# Etica della regolazione della IA

LAURA PALAZZANI*

## 1. Gli sviluppi della IA e le domande etiche

L'uomo è essere «*naturalmente tecnico*», intendendo con "tecnica" la produzione e uso di strumenti per risolvere problemi (tecnologia). Da sempre costruiamo le macchine *per rendere più facili i compiti umani* (si pensi all'aratro; alla calcolatrice)[1]. Se dunque tutto ciò è naturale, e certamente non nuovo, perché ci poniamo oggi insistentemente domande sull'uomo e la tecnica/tecnologia con riferimento alla IA?

Va detto, preliminarmente, che la IA non è un fenomeno recente (l'espressione risale al 1956), ma certamente recente è la rapida evoluzione delle tecnologie di IA nell'ambito della c.d. "nuova ondata tecnologica", caratterizzata da un progresso 'dirompente', a causa della complessità, accelerazione, ampiezza delle applicazioni, pervasività e incertezza degli scenari che si stanno delineano. Il progresso della IA, detta 'strong' - dovuto principalmente all'aumento della potenza di calcolo, alla disponibilità di enormi quantità di dati e allo sviluppo di algoritmi che correlano i dati – dischiude domande di natura etica sui confini tra intelligenza artificiale e intelligenza umana. Potrà una macchina pensare come l'uomo e sostituire l'uomo? Quali i limiti della imitazione artificiale della intelligenza naturale? Ammesso che sia tecnologicamente possibile che la macchina divenga sempre più simile all'uomo e forse anche lo sostituisca, è eticamente lecito? Come si pone il diritto nella regolazione dei confini uomo/macchina?[2].

Nell'ambito della discussione etica caratterizzata dal pluralismo delle visioni emerge una diversità e contrapposizione tra tecno-centrismo e umano-centrismo, sia nei presupposti filosofici che nelle implicazioni concrete.

## 2. Le risposte del tecno-centrismo

Il *tecno-centrismo* è la visione di chi è favorevole ad ogni intervento di *tecnologicizzazione del corpo e della mente dell'uomo*, delineando orizzonti in cui l'artificiale diviene sempre più simile al

* Ordinario di filosofia del diritto alla Lumsa; componente dell'European Group on Ethics in Science and New Technologies, Commissione europea.

[1] P. Benanti, *Le stagioni dell'IA*, in C. Caporale, L. Palazzani (a cura di), *Intelligenza artificiale*: distingue frequenter. *Uno sguardo interdisciplinare*, Edizioni CNR, Roma 2023, p. 21 e seguenti.

[2] M. Coeckelbergh, *AI Ethics*, MIT, Cambridge (Mass.) 2020; M.A. Boden, *AI: Its Nature and Future*, Oxford University Press, Oxford 2016; L. Floridi, M. Durante, *L'etica della intelligenza artificiale. Sviluppi, opportunità, sfide*, Mimesis, Raffaello Cortina, Milano 2022; L. Palazzani, *Tecnologie dell'informazione e intelligenza artificiale: sfide etiche al diritto*, Studium, Roma 2020; A. Fabris, *Etica per le tecnologie dell'informazione e della comunicazione*, Carocci, Roma 2018.

naturale e tende a "con-fondersi", per annullare la differenza tra uomo e macchina. È la visione tecno-fila di chi spinge nella direzione di uno sviluppo 'convergente' delle tecnologie, che sostituisca e superi l'umano, quale unica strada per superare i limiti biologici del corpo e i limiti neurologico-cognitivi della mente verso una perfezione trans-umana, post-umana o anche 'super-umana'. È in questo contesto che viene elaborato il concetto di 'persona elettronica', nella duplice direzione dell'uomo che si artificializza e della macchina che si umanizza antropomorficamente[3], acquisendo uno statuto morale.

Questa visione presuppone filosoficamente una concezione riduzionistica, funzionalista e non cognitivista. La concezione *riduzionistica* parte dalla considerazione materialistica dell'uomo, il cui corpo si riduce a oggetto o somma di dati o di parti anche modificabili e sostituibili meccanicamente (si pensi al robot come 'embodiment' o 'incorporazione' artificiale), e la mente a componenti neurali modificabili e sostituibili informaticamente (si pensi alla IA, che può essere impiantata in un corpo meccanico). La visione riduzionistica è strettamente connessa alla concezione del *funzionalismo* che ritiene che l'esercizio delle funzioni è ciò che conta per identificare la persona, ossia il soggetto avente dignità, a prescindere dalla natura, considerata irrilevante: ciò che conta sono le funzioni espresse o esibite, essendo indifferente se siano di un uomo o di una macchina. È la prospettiva del *non cognitivismo* che ritiene che nella natura umana non sia conoscibile e ri-conoscibile un valore: è la svalutazione del corpo biologico, considerato un 'peso', che spinge verso il superamento del biologico con la macchina, in grado di esercitare la funzione razionale e cognitiva in modo artificiale senza e oltre il 'bios'.

L'alterazione tecnologica del corpo e della mente diviene un 'dovere evolutivo'[4] che consente la 'scorciatoia tecnologica' per il miglioramento dell'umano e della specie. Un 'imperativo tecnologico' (ossia tutto ciò che è tecnologicamente possibile 'si deve' fare), la cui omissione può essere considerata colpa e negligenza nei confronti dell'evoluzionismo tecnologico. In questo scenario si aprirebbe tecno-profeticamente un orizzonte futuro che dovrebbe portare ad una liberazione dell'uomo e ad un'alterazione radicale della natura dell'uomo, sino alla totale artificializzazione dell'umano.

È la visione del *trans-umanesimo* che promuove l'abbandono del biologico e la transizione verso l'artificiale al fine di espandere le capacità umane per realizzare "vite perfette". Il movente del trans-umanismo[5] è la desiderabilità sovra-umana che arriva alla cancellazione della condizione umana stessa, percepita come limite. L'obiettivo è espresso nella Massima Centrale del Trans-umanesimo (MCT): "è

[3] P. Benanti, *The cyborg. Corpo e corporeità nell'epoca del postumano*, Cittadella, Assisi 2012.
[4] J. Harris, *Enhancing evolution, The Ethical Case of Making Better People*, *Enhancing Evolution. The Ethical Case for Making Better People*, Princeton, 2007.
[5] World Transhumanist Association, *Transhumanism Declaration*, http//humanityplus.org/learn/philosophy/transhumanist-declaration; N. Bostrom, *Superintelligence*, Oxford University Press, Oxford 2014.

etico e desiderabile utilizzare mezzi tecno-scientifici per superare la condizione umana". In questa prospettiva il potenziamento dell'uomo si realizza nella 'condizione tecno-umana' [6] dell'uomo potenziato post-umano[7]. Ciò che conta è che la macchina "funzioni" il più perfettamente possibile. la sua centralità.

La visione umano-centrica si esprime in modi diversi con considerazione che spaziano, filosoficamente, dalla fenomenologia alla ontologia.

Già a livello fenomenologico, ossia al livello di come di manifesta la macchina rispetto all'uomo, è possibile mettere in evidenza i caratteri specifici dell'uomo e i limiti della macchina. Per quanti avanzamenti abbia compiuto la ricerca di una IA che riproduca la complessa organizzazione cerebrale e che possa imitare le caratteristiche dell'intelligenza umana, sembra lontana dalla portata degli sviluppi tecnici realisticamente prevedibili. Certamente le macchine, i software e i programmi informatici e algoritmici sono oggi in grado di compiere funzioni e operazioni considerevolmente complesse comparabili – o in alcuni ambiti, anche superiori – a quelle umane: si pensi in particolare all'archiviazione di informazioni in tempi velocissimi e quantità certamente non paragonabili alla memoria umana e la capacità di calcolo. Si pensi anche ai considerevoli obiettivi raggiunti dalla IA nel ragionamento logico, nell'analisi, nella risoluzione di problemi e nella interazione linguistica. Più complessa da realizzare e dispendiosa, sia sul piano finanziario che di consumo di energia, la mobilità del robot.

Eppure, esistono differenze qualitative o essenziali tra IA e intelligenza umana: vi sono dimensioni umane che non appartengono alle macchine.

***Le macchine non sono consapevoli: l'uomo è consapevole di sé.*** Le macchine non sono in grado di riferire la somma delle proprietà/ funzioni e la serie degli atti/operazioni che compiono ad una unità o identità che ne costituisce la sintesi e permane nel tempo alle modificazioni dei caratteri e funzioni. La IA non ha un'identità nel tempo che permane alla modificazione dei caratteri e funzioni. L'uomo ha per natura, la capacità di essere consapevole di sé e di riconoscere la sua identità, come sintesi delle caratteristiche (nello spazio) e degli atti (nel tempo).

[6] Cfr. N. Bostrom, *Welcome to a World of Exponential Change*, in P. Miller, J. Wilsdon (a cura di), *Better Humans? The Politics of Human Enhancement and Life Extension*, Demos, London 2006, pp. 40-50; J. Savulescu, H. Maslen, *Moral Enhancement and Artificial Intelligence: Moral AI?*, in J. Romporti, E. Zackova, J. Kelemen (a cura di), *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*, Springer International Publishing, Switxerland 2015, pp. 79-96.
[7] R. Pepperell, *The Posthuman Condition Consciousness. Beyond the Brain*, Intellect, Bristol-Portland (OR) 2003.

***Le macchine sono automatiche: l'uomo è autonomo e libero.*** Le macchine possono mutare e adattare il loro comportamento in modo anche imprevedibile (in particolare *machine learning*, che 'impara' nella interazione con l'ambiente esterno, in modo non sempre predicibile) in base alle condizioni e alle circostanze in cui operano. Il "fatto" che tecnologicamente nell'ambito della interazione uomo-macchina possa accadere la possibilità di mutamento o adattamento imprevedibile, non significa che la macchina sia autonoma. L'autonomia, nella sua accezione originaria, può essere attribuita solo agli esseri umani: autonomia significa capacità di un agente morale di agire come essere libero secondo i valori morali e la rappresentazione di un dovere, ossia scegliendo tra le opzioni possibili quella corrispondente ai valori morali. Il termine "autonomia" non può essere applicato ad artefatti, anche se sistemi cognitivi complessi molto avanzati, in quanto non è propriamente autonomia ma semmai automatismo, indipendenza o imprevedibilità. La terminologia dei sistemi "autonomi" è ampiamente utilizzata nella letteratura scientifica e nel dibattito pubblico (ad es. macchine a guida autonoma) per riferirsi al più alto grado di automazione in termini di operazioni e decisioni e rischia di introdurre fraintendimenti sul piano filosofico.

***Le macchine calcolano: l'uomo sente.*** l'IA non è in grado di riprodurre gli aspetti emotivi e affettivi oltre il calcolo e la logica della cognizione e della decisione. Il c.d. "errore di Cartesio"[8] mette in luce la complessità della soggettività umana, la cui intelligenza è incarnata in un corpo, e dunque situata biologicamente e neurologicamente, interagendo con la dimensione della esperienza reale vissuta del soffrire e gioire, desiderare e temere, vedere, udire, toccare, odorare e gustare. L'intelligenza umana è connessa ad emozioni (intelligenza emotiva), sentimenti, può essere empatica, caratteristiche precluse alle macchine senza un corpo, o con un corpo meccanico.

***Le macchine operano sintatticamente: l'uomo interpreta i significati.*** I calcolatori e i software che supportano l'IA operano sui legami sintattici tra simboli, ma non possono considerare la dimensione semantica, ossia interpretare i significati complessi. Se la sintassi è costituita da una serie di regole formali di composizione generali applicabili ad ampie classi di elementi linguistici senza riguardo al significato, la semantica non è solo la rete di relazioni estrinseca tra termini, ma è anche la comprensione del contenuto e la connessione di senso, che può cogliere solo la mente umana.

***Le macchine predicono i comportamenti; l'uomo orienta i comportamenti.*** La macchina correla, mediante algoritmi, dati (peraltro raccolti e selezionati dall'uomo), retrospettivamente: l'uomo è anche

[8] A.R. Damasio, *L'errore di Cartesio. Emozione, ragione e cervello umano,* Adelphi, Milano 1995.

in grado di ‘generare’ informazioni nuove, proiettivamente, con l’intuizione, l’immaginazione, l’invenzione, l’innovazione, la creatività. La macchina si limita alla correzione di informazioni e alla predizione statistica di probabilità sulla base di eventi o fatti accaduti nel passato, ma è l’uomo che può comprendere il senso delle relazioni ed è in grado di delineare scenari futuri anche innovativi e impredicibili.

***Le macchine agiscono ‘come se’ avessero uno scopo; l’uomo è libero e può darsi uno scopo.*** Una macchina compie operazioni cognitive per risolvere compiti assegnati dall’uomo, ma non è in grado di auto-assegnarsi compiti, di avere aspirazioni nella direzione della crescita della conoscenza. È l’uomo, in quanto libero, che è in grado di darsi uno scopo, di avere la motivazione a cercare la verità, di avere l’inclinazione e anche la curiosità a conoscere la realtà.

***Le macchine sono costruite e autoreferenziali: l’uomo ha una identità in relazione.*** La macchina è costruita dall’uomo che la ‘nutre’ e ‘addestra’ con esempi/istanze per poi operare in modo anche indipendente e automatico ma sempre autoreferenziale. L’uomo è costitutivamente in relazione, ha la possibilità di percepire se stesso come un io e di essere riconosciuto dagli altri (simili a lui) come un tu, nella interazione e nella reciprocità. L’uomo ha la capacità di comprendere la relazionalità come dimensione del suo essere e del suo dovere, nel rapporto morale con gli altri.

***Le macchine funzionano, l’uomo esiste***. La IA si colloca nel mondo del ‘fare’ (doing) o ‘funzionare’. È un sistema meccanico, prodotto dall’uomo, che “simula” e “replica” l’intelligenza umana, o più precisamente, alcuni aspetti dell’intelligenza umana: ‘svolge funzioni’, come ad esempio il calcolo, il linguaggio, le decisioni, la risoluzione di problemi, l’organizzazione. Le macchine possono anche ridurre gli errori cognitivi e raggiungere possibili risultati prestazionali superiori sul piano della cognizione soprattutto a livello quantitativo (raccolta e correlazioni di informazioni; calcolo; assistenza cognitiva), rispetto all'uomo. L’uomo appartiene al mondo dell’‘essere’ (being) ed ‘esistere’, l’esperienza del vivere nel mondo ‘resiste’ alla tecnologia e delinea la differenza sostanziale e qualitativa tra uomo e macchina.

Il piano fenomenologico mostra la differenza uomo/macchina ‘di fatto’. Ma si potrebbe sempre obiettare che l’avanzamento tecnologico potrebbe sostituire anche quelle capacità umane che oggi ci sembrano differenti e non sostituibili.

A questo livello è importante l'integrazione alla fenomenologia della *giustificazione ontologica della centralità dell'uomo* rispetto alle macchine[9]. La specificità umana, che lo rende unico e insostituibile 'di principio' dalle macchine, è connessa alla tematizzazione della natura umana: a partire dalla concezione finalistica della natura e cognitivista della conoscenza, è possibile conoscere e riconoscere nell'uomo le finalità che ne costituiscono il fondamento della sua dignità intrinseca. Il corpo dell'essere umano è una complessità organica, un organismo vivente che è di più della somma delle parti o funzioni; le parti o funzioni interagiscono in un tutto che ne costituisce l'incarnazione della soggettività, quale centro di coordinazione ed unificazione permanente delle proprietà e degli atti. La soggettività umana ha una dignità intrinseca a prescindere dal numero di proprietà o dall'esercizio delle funzioni, in quanto 'è' qualitativamente di più della somma delle parti o della serie dei comportamenti che manifestano. L'uomo deve rimanere al centro in quanto è l'unico essere in grado di agire moralmente rispetto agli altri esseri, di darsi una norma morale, di essere motivato ad agire in vista di un fine, mentre la macchina può al massimo essere uno strumento per il fine umano, indipendente o in grado di agire in modo imprevedibile, mediante gli algoritmi.

Nella misura in cui identifichiamo fenomenologicamente e ontologicamente la specificità umana, emergono le ragioni del porre dei *limiti* alla possibilità tecnologica di costruire macchine pensanti: un limite al dominio tecnologico e all'aspirazione di 'rifare' (*remake*) la natura, in quanto la tecnologia può anche dis-umanizzare e de-umanizzare[10]. L'umano va preservato, perché 'conta' anche nell'epoca delle macchine, contro ogni forma di 'anti-umanesimo'.

Affermare, oggi, l'umano-centrismo significa accogliere le tecnologie emergenti digitali ponendosi in modo critico, ossia valutandone opportunità e limiti, ricordando che la tecnologia non è un 'destino' inevitabile. L'IA è progettata e prodotta da un essere umano: siamo noi che dobbiamo decidere come produrre l'IA. Sappiamo oggi cosa è umano e cosa vogliamo conservare di ciò che è umano in una prospettiva incentrata sull'uomo nella costruzione e progettazione della tecnologia. La "macchina" può essere una alleata dell'uomo e l'integrazione tra naturale ed artificiale può divenire *complementare*, sempre nella priorità del rispetto dell'umano, salvaguardando ciò che 'conta' dell'umano, nel quadro dei diritti umani fondamentali.

---

[9] A. Pessina, *L'essere altrove. L'esperienza umana nell'epoca della IA*, Mimesis, Milano 2023; P. Benanti, *Le macchine sapienti. Intelligenze artificiali e decisioni umane*, Marietti, Genova 2018; Id., *Human in the loop. Decisioni umane e intelligenze artificiali*, Mondadori, Milano 2022; Id., *La condizione tecno-umana: domande di senso nell'era della tecnologia*, EDB, 2022; M. De Caro, S.F. Magni, M.S. Vaccarezza, *Le sfide dell'etica*, Mondadori, Milano 2022; L. Palazzani, *Cosa resta dell'umano nell'epoca dell'IA*, in C. Caporale, L. Palazzani (a cura di), *Intelligenza artificiale*: distingue frequenter. *Uno sguardo interdisciplinare*, cit., p. 77 e seguenti.

[10] D.M. Kaplan (a cura di), *Readings in the Philosophy of Technology*, Rowman &Littlefiled Publishers, Inc., Lanham (MD) 2004.

## 3. Per un umanesimo tecnologico: i requisiti etici per una regolazione della IA

A partire dall'analisi dei documenti sulla IA elaborati, sulla base della discussione interdisciplinare e pluralista, da commissioni di esperti a livello internazionale, europeo e nazionale, emerge una sostanziale convergenza nell'adesione alla visione umano-centrica e nella identificazione dei requisiti etici minimi per la regolamentazione della IA.

Tra i documenti si ricordano a livello europeo: The High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI,* 2019[11]; European Group on Ethics in Science and New Technologies, *Statement on Artificial Intelligence, Robotics and 'Autonomous Systems'*, 2018[12] e *Future of Work. Future of Society*, 2018[13]. A livello internazionale: OECD, *Legal Instruments on Artificial Intelligence* 2020[14]; UNESCO, *Recommendation on the Ethics of AI*, 2021[15]. Recentemente, *The Bletchley Declaration by Countries Attending the AI Safety Summit*, 1-2 November 2023[16].

Nei diversi documenti, seppur con talune distinzioni, l'etica acquisisce una considerevole rilevanza e gioca il ruolo di riflessione critica per la comprensione e la valutazione dell'IA[17], riconoscendo da un lato la rilevanza di aprire opportunità tecnologiche innovative 'per' l'uomo e l'umanità e dall'altro lato l'urgenza di evitare rischi gravi ed irreversibili per l'uomo e la società e controllare, mitigare e gestire i rischi ritenuti accettabili. I requisiti etici indispensabili della visione umano-centrica della IA sono: controllo umano significativo, sicurezza, spiegabilità, non discriminazione e giustizia.

***Controllo umano significativo.*** Anche se gli esseri umani costruiscono la IA, devono mantenere il controllo e la supervisione su ciò che progettano, programmano, applicano; le macchine dovrebbero rimanere un "supporto" alle decisioni umane, che cognitivamente "assistono" le decisioni umane, ma non le "sostituiscono", se non per scopi delimitati (in particolare per compiti ripetitivi, noiosi, umilianti, pericolosi, faticosi). Le macchine non dovrebbero "competere", ma "completare" le azioni umane. La necessità di mantenere il controllo umano resta essenziale anche per evitare il possibile problema dell'eccesso di 'delega tecnologica'. Un sistema esperto che diventa ottimale nel suggerire "decisioni" all'uomo pone il rischio di diminuire l'attenzione umana con la possibile conseguenza di ridurre le abilità umane (il cosiddetto fenomeno di '*deskilling*'), riducendo la responsabilità (de-

[11] https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence
[12] https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382
[13] https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/future-work-future-society_en
[14] https://legalinstruments.oecd.org/api/print?ids=648&lang=en
[15] https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.
[16] https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
[17] Nell'ambito della medicina, Nuffield Council on Bioethics, *Artificial Intelligence in Healthcare and Research*, 2020; the Swedish National Council on Medical Ethics, *Artificial Intelligence in Healthcare*, 2020; il Comitato Nazionale per la Bioetica con il Comitato per le Biotecnologie, Biosicurezza e Scienze della Vita, *Intelligenza artificiale e medicina: aspetti etici*, 2020. Council of Europe, *Recommendation (2185/2020) on Artificial Intelligence in Health Care: Medical, Legal and Ethical Challenges Ahead*, 2020; World Health Organization (WHO), *Ethics and governance of artificial intelligence for health*, WHO guidance. Geneva: World Health Organization, 2021.

responsabilizzazione) andando verso l'artificializzazione delle scelte che possono impoverire e perfino cancellare il rapporto interpersonale (de-umanizzazione). In questo senso, è importante disciplinare la sinergia produttiva come complementarità tra uomo e macchina, ricercando modalità di "sostegno" intelligente che consentano all'uomo di avere "un controllo umano significativo" in termini di attenzione e supervisione. Si parla anche di ‘non esclusività’, ossia di uso della IA non esclusivo, ma sempre con un controllo e monitoraggio umano.

***Sicurezza e affidabilità.*** Per ogni tecnologia il design deve essere sicuro; la sicurezza è un requisito etico per ogni macchina così come lo è per prodotti farmaceutici, gli alimenti, i trasporti, ecc. Tutti i "prodotti" dell'IA devono essere confrontati, attraverso studi condotti secondo le regole delle sperimentazioni cliniche controllate. Gli studi clinici controllati rimangono il "gold standard" per la dimostrazione della sicurezza e dell'efficacia delle tecnologie. È necessaria una nuova metodologia per controllare sperimentalmente l’IA, dato il problema del meccanismo che cambia nel tempo e la convalida che richiede un costante monitoraggio e ulteriori controlli. Sarà quindi necessario dimostrare con l’aiuto di informatici e ingegneri la sicurezza e l’efficacia dell'IA per garantire che anche i danni non intenzionali possano essere minimizzati e prevenuti e per garantire la robustezza tecnica sulla base del controllo a partire dalla qualità, accuratezza e interoperabilità dei dati (sia raccolti che confrontati) e dalla qualità degli algoritmi applicati, ove il vantaggio sia misurato in termini di benefici e rischi nell'applicazione. Solo in questo modo sarà possibile dimostrare l'affidabilità di questi sistemi attraverso validazioni, monitoraggi e certificazioni che ne garantiscono l'usabilità sicura.

***Spiegabilità e trasparenza.*** I sistemi di IA dovrebbero essere, nei limiti possibili, tracciabili e spiegabili in modo trasparente, per potere ricostruire i passaggi dai dati inseriti nella macchina (input) ai risultati ottenuti dalla macchina (output). Il problema è che ciò tecnicamente non è sempre possibile: si parla di ‘opacità’[18] degli algoritmi. In alcune circostanze è impossibile anche per i programmatori di software e gli informatici spiegare come il sistema abbia raggiunto determinati risultati: il "problema della scatola nera", in cui si conoscono solo gli input e gli output, ma non è possibile conoscere ‘cosa sta in mezzo’, essendo impossibile per un essere umano (anche un esperto) analizzare l'enorme quantità di calcoli effettuati dall'algoritmo. La macchina non fornisce informazioni complete sulle correlazioni dei dati e/o sulla logica adottata per giungere a una conclusione o proporre una decisione. Ciò comporta la necessità di sviluppare una tecnologia in grado di spiegare ogni passaggio della decisione o almeno di informare da parte dei programmatori gli utenti del rischio di opacità, al fine di acquisire una

[18] L. Floridi, J. Cowls, *A Unified Framework of Five Principles for AI in Society"*, *Harvard Data Sci*ence, 2019, (1) 1; S. Wachter, B. Mittelstadt, “A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI”, *Columbia Business Law Review*, 2019, 494.

consapevolezza critica. Una spiegabilità che sia anche comprensibile anche ai non tecnici e che sempre spieghi che i risultati degli algoritmi sono correlazioni probabilistiche e non rapporti di determinazione causa-effetto.

***Non discriminazione.*** Un ulteriore requisito etico della IA è l'assenza di pregiudizi: poiché i dati sono raccolti e selezionati da esseri umani e le correlazioni sono interpretate come significative da esseri umani, è possibile (come a volte è avvenuto) che i software abbiano prodotto risultati discriminatori. L'elaborazione di categorie/tipologie/clusters (profilazione) rischia di classificare i gruppi in base a caratteristiche predefinite. Se i dati inseriti e selezionati e le correlazioni escludono alcune categorie, ad esempio sulla base della età, del genere sessuale, della etnia o del possesso di abilità cognitive, i sistemi di IA possono produrre discriminazioni. L'algor-etica esige che i dati che nutrono le macchine e gli algoritmi che correlano i dati devono essere inclusivi, equi e rappresentativi della popolazione in generale, basandosi sul principio di uguaglianza e pari dignità degli esseri umani a prescindere dalle differenze di età, genere sessuale, etnia o abilità/disabilità, senza pregiudiziali discriminazioni, inconsapevoli o consapevoli.

***Giustizia e equità di accesso.*** Nella misura in cui sia disponibile una IA sicura, affidabile, (nei limiti possibili) trasparente e non discriminatoria, tutti dovrebbero potere accedere alle opportunità offerte. Ciò è possibile solo mediante un'informazione e una educazione specifica sull'IA al fine di evitare la cosiddetta "polarizzazione delle competenze". Un adeguato progetto informativo ed educativo inclusivo deve accompagnare le persone nel processo di transizione digitale tecnologica in modo graduale. L'educazione andrebbe dunque ripensata per rafforzare la formazione di chi lavora, in modo continuo, con un'adeguazione costante alle trasformazioni tecnologiche, e la formazione di chi si inserisce nel lavoro, con una specifica attenzione a garantire condizioni per un lavoro "dignitoso" che sappia contemperare le abilità digitali con la valorizzazione delle capacità personali. È anche importante educare ad un non eccessivo affidamento alle tecnologie e una capacità di "pensiero critico" oltre le abilità digitali. Va inoltre rinnovata la formazione, introducendo corsi di etica per ingegneri, informatici e data scientist, con particolare riferimento all'etica nelle tecnologie (ethics by design / in design / for designers) e alla progettazione, metodologia e applicazione di tecnologie. Questo è l'unico modo per garantire la consapevolezza e la comprensione etica sin dall'inizio della progettazione tecnologica. È inoltre auspicabile promuovere il dibattito pubblico sugli sviluppi e sui limiti dell'IA, in modo che tutti i cittadini in modo democratico possano acquisire le basi della "alfabetizzazione della IA", promuovendo una partecipazione attiva alla discussione sociale, coinvolgendo le regioni del mondo e i

gruppi sociali svantaggiati. Questi sono i presupposti per un possibile superamento del "divario digitale" evitando l'emarginazione, la stigmatizzazione e l'esclusione di persone prive di tecnologie, competenze e motivazione ad utilizzarle, nel quadro dell'inclusività, per "non lasciare nessuno indietro".

## 4. Verso una regolazione

Nell'ambito del dibattito sulla IA, l'etica nella prospettiva umano-centrica ha acquisito e sta acquisendo sempre più un ruolo rilevante per la regolamentazione, in corso di definizione. Tutti i documenti dei comitati di esperti raccomandano ai regolatori e alla politica, sia a livello globale che continentale e nazionale, di partire dai requisiti etici indicati e di includerli nella normazione della IA, in tutto il suo 'ciclo di vita', dal disegno e progettazione all'uso e applicazione nei diversi ambiti e in considerazione di tutti i 'portatori di interessi' (industrie, centri di ricerca, informatici e ingegneri, regolatori e cittadini)[19].

La *Proposta di regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'IA (legge sulla IA) e modifica alcuni atti legislativi dell'Unione*, 21 aprile 2021, con le successive revisioni ed emendamenti ancora in corso, intende perseguire un approccio equilibrato e prudente assicurando che i cittadini europei possano beneficiare di nuove tecnologie in conformità ai valori, ai diritti fondamentali e ai principi dell'Unione, nel quadro dei diritti umani fondamentali. La regolazione si basa sulla valutazione e gestione del rischio anche sulla base della previsione e anticipazione di scenari possibili e potenziali: si parte dal principio di proporzionalità benefici/rischi, ove i rischi devono essere sempre bilanciati rispetto ai benefici per l'uomo e la società. Laddove i rischi sono gravi e irreversibili sono ritenuti inaccettabile e vanno vietato; i rischi alti medi e bassi, vanno prevenuti, monitorati e gestiti in funzione della protezione dell'uomo, in particolare dei più vulnerabili (bambini, anziani, minoranze etniche, persone con specifiche condizioni socio-economiche o culturali). I sistemi di IA identificati come a rischio inaccettabile sono quelli che minacciano la dignità umana e sociale producendo danni fisici e psico-sociali mediante ingiustificata sorveglianza di massa e individuale (al di fuori di ragioni di sicurezza), manipolazione e distorsione occulta di idee e comportamenti (es. punteggio sociale o uso di identificazione biometrica da remoto in tempo reale in

[19] Sulla regolazione della IA si veda C. Casonato, *L'intelligenza artificiale e il diritto pubblico comparato ed europeo*, DPC E online, 2022, 51, 1; C. Casonato, *Potenzialità e sfide dell'intelligenza artificiale*, in "Rivista di Biodiritto - BioLaw Journal", 2019, 1, 177-182; L. d'Avack, *La rivoluzione tecnologica e la nuova era digitale: problemi etici e giuridici,* in AA.VV., *Intelligenza artificiale. Il diritto, i diritti, e l'etica*, a cura di U. Rufolo, Giuffrè Francis Lefebure 2020, pp. 3-28; S. Amato*, Biodiritto 4.0. Intelligenza artificiale e nuove tecnologie*, Giappichelli, Torino 2020; A. Simoncini, *Verso la regolamentazione della intelligenza artificiale. Dimensioni e governo*, "Rivista di Biodiritto", 2021, 2, p. 411 e seguenti.

spazi pubblici), inganno, sfruttamento e strumentalizzazione, basate su asimmetrie di potere, in quanto forme di illegittima coercizione della autonomia, minaccia della dignità e della giustizia.

La regolazione secondo i requisiti etici di controllo umano significativo, sicurezza, spiegabilità e trasparenza nei limiti possibili, uguaglianza e non discriminazione e giustizia, consentono la gestione, eliminazione o riduzione dei rischi, che impediscono la oggettificazione dell'uomo. L'obiettivo è pianificare una governance integrata alla tecnologia (e in qualche misura anticipata rispetto all'uso), flessibile (adattabile e rivedibile in base all'evoluzione tecnologica) e sovranazionale (continentale e auspicabilmente nel contesto di una cooperazione globale) nella triangolazione scienza-etica-società. Una regolamentazione che deve partire dai dati scientifici empirici attendibili e dalla anticipazione dei possibili scenari che si possono configurare, soppesando pro e contro sul piano etico e valutando le opzioni alternative mediante anche una consultazione e partecipazione sociale dei cittadini, ponderando la decisione nel contesto di un approccio saggio e prudenziale. Nella consapevolezza che quanto più vi è un grado di esposizione nell'incertezza alla gravità ed irreversibilità del rischio/danno sull'uomo e sull'umanità, tanto più vi deve essere il riconoscimento di una responsabilità individuale e collettiva, oltre che una sostenibilità sociale e ambientale. La regolazione sarà chiamata a calibrare, nella specificità delle diverse tecnologie di IA in continua evoluzione, quali strumenti siano necessari per difendere l'uomo e i diritti fondamentali.

Nell'ambito della discussione attuale sta emergendo l'interpretazione dei nuovi diritti umani nell'epoca della IA: il diritto di non esclusività delle tecnologie e di una interazione uomo/IA che riconosca la centralità del controllo umano; il diritto di accedere alle tecnologie di cui sia dimostrata la sicurezza ed efficacia; il diritto a ricevere adeguate spiegazioni trasparenti, nella misura possibile, delle tecnologia per poterne comprendere il significato e l'impatto individuale e sociale; il diritto ad essere inclusi e non esclusi, emarginati e discriminati, dall'uso delle tecnologie; il diritto ad essere informati ed educati all'uso delle tecnologie, ad acquisire una consapevolezza critica per una valutazione autonoma (anche di rifiuto o di alternativa alle tecnologie) e il diritto a partecipare alla '*governance*' delle tecnologie emergenti, ad essere coinvolti attivamente dalle istituzioni nella definizione delle politiche e nella regolamentazione delle nuove tecnologie.