# Il diritto digitale
## Temi di informatica giuridica

a cura di
Monica Palmirani, Giovanni Sartor
Federico Galli, Salvatore Sapienza

Bologna
University Press

# IL DIRITTO DIGITALE
## Temi di informatica giuridica

a cura di
Monica Palmirani, Giovanni Sartor
Federico Galli, Salvatore Sapienza

Bologna
University Press

Title: LEGAL DESIGN AND DATA SCIENCE FOR EXPLICABLE AI IN LEGAL DOMAIN
Acronym: LEDS 4 XAIL
n. GPA: 101085576
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Education and Culture Executive Agency (EACEA). Neither the European Union nor the granting authority can be held responsible for them.



Fondazione Bologna University Press
Via Saragozza 10, 40123 Bologna
tel. (+39) 051 232 882

www.buponline.com
e-mail: info@buponline.com



ISBN 979-12-5477-771-8
ISBN on line 979-12-5477-772-5
DOI 10.30682/9791254777725

Questo volume è stato realizzato a partire da un impaginato camera-ready in formato pdf fornito dai curatori

In copertina: Summit Art Creations/Shutterstook.com

Prima edizione: marzo 2026

# INDICE

# ETICA DELL'INTELLIGENZA ARTIFICIALE

*Giorgio Bongiovanni, Claudio Novelli*

SOMMARIO: 1. Dagli androidi alle chatbot: intelligenza e agency. 2. Questioni di etica dell'IA e diritti umani. 3. Etica del digitale e dell'IA. 4. Dall'Etica alla Regolamentazione dell'IA. 5. Visioni critiche

## 1. Dagli androidi alle chatbot: intelligenza e agency

La riflessione sulle questioni etiche poste dall'Intelligenza Artificiale (IA d'ora in avanti) è quasi contemporanea alla nascita di questo ambito di ricerca. Va segnalata, tuttavia, una differenza significativa tra le analisi del primo periodo e quelle più recenti. Nel primo caso, si tratta di riflessioni che fanno riferimento a quella che è stata chiamata la *Good Old-Fashioned AI* (GOFAI) o *Intelligenza simbolica*, mentre le seconde hanno quale oggetto i sistemi basati sul *Machine Learning* (ML) e l'utilizzo dei *Large Language Models* (LLMs). Il primo tipo di IA adotta «una visione top-down dell'intelligenza» basata sulla «capacità degli esseri umani di comprendere il mondo attraverso rappresentazioni simboliche[concettuali]», il secondo utilizza, invece, «un approccio statistico [...] per trovare modelli in un mondo disordinato»[1]. L'etica legata al primo tipo di IA (old-fashioned, simbolica, cognitiva) è quella probabilmente più nota perché è stata al centro di romanzi di science fiction e di film famosi: basti pensare ai romanzi di Philip K. Dick e Isaac Asimov e ai film di d Stanley Kubrick (*2001: Odissea nello spazio* del 1968, basato su un racconto di Arthur C. Clarke) e Ridley Scott (*Blade Runner* del 1982)[2]. Tutte queste opere sono centrate su automi/androidi o macchine che, fornite di una intelligenza vicina a quella umana (cioè dotata di «comprensione,

[1] N. GOLTZ, T. DOWDESWELL (eds.), *Real World AI Ethics for Data Scientists. Practical Case Studies*, Boca Raton: CRC Press, 2023, pp. 2-3, che notano che l'approccio statistico è nato come «una risposta ai primi fallimenti dell'IA simbolica»; L. FLORIDI, *Etica delll'intelligenza artificiale*, Milano: R. Cortina, 2023, ha sintetizzato la distinzione tra i due approcci di IA come differenza tra IA *cognitiva* e IA *ingegneristica*. Per una chiara ricostruzione del passaggio dal primo al secondo approccio, N. CRISTOFOLINI, *La scorciatoia*, Bologna: Il Mulino, 2023.

[2] Si può ricordare di P.K. DICK, *Do Androids Dream of Electric Sheep?* (Il cacciatore di androidi) del 1968, da cui è tratto *Blade Runner*, e di I. ASIMOV, *Io, Robot* (I Robot) del 1950, che è una raccolta di racconti nei quali vengono sviluppate le tre leggi della robotica.

consapevolezza, acume, sensibilità, preoccupazioni, sensazioni, intuizioni, semantica, esperienza [...] significato, [...] saggezza», si ribellano agli esseri umani o cercano di imporre, anche al costo di vite umane, un dato piano. Questo tipo di IA non ha, tuttavia, alcun riscontro nello stato dell'arte tecnologico attuale, né sembra prevedibile un tale sviluppo nelle tendenze di breve-medio periodo[3].

L'attuale ricerca etica è invece centrata sul secondo tipo di IA (ingegneristica) che è un'intelligenza che cerca principalmente di replicare la capacità di «risoluzione efficace dei problemi» e dell'«esecuzione corretta dei compiti» e lo fa estraendo schemi «non ovvi e utili da grandi insiemi di dati» sulla base di una capacità di tipo statistico e sintattico che non necessita di una competenza semantica, cioè della comprensione del significato delle cose[4].

Questo passaggio ha portato a chiedersi se per il secondo tipo di IA sia effettivamente possibile parlare di intelligenza: ci sono posizioni che lo negano e altre che, anche se in forma moderata e che la distingue da quella umana, lo affermano. Da un lato, si è sostenuto che l'IA sia solo «una nuova forma dell'agire, che può affrontare con successo compiti e problemi, in vista di un obiettivo, senza alcun bisogno di essere intelligente», mentre, dall'altro, che si tratta comunque di «una forma di intelligenza, anche se non ne ha tutte le qualità», e che se una macchina «risolve un problema per cui non abbiamo alcuna teoria [...] è appropriato parlare di comportamento autonomo e "diretto a uno scopo", e anche di apprendimento e intelligenza»[5].

In ogni caso, la ricerca etica cerca di individuare i problemi che potrebbero nascere dall'IA, e le possibili soluzioni, in considerazione sia della struttura sia delle possibilità che tale intelligenza offre («capacità di percepire e classificare, formulare previsioni e proporre soluzioni, prendere decisioni e funzionare di conseguenza, eventualmente godendo di una piena autonomia operativa per alcuni compiti»)[6]. L'etica dell'IA è infatti un'*etica applicata*, cioè quella parte dell'etica che affronta pratiche e comportamenti specifici nel mondo reale[7].

---

[3] Si possono consultare dati sullo stato dell'arte tecnologico e sugli investimenti in IA su: https://ourworldindata.org/artificial-intelligence.

[4] N. GOLTZ, T. DOWDESWELL (eds.), *Real World AI Ethics for Data Scientists. Practical Case Studies*, cit., pp. 2-3; L. FLORIDI, *Etica dell'intelligenza artificiale*, cit., sottolinea il «successo sbalorditivo» della seconda e la «triste delusione» generata dalle ricerche della prima fase.

[5] L. FLORIDI, *Etica dell'intelligenza artificiale*, cit., sostiene inoltre che le nuove macchine sono «stupide come un vecchio frigorifero»; D. ANDLER, *Il duplice enigma. Intelligenza artificiale e intelligenza umana*, Torino: Einaudi, 2024; D. CRISTOFOLINI, *La scorciatoia*, cit. Una discussione importante su questa questione è quella che, negli anni '80, si è originata a partire dall'esperimento mentale della stanza cinese di J. Searle.

[6] F. FOSSA, V. SCHIAFFONATI, G. TAMBURRINI (a cura di), *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*, Roma: Carocci, 2021.

[7] L'etica applicata più nota è la bioetica, ma vi sono molti altri ambiti quali, ad. es., le etiche delle professioni.

In questo contributo, analizzeremo nel secondo paragrafo le questioni etiche (o l'intensificazione di quelle già presenti) poste dall'IA, anche alla luce del rapporto con i diritti umani. Nel terzo paragrafo, vedremo in che modo la riflessione etica (quale, ad. es. l'approccio chiamato *principlism*) li può affrontare. Nel quarto paragrafo, mostreremo come la riflessione etica ha ispirato la regolamentazione sulla IA, con particolare attenzione al contesto Europeo. Il paragrafo 5 conclude l'articolo illustrando alcune posizioni critiche nei confronti dell'etica dell'IA.

## 2. Questioni di etica dell'IA e diritti umani

In una rassegna del 2020 del *The Berkman Klein Center for Internet & Society* di Harvard[8] vengono esaminati 36 documenti (sia di organizzazioni private sia di enti pubblici o internazionali) e individuati i temi e gli ambiti in cui è necessario che l'IA trovi dei principi regolatori. Allo stesso modo, ma dalla prospettiva dei diritti umani riconosciuti a livello internazionale, si muove il documento dell'ONU sui rischi che l'IA potrebbe comportare[9]. Nel primo caso, gli aspetti messi in evidenza sono: privacy, responsabilità (*Accountability*), sicurezza (*Safety and Security*) trasparenza ed esplicabilità (*Explainability*), equità e non-discriminazione, controllo umano della tecnologia; nel secondo caso, il riferimento è a: libertà da danni fisici e psicologici, uguaglianza di fronte alla legge e protezione contro le discriminazioni, diritto alla privacy, garanzia della proprietà, libertà di pensiero, religione, coscienza e opinione, libertà di espressione e di accesso alle informazioni, diritto di partecipazione alla vita politica, diritto al lavoro e a una vita dignitosa, diritti del bambino, diritti alla cultura, all'arte e alla scienza.

Questi aspetti sono tra quelli più importanti, ma non si può dire che racchiudano tutte le criticità legate all'IA: ciò dipende, tra l'altro, dal fatto che è «la tecnologia [che] spinge (*drives*) l'etica e [che] molte questioni che ora fanno parte dell'"etica digitale" sono antecedenti alla tecnologia digitale»[10]. Una complessiva

---

[8] J. FJELD ET AL., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Center for Internet & Society, 2020, http://nrs.harvard.edu/urn3:HULInstRepos:42160420

[9] UNHUMANRIGTHS, 2023, *Taxonomy of Human Rights Risks Connected to Generative AI*, https://www.ohchr.org/sites/default/files/documents/issues/business/btech/taxonomyGenAI-Human-Rights-Harms.pdf

[10] V.C. MÜLLER, *The History of Digital Ethics*, in C. VÉLIZ (ed.), *The Oxford Handbook of Digital Ethics*, Oxford: Oxford University Press, 2023. Ciò naturalmente non significa dire che «l'innovazione digitale è alla guida, e tutto il resto rimane indietro», ma semplicemente che lo sviluppo della tecnologia può porre nuovi problemi. Sui rischi di una visione acritica della tecnologia e del suo ruolo, L. FLORIDI, *Etica dell'intelligenza artificiale*, cit.

classificazione degli aspetti critici dell'AI è comunque molto difficile[11]: è tuttavia possibile individuare, tenendo conto della distinzione tra i sistemi di IA come *oggetti*, cioè strumenti realizzati e utilizzati dall'uomo e sistemi di IA come *soggetti*, cioè gli stessi sistemi di IA, nove punti principali («main ethical issues»)[12]. Si tratta di: a) Privacy e profilazione, b) Manipolazione del comportamento, c) Opacità e Spiegabilità, d) Pregiudizi nei sistemi decisionali, e) Interazione uomo-robot, f) Automazione e occupazione, g) Machine Ethics, h) Sistemi autonomi e Responsabilità, i) Singolarità. A questi punti, va poi aggiunto quello che riguarda l'analisi del (l) Peso ambientale dell'IA. Li analizzeremo sinteticamente.

a) *Privacy e profilazione*. In questo ambito, l'IA aggiunge alcuni aspetti critici a un rischio già presente e regolato. Si tratta in particolare, in relazione alla dimensione negativa del diritto, della possibilità che l'IA offre di una più ampia raccolta e analisi dei nostri dati digitali (ottenuti via internet, social media, banche dati): ciò può avvenire, ad esempio, con l'opportunità, permessa dall'IA, di un migliore riconoscimento facciale (da telecamere di sorveglianza e fotografie) e dunque dell'acquisizione di dati non solo digitali, ma biometrici. Inoltre, l'abilità di analisi dei dati da parte dei sistemi di IA facilita e ottimizza la profilazione degli aspetti, interessi, preferenze di vita (politiche, culturali, alimentari, ecc.) dei diversi soggetti e, in questo modo, aumenta (o forse crea) la capacità di renderli un prodotto commerciale: ciò può rafforzare l'idea che quello che sembra essere «l'obiettivo principale dei social media, dei giochi e della maggior parte di Internet», sia «ottenere, mantenere e dirigere l'attenzione – e quindi la fornitura di dati». Questo, a sua volta, ha consentito di dire che «la sorveglianza è il modello di business di Internet»[13].

b) *Manipolazione del comportamento.* Anche la possibile manipolazione non è un fenomeno nuovo, è anzi antica, ma può acquisire nuove modalità attraverso i sistemi di intelligenza artificiale[14]. La manipolazione minaccia l'autonomia dei soggetti attraverso il controllo del «contenuto e la fornitura di informazioni in

---

[11] Sulle diverse classificazioni proposte, si veda S. BUIJSMAN, M. KLENK, J. VAN DEN HOVEN, *Toward a "Design for Values" Approach*, in N.A. SMUHA (ed.), *The Cambridge Handbook of the Law, Ethics and Policy of Articial Intelligence*, Cambridge: Cambridge University Press, 2025, p. 63.

[12] Questa proposta è formulata da V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, in E.N. ZALTA e U. NODELMAN (eds.), *The Stanford Encyclopedia of Philosophy*, 2023, https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/, che sostiene che su questi aspetti «there seems to be reasonable agreement».

[13] *Ibid*., che cita B. SCHNEIER, *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*, New York: W.W. Norton, 2015, e nota che «questa economia della sorveglianza e dell'attenzione viene talvolta definita "capitalismo della sorveglianza"». Il rifermento è S. ZUBOFF, *Il capitalismo della sorveglianza. Il futuro dell'umanità nell'era dei nuovi poteri*, Roma: LUISS, 2023, e alla tesi che questo tipo di capitalismo si basa sulla previsione dei comportamenti.

[14] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit.

modo tale che la persona non sia consapevole della manipolazione»[15]. In questo contesto, l'IA può potenziare la manipolazione, ad esempio, migliorando l'accuratezza delle tecniche di *nudge* o quelle di inganno. Il *nudge* (spinta gentile) interviene sull'«architettura della scelta» delle persone allo scopo di modificare il loro comportamento senza che vi siano proibizioni o incentivi[16] (ad esempio, mettendo in primo piano, in un supermercato, i prodotti che si vuole favorire). Il *nudge* sfrutta i nostri *bias* cognitivi[17] e può essere considerato una forma di paternalismo (leggero) volto a indurre scelte positive per la vita pubblica. Può tuttavia essere considerato manipolatorio in quanto «in conflitto con due componenti critiche dell'autonomia personale: il consenso informato e il libero arbitrio».

In ambito digitale, il *nudge*, grazie alla possibilità di creare spinte più precise e personalizzate, opera con strumenti volti a incentivare la presenza online quali il «caricamento automatico di nuovi contenuti», la «riproduzione automatica di contenuti musicali o video», i «premi collegati all'interazione con la piattaforma»[18]. Il secondo aspetto (inganno), connesso tra l'altro alla propaganda politica, è legato alle «faking technologies» dell'IA che possono creare, «con qualsiasi contenuto», falsi testi, foto, video e trasformare «ciò che una volta era una prova affidabile in una prova inaffidabile»[19].

c) *Opacità e Spiegabilità*. L'IA basata sul machine learning è in grado di prendere decisioni (ad esempio, sull'assegnazione di un credito, sulla scelta di un candidato per una posizione lavorativa) e di fare analisi previsionali (ad esempio, sulla libertà vigilata o la probabilità di contrarre una determinata malattia). La precisione e la rapidità con cui alcune di queste funzioni vengono svolte rende necessario garantire un livello sufficiente di trasparenza per spiegarne le ragioni (e inferenze) retrostanti.

Tuttavia, come abbiamo accennato, decisioni e previsioni avvengono estraendo «schemi (*pattern*) da un determinato set di dati» e selezionando quelli che «vengono etichettati in un modo che appare utile per la decisione»: ciò non solo limita «le opportunità di partecipazione umana», ma spesso rende impossibile «sapere come il sistema è giunto a questo output» sia per «per la persona in-

---

[15] J. WILLIAMS, *Ethical Dimensions of Persuasive Technology*, in C. VÉLIZ (ed.), *The Oxford Handbook of Digital Ethics*, cit.

[16] Si veda R. THALER, C.R. SUNSTEIN, *Nudge. La spinta gentile*, Milano: Feltrinelli, 2009.

[17] Su questi bias e la loro derivazione dai nostri meccanismi cognitivi (sistema 1 e sistema 2), si veda M. GALLETTI, *Quando l'Intelligenza Artificiale incontra le scienze comportamentali. Una riflessione sui valori morali*, in M. GALLETTI, S. ZIPOLI CAIANI (a cura di), *Filosofia dell'Intelligenza Artificiale. Sfide etiche e teoriche*, Bologna: Il Mulino, 2024.

[18] M. IENCA, E. VAYENAIN, *Digital Nudging. Exploring the Ethical Boundaries*, in C. VÉLIZ (ed.), *The Oxford Handbook of Digital Ethics*, cit.

[19] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit.

teressata» sia «per l'esperto»[20]. I sistemi che presentano bassi livelli di trasparenza, e quindi di possibilità di fornire spiegazioni, sono spesso «chiamati *black box* (scatole nere) perché operano in un modo così complesso da rendere impossibile la piena comprensione di come vengono elaborate le informazioni»[21]: sono perciò «opachi» in quanto non si è in grado di sapere «come è stato identificato un particolare schema o persino qual è lo schema». Modelli che apprendono attraverso tecniche di deep learning ricadono in questo gruppo. Ciò ha condotto a prospettare una «minaccia dell'algocrazia»[22].

d) *Pregiudizi (bias) nei sistemi decisionali*. Questo fenomeno riguarda «i modi in cui gli algoritmi possono riprodurre o addirittura amplificare i pregiudizi del passato lavorando su dati storici che riflettono pregiudizi precedenti»[23]. Ciò è avvenuto diverse volte al punto che «per alcuni anni, gli algoritmi "razzisti" e "sessisti" hanno fatto notizia». Il caso più noto è probabilmente quello dell'algoritmo COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) del sistema giudiziario americano deputato a prevedere la recidiva di chi aveva commesso crimini: questo sistema «aveva tassi di errore differenziali per i neri e i bianchi»[24] che favorivano i secondi e discriminavano i primi. Si ha perciò un pregiudizio quando la decisione è «influenzata da una caratteristica che è in realtà irrilevante per la questione»[25]. Anche se i casi più frequenti riguardano appartenenza e genere, ce ne sono molti altri, quali, ad esempio, «insegnanti licenziati a causa di oscuri sistemi di valutazione algoritmica, famiglie che non hanno ricevuto l'assistenza (insurance) perché vivevano nel quartiere sbagliato e bambini a rischio che non sono stati allontanati da ambienti sociali pericolosi a causa di problemi nei sistemi di previsione del rischio»[26]. La distorsione di questi sistemi può essere determinata dalla qualità dei dati utilizzati per

[20] *Ibid*.

[21] M. DI MICHELE, *Intelligenza Artificiale. Etica, rischi e opportunità di una tecnologia rivoluzionaria*, Santarcangelo di Romagna: Diarkos, 2023.

[22] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit., che fa riferimento a J. DANAHER, *The Threat of Algocracy: Reality, Resistance and Accommodation*, in *Philosophy & Technology*, 2016 (29/3).

[23] P. BODDINGTON, *AI Ethics. A Textbook*, Berlin, Singapore: Springer, 2023, p. 50.

[24] L. HERZOG, *Algorithmic Bias and Access to Opportunities*, in C. Véliz (ed.), *The Oxford Handbook of Digital Ethics*, cit., che indica anche casi di discriminazione verso le donne di algoritmi di Google, Amazon, Apple.

[25] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit.

[26] L. HERZOG, *Algorithmic Bias and Access to Opportunities*, cit., che sottolinea che «l'uso di tali algoritmi sta diventando sempre più diffuso» e che essi «svolgono già un ruolo in molte decisioni potenzialmente in grado di cambiare la vita, come il credito, l'alloggio, l'ammissione al college o la libertà vigilata». Un caso emblematico è lo scandalo danese degli ingenti rimborsi chiesti dal Fisco ad alcuni beneficiari di assegni familiari perché fraudolenti: l'algoritmo che aveva individuati le presunte frodi era basato su pregiudizi (sociali e razziali). Si veda B. TEN SELDAM, A. BRENNINKMEIJER, *The Dutch benefits scandal: A cautionary tale for algorithmic Enforcement*, in *EU Law Enforcement*, 2021 (April 30).

il machine learning (che possono contenere indicatori sensibili o essere scarsamente rappresentativi) o dai *pattern* ritenuti rilevanti per la decisione[27]. Ciò potrebbe anche derivare da uno «statistical bias» che si avrà, dato che un «set di dati [è] imparziale solo per un singolo tipo di problema», se tale set viene «utilizzato per un diverso tipo di problema [...] e quindi risultare distorto»[28].

e) *Interazione uomo-robot*. Il riferimento è alle «dinamiche di percezione di entrambe le parti»[29] che può essere inteso sia in senso psicologico (attaccamento, repulsione, ecc.)[30], sia in chiave etica. Da quest'ultimo punto di vista, un possibile approccio è quello basato sulla distinzione tra agenti morali e pazienti morali. La questione, in questo caso, è «come queste macchine dovrebbero comportarsi nei confronti degli esseri umani e, d'altro canto, come gli esseri umani dovrebbero comportarsi nei confronti di queste macchine». Di fronte alla questione se sia corretto dare calci a un cane robot, possiamo chiederci se quest'ultimo sia un paziente morale (cioè, schematicamente, che prova sensazioni) verso cui un agente morale umano compie atti crudeli o sia anche un agente morale (implicito o esplicito). Le possibili alternative sono quattro: i robot sono agenti e pazienti, solo agenti, solo pazienti, né agenti né pazienti[31] (v. *infra* punto h per le diverse categorie di agenti).

f) *Automazione e occupazione*. IA e robotica potrebbero, a causa dell'aumento dell'automazione (e di quello correlativo della produttività), trasformare il mercato del lavoro, rendendo di minor valore alcune capacità e obsoleti alcuni mestieri. Sebbene questa dinamica ricorra con molte innovazioni tecnologiche, rispetto a quelle del passato vi è una significativa differenza: mentre «l'automazione classica ha sostituito la forza umana», quella «digitale sostituisce il pensiero umano o l'elaborazione delle informazioni e, a differenza delle macchine, è molto economica da duplicare». Ciò potrebbe condurre a una ««««polarizzazione del lavoro"» nella quale i lavori richiesti sarebbero quelli «tecnici altamente qualificati e con alti stipendi» e quelli «di servizio poco qualificati e con bassi salari», mentre quelli in pericolo sarebbero «i lavori di media qualificazione nelle fabbriche e negli uffici, ovvero la maggior parte dei lavori [...] perché sono relativa-

---

[27] L. FLORIDI, *Etica dell'intelligenza artificiale*, cit., parla di limiti epistemici (connessioni basate su prove inconcludenti, imperscrutabili e fuorvianti) e di apofenia, cioè «vedere schemi ricorrenti (patterns) dove in realtà non ne esistono».

[28] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit.

[29] *Ibid*.

[30] L. CRACCO, *Machine ethics: stato dell'arte e osservazioni critiche*, in *Filosofia*, 2020 (65), https://-doi.org/10.13135 /2704-8195/5234, mostra che «in uno studio del 2005, è stato utilizzato un robot per interagire con bambini con disturbi dello spettro autistico».

[31] S. NYHOLM, *The Ethics of Human–Robot Interaction and Traditional Moral Theories*, in C. VÉLIZ (ed.), *The Oxford Handbook of Digital Ethics*, cit.

mente prevedibili e molto probabilmente destinati a essere automatizzati»[32]. Questa transizione e i costi sociali che ne deriveranno pongono un importante problema di giustizia distributiva.

A ciò va aggiunto che alcuni lavori sembrerebbero maggiormente tutelati rispetto al pericolo di automazione per il valore intrinseco che viene socialmente riconosciuto all'input umano. Questo riconoscimento li rende meno facilmente sostituibili dalle capacità statistico-generative dell'IA. Così, ad esempio, modelli linguistici come ChatGPT, Claude, o Deepseek, sembrano in grado di sostituire un certo tipo di funzioni lavorative (scrivere un programma informatico), ma non altre (fornire assistenza psico-terapeutica).

g) *Machine Ethics*. Questo tema fa riferimento all'«idea di programmare in qualche modo l'etica in una macchina»[33], cioè «dare alle macchine principi etici o una procedura volti al fine di individuare la soluzione dei dilemmi etici che potrebbero trovarsi di fronte, consentendo loro di funzionare e decidere in modo autonomo ed eticamente responsabile»[34]. Ciò significa considerare le macchine, «sistemi dell'IA e della robotica», come soggetti dotati «di qualche competenza morale» a cui è possibile attribuire diritti e doveri[35]: per fare ciò «potrebbe non bastare generare algoritmi nei limiti dei principi etici; sarà necessario infondere i principi morali negli algoritmi stessi»[36]. La machine ethics e la possibilità «di fornire alle macchine la capacità di prendere decisioni etiche in maniera esplicita», è tuttora un programma di ricerca «e non ha ancora prodotto, a livello pratico, nulla che possa considerarsi minimamente all'altezza delle promesse fatte inizialmente».

La ricerca si è indirizzata sia su «cosa implichi un tale tentativo e se sia desiderabile intraprenderlo», sia sui modi nei quali è possibile dotare una macchina di principi etici. Le ragioni a favore sono, tra altre, legate al fatto che «i sistemi autonomi attualmente in via di sviluppo – come le auto a guida autonoma, i ro-

---

[32] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit.

[33] L. FLORIDI, *Etica dell'intelligenza artificiale*, cit., richiama la sperimentazione del MIT *Moral-Machine* (un sito nel quale venivano proposti dilemmi etici per le auto a guida autonoma e si chiedeva ai partecipanti di indacare quale fosse la scelta corretta) quale esempio di ricerca in questa direzione.

[34] M. ANDERSON, S.L. ANDERSON, *General Introduction*, in IDD. (eds.), *Machine Ethics*, Cambridge: Cambridge University Press 2011.

[35] F. FOSSA, V. SCHIAFFONATI, G. TAMBURRINI (a cura di), *Etica dei sistemi intelligenti e autonomi: una mappa per orientarsi*, cit.

[36] L. CRACCO, *Machine ethics: stato dell'arte e osservazioni critiche*, cit., nota che «la disciplina non si occupa di *software* o di sistemi di intelligenza artificiale che "agiscono" in maniera eticamente accettabile perché le loro funzioni interne implicano di *default* un comportamento che esternamente giudicheremmo morale, senza fare riferimento a valori o teorie etiche. L'obiettivo non è, in altri termini, quello di creare macchine eticamente allineate, ma quello di fornire alla macchina una capacità di formulare giudizi morali complessi in autonomia e in linea di principio "nuovi"».

bot per l'assistenza sanitaria o i droni da guerra – dovranno prendere decisioni morali», a quello per cui è meglio fare in modo che ciò avvenga sulla base di una «capacità di ragionamento etico» e, infine, al dato per cui «le macchine con una capacità di ragionamento etico potrebbero anche migliorare il comportamento etico degli umani»[37].

La possibilità di rendere le macchine moralmente autonome ha seguito tre strategie principali: «partendo dall'alto (*top-down*), dal basso (*bottom-up*) o adottando un ibrido tra le due strategie». Le prime due strategie non sono decisive e mostrano vari problemi: la prima, che prevede di creare regole di scelta morale a partire dalle principali teorie etiche – deontologismo, consequenzialismo, etica delle virtù, intuizionismo, non fornisce indicazioni univoche e ha difficoltà di implementazione; la seconda, basata sull'apprendimento a partire da casi concreti, potrebbe seguire esempi sbagliati o non distinguere tra azione e azione corretta. La terza strategia, che sembra quella «più promettente», è quella «dei modelli in cui si combina l'apprendimento dai casi con tecniche di correzione dall'*alto*»[38].

h) *Sistemi autonomi e responsabilità*. In generale, il problema è quello del rapporto tra autonomia, in particolare etica, dei sistemi intelligenti e dei robot e responsabilità, cioè a chi si deve attribuire la responsabilità di un evento negativo. Dato che «l'autonomia nella robotica è relativa e graduale, si dice che un sistema è autonomo rispetto al controllo umano fino a un certo punto»: è perciò necessario capire «chi ha il controllo e chi è responsabile». Tuttavia, non è semplice individuare quali possano essere effettivamente considerati agenti autonomi. Una classificazione è stata proposta da Moor[39] nel 2006 che ha identificato, considerando i computer ordinari, cinque categorie: *agenti normativi*, non etici (computer), che lo sono «nel senso limitato in cui possiamo valutare la loro prestazione in termini di quanto bene svolgono i compiti assegnati»; *agenti di impatto etico* che hanno un influsso (idealmente positivo) sul mondo[40]; *agenti etici impliciti* che sono quelli che dispongono di un «software che supporta implicitamente un comportamento etico», che permette alla macchina di agire «eticamente perché le sue funzioni interne promuovono implicitamente un comportamento etico, o

[37] *Ibid*., nota che anche se «l'impresa della *machine ethics* sembra avere dei motivi *prima facie* validi per essere perseguita», sono molti anche quelli negativi.

[38] *Ibid*.

[39] J.H. Moor, *The Nature, Importance, and Difficulty of Machine Ethics*, in M. Anderson, S.L. Anderson (eds.), *Machine Ethics*, cit.

[40] *Ibid*., indica quale esempio i «fantini robot» che hanno sostituito i ragazzi, liberandoli dalla schiavitù, per guidare i cammelli nelle corse in Qatar.

almeno evitano un comportamento non etico»;[41] *agenti etici espliciti* quelli che sarebbero «in grado di formulare giudizi etici plausibili e giustificarli» ed essere autonomi «in quanto in grado di gestire situazioni di vita reale» imprevedibili e che implicano dilemmi etici[42]; *agente etico completo* che è quello che «può formulare giudizi etici espliciti e può giustificarli in modo ragionevole» e può esserlo solo «un essere umano adulto».

Di fronte alla possibilità di avere macchine e robot (come le auto a guida autonoma e i robot militari) in grado di essere agenti etici espliciti e capaci di affrontare dilemmi morali come il *trolley problem* per le auto[43], il problema è «come dovrebbero essere distribuiti responsabilità e rischi nel complesso sistema in cui operano i veicoli». Si tratta di un problema che richiederà l'adattamento e la modifica dei nostri «attuali schemi concettuali» (etici, giuridici, tecnici) quale potrebbe essere l'attribuzione della personalità giuridica ai robot o l'individuazione di una responsabilità distribuita[44].

i) *Singolarità*. Si tratta della possibilità, prefigurata da Irving John Good nel 1965[45], che «l'intelligenza artificiale raggiunga un livello di intelligenza umano (IA generale)», che, a questo punto, «i sistemi di IA sviluppino autonomamente sistemi che, superando il livello di intelligenza umano, diventino superintelligenti» (o ultraintelligenti) e che, a loro volta, diano vita a un ulteriore e rapido sviluppo di sistemi ancora più intelligenti. Questo percorso e la svolta che include (essere fuori dal controllo umano) è la singolarità tecnologica il cui sviluppo sarebbe difficile da prevedere[46]. Questo «ciclo di auto-miglioramento ricorsivo, in cui ogni nuova iterazione crea una versione più intelligente di se stessa, culminando nella creazione di una superintelligenza», porta con sé, per alcuni studiosi, il timore che «i sistemi futuri, se non gestiti correttamente, [possano] rappre-

---

[41] *Ibid*., quale l'Autopilota di un aereo che contribuisce al fatto che l'aereo arrivi a destinazione in sicurezza e tempo.

[42] *Ibid*., p. 16, si chiede: queste macchine saranno in grado di «fare scelte etiche (do ethics) in un modo in cui, ad esempio, un computer può giocare a scacchi?».

[43] Questo dilemma etico, proposto da P. Foot, è nella versione originale (ve ne sono diverse) la scelta tra far morire 5 persone che saranno investite da un vagone fuori controllo o salvarle deviando il vagone su un altro binario uccidendone però una che si trova sulla deviazione.

[44] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit., che rinvia M. TADDEO, L. FLORIDI, *How AI Can Be a Force for Good*, in *Science*, 2018 (361), sostiene questo tipo di responsabilità alla luce del fatto che «gli effetti delle decisioni o delle azioni basate sull'intelligenza artificiale sono spesso il risultato di innumerevoli interazioni tra molti attori, tra cui designer, sviluppatori, utenti, software e hardware».

[45] I.J. GOOD, *Speculations Concerning the First Ultraintelligent Machine*, in F.L. ALT, M. RUBINOFF (eds.), *Advances in Computers 6*, New York, London: Academic Press, 1965.

[46] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit. K. VOLD, D.R. HARRIS, *How Does Artificial Intelligence Pose an Existential Risk?*, in C. VÉLIZ (ed.), *The Oxford Handbook of Digital Ethics*, cit., caratterizzano la superintelligenza come «qualsiasi intelletto che superi di gran lunga le prestazioni cognitive degli esseri umani in praticamente tutti i domini di interesse».

sentare una minaccia esistenziale» per il genere umano. Si tratterebbe di un percorso progressivo nel quale ci si troverebbe di fronte a diversi rischi. Il primo è quello della perdita di controllo: «l'opacità e l'imprevedibilità di questi sistemi» e il fatto di essere «strategicamente avvantaggiati (più intelligenti) li rendono potenzialmente pericolosi» in quanto, senza controllo, potrebbero «impazzire prima di quanto pensiamo e prima di avere la possibilità di intervenire». Il secondo è relativo ai possibili «danni per ragioni strumentali» nei quali la «considerazione per la vita umana» potrebbe essere superata dalla necessità di raggiungere un determinato obiettivo o da altre considerazioni[47]. Il terzo è quello del *disallineamento* dei valori, cioè che, nonostante l'IA sia programmata a fini benevoli «o che i suoi valori siano allineati in modo affidabile con i nostri», è possibile che li interpretino in modo distorto o contrario. Questo sia perché è molto difficile definire in modo chiaro i valori, sia in quanto sono suscettibili di diverse interpretazioni[48].

l) *Ambiente*. Uno dei temi che non è sempre al centro dell'attenzione è «se lo sviluppo dell'IA sia sostenibile dal punto di vista ambientale: come tutti i sistemi informatici, i sistemi di IA producono rifiuti molto difficili da riciclare e consumano grandi quantità di energia, soprattutto per l'addestramento dei sistemi di apprendimento automatico»[49]. ChatGP3, ad esempio, avrebbe prodotto 550 milioni di tonnellate di anidride carbonica in fase di addestramento e richiesto qualcosa come 3,5 milioni di litri d'acqua[50]. In generale, «secondo l'Agenzia internazionale dell'energia, i centri dati, l'intelligenza artificiale e le criptomonete sono responsabili del 2 per cento del consumo mondiale di elettricità, un dato che potrebbe raddoppiare entro il 2026». Allo stesso modo, secondo una stima riportata dal Financial Times «entro il 2027 la crescita dell'IA [potrebbe] produrre un aumento del prelievo idrico compreso tra 4,2 e 6,6 mi-

---

[47] K. VOLD, D.R. HARRIS, *How Does Artificial Intelligence Pose an Existential Risk?*, cit., rinviano a S. HAWKING, *Brief Answers to the Big Questions*, London: Hodder & Stoughton, 2018: «[I]l vero rischio con l'IA non è la cattiveria, ma la competenza [...] Probabilmente non sei un malvagio odiatore di formiche che calpesta le formiche per cattiveria, ma se sei responsabile di un progetto di energia verde idroelettrica e c'è un formicaio nella regione da allagare, tanto peggio per le formiche. Non mettiamo l'umanità nella posizione di quelle formiche».

[48] K. VOLD, D.R. HARRIS, *How Does Artificial Intelligence Pose an Existential Risk?*, cit., mettono in luce ulteriori problemi, quali: la possibilità che si scateni, senza la necessaria prudenza, una corsa all'IA da parte dei diversi gruppi in vista dei vantaggi strategici che potrebbe avere l'IA avanzata; la militarizzazione dell'IA in riferimento al suo uso da parte di attori malintenzionati, il rischio di armi letali completamente autonome e quello della presenza di un'IA malintenzionata.

[49] V.C. MÜLLER, *Ethics of Artificial Intelligence and Robotics*, cit.

[50] L. TECLEME, *Qual è l'impatto ambientale dell'intelligenza artificiale*, 2024, https://valori.it/impatto-ambientale-intelligenza-artificiale/

liardi di metri cubi all'anno, più o meno la metà di quanta ne consuma il Regno Unito»[51].

## 3. Etica del digitale e dell'IA

L'etica dell'IA è una ramificazione dell'etica del digitale, ossia «quel settore dell'etica che studia e valuta i problemi morali relativi a dati e informazioni», e che comprende tre settori di ricerca (etica dei dati, degli algoritmi e delle pratiche) tra loro strettamente intrecciati. Si tratta di un'indagine intrapresa per «formulare e supportare soluzioni moralmente buone, per esempio buone condotte o buoni valori» con l'obiettivo di rendere «la regolazione digitale e la governance digitale» non solo «sostenibil[i] [o accettabili], ma anche socialmente preferibil[i] (equ[e])»[52]

Come già indicato, l'etica dell'IA è un'etica *applicata* (applied ethics) che cerca di dare risposte alle questioni etiche o a specifici dilemmi morali che si presentano in determinati ambiti di attività o di ricerca. L'etica digitale, per lungo tempo, non si è trovata di fronte a veri e propri dilemmi morali come avviene, ad esempio, nella bioetica (eutanasia, gestazione per altri), ma piuttosto problemi generali (anche futuribili) e, di conseguenza, cerca di indicare linee di scelta volte a rendere l'IA, come visto, socialmente vantaggiosa. I compiti dell'etica dell'IA sono perciò quelli di individuare «quali sono i problemi etici rilevanti» e di formulare principi e metodi per giungere a soluzioni eticamente valide. In sostanza, l'etica digitale applicata ha cercato di individuare alcuni principi e un metodo per giungere alla soluzione (spesso da tradurre in termini giuridici) dei diversi problemi che l'IA pone (come abbiamo visto, violazioni della privacy, manipolazione del comportamento, opacità, discriminazioni, equità sociale, desiderabilità di macchine etiche, rischi esistenziali).

Per capire come l'etica dell'IA individua e tratta queste situazioni, è necessario partire dalle riflessioni che vengono sviluppate negli altri livelli della ricerca etica. L'etica applicata è infatti solo una parte dell'indagine etica che comprende anche l'etica *normativa* e la *metaetica*[53]. Le prime due sono relative, in modi diversi, alle scelte dei soggetti (etica applicata e normativa), mentre la terza analizza le

---

[51] G. CRESCENTE, *Il peso dell'intelligenza artificiale sull'ambiente*, 2024, https://www.in ternazionale.it/notizie/gabriele-crescente/2024/03/22/intelligenzaartificialeambiente, nota che questa «crescita sta già mettendo in crisi le reti elettriche di alcuni paesi, come l'Irlanda, che dopo aver cercato per anni di attirare i giganti del settore dell'informatica, ha recentemente deciso di limitare le autorizzazioni per nuovi centri dati».

[52] L. FLORIDI, *Etica dell'intelligenza artificiale*, cit.

[53] V.C. MÜLLER, *The History of Digital Ethics*, cit., sottolinea che «i campi della filosofia pratica sono trattati in genere come il cugino povero e un po' imbarazzante che deve lavorare per vivere piuttosto che avere soldi in banca».

basi di queste scelte. La prima, come visto, riguarda scelte pratiche (ad es., «Il suicidio assistito è eticamente accettabile?», la machine ethics è auspicabile?), la seconda, le impostazioni generali di tali scelte (ad es., è necessario promuovere il miglior risultato, vanno seguiti i principi che sono universalizzabili, è necessario essere virtuosi per fare scelte corrette), la terza, i fondamenti di tali scelte (ad es., «quando pronunciamo giudizi etici come "uccidere innocenti è sbagliato", stiamo esprimendo le nostre convinzioni o stiamo facendo qualcos'altro?» Vi sono fatti etici?)[54].

In generale, le teorie etiche normative sono quelle che cercano di stabilire valori di riferimento per guidare giudizi e azioni, con ciò stabilendo cosa è giusto e cosa è sbagliato. Vi sono tre correnti principali: il *consequenzialismo*, che ha quale riferimento «i risultati delle azioni»; le *teorie deontologiche*, focalizzate «sulla natura delle azioni»; l'*etica delle virtù*, centrata sul «carattere morale dell'agente». Per il consequenzialismo, è moralmente giusto ciò che produce «i migliori risultati complessivi»[55], perciò «un atto o una regola vanno valutati in relazione agli effetti che producono, ovvero in base alla loro capacità di realizzare un qualche bene identificabile». La corrente più importante di questo approccio è l'*utilitarismo* che sostiene che «l'oggetto di valore da tenere in considerazione è l'utilità, che può essere intesa, a seconda delle diverse versioni utilitariste, come piacere, come felicità o come soddisfazione di preferenze» (welfarismo) e che «la valutazione delle conseguenze», cioè se una scelta conduce alla massimizzazione dell'utilità per il maggior numero di persone (benessere collettivo definito come somma delle utilità) deve essere «formulata da un osservatore imparziale in termini di utilità aggregata» (ordinamento somma)[56].

Le teorie deontologiche di derivazione kantiana cercano di determinare quali possono essere le «regole fisse e generali di azione morale» che, «formulate a un livello di elevata astrazione», hanno «un'applicabilità universale». Ciò può essere fatto, ad esempio, in base all'imperativo categorico kantiano espresso nei termini del principio di universalizzazione («agisci solo secondo quella massima che puoi allo stesso tempo volere che diventi una legge universale») o come rispetto universale per la dignità e l'autonomia morale delle persone («non trattare l'umanità mai semplicemente come un mezzo per raggiungere un fine, ma sempre allo stesso tempo come un fine»).

---

[54] Si veda, T. MCPHERSON, D. PLUNKETT, *Introduction. The Nature and Explanatory Ambitions of Metaethics*», in IDD. (eds.), *The Routledge Handbook of Metaethics*, Abington: Routledge, 2018, pp. 2 ss.

[55] P. BODDINGTON, *AI Ethics. A Textbook*, cit., pp. 231-232.

[56] B. CASALINI, L. CINI, *Giustizia, uguaglianza e differenza*, Firenze: Firenze University Press, 2012, p. 14.

L'etica della virtù ha le sue radici nella filosofia greca (in particolare Aristotele) e si basa su tre concetti ne derivano: arête (eccellenza o virtù), phronesis (saggezza pratica, prudenza) ed eudaimonia (fioritura umana o felicità). A differenza degli approcci consequenzialisti e deontologici, questo orientamento non valuta le azioni (in relazione alle conseguenze o ai doveri), ma si basa «sulle caratteristiche morali dei singoli agenti [...] (come l'onestà, l'umiltà, il coraggio, la compassione)» e l'utilizzo della saggezza pratica (esperienza di vita e capacità di capire i diversi aspetti di una situazione) nelle scelte etiche[57].

La *metaetica*, invece, studia la natura generale dell'etica ponendosi quattro domande basilari: a) cosa facciamo e a cosa ci riferiamo quando utilizziamo le espressioni etiche quali giusto o sbagliato (indagine semantica); b) qual è la natura, se esiste, dei fatti e proprietà morali a cui ci riferiamo, se sono cioè fatti o desideri (indagine metafisica); c) quali stati mentali abbiamo quando usiamo espressioni morali, se si tratta di credenze o attitudini conative (indagine psicologico-cognitiva); d) ammesso che esistano, come possiamo conoscere le entità morali a cui ci riferiamo (indagine epistemologica)[58].

Le etiche normative sembrano essere un punto di partenza importante per la soluzione delle questioni etiche legate all'IA. Un'etica consequenzialista, ad esempio, permette di valutare «l'introduzione di nuove tecnologie [...] in base ai benefici e ai danni complessivi» che può comportare. Questi approcci si prestano cioè all'utilizzo «di metodi computazionali di analisi quantitativa in cui [...] i potenziali danni e benefici di un particolare corso d'azione possono essere conteggiati [...], opportunamente adattati per scala e probabilità di impatto e ottimizzati per un risultato netto vantaggioso». Vi sono poi ampie applicazioni delle teorie deontologiche, ad esempio, nei diversi codici di etica dell'IA che prescrivono regole e doveri di applicazione generale, come rispettare la privacy degli individui, non generare danni, cercare di dar vita a sistemi equi, ecc. L'etica della virtù sembra poi fornire un utile strumento preparatorio, e di metodo, che indica di basarsi su «tratti specifici di eccellenza morale» nel proprio comporta-

[57] S. VALLOR, *Virtues in the Digital Age*, in C. VÉLIZ (ed.), *The Oxford Handbook of Digital Ethics*, cit.; R. HURSTHOUSE, G. PETTIGROVE, *Virtue Ethics*, in E.N. ZALTA e U. NODELMAN (eds.), *The Stanford Encyclopedia of Philosophy*, cit., 2023.

[58] L'analisi metaetica è fondamentale per questioni quali quelle dell'allineamento dei valori (supra punto 2i). Si possono distinguere due grandi gruppi di teorie metaetiche: il *cognitivismo* e il *non cognitivismo*. Per il primo, le cui due principali correnti sono il *naturalismo* e il *non naturalismo*, vi sono fatti etici e perciò valori che possiamo conoscere; per il secondo, la cui principale corrente è l'*espressivismo*, questi fatti non esistono, i giudizi etici esprimono desideri e i valori sono relativi.

mento digitale e di esercitare la prudenza pratica nelle «dinamiche morali degli ambienti digitali»[59].

Tuttavia, questi approcci di etica normativa hanno mostrato una serie di importanti problemi. Il consequenzialismo corre il «rischio di trascurare danni e benefici non quantificabili», di non tenere conto, in particolare nell'utilitarismo, delle «richieste di giustizia e dei diritti morali dell'individuo», di ignorare la difficoltà del calcolo delle conseguenze, specialmente nel campo digitale caratterizzato da «complessità, innovazione, e interdipendenza sistemica».

Le teorie deontologiche hanno il problema della difficoltà di applicazione dei principi generali individuati, quella di stabilire «un metodo rigoroso e affidabile per conciliare doveri contrastanti o per risolvere casi in cui esiste più di un modo per seguire tutte le regole o, molto più comunemente, nessun modo per seguirle tutte», quella della «relativa insensibilità al contesto morale e alla particolarità» dei diversi casi[60]. L'etica della virtù non riesce a produrre in qualche modo principi codificabili, a fornire un resoconto adeguato di quale deve essere l'azione giusta (si può compiere un'azione giusta senza essere virtuosi e una persona virtuosa può compiere l'azione sbagliata), vi può essere un conflitto tra diverse virtù, le diverse culture possono sostenere virtù diverse[61].

Una proposta ulteriore per sviluppare l'etica dell'IA è il *principialismo* (principlism). Si tratta di un approccio che è stato sviluppato nell'ambito della bioetica da Beauchamp e Childress[62] e che una serie di autori ha proposto di utilizzare per l'IA[63]. È una proposta che individua quattro principi etici (beneficenza, non maleficenza, autonomia, giustizia) che sono considerati la chiave per superare le difficoltà di applicazione delle etiche normative alle diverse questioni pratiche. I quattro principi vengono infatti visti come principi *intermedi* che «non sono concepiti come assiomi o principi supremi – come ad esempio l'imperativo kantiano, o la formula della massimizzazione della felicità nell'utilitarismo –, ma [...] come principi che intermediano tra quelli "supremi" e la pratica dei casi concre-

---

[59] P. BODDINGTON, *AI Ethics. A Textbook*, cit.; S. VALLOR, *Virtues in the Digital Age*, cit., che aggiunge che «nel contesto più ristretto dell'etica aziendale e organizzativa», vi sono «forme correlate di analisi costi-benefici (CBA) adattate alla "gestione del rischio" etica».

[60] S. VALLOR, *Virtues in the Digital Age*, cit., che sottolinea, per il deontologismo, il problema di «stabilire quali doveri sono più importanti da soddisfare», di cosa fare «quando il modo migliore per seguire una regola (ad esempio, rispettare la privacy degli utenti) sembra a tutti gli effetti incompatibile con un'altra (come garantire la trasparenza o promuovere l'equità)».

[61] R. HURSTHOUSE, G. PETTIGROVE, *Virtue Ethics*, cit.

[62] T.L. BEAUCHAMP, J.F. CHILDRESS, *Principles of Biomedical Ethics*, Oxford: Oxford University Press, 2019.

[63] L. FLORIDI ET AL., *AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, in *Minds and Machines,* 2018 (28).

ti»[64]. L'idea di fondo è «che gli aspetti moralmente rilevanti di molte situazioni del mondo reale non possono essere ridotti a un singolo, supremo principio morale, ma devono essere catturati da una pluralità di principi irriducibili». I quattro principi derivano (o fanno parte) dalla *common morality*, cioè sono «requisiti morali universali e fondamentali che tutte le persone imparziali e moralmente impegnate accettano». In questa impostazione, i principi, che per essere utilizzati devono essere specificati rispetto ai casi, sono *prima facie*, si applicano cioè a meno che non entrino in conflitto, in una determinata situazione, con qualcuno degli altri principi (ad esempio, in bioetica, autonomia vs. beneficenza). I principi sono «ragioni morali a favore o contro certi atti in ogni caso a cui si applicano»: la soluzione avviene attraverso un *bilanciamento* nel quale vengono valutati il peso e l'importanza dei diversi principi. La possibilità di soluzioni contrastanti e la necessità di giustificare il risultato richiede, per il principialismo, un metodo di valutazione specifico, l'equilibrio riflessivo che è un requisito di coerenza tra scelte, principi e intuizioni morali[65].

L'adozione del principialismo per l'etica dell'IA è motivata da tre aspetti preminenti: «l'enorme volume di principi proposti» nel campo dell'IA dai vari enti e strutture che «rischia di risultare soverchiante e fuorviante», «un elevato grado di sovrapposizione»[66] tra i principi proposti, il fatto che tra «le aree dell'etica applicata, la bioetica è quella che più assomiglia all'etica digitale nell'affrontare ecologicamente nuove forme di agenti, pazienti e ambienti»[67]. I quattro principi bioetici sembrano per questo adattarsi «bene alle nuove sfide etiche poste dall'IA»: ad essi, va affiancato, per le peculiarità dell'IA, un «ulteriore nuovo principio: l'esplicabilità, intesa come incorporazione sia di intelligibilità che di responsabilità».

Da questa prospettiva, l'etica dell'IA si baserebbe dunque su cinque principi, beneficenza, non maleficenza, autonomia, giustizia, esplicabilità. Li vediamo brevemente:

– *beneficenza*: fare del bene può essere sintetizzato nel rispetto della dignità umana e della sostenibilità. Per quest'ultimo aspetto, «la tecnologia dell'IA deve essere in linea con […] l'assicurare le precondizioni di base per la vita sul nostro

---

[64] G.N. PINTO, *I fondamenti epistemologici della bioetica*, in *L'Altro Diritto*, 2019 (3).

[65] T.L. BEAUCHAMP, O. RAUPRICH, *Principlism*, in H. TEN HAVE (ed.), *Encyclopedia of Global Bioethics*, Cham: Springer, 2016.

[66] L. FLORIDI, *Etica dell'intelligenza artificiale*, cit., che nota che «l'enorme volume di principi proposti [...] rischia di diventare soverchiante e fuorviante, sollevando due potenziali problemi. O i vari insiemi di principi etici per l'IA sono simili, portando a inutili ripetizioni e ridondanze, oppure, se differiscono in modo significativo, sono suscettibili di ingenerare confusione e ambiguità».

[67] L. FLORIDI ET AL., *AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, cit.

pianeta, la continua prosperità per l'umanità e la conservazione di un buon ambiente per promuovere il benessere delle persone e del pianeta»;

– *non maleficenza*: si pone «contro le varie conseguenze negative derivanti dall'uso eccessivo o improprio delle tecnologie» che comprende: «la prevenzione delle violazioni della privacy», «le minacce di una corsa agli armamenti» basati sull'IA, «l'automiglioramento ricorsivo dell'IA» e richiede di operare «all'interno di limiti sicuri»;

– *autonomia*: nel contesto dell'IA, questo principio «significa trovare un equilibrio tra il potere decisionale che ci riserviamo e quello che deleghiamo agli agenti artificiali» e implica la possibilità che l'autonomia di questi ultimi sia «resa intrinsecamente reversibile, qualora l'autonomia umana debba essere protetta o ristabilita»;

– *giustizia*: significa «promuovere la prosperità, preservare la solidarietà, evitare l'iniquità», eliminare **«**tutti i tipi di discriminazione» e il **«**rischio di distorsioni (bias) negli insiemi di dati utilizzati per addestrare i sistemi di IA»;

– *esplicabilità*: è un principio che «completa gli altri quattro: affinché l'IA sia benefica e non malefica, dobbiamo essere in grado di comprendere il bene o il danno che sta effettivamente facendo alla società e in quali modi; affinché l'IA promuova e non limiti l'autonomia umana, la nostra "decisione su chi dovrebbe decidere" deve essere informata dalla conoscenza di come l'IA agirebbe al nostro posto [...] affinché l'IA sia giusta, dobbiamo sapere chi ritenere eticamente o legalmente responsabile in caso di un esito grave e negativo»[68]. Persegue gli stessi obiettivi, quale corrente dell'etica applicata, l'approccio dell'etica umanistica (*humanistic ethics*). Questa impostazione ha quale punto di partenza, come il principialismo, le insufficienze e difficoltà dell'etica normativa e pone come obiettivo dell'innovazione tecnologica e dell'impatto dell'IA «lo sviluppo e la fioritura umana» (del singolo e di tutti) nella direzione di una «prosperità condivisa e del bene comune». Allo stesso modo, viene richiamata la bioetica come disciplina di riferimento e l'utilizzo di un metodo ex ante volto a prevedere i rischi dell'IA. I pricipi di base di questo aproccio sono, in estrema sintesi, la pluralità dei valori da considerare, l'importanza delle procedure di scelta, la partecipazione alla definizione del «benessere umano e della moralità»[69].

---

[68] L. FLORIDI, *Etica dell'intelligenza artificiale*, cit.

[69] R. FIORAVANTE, A.VACCARO, *Introduction*; IDD., *Humanism, Dehumanization, Integral Human Development: An Overview of Boundaries and Capabilities of Human-Centered Artificial Intelligence in Business and Organizations*; J. TASIOULAS, *Artificial Intelligence, Humanistic Ethics*, tutti in R. FIORAVANTE, A. VACCARO (eds.) *Humanism and Artificial Intelligence*, Cham: Springer, 2025.

## 4. Dall'Etica alla regolamentazione dell'IA

I principi etici e gli approcci sopra descritti hanno spesso influenzato le regolamentazioni sull'IA discusse e adottate negli ultimi anni. L'Unione Europea (UE) ha svolto un ruolo pionieristico in questo ambito, producendo numerosi documenti, soprattutto negli ultimi cinque anni. Tuttavia, la debolezza del settore tecnologico europeo ha suscitato scetticismi riguardo alla necessità di introdurre nuovi principi, norme e standard giuridici. Nonostante ciò, considereremo l'UE come un osservatorio privilegiato, in quanto rappresenta un interessante laboratorio di regolamentazione dell'IA.

### *4.1. Le linee guida etiche sull'AI (AI HLEG): un approccio principialista*

L'approccio europeo all'IA è incentrato sul rispetto della centralità dell'essere umano nello sviluppo e nell'impiego di questa tecnologia, un valore spesso definito come *human-centric A.I.*[70]. Il primo documento giuridico della strategia europea in materia è stato l'Ethics Guidelines for Trustworthy Artificial Intelligence, pubblicato nel 2019 dal gruppo di esperti indipendenti di alto livello. Queste linee guida, di natura non vincolante, stabiliscono che un'IA può essere considerata affidabile solo se basata su quattro principi fondamentali: rispetto dell'autonomia umana, prevenzione dei danni, giustizia (intesa come fairness) e spiegabilità.

Per tradurre questi principi in pratica, le linee guida individuano sette requisiti specifici che i sistemi di IA devono dimostrare di avere: 1) Intervento e sorveglianza umani; 2) Robustezza tecnica e sicurezza; 3) Riservatezza e governance dei dati; 4) Trasparenza (intesa come tracciabilità e spiegabilità dei processi); 5) Diversità, non discriminazione ed equità; 6) Benessere sociale e ambientale; 7) Responsabilità (intesa come verificabilità, segnalazione degli effetti negativi potenziali e loro minimizzazione). A supporto dell'attuazione di questi requisiti, le linee guida includono una lista di valutazione sotto forma di questionario, destinata agli operatori del settore, con l'obiettivo di rendere i criteri più operativi e applicabili nella pratica.

L'approccio etico delineato in queste linee guida si ispira al principialismo, come evidenziato dall'assenza di una gerarchia tra i sette principi fondamentali. Questa scelta implica che, in situazioni di conflitto o interferenza tra i principi – ad esempio, tra robustezza e benessere ambientale – sarà necessario operare un bilanciamento, valutando attentamente le specificità del contesto di applicazione.

---

[70] A citare questo principio è lo stesso portale dell'UE: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

Un ulteriore aspetto rilevante sottolineato nel documento è la reale portata di questi principi, che assumono piena pregnanza nei casi in cui siano direttamente o indirettamente coinvolti esseri umani. Al contrario, nelle applicazioni puramente industriali, la loro rilevanza potrebbe risultare attenuata. Tuttavia, benché questa sia una precisazione ragionevole, la distinzione tra ambiti di applicazione umani e industriali non è sempre netta e facilmente definibile.

### *4.2. L'AI Act: proporzionalismo ponderato*

La strategia europea assume contorni decisamente più concreti con l'Artificial Intelligence Act (AIA), il regolamento più dibattuto degli ultimi anni in questo ambito. A differenza delle linee guida precedenti, l'AIA introduce un quadro normativo articolato e giuridicamente vincolante, caratterizzato però da un'elevata complessità nell'implementazione. Il regolamento disciplina in modo estensivo l'intero ciclo di vita dell'IA – dalla progettazione allo sviluppo fino all'utilizzo – superando un approccio frammentario legato a singoli aspetti, come la privacy o la responsabilità civile.

Nei prossimi anni, chiunque voglia produrre o commercializzare tecnologie basate sull'IA all'interno dell'Unione Europea dovrà conformarsi agli standard stabiliti dall'AIA e dai relativi atti implementativi. In questo senso, il regolamento può essere considerato una sorta di «Costituzione europea sull'IA», destinata ad avere un impatto cruciale sulla strategia europea. La sua efficacia sarà determinante per rendere l'UE un mercato attrattivo per l'IA e per esercitare un'influenza normativa a livello globale, secondo il cosiddetto Brussels effect, ovvero la capacità dell'UE di imporre i propri standard su scala internazionale.

Alcuni segnali preliminari di questo effetto sono riscontrabili in alcune proposte legislative avanzate negli Stati Uniti a livello dei singoli stati, sebbene l'approccio federale americano, come vedremo, si differenzi sensibilmente da quello europeo. È sicuramente possibile che l'impatto dell'AIA sulla legislazione globale sia molto più ridimensionato di quello che ci si aspetta, ma per il momento sembra aver orientato il dibattito su alcune questioni nodali (e.g., la valutazione del rischio).

L'AIA introduce un quadro normativo che classifica i sistemi di IA in quattro categorie di rischio: inaccettabile, alto, limitato e minimo. In base a questa classificazione, il legislatore adotta un approccio proporzionato, distribuendo gli oneri normativi in modo da prevenire o mitigare i potenziali impatti negativi dell'IA. I sistemi ritenuti a rischio inaccettabile, ovvero quelli che minacciano i valori fondamentali europei e i diritti fondamentali – come nel caso dei sistemi di social scoring – sono vietati del tutto. Per le categorie di rischio inferiore, invece, vengono introdotti obblighi di conformità progressivi, che impongono

oneri crescenti agli sviluppatori e, in misura ancora maggiore, agli utilizzatori dei sistemi di IA, con l'obiettivo di ridurre i rischi e prevenire eventuali danni.

I sistemi ad alto rischio sono soggetti agli obblighi più stringenti, che includono: a) il rispetto di rigorosi requisiti sulla qualità dei dati di addestramento, b) la rendicontazione trasparente delle informazioni rilevanti, c) l'obbligo per i *deployers* (chi utilizza il sistema) di effettuare una valutazione d'impatto sui diritti fondamentali.

Tuttavia, questa ultima disposizione ha sollevato diverse perplessità. Da un lato, emergono dubbi metodologici riguardanti l'effettiva coerenza e armonizzazione delle valutazioni tra i diversi attori coinvolti[71]. Dall'altro, si rileva una questione più profonda di natura istituzionale: affidare il bilanciamento tra diritti fondamentali, una funzione tradizionalmente riservata agli organi giurisdizionali, a soggetti privati potrebbe compromettere la coerenza e l'imparzialità del processo.

Queste criticità sollevano interrogativi sull'efficacia e sull'equità del quadro normativo proposto, rendendo necessario un attento monitoraggio della sua implementazione pratica.

La classificazione dei sistemi di IA in categorie di rischio è basata su criteri computazionali e, soprattutto, sugli ambiti di utilizzo che vengono delineati in termini piuttosto generici, includendo settori come l'educazione, la giustizia, i processi democratici e i servizi essenziali. Se un sistema di IA viene sviluppato o utilizzato in questi ambiti, viene automaticamente classificato come ad alto rischio, indipendentemente dalle specificità del caso d'uso.

Tuttavia, per affrontare contesti più complessi e variegati, come quelli legati all'IA ad uso generale, inclusa l'IA generativa, l'AIA introduce la nozione di «significatività del rischio», un criterio che impone di valutare in modo più dettagliato l'impatto potenziale della tecnologia. In questo contesto, si considerano fattori come i soggetti esposti e la magnitudine del rischio, permettendo un'analisi più contestualizzata e adattabile.

L'approccio basato sul rischio dell'AIA è uno dei principali punti di forza della strategia regolatoria dell'UE. Questo approccio consente, infatti, di: stabilire un ordine di priorità valoriale e di obiettivi, definendo chiaramente quali ambiti richiedano maggiore attenzione e intervento; responsabilizzare i policymaker, fornendo un quadro strutturato per bilanciare innovazione e protezione dei diritti fondamentali; valutare i costi sociali associati all'impiego dell'IA, assicurando un bilanciamento tra benefici e potenziali impatti negativi. Le normati-

[71] C. NOVELLI, G. GOVERNATORI, A. ROTOLO, *Automating Business Process Compliance for the EU AI Act*, in G. SILENO ET AL. (eds.), *Legal Knowledge and Information Systems*, Amsterdam: IOS Press, 2023.

ve basate sul rischio offrono, inoltre, strumenti utili per affrontare l'incertezza tecnologica, fornendo risposte sia qualitative che quantitative attraverso metodologie consolidate, come l'analisi costi-benefici. Questo approccio fornisce una base razionale per la gestione delle complessità legate all'adozione dell'IA, garantendo flessibilità e adattabilità alle diverse evoluzioni tecnologiche e sociali.

Per tutte queste caratteristiche, l'AIA sembra ispirarsi a un approccio etico di matrice consequenzialista, in quanto pone al centro dell'attenzione le conseguenze – positive o negative – derivanti dall'uso dell'IA, con particolare riguardo alla possibilità di ledere diritti fondamentali. Tuttavia, l'AIA non può essere considerato interamente consequenzialista o utilitaristico, poiché riconosce un valore intrinseco a determinati principi morali, introducendo così un elemento deontologico e, nel bilanciamento, principialista. Questo equilibrio tra conseguenze e principi diversi richiama una prospettiva proporzionalista, in cui la deviazione da determinati principi è giustificabile solo in presenza di circostanze che lo rendano strettamente necessario[72].

Secondo questa impostazione, per giustificare moralmente un'azione che comporta potenziali danni, è indispensabile dimostrare – ed è qui che emerge l'aspetto più consequenzialista dell'approccio – che i benefici risultanti siano quantitativamente e qualitativamente superiori o almeno proporzionati ai danni. L'AIA, in questo senso, cerca di bilanciare il principio di precauzione, volto a proteggere individui e gruppi vulnerabili, con interessi individuali (ad esempio, la tutela dei diritti dei consumatori) e collettivi (come la promozione dell'innovazione tecnologica). Un esempio tipico di questo bilanciamento è rappresentato dal trade-off tra sicurezza e privacy: se un sistema di IA migliora la sicurezza pubblica, è accettabile una riduzione della privacy solo se il beneficio ottenuto è proporzionato e giustificato rispetto al sacrificio imposto.

Uno dei limiti principali del proporzionalismo, e quindi dell'impianto etico dell'AIA, risiede nelle difficoltà legate alla quantificazione e commensurabilità dei diversi fattori coinvolti. Bilanciare elementi di natura etica, economica e sociale richiede criteri abbastanza chiari che spesso sono difficili da definire e applicare in modo uniforme, rendendo il processo vulnerabile a interpretazioni soggettive. Inoltre, l'assenza di parametri univoci rischia di generare discrepanze nelle valutazioni di conformità.

Nonostante queste criticità, il bilanciamento tra diritti e interessi è una pratica consolidata nel diritto, che legislatori e giudici esercitano abitualmente per ri-

---

[72] B. HOOSE, *Proportionalism: The American Debate and its European Roots*, Washington, DC: Georgetown University Press, 1987. L'utilizzo dei principi intermedi (beneficenza, non maleficenza ecc.) consente, nei bilanciamenti, di arrivare a compromessi nei casi in cui non vi sia accordo sui valori fondamentali in conflitto.

solvere conflitti tra principi di pari rango, come quelli di natura costituzionale[73]. Analogamente, un'applicazione e un'interpretazione dell'AIA ispirata a questa logica potrebbero favorire decisioni coerenti e trasparenti, contribuendo a una regolamentazione dell'IA che sia al tempo stesso efficace e rispettosa dei valori fondamentali dell'Unione Europea.

### *4.3. Il resto del mondo*

Come anticipato, la strategia di regolamentazione adottata dall'Unione Europea è attualmente la più articolata e complessa. Tuttavia, diversi paesi stanno progressivamente cercando di introdurre proprie normative e standard per disciplinare la commercializzazione dell'IA. Alcuni di essi, come il Regno Unito e il Canada, sembrano allinearsi all'approccio europeo basato sui principi, enfatizzando aspetti etici e valoriali nella gestione dell'IA. Altri, invece, adottano strategie differenti, come nel caso degli Stati Uniti, il principale attore globale nello sviluppo di questa tecnologia[74], la cui impostazione normativa merita un approfondimento specifico.

Negli Stati Uniti, a livello federale, l'intervento più significativo finora è stato l'Executive Order (EO) emanato dall'amministrazione Biden, che delineava un primo quadro di principi guida per la regolamentazione dell'IA. Tuttavia, il provvedimento, privo di carattere vincolante per gli Stati federati, è stato revocato poche ore dopo l'insediamento dell'amministrazione Trump, evidenziando la fragilità di un approccio privo di una solida base legislativa[75].

A livello statale, invece, si osservano sviluppi più concreti, con la California che emerge come il principale laboratorio normativo per l'IA negli Stati Uniti. Lo Stato ha introdotto diverse iniziative legislative, focalizzandosi su temi come la trasparenza degli algoritmi, la protezione dei dati e la responsabilità d'uso, cercando di bilanciare lo sviluppo tecnologico con la tutela dei diritti dei cittadini.

In sintesi, mentre l'UE punta a una regolamentazione dettagliata e preventiva, volta a garantire un elevato livello di protezione dei diritti fondamentali, gli

---

[73] Alcune metodologie di bilanciamento sono presenti in R. ALEXY, *On Balancing and Subsumption. A Structural Comparison*, in *Ratio Juris*, 2003 (16), https://doi.org/ 10.1046/j.0952-1917.2003.00244.x, e in A. BARAK, *Proportionality: Constitutional Rights and Their Limitations*, Cambridge: Cambridge University Press, 2012. Viene esplorata l'applicazione di uno di questi due metodi in C. NOVELLI, F. CASOLARI, A. ROTOLO ET AL., *AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act,* in *Digital Society*, 2024 (3), https://doi.org/10.1007/s44206-024-00095-1.

[74] L'Artificial Intelligence Index Report della Stanford University segnala che negli USA sono stati sviluppati 51 dei modelli di IA più rilevanti contro i 21 dell'UE e i 15 della Cina. Fonte: https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI AI-Index-Report-2024.pdf.

[75] C. NOVELLI, A. GUR, L. FLORIDI, *Two Futures of AI Regulation under the Trump Administration*, 2025, http://dx.doi.org/10.2139/ssrn.5198926

Stati Uniti, pur con approcci differenti a livello federale e statale, privilegiano un modello più flessibile e adattivo, volto a favorire l'innovazione e la competitività nel mercato globale dell'IA. Talvolta, come con l'amministrazione Trump, questa flessibilità è esasperata e diventa una vera e propria svolta verso la deregolamentazione.

Più nel dettaglio, l'EO americano rappresenta l'emblema opposto dell'approccio europeo. È, infatti, un esempio di soft law che promuove l'autoregolamentazione da parte dell'industria dell'IA senza imporre limiti stringenti o standard vincolanti. A differenza dell' AIA europeo, che è esplicitamente basato su principi etici e diritti fondamentali, l'EO adotta un approccio market-driven, orientato principalmente alla concorrenza e all'innovazione[76]. Questa impostazione è evidente anche nella totale assenza di sanzioni per il mancato rispetto delle linee guida, lasciando così ampio margine di discrezione agli operatori del settore.

Di conseguenza, l'approccio regolatorio delineato dall'EO non può essere ricondotto a quello dell'"AI Ethics", ossia, come visto, un quadro normativo che promuove valori etici fondamentali. Al contrario, si inserisce nella prospettiva della "AI Safety", che si concentra esclusivamente sulla prevenzione di malfunzionamenti tecnici e danni non intenzionali, senza definire principi etici specifici da proteggere o promuovere. In questo senso, l'EO non si preoccupa delle potenziali discriminazioni o degli impatti sociali più ampi dell'IA, ma piuttosto della sua affidabilità tecnica e della sicurezza operativa.

Un'unica eccezione a questa impostazione emerge nel settore dell'occupazione, in cui l'EO riconosce la necessità di proteggere i lavoratori dai rischi connessi alla sorveglianza tramite IA e dalla possibile disoccupazione derivante dall'automazione. Tuttavia, anche in questo ambito, il documento si limita a formulare raccomandazioni generali senza introdurre misure concrete di tutela.

Sebbene l'EO sembri adottare un approccio privo di un'esplicita presa di posizione etica, si potrebbe obiettare che, implicitamente, sottoscriva una visione morale specifica. Questa visione risulta essere un'ibridazione tra consequenzialismo e deontologia. Dal consequenzialismo, eredita l'attenzione alla minimizzazione delle conseguenze negative e l'applicazione del principio di precauzione, mirato a evitare danni futuri. Tuttavia, il fatto che la sicurezza e la libertà di innovare e auto-regolamentarsi siano considerate valori imprescindibili, non

[76] Per una sintesi dei più rilevanti elementi di differenza tra la regolamentazione europea basata sui principi e quella americana orientata al mercato, si veda M. WÖRSDÖRFER, *Biden's Executive Order on AI and the E.U.'s AI Act: A Comparative Computer-Ethical Analysis*, in *Philosophy & Technology*, 2024 (37).

negoziabili, rivela la presenza di una componente deontologica, che conferisce a questi principi uno status quasi assoluto.

In definitiva, l'EO riflette un approccio apparentemente pragmatico e privo di riferimenti etici espliciti, ma che, in realtà, incorpora una visione morale sottostante, fondata sull'equilibrio tra precauzione e libertà di mercato.

Su questa scia, nel luglio 2025, l'amministrazione Trump ha anche inaugurato un Action Plan[77] con l'intento di promuovere le infrastrutture collegate all'IA (ad esempio, i data centers), incoraggiare la competività dell'industria americana, anche sulla base processi di certificazione velocizzata e finanziamenti federali. Tuttavia, il supporto federale verrà garantito soltanto a quelle aziende che non mostrano pregiudizi ideologici. Il riferimeno negativo (ufficioso) è a quelle aziende che hanno aderito a policy che la società americana conservatrice etichetta come «woke» e che spesso si traducono in un approcio etico alla regolamentazione dell'IA (AI Ethics), opposto ad un approccio incentrato sulla mera sicurezza da malfunzionamento (AI Safety).

Un atteggiamento regolatorio che similmente privilegia l'innovazione è emerso anche in Giappone con l' "Act on Promotion of Research and Development and Utilization of Artificial Intelligence-Related Technologies" (2025). Il regolamento ricorre a norme di soft law, stabilendo obiettivi e line guide strategiche generali e insistendo sull'esigenza che le aziende cooperino tra di loro e con il governo centrale, ma senza prescrivere sanzioni o requisiti obbligatori per la progettazione dei sistemi di IA.

L'approccio della Cina alla regolamentazione dell'IA, sebbene meno articolato rispetto ad altri modelli, presenta alcune somiglianze con la strategia dell'Unione Europea, piuttosto che con quella adottata finora dagli Stati Uniti. Sebbene sia la Cina che gli Stati Uniti abbiano dimostrato una certa flessibilità e apertura verso il business[78], anche in Cina come in UE la regolamentazione è in gran parte esplicitamente valoriale, sebbene i principi di riferimento siano radicalmente diversi. Mentre l'UE fonda il proprio impianto normativo sui valori liberali e sullo stato di diritto, la strategia cinese intende tutelare anzitutto i valori centrali del Partito Comunista Cinese, ai quali le tecnologie di IA devono rigorosamente conformarsi.

Un aspetto distintivo della regolamentazione cinese è l'obbligo per i modelli di IA generativa di rispettare precise direttive ideologiche, evitando la genera-

[77] Trump White House, *White House Unveils America's AI Action Plan*, 2025, https://www. whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-act ion-plan/.

[78] A. Huyue Zhang, *The Promise and Perils of China's Regulation of Artificial Intelligence*, in *Columbia Law School*, April 2024, https://www.law.columbia.edu/sites/default/files/202404/The%20Promise%20and%20Perils%20of%20Chinese%20AI%20Regulationpdf

zione di contenuti considerati politicamente sensibili o in contrasto con la narrazione ufficiale. Questo vincolo evidenzia un controllo centralizzato sull'IA, volto a preservare la stabilità politica e sociale secondo le priorità governative.

Un'altra differenza significativa rispetto all'approccio europeo riguarda il processo di autorizzazione. In Cina, qualsiasi modello di IA destinato alla commercializzazione o all'uso pubblico deve ottenere una approvazione preventiva da parte delle autorità governative, in particolare dalla Cyberspace Administration of China (CAC), secondo quanto stabilito dalle "Misure ad Interim per la gestione dei servizi di IA generativa". Questo sistema si discosta dal modello europeo, che prevede valutazioni e certificazioni da parte di agenzie indipendenti, garantendo così una maggiore separazione tra regolatore e soggetti regolati.

## 5. Visioni critiche

La discussione critica sull'etica dell'IA può essere ricondotta a due visioni principali: una di tipo costruttivo e una di tipo distruttivo.

Nel primo gruppo, rientrano quelle critiche che denunciano l'apparato concettuale e l'imprecisione terminologica dell'etica dell'IA, senza per questo rigettare in maniera globale l'area di ricerca, ma anzi cercando di suggerire dei miglioramenti. La terminologia finora utilizzata sarebbe in grado di evocare intuizioni ingannevoli, capaci di spostare la sensibilità etica e l'attenzione politica verso false credenze o di sopravvalutare profili di secondaria importanza. Questo tipicamente accadrebbe quando si ricorre a metafore o altre forme di linguaggio non letterale. Ad esempio, espressioni come «intelligenza (dell'IA)» o «"allucinazioni" (dei modelli linguistici)» cercherebbero di antropomorfizzare il fenomeno[79]. Bisognerebbe evitare, secondo i critici, di accostare i due campi semantici, come appunto l'intelligenza degli esseri umani a quella forma di ragionamento epressa dai sistemi di IA, che svolgono operazioni di mera associazione statistica.

Talvolta, le critiche costruttive suggeriscono di ridimensionare le aspettative (e le paure) che alimentano i dibatti etici sull'IA, invitando a trattare questa tecnologia in maniera più neutrale, antropomorfizzandola meno e guardandola più come una infrastruttura (alla pari dell'elettricità). In questo modo, la parte costruttiva di queste critiche suggerisce di spostare l'attenzione della riflessione etico-giuridica verso le risorse materiali e umane che permettono alla tecnologia

[79] L. FLORIDI, A.C. NOBRE, *Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences*, in *Minds & Machines*, 2024 (34), https://doi.org/10.1007/s11023-024-09670-4.

di esistere e funzionare. Infatti, sarebbero sistematicamente sottovalutate una serie di questioni – e relazioni di potere – relative a come vengono raccolti i dati su cui l'IA viene addestrata, all'impatto ambientale della produzione e conservazione dei dati e allo sfruttamento di lavoratori coinvolti nel processo di produzione (e anche successivamente). Ciò sarebbe dovuto, in parte, all'erroneo apparato concettuale che supporta l'analisi etica in questo ambito.

Un'interessante critica costruttiva è anche quella che stigmatizza l'eccessivo ricorso all'IA per la risoluzione di problemi, siano essi di carattere tecnico o sociale. Il bersaglio polemico di questa critica può essere definito come "tecno-ottimismo" o "tecno-soluzionismo"[80]. In particolare, ciò che viene avvertito come dannoso per la riflessione etica è lo scarso spirito critico che dovrebbe guidare la scelta strategica di ricorrere all'IA. L'assenza di tale spirito critico porterebbe, in ultima istanza, a sottovalutare la componente ideologico-politica della tecnologia. L'IA diventerebbe una sorta di panacea anche per situazioni che potrebbero essere affrontate senza di essa, e con meno complicazioni. Pertanto, chi sviluppa questa critica sostiene che l'etica dell'IA debba interessarsi in primo luogo (e maggiormente) dell'effettiva necessità di utilizzare questa tecnologia, valutandone il vantaggio in comparazione con soluzioni alternative (che possono contemplare altre tecnologie digitali). In alcuni casi, l'utilizzo dell'IA potrebbe risultare inutile – o addirittura dannoso – perché i problemi che si intendono risolvere fanno parte di una dimensione che precede causalmente quella su cui interviene l'IA. Il problema della discriminazione nelle predizioni dell'IA è in questo senso emblematico. Esso rivela spesso che la discriminazione riguarda i dati di addestramento oppure il modo in cui i dati sono stati raccolti e processati. Magari sono dati inaffidabili o parziali. In situazioni come questa, impiegare acriticamente l'IA non solo risolve il problema, ma rischia di amplificarne la pericolosità.

Nel secondo gruppo rientrano invece le critiche che, contestando l'etica dell'IA su diversi fronti (soprattutto metodologici), ne mettono in dubbio l'intero impianto e lo statuto di riflessione morale indipendente. Spesso, queste critiche la bollano come semplice «ethics washing», un esercizio retorico asservito agli interessi di potenti attori privati, in primis le Big Tech. In realtà, l'ethics washing non è prerogativa esclusiva delle compagnie private, ma emrgerebbe anche dai documenti legislativi o di indirizzo di attori pubblici, come stati nazionali e sovranazionali (democratici e non). Secondo alcuni, come James Steinhoff, la subordinazione a interessi privati è frutto di un monopolio intellettuale – sia culturale sia economico – che le grandi aziende esercitano per centralizza-

[80] J.C. Heilinger, *The Ethics of AI Ethics. A Constructive Critique*, in *Philosophy & Technology,* 2022 (35), https://doi.org/10.1007/s13347-022-00557-9.

re l'innovazione[81]. Da tale prospettiva, l'etica dell'IA sarebbe parte integrante di un'agenda capitalistica, con la ricerca affidata in subappalto a organizzazioni e ricercatori (universitari o indipendenti) finanziati in modo diretto o indiretto dalle Big Tech. Ciò influenzerebbe non solo l'elaborazione di principi e linee guida – considerate semplici operazioni di marketing e pubbliche relazioni – ma anche i tentativi di tradurre questi stessi principi in standard pratici, i quali finirebbero sempre subordinati alla necessità di sviluppare rapidamente nuovi modelli e trarne profitto.

Di conseguenza, secondo questa critica «distruttiva», non ci sarebbe una via d'uscita per salvare l'etica dell'IA come vera disciplina e realizzare gli obiettivi che dichiara di perseguire. Finché rimane nei confini della logica capitalistica, non potrà mai costituirsi come riflessione etica indipendente. L'unica strada sarebbe una critica organica ai vincoli strutturali del capitalismo, che richiederebbe però un radicale cambio di passo: lasciare il classico linguaggio etico per una prospettiva più ampia di critica socio-economica.

[81] J. STEINHOFF, *AI Ethics as Subordinated Innovation Network*, in *AI & Society*, 2024 (39), https://doi.org/10.1007/s00146-023-01658-5