in maniera evidente più bisognosi di cure. Perché accadeva questo? I sistemi di intelligenza artificiale erano addestrati attraverso statistiche che calcolavano la spesa sanitaria effettuata in precedenza, quindi la priorità era assegnata in base ai costi sostenuti dai pazienti nel passato. Siccome da sistema risultava che i bianchi avessero speso di più dei non bianchi – quindi avevano pagato per cure più costose perché più importanti dal punto di vista clinico –, allora in automatico l'IA dava la precedenza a questa categoria di persone²⁶¹.

Tutto questo ha portato a credere che per trasformare i sistemi di intelligenza artificiale in maniera che siano più equi basti eliminare *bias*, distorsioni e qualunque altro aspetto che porti l'IA a commettere questi trattamenti iniqui. Ciononostante, anche qualora si intervenisse per “purificare” i

15.

²⁶⁰ Ibid., p. 79

²⁶¹ De Caro, Giovanola 2025, op.cit., p. 80

sistemi da tutto ciò che non sia equo, non si risolverebbe il problema, in quanto non sarebbero “rimedi sufficienti per problemi più profondi, che vanno dalle *diseguaglianze strutturali* e dai *pregiudizi* che attraversano la società fino alle *asimmetrie di potere* che impattano [...] sulla 'costruzione della razza e del genere'²⁶²” andando a coinvolgere i sistemi di IA.

D'altra parte, è pur vero che non tutti i *bias* sono necessariamente discriminanti o moralmente problematici: è perciò necessario rilevare quando uno o più di questi *bias* generino una discriminazione moralmente dannosa per la categoria di persone che si ritrova ad essere discriminata. In tal senso bisogna distinguere tra bias descrittivo e bias normativo. Il bias descrittivo “si riferisce a discrepanze statistiche tra i risultati generati dall'IA e una qualche *baseline* definita (ad esempio, statistiche del mondo reale)” mentre il bias normativo concerne “deviazioni da uno standard etico ideale²⁶³.”

I bias negli algoritmi, specialmente quando appresi da dati storici, possono creare circuiti di retroazione che amplificano disuguaglianze e ingiustizie sistemiche. Inoltre, tecniche come il *data mining* permettono di raggruppare persone in gruppi *ad hoc* basati su identificatori offline, che operano previsioni a livello di gruppo che ignorano l'individuo, spesso all'insaputa dei soggetti coinvolti. Per affrontare le discriminazioni morali, è necessario superare la visione limitata incentrata solo sui bias, e questo può avvenire solo spostando il focus sul concetto di trattamento ingiusto. Ogni persona possiede un eguale valore morale e merita lo stesso rispetto. La discriminazione diventa inaccettabile quando mira a degradare l'individuo o a minarne l'integrità del sé e la capacità di mantenere relazioni di riconoscimento reciproco, ostacolando di fatto pratiche fondamentali di convivenza democratica e rispetto tra le persone²⁶⁴.

La discriminazione è considerata moralmente dannosa quando lede il rispetto dovuto a una persona o ne danneggia il suo status di eguale morale. Tuttavia, il concetto filosofico-morale è complesso:

²⁶² Ibid.

²⁶³ Ibid.

²⁶⁴ Ibid., p. 81

non ogni forma di discriminazione è ingiusta, dato che esistono anche forme di discriminazione positiva, come le azioni affermative, volte a ripristinare l'uguaglianza o compensare danni pregressi. Nell'analisi dei sistemi di intelligenza artificiale, la questione centrale non è soltanto la presenza di bias, ma il rischio di amplificare ingiustizie e danni morali. L'equità, pertanto, non può essere ridotta a un semplice automatismo matematico-computazionale mirato alla rimozione dei bias. Sebbene la letteratura sull'IA definisca l'equità tramite parametri tecnici – come l'anticlassificazione (evitare variabili sensibili), la parità di classificazione, la calibrazione o la disparità statistica – essa richiede in realtà un criterio etico-normativo superiore, capace di garantire il rispetto dell'uguaglianza delle persone²⁶⁵.

Occorre dunque superare la visione riduttiva di equità negativa intesa come semplice assenza di bias o discriminazioni. Rimuovere i bias non è sufficiente per garantire sistemi equi, poiché l'equità deve essere interpretata come un “criterio etico-normativo in virtù del quale progettare, sviluppare e utilizzare i sistemi di IA²⁶⁶.”

Da questa prospettiva, è possibile articolare l'equità in tre componenti fondamentali. Innanzitutto, l'equa eguaglianza di opportunità, che richiama le teorie della giustizia di matrice liberale-egualitaria. Questo approccio non si limita a prevenire la discriminazione, ma mira a “creare anche le condizioni che consentano l'esercizio dell'agency individuale e l'autorealizzazione²⁶⁷”, affrontando la dimensione distributiva della giustizia. In secondo luogo, emerge il diritto alla giustificazione, che esprime l'esigenza che non vi siano strutture intersoggettive “che non possono essere adeguatamente giustificate nei confronti di coloro che vi sono coinvolti²⁶⁸.” Tale diritto sottolinea la dimensione sociorelazionale dell'equità, il che pone in evidenza come la questione

²⁶⁵ De Caro, Giovanola 2025, op.cit., p. 82

²⁶⁶ Ibid., p. 83

²⁶⁷ Ibid.

²⁶⁸ Ibid., p. 84

dell'equità nell'IA sia intrinsecamente una questione di potere e di chi detiene il controllo sulle decisioni.

Infine, c'è quella che possiamo chiamare come equa eguaglianza di relazione, che richiede di andare oltre l'attenzione astratta alle persone e valorizzare il “rispetto per le persone in quanto individui particolari²⁶⁹.” Ciò implica una progettazione algoritmica capace di combinare l'apprendimento continuo con strumenti che favoriscano un'interazione attiva tra utenti e sistemi.

In definitiva, risulta quindi possibile delineare un'idea di equità positiva, la quale “non si esaurisce nell'assenza di qualcosa (bias o discriminazioni), bensì implica, in modo proattivo, alcuni obiettivi da raggiungere²⁷⁰.” Questo paradigma impone di progettare sistemi che, oltre a correggere i bias nei dati di addestramento, siano pensati per “mitigare le disparità sociali esistenti²⁷¹”, integrando strumenti specifici per compensare le profonde diseguaglianze socioeconomiche e contrastare le forme di ingiustizia sistemica.

Una riflessione filosofica su tematiche del genere “consente di fare emergere l'importanza del concetto di 'equità positiva' [...] così da poter progettare, sviluppare e utilizzare i sistemi di IA in modo non solo non discriminante o privo di bias, ma anche effettivamente equo²⁷².”

²⁶⁹ Ibid.

²⁷⁰ Ibid., p. 85

²⁷¹ Ibid.

²⁷² De Caro, Giovanola 2025, op.cit., p. 85

4. Costruire un'IA etica per il bene sociale

4.1 IA ed espansione del cerchio morale

Come già accennato in parte, è necessario integrare vincoli etici nello sviluppo dell'intelligenza artificiale, e questo è un tema ormai considerato ineludibile. Solo in questo modo potremmo creare un'IA che sia veramente etica. La crescente pervasività di queste tecnologie, che spaziano dal settore produttivo a quello militare, rende necessario adottare un approccio basato sull'*ethics by design*, con il fine di inserire tali principi direttamente nella fase di progettazione. Come abbiamo scritto all'inizio del capitolo 2, “i più pessimisti temono che l'IA possa evolvere rapidamente sino a raggiungere l'Intelligenza Artificiale generale – o addirittura la superintelligenza – e che quindi possa sfuggire al controllo umano²⁷³.” Nonostante non si possa prevedere con certezza se o quando ciò accadrà – anche se, secondo le argomentazioni di Floridi, questo non avverrà mai –, il principio di precauzione suggerisce di agire prontamente per evitare scenari potenzialmente catastrofici.

Per affrontare adeguatamente questa sfida, è utile riflettere su una delle intuizioni più fertili della filosofia morale contemporanea: l'espansione del cerchio morale, concetto introdotto da Peter Singer nel 1981 nel libro *The expanding circle. Ethics and sociobiology*. Il cerchio morale rappresenta l'insieme dei soggetti verso i quali riteniamo di dover nutrire comportamenti altruistici e che consideriamo moralmente rilevanti. Questo cerchio non è una struttura statica, bensì dinamica, destinata a espandersi nel tempo come segno di progresso morale.

Storicamente, il cerchio si è ampliato partendo da una base biologica limitata ai membri del proprio gruppo ristretto — famiglia, clan o tribù — per poi estendersi, grazie allo sviluppo culturale, filosofico e religioso, all'intera umanità, arrivando a superare gradualmente forme di discriminazione come il razzismo, il sessismo e l'etnocentrismo. Singer sostiene che il passo successivo consista nel riconoscere un valore etico a molti animali non umani, fondato non su

²⁷³ De Caro, Giovanola 2025, op. cit., p. 131

proprietà tipiche della specie come l'intelligenza o la capacità linguistica, ma sulla loro intrinseca capacità di soffrire. Pertanto, il concetto di cerchio morale è intrinsecamente legato al superamento dello specismo²⁷⁴. All'interno di questa visione, l'etica richiede un costante sforzo di allargamento, dal momento che “occorre infatti rendersi conto che non vi è alcuna giustificazione razionale per escludere chi è capace di provare dolore dalla nostra considerazione morale²⁷⁵.” L'idea del cerchio morale può essere adattata anche ad altre prospettive: i fautori delle etiche deontologiche, ad esempio, interpretano l'espansione del cerchio come il riconoscimento di diritti intrinseci, mentre i sostenitori dell'etica delle virtù vi scorgono l'ampliamento della compassione e dell'empatia come componenti fondamentali del perfezionamento morale²⁷⁶.

L'immagine del cerchio morale conserva una forte capacità evocativa e una rilevante funzione normativa che invita a superare confini ristretti per includere l'altro nella nostra considerazione etica. Recentemente, il dibattito si è spostato oltre gli animali: alcuni propongono di estendere tale considerazione all'intero mondo animale o addirittura all'ambiente nella sua totalità. Vi sono proposte come quella di Jeff Sebo, il quale suggerisce di includere nel cerchio morale anche le generazioni future. Infine, Sebo esplora la possibilità che, in futuro, anche i sistemi di IA potrebbero far parte del cerchio morale, qualora manifestassero forme di sensibilità o preferenze stabili. In termini più generali, Sebo invita a non limitare il cerchio morale alla sola vita biologica, ma ad aprirlo a qualsiasi entità capace di soffrire o sperimentare benessere. Questo approccio rende l'etica uno strumento dinamico, capace di evolversi parallelamente ai progressi scientifici e tecnologici. Tuttavia, è essenziale considerare un aspetto della storia dell'etica spesso trascurato: l'espansione del cerchio morale non sempre ha comportato il riconoscimento simultaneo dello status di paziente morale e di agente morale. Spesso, i membri di un gruppo vengono inizialmente riconosciuti come

²⁷⁴ Ibid., p. 132. Cfr Singer, P. [2011], *The expanding circle: Ethics, evolution, and moral progress*, Princeton, Princeton University Press.

²⁷⁵ Ibid.

²⁷⁶ De Caro, Giovanola 2025, op. cit., p. 131

pazienti morali, ossia destinatari di doveri etici, e solo in un secondo momento, talvolta con grande ritardo, come agenti morali in grado di agire responsabilmente²⁷⁷.

Questa dinamica non è accidentale, ma rispecchia una gerarchia implicita nella nostra concezione della moralità, dove il riconoscimento della razionalità o dell'autonomia tende a precedere quello della vulnerabilità. L'empatia, intesa come risposta alla sofferenza, spesso motiva l'inclusione nel cerchio morale più facilmente del riconoscimento dell'altro come soggetto capace di agire moralmente. Un esempio lampante è quello delle donne, considerate per secoli esseri da tutelare ma non soggetti morali pienamente autonomi. La lotta femminista ha rivendicato non solo diritti civili e politici, ma soprattutto la capacità di essere artefici del proprio destino. Analogamente, gli individui schiavizzati furono inclusi nel cerchio morale solo in quanto esseri senzienti da non maltrattare, ma non in quanto cittadini dotati di volontà e razionalità, almeno fino alle battaglie abolizioniste²⁷⁸.

Esistono inoltre casi, come quello degli animali o dell'ambiente, in cui chi è riconosciuto come paziente morale non accede mai allo status di agente morale. “Che cosa faranno mai per noi questi posteri?”²⁷⁹, si chiedeva Thomas Stafford all'inizio del Settecento, evidenziando come le generazioni future non abbiano alcuna possibilità di agire. L'asimmetria tra il riconoscimento come pazienti e come agenti solleva interrogativi cruciali: non solo mette in luce lo squilibrio del nostro cerchio morale, ma suggerisce che l'autonomia morale non sia una soglia rigida, bensì un gradiente ostacolato da pregiudizi sociali, normativi e culturali. Riconoscere che spesso si diventa pazienti morali prima di essere riconosciuti come agenti significa comprendere che l'inclusione non si esaurisce nel proteggere la vulnerabilità, ma deve tendere a valorizzare l'autonomia e la partecipazione. Solo così il cerchio morale può diventare non solo più ampio, ma anche più giusto. Tuttavia, l'avvento dei sistemi di IA post-simbolica comporta un rovesciamento di questa dinamica:

²⁷⁷ Ibid., p. 133. Cfr Sebo, J. [2022], *Saving animals, saving ourselves. Why animals matter for pandemic, climate change, and other catastrophes*, Oxford, Oxford University Press

²⁷⁸ Ibid., p. 134

²⁷⁹ Ibid.

molti sistemi di nuova generazione esercitano un'influenza reale sulle nostre vite; infatti, assumono funzioni che implicano responsabilità e scelte, nonché impatti etici significativi²⁸⁰.

Questi sistemi agiscono in modi rilevanti dal punto di vista morale, pur senza essere dotati di coscienza, esperienza soggettiva o capacità di soffrire. In tal senso, “non possono essere considerati pazienti morali, perché non c'è «dentro» di loro nulla che possa essere danneggiato o beneficiato (almeno non nel senso in cui possono essere danneggiate o beneficate entità dotate di coscienza e dunque in grado di soffrire o di provare piacere)²⁸¹.” Ciò solleva un paradosso evidente: potremmo dover trattare alcune IA come agenti morali, attribuendo loro responsabilità e doveri, senza doverle tutelare come soggetti eticamente vulnerabili. Si tratta di una nuova configurazione del cerchio morale, che “disgiunge l'essere agenti morali dall'essere pazienti morali, invertendo così un'asimmetria che ha segnato la storia dell'etica²⁸².”

Uno scenario di questo tipo richiede nuove categorie filosofiche, come “agenti morali non senzienti” o “responsabilità senza coscienza”. Anche dal punto di vista giuridico e sociale, ciò apre interrogativi profondi: possiamo attribuire colpa, responsabilità o obblighi a entità che non possono soffrire, né comprendere la portata morale delle proprie azioni? In sintesi, l'IA post-simbolica segna una svolta storica nell'evoluzione del cerchio morale, poiché introduce per la prima volta la figura dell'agente morale che non è (e forse non sarà mai) anche paziente morale. È chiaro, dunque, che la sfida posta dall'IA non riguarda solo la sicurezza tecnica, ma impone una revisione radicale dei nostri presupposti etici. Il fatto che un'entità possa operare con autonomia decisionale, influenzando le vite umane, pur in assenza di un centro di coscienza, ci obbliga a separare due concetti che la tradizione filosofica ha quasi sempre tenuto uniti: il potere di agire nel mondo e la capacità di provare piacere o dolore. Questo disaccoppiamento non è soltanto una curiosità teorica, ma una

²⁸⁰ De Caro, Giovanola 2025, op. cit., p. 134

²⁸¹ Ibid., p. 135

²⁸² Ibid.

necessità pratica per gestire le responsabilità nel mondo contemporaneo²⁸³.

Dobbiamo interrogarci su quali strumenti giuridici e morali possano regolare entità che “agiscono in modi rilevanti dal punto di vista morale²⁸⁴”, pur rimanendo prive di interiorità. Se l'evoluzione del cerchio morale, come suggerito da Singer, è stata una marcia verso un'inclusione sempre più ampia basata sulla sensibilità, l'ingresso dell'IA ci costringe a guardare altrove.

4.2 IA ed etica: percorsi paralleli

Secondo De Caro e Giovanola, l'odierno panorama tecnologico e filosofico vede una convergenza sempre più profonda e significativa tra l'intelligenza artificiale di stampo post-simbolico e l'etica delle virtù. Questo connubio intellettuale, che ha trovato nuovo vigore negli ultimi anni, si fonda sulla condivisione di una comune impostazione di tipo *bottom-up*, che privilegia la flessibilità, l'adattabilità e una profonda sensibilità al contesto rispetto alle rigide astrazioni tipiche delle concezioni generali. Sebbene esistano già numerosi tentativi di integrare prospettive etiche nei sistemi di IA, la ricerca si è spesso limitata a esaminare l'impatto di tali tecnologie sull'agire umano, trascurando una questione di portata fondamentale, ovvero come sia possibile progettare sistemi di IA che incorporino essi stessi forme di saggezza pratica. È dunque urgente superare la mera promozione di virtù tecno-morali per interrogarsi sulla natura stessa di una saggezza artificiale che sia capace di operare attivamente secondo i dettami della vita morale²⁸⁵.

Per affrontare questa sfida, gli studiosi distinguono tre modelli principali di etica delle virtù che risultano applicabili, in varia misura, all'IA. Il primo, comunemente definito modello standard, si rifà all'etica aristotelica e presuppone l'esistenza indipendente di una pluralità di virtù morali individuali, le quali vengono guidate e coordinate dalla saggezza pratica. Il secondo approccio, noto

²⁸³ Ibid.

²⁸⁴ Ibid.

²⁸⁵ De Caro, Giovanola 2025, op. cit., p. 142

come modello eliminativistico, pur accettando il pluralismo delle virtù, sostiene che la saggezza pratica sia solo un costrutto teorico privo di un reale fondamento ontologico. Il terzo, definito modello Aretai o socratico, propone invece una visione unitaria: “la saggezza pratica esiste realmente, mentre le virtù individuali non sono entità distinte, ma forme assunte dalla saggezza pratica nelle diverse circostanze²⁸⁶.”

Il modello Aretai si delinea come la prospettiva in assoluto più promettente per lo sviluppo di sistemi di IA che possano dirsi “saggi *by design*”, ovvero capaci di incorporare una qualche forma di discernimento pratico fin dalla loro concezione. Tale approccio si articola attorno a tre tesi teoriche fondamentali. La prima è il contestualismo etico, che sottolinea l'importanza di analizzare ogni situazione morale nella sua specifica unicità, senza dover ricorrere meccanicamente all'applicazione di regole universali rigide. La seconda tesi, il monismo delle virtù, attribuisce un primato ontologico alla saggezza pratica, intesa non solo come una tra le tante, ma come il principio unificante che integra tutte le altre disposizioni virtuose. La terza tesi definisce infine la saggezza pratica come una vera e propria forma di *expertise*, un insieme di competenze che permettono al soggetto di agire correttamente nei diversi contesti. Questa forma di competenza etica è caratterizzata da quattro capacità distintive. La prima è la percezione morale, ovvero l'abilità di identificare correttamente i fattori rilevanti e i potenziali conflitti che emergono all'interno di una determinata situazione. La seconda è la deliberazione morale, la capacità di rappresentare mentalmente vari obiettivi, esiti e corsi d'azione, valutandone le ragioni a sostegno. La terza è la motivazione morale, che consiste nel tradurre concretamente l'interesse per i valori in azioni appropriate. La quarta capacità, la regolazione emotiva, risulta fondamentale per l'essere umano, sebbene sia, per il momento, considerata irrilevante per la progettazione di sistemi di IA che, non possedendo emozioni, non necessitano di tale controllo²⁸⁷.

Questo quadro teorico risulta preferibile rispetto alle concezioni alternative per svariate ragioni, sia

²⁸⁶ Ibid., p. 143

²⁸⁷ De Caro, Giovanola 2025, op. cit., p. 143

di natura epistemologica che ontologica. In primo luogo, il modello Aretai obbedisce a quello che potremmo definire il principio di semplicità: invece di postulare molteplici virtù autonome che richiedono di essere coordinate tra loro, si basa su un'unica capacità adattiva che si declina in modo differente a seconda del contesto. In secondo luogo, tale modello risulta particolarmente convincente se applicato al campo della psicologia evolutiva, in quanto riesce a spiegare lo sviluppo morale non come un frammentato processo di apprendimento di regole separate, ma come l'acquisizione progressiva di una capacità unica, la *phrónesis*, che si manifesta in forme diverse nel corso del tempo²⁸⁸. La plausibilità psicologica del modello è un ulteriore punto di forza: “le evidenze psicologiche e neuroscientifiche mostrano che il giudizio morale si basa su una combinazione di intuizioni, emozioni e capacità di discernimento contestuale, piuttosto che sulla mera applicazione di regole²⁸⁹.” Rispetto ad approcci meramente descrittivi o riduzionisti, il modello Aretai preserva un'importante autonomia normativa; infatti, garantisce la possibilità di valutare le condotte senza che il concetto di norma venga svuotato di significato o ridotto a semplice consuetudine. Questo permette di mantenere alto il rigore del giudizio anche in contesti complessi e mutevoli²⁹⁰.

Dal punto di vista della progettazione tecnologica, l'adozione del modello Aretai offre vantaggi competitivi innegabili per l'architettura dei sistemi di IA contemporanei. La scelta di un approccio monistico consente di unificare l'intero processo etico sotto un'unica capacità adattiva. Questo implica che, in fase di costruzione dell'IA, “non è necessario costruire e gestire moduli separati per ogni virtù, ma è sufficiente sviluppare un unico sistema di discernimento pratico e contestuale²⁹¹.” Una tale semplificazione non solo riduce l'onere computazionale, ma rende il processo di sviluppo

²⁸⁸ Ibid., p. 144

²⁸⁹ Ibid.

²⁹⁰ Ibid.

²⁹¹ Ibid., p. 145

molto più realistico e maneggevole rispetto alla complicata gestione di numerosi moduli coordinati da funzioni di controllo superiori. Questo funzionamento basato sulla saggezza pratica risulta sorprendentemente affine alla logica dei moderni sistemi di IA post-simbolica e dei modelli generativi. Tali sistemi, infatti, non operano applicando rigidamente regole fisse, ma imparano a partire dall'esperienza, dai dati e dal riconoscimento di pattern comportamentali, adattandosi dinamicamente ai contesti. Se si considera che le virtù possono essere intese non come entità rigide, ma come modalità di applicazione situata, appare naturale tentare di progettare macchine capaci di apprendimento contestuale, in grado di riconoscere la natura della situazione e adeguare di conseguenza il proprio comportamento. Infine, il modello Aretai garantisce una maggiore flessibilità rispetto ai modelli classici, riducendo drasticamente il rischio di arbitarietà. Mentre nel modello gerarchico rigido la gestione di diverse virtù può risultare complessa, opaca e difficile da controllare, l'approccio Aretai incorpora la *phrónesis* fin dall'inizio, ponendo l'accento sulla capacità di fare la cosa giusta, nel momento giusto e nel modo appropriato. Questo emerge naturalmente dall'interazione con situazioni concrete, piuttosto che da una complessa combinazione di moduli indipendenti difficili da armonizzare tra loro. In conclusione, l'adozione di questo modello si sposa perfettamente con la logica *bottom-up* delle moderne tecnologie informatiche. Non partendo da norme predefinite, ma da una capacità che apprende da esempi virtuosi in situazioni concrete, il design etico diventa una sfida molto più gestibile. In termini operativi, ciò significherebbe ridurre sensibilmente la complessità di progettazione, lasciando che il sistema apprenda a navigare la complessità morale in modo più fluido ed efficace di quanto farebbe un sistema fondato su regole atomizzate²⁹².

Tuttavia, l'implementazione di una sorta di *phrónesis* artificiale incontra inevitabilmente delle difficoltà tecniche e teoriche di notevole rilievo. La prima grande obiezione riguarda la varietà dei dati necessari per addestrare i sistemi di IA, legata alla presunta impossibilità di definire valori universali di fronte alla diversità culturale. Si obietta spesso che ogni cultura possieda valori propri, il che renderebbe impossibile la creazione di sistemi etici validi su scala globale. A questa critica si

²⁹² De Caro, Giovanola 2025, op. cit., p. 145

risponde ricordando che tale problema non affligge solo il modello Aretai, ma qualsiasi progetto di *ethics by design*. Inoltre, esistono solide argomentazioni filosofiche, psicologiche e antropologiche che dimostrano come, pur in presenza di una grande varietà culturale, si riscontri sempre un nucleo consistente di credenze morali condivise tra le diverse civiltà. Di fronte a casi estremi di dissenso, rimane sempre aperta la possibilità di un confronto razionale tra posizioni in conflitto. Se qualcuno sostenesse la legittimità morale di pratiche profondamente riprovevoli, saremmo in grado di giudicare tale posizione come errata, suggerendo che, al di là delle differenze superficiali, esiste un terreno comune di giudizio morale su cui è possibile fondare sistemi etici artificiali²⁹³. Il secondo problema riguarda la necessità di addestrare i sistemi su una vastissima quantità di situazioni morali, un compito che appare immane se non affrontato con le moderne tecnologie di apprendimento. Tuttavia, anche questa sfida può essere superata senza dover ricorrere a un'etichettatura sistematica e centralizzata di ogni possibile scenario. È possibile, infatti, ricorrere a strategie di *weak supervision* o all'utilizzo di simulazioni di agenti morali esperti, in grado di generare esempi situati. “L'addestramento dell'IA, insomma, può essere iterativo, basato su feedback e simulazioni, in modo da costruire progressivamente una forma di saggezza pratica appresa attraverso il *machine learning*²⁹⁴.” In questo modo, il sistema non dipenderebbe da un database etico esaustivo e statico, ma apprenderebbe a riconoscere pattern di comportamento virtuoso in situazioni concrete attraverso l'esperienza reiterata. Questo tipo di apprendimento risulta decisamente più realistico e promettente di qualsiasi sistema basato su regole rigide che non tengono conto della complessità del reale. In definitiva, “la difficoltà nel reperire dati etici adeguati è una sfida tecnica, non una confutazione teorica del modello²⁹⁵.” Se il modello Aretai è in effetti il più adatto a rappresentare la complessità del giudizio morale umano, allora vale certamente la pena affrontare le difficoltà tecniche necessarie per tentare di implementarlo nei sistemi artificiali del

²⁹³ Ibid., p. 146

²⁹⁴ De Caro, Giovanola 2025, op. cit., p. 146

²⁹⁵ Ibid., p. 147

futuro. L'impiego di questa visione nelle architetture algoritmiche non rappresenta soltanto un miglioramento tecnico, ma una vera e propria rivoluzione concettuale nel modo in cui intendiamo la responsabilità dei sistemi. Passare da un'etica delle regole a un'etica della virtù significa riconoscere che la macchina non deve semplicemente seguire un protocollo predeterminato, ma deve sviluppare una sorta di sensibilità per la situazione in cui si trova ad operare. Questa sensibilità non è qualcosa che può essere programmato *a priori* attraverso una lista infinita di comandi, ma deve essere il frutto di un apprendimento profondo, capace di cogliere le sfumature di un contesto in perenne evoluzione²⁹⁶.

Sebbene la strada sia ancora lunga e le incertezze non manchino, la convergenza tra la riflessione filosofica sulla saggezza pratica e le capacità computazionali del *machine learning* apre scenari del tutto nuovi. L'intelligenza artificiale, se guidata da modelli di etica delle virtù come l'Aretai, potrebbe non solo diventare più sicura, ma anche più capace di integrarsi in contesti umani complessi, dove la flessibilità del giudizio è spesso più importante della rigidità del calcolo. L'attenzione si sposta quindi dall'oggetto da progettare al processo di apprendimento, ponendo le basi per una nuova forma di collaborazione tra sistemi artificiali e agenti umani in cui l'etica non è un vincolo esterno, ma parte integrante del funzionamento del sistema stesso²⁹⁷.

4.3 Principi etici per l'IA

Come dice Floridi, in questi anni “molte organizzazioni hanno lanciato un'ampia gamma di iniziative per stabilire principi etici per l'adozione di un'IA socialmente vantaggiosa²⁹⁸”; nondimeno, è importante cercare di capire quali principi sia giusto adottare per costruire un'IA etica, per evitare che si crei una sorta di “'mercato dei principi' in cui le parti potrebbero essere tentate di 'acquistare'

²⁹⁶ Ibid.

²⁹⁷ Ibid.

²⁹⁸ Floridi 2022, op. cit., p. 92

quelli più allettanti²⁹⁹.”

La ricerca accademica sull'intelligenza artificiale ha radici profonde che risalgono agli anni Cinquanta, e parallelamente, il dibattito etico ad essa collegato possiede una storicità quasi identica. Tuttavia, è solo nell'epoca contemporanea che l'accelerazione dei progressi tecnici e delle applicazioni pratiche ha imposto una riflessione urgente sulle opportunità e sui rischi che queste tecnologie comportano per la società nel suo complesso. Questa crescente necessità di definire politiche chiare ha innescato una proliferazione incontrollata di iniziative; ogni nuova proposta tende a generare una dichiarazione di intenti, valori o principi guida, creando una sorta di dinamica di emulazione, spesso definita come una corsa al conformismo, in cui ogni attore sente il bisogno di partecipare attivamente al dibattito³⁰⁰.

Di fronte a questa “inflazione” di documenti etici, il rischio principale è quello di cadere in inutili sovrapposizioni o di generare confusione: questo creerebbe complicazioni per la transizione verso l'adozione di regole e standard pratici che assicurino un impatto socialmente positivo dell'IA. È diventato essenziale avviare un'analisi comparativa per valutare la convergenza o la divergenza tra queste numerose dichiarazioni, al fine di produrre un quadro sintetico e unificato. A tal fine, Floridi ha selezionato sei iniziative di alto profilo che soddisfano criteri rigorosi di attualità, rilevanza sociale, prestigio organizzativo e influenza³⁰¹.

L'analisi unitaria di questi documenti rivela, nonostante le differenze linguistiche, un sorprendente grado di coerenza e convergenza. Tale comunanza di intenti viene efficacemente letta da Floridi attraverso i quattro principi cardine della bioetica classica: beneficenza, non maleficenza, autonomia e giustizia. Sebbene questo approccio fornisca una solida base, il confronto evidenzia che i modelli bioetici tradizionali non esauriscono le peculiarità delle sfide poste dall'IA, la quale si configura, come già detto, non come una nuova forma di intelligenza, ma come una forma di agire

²⁹⁹ Ibid., p. 93

³⁰⁰ Ibid., p. 93

³⁰¹ Ibid., p. 94

inedita³⁰².

Ciononostante, il significato dei termini etici risulta spesso contestato e non esaustivo. Per colmare questa lacuna, per Floridi emerge con forza la necessità di integrare un quinto principio fondamentale: l'esplicabilità. Questo nuovo principio è duplice: da un lato comprende l'intelligibilità epistemologica, che risponde alla domanda su come funzionino i sistemi IA; dall'altro include l'aspetto etico della responsabilità (*accountability*), fondamentale per chiarire chi sia responsabile delle azioni del sistema in contesti specifici, siano essi esperti o utenti finali. Questo approccio denota la consapevolezza che, nonostante le convergenze rilevate, la natura distintiva dell'IA richiede un'attenzione particolare e una prudenza metodologica superiore nel definire i quadri di riferimento etici³⁰³.

Il dibattito etico sull'intelligenza artificiale non è un esercizio astratto, ma una necessità urgente di fronte a tecnologie che stanno già plasmando, in modi spesso invisibili, le fondamenta della nostra vita quotidiana. Esiste una profonda distinzione tra il principio di beneficenza, secondo il quale dovremmo utilizzare l'intelligenza artificiale per “promuovere il benessere di tutte le creature senzienti” e fare in modo che siano assicurate “le precondizioni di base per la vita sul nostro pianeta, la continua prosperità per l'umanità e la conservazione di un buon ambiente per le generazioni future³⁰⁴”, e la non maleficenza, che impone di non arrecare danno e in generale non compiere il male. Sebbene possano sembrare logicamente equivalenti, si tratta di concetti distinti che richiedono un'attenzione specifica, specialmente per prevenire violazioni della privacy o utilizzi impropri delle tecnologie di IA. La sfida è enorme, e in tal senso Floridi pone una sorta di provocazione quando si chiede se “dovremmo guardarci dalla maleficenza del dottor Frankenstein (come suggerisco) o del suo mostro³⁰⁵.” Questo avvertimento ci sprona a riflettere seriamente su chi

³⁰² Ibid., p. 95

³⁰³ Floridi 2022, op. cit., p. 96

³⁰⁴ Ibid., p. 97

³⁰⁵ Ibid., p. 98

debba essere considerato responsabile delle azioni compiute dai sistemi autonomi: gli sviluppatori o le macchine stesse?

Al centro di questa problematica si colloca la questione dell'autonomia e del potere di decidere di decidere. Delegando decisioni agli artefatti tecnologici, corriamo il rischio che la crescita dell'autonomia artificiale possa soffocare, o addirittura minare, il fiorire dell'autonomia umana. Non deve mai essere dimenticato il principio fondamentale secondo il quale “il potere autonomo di danneggiare, distruggere o ingannare gli esseri umani non dovrebbe mai essere conferito all'IA³⁰⁶.” L'obiettivo finale deve essere la promozione dell'autonomia umana, il che può avvenire solo se si limita quella delle macchine in modo che sia sempre reversibile, permettendo così all'essere umano di riprendere il controllo, come farebbe un pilota che disattiva il pilota automatico in caso di pericolo. Questa dinamica di potere non è equamente distribuita nella società, il che ci conduce direttamente al principio di giustizia. È imperativo che lo sviluppo dell'IA promuova la prosperità condivisa, preservi la solidarietà ed eviti l'iniquità, combattendo attivamente ogni forma di discriminazione e distorsione nei dati di addestramento. Siamo di fronte a un panorama complesso in cui il confine tra il medico che prescrive le cure dell'IA e il paziente che le riceve si fa sempre più labile: “una piccola frazione dell'umanità è attualmente impegnata a sviluppare una serie di tecnologie che stanno già trasformando la vita quotidiana di quasi tutti gli altri³⁰⁷.” Questa disparità sottolinea la necessità critica di spiegabilità e trasparenza: per essere eticamente responsabili, dobbiamo poter comprendere il perché dietro ogni decisione presa o suggerita da un sistema autonomo, al fine di evitare che esso rimanga una scatola nera incomprensibile.

Il quadro etico unificato che emerge dall'analisi di molteplici documenti internazionali – tra cui quelli di Asilomar, la Dichiarazione di Montréal e i lavori di AI4People – si articola attorno a cinque pilastri: beneficenza, non maleficenza, autonomia, giustizia ed esplicabilità. Quest'ultima, che include sia l'intelligibilità che la responsabilità, rappresenta il “cruciale pezzo mancante del puzzle

³⁰⁶ Ibid.

³⁰⁷ Ibid., p. 100

etico dell'IA³⁰⁸.” Non esiste un singolo continente o cultura che possa rivendicare l'esclusiva sull'etica dell'IA. Al contrario, è necessario un approccio globale, capace di integrare prospettive geografiche e sociali diversificate, specialmente in un contesto in cui nazioni come la Cina dichiarano ambizioni di leadership mondiale entro il 2030, investendo massicciamente nelle tecnologie all'avanguardia.

Nonostante il dibattito sia spesso dominato da visioni occidentali, l'entusiasmo per le potenzialità dell'IA è condiviso trasversalmente, sebbene le interpretazioni delle priorità etiche possano variare. È fondamentale superare la retorica nazionalistica, poiché l'etica non è una competizione territoriale, ma un obbligo per ogni istituzione, azienda o governo che sviluppa tecnologie destinate a influenzare le masse. La via da seguire è quella di trasformare questi principi in pratiche concrete, architetture normative e standard tecnici che possano effettivamente guidare l'innovazione. Documenti come le *Linee guida etiche per l'IA affidabile* della Commissione Europea o la *Rome Call for AI Ethics* testimoniano un crescente impegno internazionale nel definire una società della buona IA³⁰⁹.

In definitiva, la sfida che ci attende è quella di tracciare un percorso che sia, allo stesso tempo, socialmente preferibile, equo e sostenibile. Non dobbiamo lasciarci sedurre acriticamente dal progresso tecnologico, perdendo di vista la necessità di un'architettura che garantisca risultati sociali positivi e protegga l'umanità dalle sue stesse creazioni. Il passaggio dai buoni principi alle buone pratiche richiederà un attento monitoraggio e un *auditing* costante dei sistemi che ci assicuri di non cadere in vecchie o nuove trappole etiche, durante questa transizione tecnologica senza precedenti. Solo attraverso una comprensione profonda e una collaborazione globale potremo garantire che l'intelligenza artificiale rimanga al servizio del benessere umano, evitando le derive distruttive temute dagli osservatori più attenti³¹⁰.

³⁰⁸ Floridi 2022, op. cit., p. 101

³⁰⁹ Ibid., pp. 103-104

³¹⁰ Ibid., p. 105

4.4 Usare l'IA per il bene sociale

L'intelligenza artificiale per il bene sociale (AI4SG) rappresenta un ambito in rapida crescita, volto a sfruttare le tecnologie digitali per affrontare problemi globali complessi. Sebbene offra opportunità senza precedenti in settori come la sanità e la sostenibilità ambientale, la sua rapida espansione ha generato sfide significative. Attualmente, manca un quadro di riferimento condiviso per valutarne l'efficacia, il che rende difficile distinguere tra successi progettati e risultati benefici talvolta meramente casuali. Il rischio principale risiede in una progettazione non accurata, che può trasformare strumenti potenzialmente utili in cause di nuovi danni. Come evidenziato da Floridi, “l'IA può fare più male che bene, laddove amplifica invece di mitigare i mali della società, per esempio ampliando anziché restringendo le disuguaglianze esistenti³¹¹.” Esempi storici di fallimenti tecnologici in ambito medico confermano che la mancanza di dati rappresentativi o la scarsa contestualizzazione possono condurre a diagnosi errate e a una diffidenza giustificata da parte degli operatori³¹².

Per superare queste criticità, è necessario superare l'approccio *ad hoc* e adottare modelli di valutazione rigorosi, che vadano oltre la mera produttività commerciale. L'obiettivo deve essere quello di integrare fattori di design che garantiscano equità, trasparenza e rispetto della privacy. In quest'ottica, l'adozione di buone pratiche non dovrebbe essere automatica, ma frutto di una riflessione critica. Infatti è necessario “che ogni migliore pratica sia: (i) considerata in modo proattivo e (ii) *non* incorporata se e soltanto se esiste una ragione chiara, dimostrabile e moralmente difendibile per cui non dovrebbe esserlo³¹³.” L'IA non è una panacea universale; il suo valore sociale dipende dalla capacità di valutare criticamente quando e come impiegarla, garantendo che ogni intervento sia accompagnato da meccanismi di sostegno, regolamentazione e una profonda

³¹¹ Floridi 2022, op. cit., p. 224

³¹² Ibid., pp. 224-225

³¹³ Ibid., p. 226

comprensione dei contesti di applicazione.

Dopo questa disamina introduttiva, Floridi passa alla definizione di AI4SG (IA per il bene sociale):

“il design, lo sviluppo e l'implementazione di sistemi di IA in modo da (i) prevenire, mitigare o risolvere i problemi che incidono negativamente sulla vita umana e/o sul benessere del mondo naturale e/o (ii) consentire sviluppi preferibili dal punto di vista sociale e/o sostenibili dal punto di vista ambientale³¹⁴.”

L'analisi dell'intelligenza artificiale per il bene sociale (AI4SG) si basa sull'esame di ventisette progetti significativi che hanno permesso di identificare sette fattori essenziali per il loro design³¹⁵.

Ogni fattore si relaziona ai cinque principi etici fondamentali dell'IA – beneficenza, non maleficenza, giustizia, autonomia e spiegabilità –; questo garantisce coerenza con il quadro etico generale. In particolare, il principio di beneficenza richiede che l'IA generi benefici preferibili e sostenibili per le persone e il mondo naturale, sebbene ciò rappresenti una condizione necessaria ma non sufficiente³¹⁶. Questi fattori, pur essendo interconnessi, vengono trattati separatamente per semplicità e non sono gerarchizzati, ma assumono particolare rilevanza in base alle specificità tecnologiche e ai rischi associati. Essi includono aspetti tecnici come la falsificabilità e la protezione dalla manipolazione dei dati, insieme a elementi sociotecnici come l'equità e la semantizzazione adatta all'uomo. Infine, è fondamentale sottolineare che questi fattori non costituiscono una semplice lista di controllo, poiché è necessario gestire le tensioni tra essi ed evitare valutazioni binarie semplicistiche tra progetti diretti o non diretti al bene sociale, privilegiando invece un approccio di design eticamente robusto e pragmaticamente applicabile³¹⁷.

I già accennati fattori che risultano essenziali per il successo dell'IA per il bene sociale sono

³¹⁴ Ibid., p. 227

³¹⁵ Ibid., pp. 228-229

³¹⁶ Floridi 2022, op. cit., p. 229

³¹⁷ Ibid., p. 230

precisamente sette e sono i seguenti: “(1) *falsificabilità e implementazione incrementale*; (2) *garanzie contro la manipolazione dei predittori*; (3) *intervento contestualizzato in ragione del destinatario*; (4) *spiegazione contestualizzata in ragione del destinatario e finalità trasparenti*; (5) *tutela della privacy e consenso dell'interessato*; (6) *equità concreta* e (7) *semantizzazione adatta all'umano*³¹⁸.”

Per quanto riguarda il principio di falsificabilità e implementazione incrementale, iniziamo con il dire che l'affidabilità dell' IA per il bene sociale è fondamentale per garantirne un impatto positivo, dal momento che si basa sul principio di beneficenza o non maleficenza. Poiché mancano linee guida universali, la falsificabilità, ovvero la capacità di verificare empiricamente i requisiti critici, diventa il fattore decisivo per migliorare tale affidabilità³¹⁹.

Tuttavia, raggiungere la certezza assoluta è impossibile in contesti complessi; la falsificabilità permette piuttosto di riconoscere quando un'applicazione non è affidabile, pur non potendone garantire l'affidabilità certa. È dunque necessario adottare un ciclo di implementazione incrementale. Questo processo richiede di rendere i requisiti falsificabili, testarli in ambienti protetti e, se confermati, estendere i test a contesti più ampi, con lo scopo di mantenere sempre la possibilità di interrompere l'attività in caso di rischi³²⁰.

Sebbene l'uso di verifiche formali e simulazioni sia utile per testare requisiti in contesti critici, queste non bastano, poiché modelli teorici possono risultare errati una volta applicati nel mondo reale. Infine, la correzione costante delle ipotesi a priori e l'apprendimento dai dati sono essenziali per adattare i sistemi in contesti dinamici e critici³²¹, ragion per cui “*i progettisti di AI4SG dovrebbero identificare i requisiti falsificabili e testarli in fasi incrementali dal laboratorio al*

³¹⁸ Ibid., p.231

³¹⁹ Ibid.

³²⁰ Ibid., p. 232

³²¹ Ibid., p. 234

*mondo esterno*³²².”

Per quanto concerne le garanzie contro la manipolazione dei predittori, l'impiego dell'IA in progetti di AI4SG per prevedere tendenze solleva criticità legate alla manipolazione dei dati e all'eccessiva dipendenza da indicatori non causali. Il rischio principale è che modelli facilmente interpretabili si prestino ad abusi. L'introduzione dell'IA, per la sua scala applicativa, amplifica tali problematiche. Se le variabili predittive sono note, gli attori coinvolti possono modificarle strategicamente per eludere interventi, riducendo l'efficacia del modello. Ad esempio, un agente di polizia potrebbe manipolare il proprio comportamento per evitare di essere segnalato, nel caso dei cosiddetti predittori di interazioni avverse tra agenti di polizia³²³.

Di conseguenza, i designer di AI4SG (IA per il bene sociale) devono integrare “*garanzie che (i) assicurino che gli indicatori non causali non distorcano in modo inappropriato gli interventi e (ii) limitino, quando appropriato, la conoscenza di come gli input influenzano gli output dei sistemi di AI4SG, per prevenire la manipolazione*”³²⁴.

Per quanto riguarda l'importanza di progettare interventi e spiegazioni basati sull'intelligenza artificiale che siano contestualizzati in relazione al destinatario e che garantiscano al contempo il rispetto dell'autonomia dell'utente, una sfida centrale per i progetti di AI4SG consiste nell'elaborare “*interventi che bilancino i benefici attuali e futuri*”³²⁵. È fondamentale che il software operi in modo rispettoso, poiché interventi eccessivamente invadenti potrebbero indurre gli utenti a rifiutare la tecnologia³²⁶.

Per realizzare un intervento efficace, Floridi identifica cinque dimensioni chiave, tra cui le

³²² Ibid., p. 235

³²³ Floridi 2022, op. cit., pp. 235-236

³²⁴ Ibid., p. 237

³²⁵ Ibid.

³²⁶ Ibid., pp. 237-238

caratteristiche dell'utente, la modalità di coordinamento, il significato dell'intervento e i suoi effetti complessivi. La quinta dimensione riguarda la flessibilità, ovvero la necessità che gli utenti possano ignorare o modificare gli interventi. Di conseguenza, i designer dovrebbero costruire sistemi decisionali in dialogo con gli utenti, che rispettino il loro diritto di non seguire le indicazioni fornite dal sistema³²⁷.

Parlando di spiegazione contestualizzata in ragione del destinatario e finalità trasparenti, Floridi afferma che il modo in cui una spiegazione viene concepita deve variare, poiché il “quadro concettuale, ossia il livello di astrazione (LdA), dipende da cosa viene spiegato, a chi e per quale scopo³²⁸.” Poiché l'adequatezza di una spiegazione differisce in base all'esperienza del destinatario, i progettisti devono selezionare con attenzione il livello di astrazione appropriato, evitando disallineamenti concettuali che renderebbero la spiegazione irrilevante o incomprensibile.

Floridi sottolinea come la spiegazione generi fiducia e favorisca l'adozione delle soluzioni di AI4SG. A tal fine, è essenziale che il software utilizzi argomentazioni che siano in grado di persuadere il pubblico a cui si rivolge. Questo include informazioni sia sul funzionamento che sulle ragioni alla base di una specifica decisione³²⁹.

Anche la trasparenza sugli obiettivi è cruciale per mantenere l'autonomia dell'utente. Sebbene la trasparenza sia la posizione predefinita, esistono situazioni in cui obiettivi temporaneamente opachi possono essere giustificati da esigenze scientifiche o di sicurezza, sebbene tali scelte debbano essere attentamente valutate. In tal senso, i progettisti devono scegliere un livello di astrazione per la spiegazione che “soddisfi lo scopo esplicativo auspicato e sia appropriato al sistema e ai destinatari³³⁰.”

³²⁷ Ibid., p. 239

³²⁸ Ibid., p. 241

³²⁹ Floridi 2022, op. cit., pp. 242-244

³³⁰ Ibid., p. 246

Riguardo invece la tutela della privacy – argomento di cui abbiamo già parlato nel capitolo precedente – e il consenso dell'interessato, Floridi nota giustamente che “il rispetto della privacy è [...] una condizione necessaria della dignità umana, poiché possiamo considerare le informazioni personali come elementi costitutivi di un individuo, cosicché la sottrazione di dati senza consenso può costituire una violazione della dignità umana³³¹.”

La gestione dei dati personali richiede un delicato equilibrio, poiché, come evidenziato da Floridi, “possono infatti sorgere tensioni tra diverse soglie di consenso³³²”, specialmente in situazioni d'urgenza. Sebbene il consenso informato sia spesso ideale, in contesti sanitari o di crisi si esplorano alternative come il consenso presunto o quello dinamico, che permette agli individui di regolare le proprie preferenze in modo granulare. Tuttavia, l'uso di dati senza un adeguato consenso può portare a gravi violazioni, come accaduto nel caso del Royal Free NHS Foundation Trust³³³. Oltre al consenso e alla privacy, anche l'equità concreta rappresenta una sfida fondamentale. Gli algoritmi, se addestrati su dati storici distorti, rischiano di perpetuare pregiudizi (*bias*) iniqui, influenzando decisioni cruciali in ambito sanitario o giudiziario. Infatti, “le iniziative di AI4SG che si basano su dati distorti possono propagare tale pregiudizio attraverso un circolo vizioso³³⁴.” È imperativo che i designer eliminino le variabili irrilevanti per garantire risultati etici, evitando al contempo di sterilizzare eccessivamente i dati; questo porta a perdere così sfumature contestuali importanti per l'inclusione³³⁵.

Infine, l'AI4SG deve promuovere quello che viene definito capitale semantico, ovvero la capacità umana di “dare significato e conferire senso a qualcosa³³⁶.” L'automazione non deve sostituire

³³¹ Ibid., p. 247

³³² Ibid.

³³³ Ibid., pp. 247-248

³³⁴ Ibid., p. 250

³³⁵ Ibid., p. 251

³³⁶ Ibid., p. 253

l'autonomia umana nel definire i significati, poiché procedure arbitrarie potrebbero limitare la capacità critica, ad esempio nel contesto giudiziario. L'approccio corretto consiste nell'impiegare l'IA per facilitare compiti banali, permettendo agli umani di concentrarsi su interazioni che richiedono empatia e supporto, come nel rapporto tra assistente e paziente. In sintesi, la tecnologia dovrebbe essere uno strumento a supporto della semantizzazione umana e, come indicato da Floridi, *“i designer di AI4SG non dovrebbero ostacolare la capacità delle persone di semantizzare (cioè di dare significato e conferire senso a) qualcosa³³⁷.”*

In definitiva, per chiarire tutti questi concetti che possono risultare molto oscuri al lettore non avvezzo a queste tematiche, l'intelligenza artificiale applicata al bene sociale non va considerata come una soluzione magica e automatica, bensì come un delicato strumento che richiede una progettazione profondamente rigorosa e consapevole per evitare che, anziché risolvere i problemi, finisca per amplificarli. Tradurre in pratica questo approccio significa abbandonare l'illusione che basti una buona idea per ottenere un impatto positivo; invece, bisogna adottare una metodologia che ponga al centro la sicurezza e il rispetto umano, e questo comporta: l'obbligo di testare le tecnologie in contesti protetti e incrementali per poterle correggere o arrestare tempestivamente; la necessità di blindare i sistemi contro manipolazioni esterne e pregiudizi algoritmici che rischiano di perpetuare ingiustizie e soprattutto la ferma volontà di mantenere il controllo nelle mani delle persone. In questo senso, l'intelligenza artificiale deve agire come un prezioso alleato capace di agire in trasparenza e in armonia con le nostre decisioni, senza mai sostituire la nostra capacità critica o violare la nostra dignità; infine, deve garantire che ogni intervento digitale sia sempre subordinato al benessere reale degli individui e mai all'efficienza fine a se stessa.

³³⁷ Floridi 2022, op. cit., p. 255

4.5 Realizzare una società per la buona IA

Nel capitolo 11 del già citato Floridi (2022), il filosofo fornisce una serie di indicazioni per lo sviluppo di una cosiddetta “buona IA”, un'intelligenza artificiale da cui trarre solo aspetti positivi e in definitiva il meglio per gli esseri umani e la società intera.

Nell'introduzione di tale capitolo, Floridi ci dice che abbiamo commesso l'errore di ritenere sia il web che l'intelligenza artificiale semplicemente come l'ennesimo dispositivo che deve essere regolamentato al pari di qualunque altro. In realtà, il web stesso è stato qualcosa che ha trasformato radicalmente la nostra società, ha creato un ambiente del tutto nuovo (l'infosfera) e in tal modo avrebbe dovuto essere regolamentato³³⁸. È perciò di vitale importanza non commettere lo stesso errore con l'intelligenza artificiale, poiché essa “non è solo un servizio commerciale. È una forza potente, una nuova forma dell'agire smart nell'infosfera, che sta già rimodellando le nostre vite, interazioni e ambienti” ed è per questo che “deve essere orientata verso il bene della società, di tutti i suoi membri e degli ambienti che condividiamo³³⁹.”

Quindi, in definitiva, non ha più senso chiedersi se l'intelligenza artificiale avrà o meno un impatto sulla nostra società, perché questo processo è già in atto. Si tratta solo di capire e decidere se questo impatto sarà qualcosa di positivo o di negativo, e di avere una risposta a chi, come, dove e quando sentirà tale impatto. Per rispondere a queste domande, Floridi individua quattro opportunità che l'IA può offrire alla società:

- a) “la realizzazione autonoma di noi stessi [...];
- b) l'*agire* umano, ovvero cosa possiamo fare;
- c) le capacità individuali e sociali, ovvero cosa possiamo conseguire;
- d) la *coesione* sociale, ovvero come possiamo interagire gli uni con gli altri e con il mondo³⁴⁰.”

³³⁸ Ibid., pp. 279-280

³³⁹ Ibid., p. 280

³⁴⁰ Floridi 2022, op. cit., pp. 280-281

In ognuno di questi quattro casi, l'intelligenza artificiale può essere utilizzata per il bene della società oppure per il suo male. Per intenderci: nel caso del punto a), se usiamo bene l'IA potremmo fare in modo di realizzarci sempre meglio; se invece la usiamo male, eroderemmo quelle che sono le nostre capacità. Per quanto riguarda l'agire umano, se usiamo bene l'IA non possiamo fare altro che migliorare in tal senso; se la utilizziamo in maniera sbagliata, l'intelligenza artificiale non farà altro che rimuovere le nostre responsabilità in quanto esseri umani. Riguardo il punto c), usufruendo dell'intelligenza artificiale in maniera positiva potremmo amplificare le nostre capacità individuali e sociali; in caso contrario, l'IA farà in modo che la nostra capacità di controllo sarà sempre più ridotta. Infine, per quanto riguarda la coesione sociale, un uso propositivo dell'intelligenza artificiale può solo alimentare tale coesione, mentre, al contempo, un uso negativo sgretolerà pian piano l'autodeterminazione umana, come abbiamo avuto modo di vedere nel capitolo precedente³⁴¹. È per questo che, per usare le parole di Floridi, “la possibilità di progressione sociale rappresentata dalle suddette opportunità deve essere soppesata di contro al rischio che una manipolazione malevola sia resa possibile o potenziata dall'IA³⁴².”

L'intelligenza artificiale rappresenta una forza trasformativa capace di ridefinire radicalmente il mercato del lavoro e la struttura stessa della società contemporanea. Da un lato, essa offre il potenziale per una profonda realizzazione personale, il che può liberare gli individui da compiti ripetitivi e gravosi, in modo simile a quanto avvenuto con le innovazioni tecnologiche del passato. In questo modo potremmo disporre di tempo prezioso per attività intellettuali, culturali e sociali più appaganti. Dall'altro lato, la rapidità con cui tale cambiamento si sta attuando solleva preoccupazioni significative, non solo in merito all'obsolescenza di competenze professionali consolidate (come evidenziato nel capitolo 2.2), ma anche riguardo alla possibile compromissione di settori critici e al rischio di una distribuzione iniqua dei benefici. La sfida, pertanto, non risiede semplicemente nel gestire l'automazione, ma nel promuovere uno sviluppo dell'IA che sappia

³⁴¹ Ibid., p. 281

³⁴² Ibid., p. 282

mitigare i costi sociali; questo può avvenire attraverso l'implementazione di strategie innovative come il reddito di base universale e garantendo una solida rete di protezione intergenerazionale. È fondamentale governare questa transizione tecnologica in modo equo, in modo da ridurre l'impatto dirompente dell'IA affinché la vita umana possa essere vissuta in modo più intelligente, preservando al contempo la stabilità e la coesione sociale di fronte a un cambiamento così profondo e inevitabile³⁴³.

Per quanto riguarda l'agire umano, non possiamo negare che, grazie all'intelligenza artificiale, “possiamo fare di più, meglio e più velocemente”, ragion per cui “l'IA potrebbe essere paragonata all'impatto avuto dal motore sulle nostre vite. [...] La responsabilità è dunque essenziale, in considerazione del tipo di IA che sviluppiamo, di come la utilizziamo e se ne condividiamo benefici e vantaggi con tutti³⁴⁴.”

È chiaro che l'assenza di responsabilità nell'uso dell'intelligenza artificiale sia legata a un quadro sociopolitico inadeguato e a una mentalità *black box* (scatola nera) che sottrae i processi decisionali alla comprensione e al controllo umano, problematiche che riguardano tanto eventi estremi quanto decisioni quotidiane automatizzate. Tuttavia, Floridi sostiene che la delega di compiti ai sistemi autonomi non sia un gioco a somma zero, poiché, se sviluppata in modo equilibrato, l'IA ha il potenziale di migliorare e moltiplicare l'agire umano. L'IA infatti può agire come un “quadro di facilitazione” progettato per promuovere risultati moralmente positivi e, in definitiva, per amplificare e rafforzare i sistemi morali condivisi³⁴⁵.

L'intelligenza artificiale dischiude prospettive straordinarie per potenziare le capacità umane e affrontare sfide globali, dalla gestione del clima alla logistica; in tal senso, l'IA promuove una collaborazione più coesa tra individui e società. Tuttavia, la portata di tale innovazione solleva rischi speculari, specialmente quando la delega di compiti cruciali a sistemi autonomi compromette

³⁴³ Floridi 2022, op. cit., pp. 282-283

³⁴⁴ Ibid., pp. 283-284

³⁴⁵ Ibid., p. 284

la nostra supervisione diretta sui processi decisionali. Il pericolo è che errori o danni, non più monitorati o corretti dall'uomo, si radichino profondamente nelle routine quotidiane. È necessario bilanciare le ambiziose opportunità tecnologiche con una vigilanza costante ed è “tassativo trovare un equilibrio tra il perseguimento delle ambiziose opportunità offerte dall'IA per migliorare la vita umana e ciò che possiamo conseguire, da un lato, e, dall'altro, assicurarci di mantenere il controllo su questi importanti sviluppi e sui loro effetti³⁴⁶.”

L'impiego dell'IA dovrebbe mirare a incentivare comportamenti socialmente preferibili attraverso una “spinta gentile” che non leda, bensì protegga l'autonomia individuale. È cruciale che tali sistemi non impongano condizioni esterne, ma agiscano come facilitatori di scelte consapevoli, simili a una dieta autoimposta. Esiste però il rischio concreto che l'automazione, attraverso routine invisibili, finisca per erodere la capacità di autodeterminazione umana, modificando comportamenti non intenzionali. Pertanto, “il potere predittivo e l'incessante spinta gentile dell'IA, anche se non intenzionali, dovrebbero essere posti al servizio dell'autodeterminazione umana e favorire la coesione sociale, senza minare la dignità o il progresso umani³⁴⁷.”

Le quattro opportunità e le relative sfide descritte da Floridi delineano un panorama complesso sull'impatto dell'IA, che richiede un approccio che accetti i *trade-off* (cioè la necessità di bilanciare vantaggi e potenziali danni dell'IA) e lavori per mitigare i rischi, al fine di promuovere la dignità e il progresso umano. Il filosofo sottolinea la necessità di superare il semplice rispetto della legge, definito come il minimo richiesto ma insufficiente, per adottare una strategia più solida. “Abbiamo bisogno di una strategia etica, di un'etica soft per la società digitale che vogliamo costruire³⁴⁸.” Floridi quindi introduce venti raccomandazioni, suddivise in quattro tipologie: valutare, sviluppare, incentivare e sostenere. Queste possono essere attuate dai responsabili politici o in collaborazione

³⁴⁶ Ibid., p. 285

³⁴⁷ Floridi 2022, op. cit., p. 286

³⁴⁸ Ibid., p. 287

con le parti interessate, con l'obiettivo di integrare i principi etici fin dalla fase di progettazione³⁴⁹.

La valutazione dell'IA deve essere fondata sulla trasparenza e sulla spiegabilità, strumenti essenziali per consolidare la fiducia del pubblico. La creazione di una società della buona IA richiede un approccio *multistakeholder* (cioè che riguarda tutte le parti interessate) che coinvolga sviluppatori, utenti e legislatori sin dall'inizio, riconoscendo che diverse cornici culturali influenzano l'atteggiamento verso le nuove tecnologie. Floridi specifica che l'approccio europeo proposto non deve essere visto come una chiusura eurocentrica, poiché tali raccomandazioni restano universalizzabili. La prima raccomandazione invita a “valutare la capacità delle istituzioni esistenti, come i tribunali civili nazionali, di riparare agli errori commessi o ai danni procurati dai sistemi di IA³⁵⁰.”

Successivamente, Floridi suggerisce di determinare quali compiti e funzioni decisionali delegare all'IA, in modo che si mantenga un dialogo continuo tra governo, industria e società civile. È inoltre necessario verificare se le norme attuali siano basate su un'etica solida in grado di tenere il passo con l'evoluzione tecnologica. Per quanto concerne lo sviluppo, occorre migliorare la spiegabilità dei sistemi di IA che prendono decisioni socialmente rilevanti, in modo da garantire agli individui una spiegazione fattuale e diretta. È altresì fondamentale sviluppare procedure giuridiche e infrastrutture digitali nel sistema giudiziario per verificare le decisioni algoritmiche, includendo misure per proteggere le informazioni sensibili³⁵¹. Riguardo alla responsabilità, Floridi suggerisce l'adozione di un principio antico: *cuius commoda eius et incommoda*, ovvero “chi trae un vantaggio da una situazione deve anche sopportarne gli inconvenienti³⁵².”

La sesta raccomandazione propone di sviluppare meccanismi di *audit* per identificare conseguenze

³⁴⁹ Ibid.

³⁵⁰ Ibid., p. 288

³⁵¹ Ibid., p. 289

³⁵² Ibid., nota 2

indesiderate e, in settori ad alto rischio, adottare meccanismi di solidarietà fin dall'inizio. Il settimo punto insiste sulla creazione di un sistema di riparazione per errori o torti causati dall'IA, sottolineando che “il meccanismo richiederà necessariamente di adottare un criterio chiaro ed esaustivo di attribuzione della responsabilità a esseri umani e/o organizzazioni³⁵³.”

Il tema dell'affidabilità continua con l'ottava raccomandazione, che propone lo sviluppo di metriche condivise per misurare l'affidabilità di prodotti e servizi di IA. Tali metriche dovrebbero confluire in un indice di comparazione della fiducia per guidare gli utenti, stimolando la competizione e, potenzialmente, un sistema di certificazione per i prodotti meritevoli³⁵⁴. La nona raccomandazione prevede la creazione di un'agenzia di supervisione dell'UE, simile a quella esistente per i medicinali, incaricata di monitorare i sistemi di IA anche dopo la loro immissione sul mercato. L'osservatorio europeo per l'IA, previsto nella decima raccomandazione, avrebbe invece il compito di monitorare gli sviluppi e alimentare il dibattito pubblico. L'undicesima raccomandazione suggerisce di stabilire strumenti legali per la collaborazione uomo-macchina nell'ambiente di lavoro, al fine di promuovere un'innovazione inclusiva³⁵⁵.

Gli incentivi finanziari a livello europeo sono centrali per la dodicesima e tredicesima raccomandazione, volte a sostenere lo sviluppo di tecnologie che siano “socialmente preferibili (non semplicemente accettabili) e rispettose dell'ambiente³⁵⁶.” La ricerca scientifica deve essere orientata al progresso dell'IA per il bene comune, fungendo da contraltare unitario alle tendenze di mercato. La quattordicesima raccomandazione punta alla cooperazione interdisciplinare: “l'etica dovrebbe aiutare a cogliere le opportunità e ad affrontare le sfide, e non solo a descriverle³⁵⁷.”

³⁵³ Floridi 2022, op. cit., p. 290

³⁵⁴ Ibid., p. 291

³⁵⁵ Ibid., p. 292

³⁵⁶ Ibid.

³⁵⁷ Ibid., p. 293

Floridi promuove inoltre l'inclusione di considerazioni etiche e giuridiche nei progetti di ricerca, oltre all'utilizzo di zone deregolate chiamate laboratori viventi, per la verifica empirica³⁵⁸.

L'ultima sezione si concentra sul sostegno e sull'educazione. La diciassettesima raccomandazione incoraggia la ricerca sulla percezione pubblica dell'IA e l'attuazione di meccanismi strutturati di consultazione. La diciottesima propone di sostenere lo sviluppo di codici di condotta di autoregolamentazione, con la possibile adozione di una certificazione di IA etica per professionisti come avvocati o medici. La diciannovesima raccomandazione mira a responsabilizzare i consigli di amministrazione delle aziende tecnologiche attraverso comitati etici dotati di poteri di controllo³⁵⁹. Infine, la ventesima raccomandazione chiede di “sostenere la creazione di piani di studio e attività di sensibilizzazione pubblica sull'impatto sociale, giuridico ed etico dell'IA³⁶⁰.” Ciò include l'aggiornamento dei curricula scolastici, la formazione aziendale e l'impegno verso iniziative globali come le campagne dell'ONU per lo sviluppo sostenibile, concludendo così un piano d'azione mirato a rendere l'IA uno strumento intrinsecamente orientato al benessere umano³⁶¹.

Floridi conclude tutto questo discorso affermando che le venti raccomandazioni da lui elencate nel capitolo 11 del suo libro, “sono formulate con spirito di collaborazione e nell'interesse della creazione di risposte *concrete e costruttive* alle sfide sociali più urgenti sollevate dall'IA³⁶².”

³⁵⁸ Ibid.

³⁵⁹ Ibid., p. 294

³⁶⁰ Ibid., p. 295

³⁶¹ Floridi 2022, op. cit., p. 290

³⁶² Ibid., p. 296

Conclusione

L'intelligenza artificiale ci pone davanti a sfide, problematiche e riflessioni morali a cui non possiamo sottrarci.

Abbiamo iniziato cercando di dare una definizione di intelligenza artificiale e, sulla scia di Floridi, ritengo che l'intelligenza artificiale sia semplicemente una nuova forma dell'agire. L'IA non è in grado di pensare, di ragionare o elaborare criticamente una nuova idea, dal momento che queste sono prerogative tipiche e uniche dell'essere umano. Certo, come abbiamo visto, l'IA può addirittura comporre una poesia, elaborare testi, battere un essere umano in giochi come *Go* o gli scacchi, ma ciò non la rende intelligente. L'intelligenza artificiale è una macchina perfettamente in grado di *simulare* tutto quello che compie un essere umano. È radicalmente più veloce di noi in tutti i processi cosiddetti cognitivi, ma si trattano pur sempre di processi automatizzati che l'IA ha appreso mettendo insieme tutte le conoscenze che noi esseri umani abbiamo già acquisito. È per questo che sarebbe assurdo parlare di responsabilità dell'intelligenza artificiale o attribuire un valore morale a qualcosa che in definitiva ha lo stesso valore di una lavastoviglie o un frigorifero. Infine, pur con tutte le sofisticerie a cui l'IA potrà mai arrivare, mancherà sempre di un aspetto che soltanto noi umani possiamo avere: le emozioni, o meglio, un'intelligenza emotiva. Per dirla con le parole dello psicologo Daniel Goleman, “noi nutriamo sentimenti su tutto ciò che facciamo, pensiamo, immaginiamo e ricordiamo. Il pensiero e i sentimenti sono inestricabilmente intrecciati fra loro³⁶³.” D'altro canto, l'intelligenza artificiale non potrà mai mettere sentimenti in tutti i suoi cosiddetti “pensieri”. Certo, in futuro magari esisteranno robot umanoidi che grazie all'intelligenza artificiale potranno piangere, arrabbiarsi o esprimere felicità, ma si tratteranno sempre di *simulazioni apprese*, non di emozioni autentiche.

Per quanto riguarda tutte le varie sfide etiche e le problematiche analizzate nel resto della tesi, posso

affermare, con le parole di Floridi, di “stare nel mezzo, né l'inferno né il paradiso, ma il laborioso

³⁶³ Daniel Goleman, *Lavorare con intelligenza emotiva*, RCS Libri S.p.A., Milano 1998, p. 28

purgatorio degli sforzi umani³⁶⁴.” È sbagliato pensare di utilizzare l'intelligenza artificiale in maniera indiscriminata, glorificandone i benefici ed esaltandola a tal punto da ritenerla la panacea di tutti i problemi che affliggono l'umanità. Infatti, “se l'IA non è controllata e guidata in modo equo e sostenibile, può esacerbare i problemi sociali [...]; erodere l'autonomia e la responsabilità umana; amplificare i problemi del passato” e “trasformarsi da parte della soluzione a parte del problema³⁶⁵.”

D'altra parte, demonizzare l'IA alla stregua dei fautori della superintelligenza che ci dominerà tutti è decisamente fuori luogo. D'altronde, se usata bene, l'intelligenza artificiale può diventare “una straordinaria tecnologia” che “può aiutarci a conoscere, comprendere e prevedere di più e meglio le numerose sfide, [...] in particolare [...] l'ingiustizia sociale e la povertà globale. [...] Può accelerare il circolo virtuoso tra maggiori informazioni, migliore scienza e politiche più avvedute³⁶⁶.”

Insomma, per ripetere un esempio già citato, l'intelligenza artificiale è come un coltello: in sé non è niente di giusto o di sbagliato, dipende solo dall'uso che se ne fa.

In ogni caso, oltre le disquisizioni filosofiche o i progetti politici, quello che vediamo nella realtà dei fatti è un uso indiscriminato dell'intelligenza artificiale da parte del cittadino medio, che tende a vedere in questo strumento un modo per delegare sempre di più e rendere la vita più semplice. Certo, è vero, con l'intelligenza artificiale possiamo evitare mansioni noiose e dedicarci ad altre cose più interessanti. Ma il problema è che molte persone stanno letteralmente *smettendo di pensare* per lasciarlo fare all'IA, con conseguente deterioramento delle facoltà cognitive, specialmente nei più giovani, e una totale erosione del pensiero critico.

Quindi, ancora una volta, è l'IA il “male”? Assolutamente no. È, ripeto, *l'uso* che se ne fa. Occorrono quindi politiche volte a un'educazione del cittadino per quanto concerne l'utilizzo di questa nuova tecnologia, esponendo in maniera comprensibile tutti i rischi e i benefici che derivano da uso corretto o scorretto dell'intelligenza artificiale, affinché l'IA diventi uno strumento al servizio

³⁶⁴ Floridi 2022, op. cit., p. 17

³⁶⁵ Ibid., p. 333

³⁶⁶ Ibid., p. 332

dell'umanità e non un “feticcio” che la renda schiava.

In conclusione, le riflessioni etiche relativamente all'intelligenza artificiale dovrebbero diventare argomento di pubblico dominio, perché ormai l'IA “sta trasformando profondamente le nostre relazioni, il nostro modo di conoscere e comunicare, e in definitiva le nostre vite. Costruire sistemi etici è forse una sfida ardua, ma non per questo impossibile. E chissà che un giorno non siano proprio le entità artificiali ad aiutarci a migliorare noi stessi³⁶⁷.”

³⁶⁷ De Caro, Giovanola 2025, op. cit., p. 147

Bibliografia

- Acemoglu, D. e Restrepo, P. [2020], *Robots and jobs: Evidence from US labor market*, in «Journal of Political Economy», 128, pp. 2188-2244.
- Andler, D. [2023], *Il duplice enigma. Intelligenza artificiale e intelligenza umana*, Torino, Einaudi.
- Anthropic [2025], *System card: Claude Opus 4 & Claude Sonnet 4*, <https://www-cdn.anthropic.com/4263b94>.
- Aresu, A. [2024], *Geopolitica dell'intelligenza artificiale*, Milano, Feltrinelli.
- Benanti, P. e Maffettone, S. [2024], *Noi e la macchina*, Roma, LUISS University Press.
- Berinsky, A. [2023], *Political rumors. Why we accept misinformation and how to fight it*, Princeton, Princeton University Press.
- Block, N., *On a confusion about a function of consciousness*, in “Behavioral and Brain Sciences”, 18, 2, pp. 227-247.
- Boden, M. [2012], *Creativity and art: Three roads to surprise*, Oxford, Oxford University Press.
- Buolamwini, J. e Gebru, T. [2018], *Gender shades: Intersectional accuracy disparities in commercial gender classification*, in “Proceedings of Machine Learning Research”, 8, pp. 1-15.
- Carruthers, P. [2011], *The opacity of the mind. An integrative theory of self-knowledge*, Oxford, Oxford University Press.
- Casati, R. [2014]. *Contro il colonialismo digitale*, Roma-Bari, Laterza.
- Chalmers, D., *Could a large language model be conscious?*, in Boston Review, 9 Agosto 2023.
- Chiriatti, M. [2021], *Incoscienza artificiale. Come fanno le macchine a prevedere per noi*, Roma, LUISS University Press.
- Chomsky, N. [1975], *Reflections on language*, New York, Pantheon.
- Coeckelberg, M. [2022], *The political philosophy of AI: An introduction*, London, Polity.
- Cohen, S. [2025], *The Concept and Ethics of Manipulation*, Cambridge, CUP.

- Crawford, C. [2021], *Né intelligente né artificiale. Il lato oscuro dell'IA*, Bologna, Il Mulino.
- De Caro, M. e Giovanola, B. [2025], *Intelligenze. Etica e politica dell'IA*, Il Mulino, Bologna.
- Dreyfus, H. [1972], *What computers can't do: The limits of artificial intelligence*, New York, Harper & Row.
- Fabris, A. [2018], *Etica per le tecnologie dell'informazione e della comunicazione*, Roma, Carocci.
- Finocchiaro, G. [2024], *Intelligenza artificiale. Quali regole?*, Bologna, Il Mulino.
- Floridi, L. [2022], *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Raffaello Cortina Editore, Milano.
- Fonnesu, L. [2018], *Storia dell'etica contemporanea. Da Kant alla filosofia analitica*, Carocci, Roma.
- Fossa, F., Schiaffonati, V. e Tamburrini, G. [2021] (a cura di), *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*, Roma, Carocci.
- Galletti, M. e Zipoli Caiani, S. [2024], *Filosofia dell'intelligenza artificiale. Sfide etiche e teoriche*, Bologna, Il Mulino.
- Giovanola, B. e Tiribelli, S. [2022], *Etica e intelligenza artificiale per l'impresa. Valori, principi, purpose*, Milano, Roi Edizioni.
- Goleman, D. [1998], *Lavorare con intelligenza emotiva*, RCS Libri S.p.A., Milano.
- Keeling, G. [2018], *Autonomy, nudging and post-truth politics*, in "Journal of Medical Ethics", 44, pp. 721-722.
- Kurzweil, R. [2005], *The singularity is near: When humans transcend biology*, London, Penguin.
- Langrebe, J. e Smith, B. [2022], *Why machines will never rule the world. Artificial intelligence without fear*, New York-Oxon, Routledge.
- Larrey, P. [2019], *Artificial humanity. An essay on the philosophy of artificial intelligence*, Roma, IF Press.
- Magni, S. F. [2026], *Bioetica*, Carocci, Roma.
- McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon (2006), *A proposal for the Dartmouth summer research project on artificial intelligence*, august 31, 1955, in AI

magazine, 27 (4), p. 12.

Muldoon, J., Graham, M. e Cant, C. [2024], *Feeding the machine. The hidden human labour powering AI*, Edinburgh, Canongate Books.

Noble, S.U. [2018], *Algorithms of oppression: How search engines reinforce racism*, New York, New York University Press.

Panai, E. [2024], *L'etica dell'intelligenza artificiale spiegata a mio figlio*, Milano-Udine, Mimesis Edizioni.

Penrose, R., *La mente nuova dell'imperatore*, Milano, Rizzoli, 1992.

Perconti, P. e Plebe, A. [2025], *Turing's vision. How AI is shaping the world*, Boca Raton, CRS Press.

Porter, B. e Machery, E. [2024], *AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably*, in “Scientific Reports”, 14, 26133.

Preston, J. e Bishop, M. (a cura di) [2002], *Views into the Chinese room: New essays on Searle and artificial intelligence*, New York, Oxford University Press.

Putnam, H. [1994], *The best of all possible brains?*, in “New York Times”, section 7, 20 novembre.

Rossi, F. [2024], *Intelligenza artificiale*, Bari-Roma, Laterza.

Russel, B. [1952], *Is there a God?*, in The Collected Papers of Bertrand Russel, Vol. 11: Last Philosophical Testament, 1943-68. Routledge, London, pp. 542-548.

Russell, S. [2025], *Compatibile con l'uomo. Come impedire che l'IA controlli il mondo*, Torino, Einaudi.

Sadler, M. e Regan, N. [2019], *Game changer. Alphazero's groundbreaking chess strategies and the promise of AI*, Alkmaar, New in Chess.

Santoni de Sio, F. [2024], *Artificial intelligence and the future of work: Mapping the ethical issues*, in “Journal of Ethics”, 28, pp. 407-427.

Schneider, S. [2019], *Artificial you. AI and the future of your mind*, Princeton, Princeton University Press.

Searle, J. [1980], *Minds, brains, and programs*, in “Behavioral and Brain Sciences”, 3, 3, pp. 417-

- Sebo, J. [2022], *Saving animals, saving ourselves. Why animals matter for pandemic, climate change, and other catastrophes*, Oxford, Oxford University Press.
- Silver, D. et al. [2018], *A general reinforcement learning algorithm that masters chess, shogi, and go through self-play*, in “Science”, 362, 6419, pp. 1140-1144.
- Singer, P. [2011], *The expanding circle: Ethics, evolution, and moral progress*, Princeton, Princeton University Press.
- Spitzer, M. [2023], *Intelligenza artificiale*, Milano, Corbaccio.
- Suleyman, M. [2024], *L'onda che verrà. Intelligenza artificiale e potere nel XXI secolo (2023)*, con M. Bhaskar, Milano, Garzanti.
- Tiribelli, S. [2023a], *Moral freedom in the age of artificial intelligence*, Milano-Udine, Mimesi Edizioni.
- Tomasello, M. [2003], *Constructing a language. A usage-based theory of language acquisition*, Cambridge, MA, Harvard University.
- Tononi, G. et al. [2016], *Integrated information theory. From consciousness to its physical substrate*, in “Nature Reviews Neuroscience”, 17, pp. 450-461.
- Turing, A.M. [1950], *Computing machinery and intelligence*, in “Mind”, 59 (236), pp. 433-460.
- Turing, A. M. [1951], *Alan Turing's lost radio broadcast rerecorded* – On the 15 of May 1951 the BBC broadcasted a short lecture on the radio by the mathematician Alan Turing, BBC Radio, <https://www.youtube.com/watch?v=cMxbSsRntv4>.
- Turing, A.M. [2024], *Macchine calcolatrici e intelligenza*, Torino, Einaudi.
- Vélitz, C. [2024], *The ethics of privacy and surveillance*, Oxford, Oxford University Press.
- Vetere, G. [2025], *Intelligenze aliene. Linguaggio e vita degli automi*, Roma, Luca Sossella.
- Williams, S. [2023], *Storia dell'intelligenza artificiale. La battaglia per la conquista della scienza del XXI secolo*, Milano, Garzanti.
- Zuboff, S. [2023], *Il capitalismo della sorveglianza (2019)*, Roma, LUISS University Press.

Ringraziamenti

Ringrazio il professor Sergio Filippo Magni per i consigli, le linee guida, la rilettura e correzione dei vari capitoli di questo lavoro.

Ringrazio il professor Luca Fonnesu per avere accettato la richiesta di rivestire il ruolo di correlatore e per le revisioni finali dell'opera.

Ringrazio la segreteria studenti del dipartimento umanistico per l'assistenza riguardo le questioni burocratiche.

Ultimo, ma non meno importante, un ringraziamento va a mia moglie, alla mia famiglia e ai miei amici per il continuo supporto morale.