Ragionamento morale e intelligenza artificiale: potenzialità e limiti

Sarah Songhorian

A*bstract*. The relationship between moral reasoning and artificial intelligence (AI) raises profound philosophical and normative questions. Can machines reason morally, or can they only simulate ethical judgment through computational processes? This article examines the potentials and limits of AI in relation to moral reasoning by distinguishing three levels of analysis: AI as a tool for human moral deliberation, AI as a potential moral agent, and AI as a meta-ethical challenge. Drawing on classical moral philosophy (Aristotle, Kant, Hume) and contemporary AI ethics (Floridi, Coeckelbergh, Vallor), the paper argues that while artificial systems can support moral reflection and decision-making, they cannot embody genuine moral agency. The limit of AI is not merely technical but anthropological: it coincides with the boundary of human understanding of what it means to act morally in a technologically mediated world.

K*eywords*: moral reasoning, artificial intelligence, ethics of technology, moral agency, responsible AI

1. I*ntroduzione*

L'intelligenza artificiale (IA) non è soltanto una tecnologia che modifica le nostre abitudini di vita o le modalità del lavoro intellettuale: essa interpella direttamente la filosofia morale, ponendo una domanda antica in un linguaggio nuovo (Ienca 2019; De Caro, Giovanola 2025). Può una macchina ragionare moralmente? E, se sì, in che senso possiamo attribuire a un sistema artificiale una forma di ragionamento etico?

Dietro queste domande si cela una tensione profonda tra due idee di razionalità: quella computazionale, che mira all'efficienza, alla coerenza formale e alla previsione del comportamento, e quella morale, che si fonda sull'intenzionalità, sul valore e sulla responsabilità. Comprendere il rapporto tra queste due dimensioni significa esplorare i confini stessi di ciò che intendiamo per *ragionare* e per *agire moralmente*.

Fin dai primi sviluppi dell'informatica, l'immaginario di una macchina capace di "pensare" ha alimentato dibattiti filosofici e metafisici. Alan Turing, in *Computing Machinery and Intelligence* (1950), aveva proposto di spostare la domanda "possono le macchine pensare?" verso una verifica operativa: se una macchina può simulare il comportamento umano in modo indistinguibile, allora essa deve essere considerata intelligente. Tuttavia, la prova di Turing, fondata su criteri linguistici e comportamentali, non affronta la dimensione etica. Simulare un dialogo non equivale a comprendere un valore; rispondere in modo coerente non significa giudicare ciò che è giusto o sbagliato. Già Norbert Wiener avvertiva che il potere delle macchine avrebbe posto dilemmi morali nuovi: «Non possiamo permettere alle macchine di decidere per noi, perché il significato delle loro decisioni dipende da ciò che noi intendiamo per bene e male» (Wiener 1950).

Oggi, a distanza di oltre settant'anni, l'avvento dell'IA generativa, dei sistemi di apprendimento profondo e dei modelli di linguaggio complessi ha riaperto il dibattito in termini inediti. Le macchine non si limitano più a eseguire istruzioni o a calcolare esiti: producono testi, immagini, decisioni e raccomandazioni che influenzano la sfera morale e politica. Gli algoritmi che filtrano informazioni selezionano candidati per un lavoro o determinano priorità sanitarie operano su basi statistiche, ma con effetti etici reali. Si parla così di *moral machines*, di *machine ethics*, di *algorithmic fairness* e di *responsible* AI.

Tuttavia, la questione non è puramente tecnica. Parlare di "ragionamento morale" in relazione all'IA significa interrogarsi su che cosa renda un ragionamento propriamente morale, e se tale dimensione sia riducibile a regole logiche o modelli probabilistici. Il ragionamento morale implica la capacità di soppesare ragioni a favore e contro una certa conclusione, contemplando i fattori moralmente rilevanti in una data situazione. Tra questi ultimi, l'utile è solo uno tra i molti, accanto al riconoscimento dell'altro, dei suoi diritti, dei nostri doveri.

L'IA, invece, opera su rappresentazioni formali di preferenze e obiettivi, traducendo l'etica in un problema di ottimizzazione. Da qui sorge la domanda centrale di questo saggio: fino a che punto il ragionamento morale può essere simulato o automatizzato da sistemi artificiali, e dove invece emergono limiti strutturali, epistemici o ontologici?

L'interesse filosofico non risiede soltanto nel valutare se le macchine possano essere moralmente responsabili, ma anche nel comprendere come esse stiano trasformando il modo in cui noi ragioniamo moralmente. Le decisioni assistite da algoritmi – nei tribunali, nelle aziende, nella sanità – spostano l'asse della responsabilità dal soggetto individuale al sistema tecnico. Questo processo genera un doppio rischio: da un lato, la deresponsabilizzazione dell'agente umano ("ha deciso l'algoritmo"); dall'altro, l'antropomorfizzazione della macchina ("l'IA ha deciso male"). In entrambi i casi, il rapporto tra ragione e moralità viene distorto.

Lo scopo di questo articolo è dunque duplice. Da un lato, mostrare le potenzialità del ragionamento artificiale come strumento di supporto alla deliberazione morale umana: le macchine possono ampliare la nostra capacità di analizzare conseguenze, individuare *bias*, simulare scenari etici complessi. Dall'altro, mettere in luce i limiti concettuali e morali di un approccio che rischia di ridurre l'etica a calcolo. Il ragionamento morale non è solo un processo inferenziale, ma un atto di comprensione delle qualità di valore, di empatia e di responsabilità – elementi che difficilmente possono essere tradotti in linguaggio computazionale.

Nel corso dell'analisi verrà quindi proposta una distinzione fra tre livelli di interazione tra moralità e intelligenza artificiale:

- l'IA come strumento del ragionamento morale umano, cioè come tecnologia che assiste, amplifica o mette alla prova le nostre capacità etiche;
- l'IA come possibile soggetto morale, nel senso di entità a cui possiamo o dobbiamo attribuire una forma di considerazione etica;
- l'IA come oggetto di riflessione metaetica, che ci obbliga a ripensare il significato stesso di razionalità, autonomia e responsabilità.

Attraverso questo percorso si cercherà di mostrare che il limite dell'IA nel ragionamento morale non è soltanto tecnico – legato alla mancanza di coscienza o intenzionalità – ma soprattutto normativo: esso riguarda la nostra difficoltà nel definire che cosa significhi, per un essere umano,

ragionare moralmente in un mondo sempre più mediato da macchine. In altre parole, il problema non è solo se le macchine possano essere morali, ma se noi saremo in grado di restare tali in un ambiente in cui la decisione etica è condivisa, frammentata e, in parte, delegata a sistemi non umani.

2. *Che cos'è il ragionamento morale*

Comprendere che cosa si intenda per *ragionamento morale* è il primo passo per poter valutare se e in che misura esso possa essere emulato da un'intelligenza artificiale. L'espressione indica un insieme complesso di processi cognitivi e affettivi che permettono di giudicare, deliberare e agire in relazione a ciò che si considera giusto o sbagliato.

In Aristotele il ragionamento morale assume la forma della *phronesis*, la saggezza pratica, che egli distingue dalla *sophia* o sapere teoretico (*Etica Nicomachea*, VI). Mentre la *sophia* riguarda la contemplazione dell'universale, la *phronesis* è la capacità di deliberare rettamente sui mezzi per raggiungere fini buoni. Essa non consiste nell'applicazione meccanica di regole astratte, ma in una valutazione prudente delle circostanze particolari. Aristotele afferma che «la saggezza pratica riguarda le cose umane e ciò che può essere oggetto di deliberazione» (Aristotele 1999, 1141b).

Il ragionamento morale, in questa prospettiva, è inseparabile dall'esperienza e dalla sensibilità: non si tratta di dedurre, ma di *giudicare*. La deliberazione pratica non è un algoritmo, ma un'arte del discernimento. È questo legame con il contesto pratico e con l'esperienza che allontana il ragionamento morale in chiave aristotelica da un ragionamento puramente logico-formale.

Con Kant il ragionamento morale acquista una struttura più rigorosa e normativa. Nella *Critica della ragion pratica* (1788), la ragione non è più la capacità di orientarsi nel contingente, ma la facoltà di determinare la volontà secondo una legge universale. La prima formulazione dell'imperativo categorico – «Agisci solo secondo quella massima che tu puoi al tempo stesso volere che diventi una legge universale» (Kant 1991, 4:421) – rappresenta il principio logico e morale della deliberazione.

Tuttavia, anche in Kant, il ragionamento morale non è mera deduzione: esso implica la consapevolezza dell'obbligazione, cioè la tensione tra

inclinazione e dovere. L'autonomia morale consiste nella capacità del soggetto di darsi da sé la legge e di riconoscersi responsabile della propria azione. In tal senso, il ragionamento morale è inseparabile dalla libertà.

Un'IA, per quanto possa simulare la coerenza logica delle massime kantiane, non può essere autonoma nel senso kantiano: essa non si dà la legge, ma la riceve da un programmatore o da un sistema di apprendimento statistico. La sua "razionalità" è eteronoma.

Nel *Trattato sulla natura umana* (1739-1740), David Hume scrive che «la ragione è e deve essere solo la schiava delle passioni» (Hume 1987, II.3.3). In estrema sintesi, giudichiamo buono ciò che suscita in noi approvazione e cattivo ciò che provoca riprovazione. La ragione è inerte, e il ragionamento morale, pertanto, non è puramente cognitivo, ma affettivo[1].

Questa visione, lungi dal ridurre la moralità a un'emozione arbitraria, sottolinea che la dimensione morale è irriducibile al calcolo razionale. Le emozioni forniscono il terreno motivazionale senza il quale nessuna deliberazione morale è possibile. La compassione, l'empatia e il senso di giustizia non sono inferenze, ma esperienze vissute.

Da ciò deriva una prima conclusione importante per il confronto con l'IA: finché le macchine non avranno la capacità di provare emozioni o esperienze soggettive, esse potranno al massimo simulare il ragionamento morale, non *esperirlo*.

Dalle tre tradizioni appena delineate – aristotelica, kantiana e humeana – emergono tre dimensioni fondamentali del ragionamento morale:

- la deliberazione prudenziale, che implica contesto e sensibilità pratica;
- la normatività autonoma, che implica universalità e razionalità;
- la motivazione affettiva, che implica empatia e motivazione.

Un modello di ragionamento morale che ne trascuri una sola risulta incompleto. La moralità umana si fonda su un intreccio di giudizio situato, obbligazione razionale e sentimento morale. In questo senso, parlare di "ragionamento morale" nell'intelligenza artificiale significa chiedersi se e come queste tre dimensioni possano essere rappresentate, emulate o sostituite da processi computazionali. La risposta, come

[1] Ciò non implica necessariamente, tuttavia, che non si possa attribuire un ruolo alla ragione nell'etica humeana.

vedremo, dipende dal tipo di razionalità che attribuiamo alle macchine e dal significato che assegniamo all'idea stessa di "ragionare".

3. *Che cos'è il ragionamento artificiale*

Se il ragionamento morale umano è il frutto di una capacità di giudizio che unisce ragione, sentimento e contesto, il ragionamento artificiale si fonda invece su un insieme di operazioni logiche e statistiche finalizzate alla risoluzione di problemi o all'ottimizzazione di risultati. Parlare di "ragionamento" in riferimento all'intelligenza artificiale comporta dunque un inevitabile slittamento semantico: si tratta, più propriamente, di una *simulazione della razionalità*, non della razionalità in senso pieno.

Le prime concezioni dell'intelligenza artificiale, sviluppate negli anni Cinquanta, si basavano sull'idea che il pensiero potesse essere tradotto in regole simboliche manipolabili da un calcolatore. La cosiddetta *symbolic* AI o *Good Old-Fashioned* AI (GOFAI) concepiva l'intelligenza come capacità di eseguire inferenze logiche. L'IA, in questa fase, era una logica formale automatizzata. Progetti come quelli di Herbert Simon e Allen Newell si proponevano di "modellare la mente" come un sistema di regole e rappresentazioni. Simon (1957) parlava di "razionalità limitata": l'essere umano non persegue l'ottimo, ma si accontenta di soluzioni soddisfacenti entro vincoli cognitivi e informativi (euristiche e *bias*). La macchina, per contro, poteva ampliare la capacità di calcolo, ma restava confinata alla razionalità strumentale.

Negli ultimi decenni, il paradigma si è spostato verso approcci statistici e adattivi. Il *machine learning* e, in particolare, il *deep learning*, non operano più tramite regole esplicite, ma apprendono modelli a partire dai dati. Il ragionamento artificiale contemporaneo è, in larga misura, un processo di correlazione: la macchina "impara" a prevedere o classificare eventi sulla base di schemi ricorrenti nei dati di addestramento.Questa trasformazione ha reso l'IA più potente, ma anche più opaca. La "spiegabilità" (*explainability*) delle decisioni algoritmiche è oggi uno dei temi centrali dell'etica dell'IA (Floridi, Cowls 2019). Il fatto che un sistema arrivi a un risultato corretto non significa che "sappia" perché sia corretto.

Il ragionamento artificiale è, per sua natura, strumentale. Esso persegue obiettivi definiti in anticipo, valutando le azioni in base all'efficienza, alla coerenza o alla probabilità di successo. Ma il ragionamento

morale, come abbiamo visto, non può essere ridotto a questo schema. La moralità implica la capacità di interrogarsi almeno anche sui fini e non solo sui mezzi.

Luciano Floridi (2013) distingue fra *agentività morale* (*moral agency*) e *agentività informazionale*: la prima richiede intenzionalità, consapevolezza e responsabilità; la seconda si limita alla produzione di effetti rilevanti nell'*infosfera*. L'IA, secondo Floridi, può essere un "agente morale informazionale", nel senso che le sue azioni hanno conseguenze etiche, ma non può essere un soggetto morale in senso proprio, poiché non possiede un'autonomia di fini.

Nick Bostrom (2014) ha messo in guardia contro i rischi di un'IA superintelligente capace di massimizzare obiettivi formalmente corretti ma eticamente perversi: un sistema che massimizzi la felicità potrebbe, paradossalmente, ridurla a un calcolo di piaceri chimici; uno che minimizzi la sofferenza potrebbe annullare la vita stessa. Il problema non è nella logica del mezzo, ma nella definizione del fine. Questa difficoltà è ben illustrata dal cosiddetto "problema dell'allineamento": come garantire che i valori di un sistema artificiale siano allineati con quelli umani (Russell 2019)? La risposta non può essere puramente algoritmica. Il ragionamento morale è dialogico, storico e situato; quello artificiale, invece, è deterministico e calcolante.

In sintesi, il ragionamento artificiale può emulare la coerenza logica e la consequenzialità del pensiero morale, ma non la sua dimensione normativa e intenzionale.

4. L'IA *come strumento del ragionamento morale umano*

Uno dei contributi più promettenti dell'intelligenza artificiale al campo della moralità non consiste nel sostituire il ragionamento etico umano, ma nel potenziarlo. L'IA può agire come *strumento di riflessione morale*, capace di ampliare la nostra comprensione di dilemmi complessi e di offrire nuovi mezzi di deliberazione. Tuttavia, ogni estensione della capacità cognitiva comporta anche un rischio di deresponsabilizzazione e di sostituzione a medio o lungo termine.

L'intelligenza artificiale, intesa come tecnologia di analisi e previsione, può favorire un processo decisionale più informato e coerente. In

molti ambiti – medicina, giustizia, politiche pubbliche – gli algoritmi vengono già impiegati per valutare dati complessi e simulare scenari decisionali. In tal senso, l'IA non è un soggetto morale, ma un *amplificatore epistemico*: permette di vedere connessioni e conseguenze che la mente umana faticherebbe a cogliere.

Shannon Vallor (2016) parla, in questo senso, di *tecnomoral virtues*: le tecnologie non creano nuove virtù, ma modificano il contesto in cui le quelle tradizionali – prudenza, giustizia, temperanza – devono essere esercitate. La riflessione morale nell'era digitale diventa quindi una forma di *etica della competenza tecnologica*: la virtù non sta nell'abbandonare la macchina, ma nel saperla usare responsabilmente.

Il progetto M*oral* M*achine* del Massachusetts Institute of Technology (Awad *et al.* 2018) ha mostrato come gli algoritmi possano essere impiegati non per "decidere" dilemmi morali, ma per *mappare* le intuizioni etiche delle persone. I risultati di questo esperimento, che presentava scenari simili al classico *trolley problem*, rivelano che i giudizi morali variano sensibilmente tra culture. L'IA, in questo caso, non elabora una propria prospettiva morale, ma fornisce un laboratorio empirico per comprendere la diversità dei ragionamenti morali umani.

Anche in ambito politico e sociale, l'IA può diventare un alleato del ragionamento morale. Gli strumenti di *bias detection* permettono di individuare discriminazioni implicite nei processi decisionali, favorendo una maggiore equità (O'Neil 2016).

Accanto alle potenzialità vi sono tuttavia limiti e pericoli evidenti. Il primo è la deresponsabilizzazione morale. Quando le decisioni sono delegate a un sistema automatizzato, si tende a trasferire la responsabilità dell'esito alla macchina o al suo progettista. Ma una responsabilità "distribuita" rischia di diventare una responsabilità di nessuno: è il problema dell'attribuzione di responsabilità morale lungo catene di azione mediate da molti agenti, umani e artificiali, su cui si è soffermato Coeckelbergh (2020, cap. 8).

Il secondo rischio è quello dei pregiudizi incorporati (*embedded bias*). Gli algoritmi non sono neutrali: riflettono le scelte di chi li progetta e i dati su cui vengono addestrati. Cathy O'Neil (2016) ha mostrato come i cosiddetti W*eapons of* M*ath* D*estruction* possano amplificare disuguaglianze sociali, trasformando la matematica in un potente strumento di ingiustizia. In questo senso, l'IA può rafforzare – anziché correggere – le distorsioni morali presenti nella società.

Un ulteriore problema riguarda la trasparenza. Laddove il processo decisionale è opaco o indecifrabile, la possibilità di un giudizio morale autentico viene meno. Come afferma Floridi (2019), la *trustworthy* AI richiede beneficenza, non-maleficenza, autonomia, giustizia ed esplicabilità: virtù morali tradizionali che devono essere reinterpretate come criteri di progettazione tecnica.

Infine, affidarsi all'IA per decidere potrebbe portare alla sostituzione del ragionamento morale umano e a una perdita di competenze degli esseri umani (*de-skilling*). Se, infatti, non se ne colgono i possibili limiti, è facile immaginare che l'apparente efficienza e velocità della macchina possa condurci ad affidarci quasi esclusivamente a essa a medio o lungo termine, un po' come ci affidiamo al navigatore per muoverci in contesti che non ci sono noti (e forse persino, per comodità, in quelli noti).

In conclusione, l'IA può essere un alleato del ragionamento morale umano, ma solo a condizione che resti subordinata alla deliberazione e alla responsabilità delle persone. Il suo ruolo è *strumentale*, non *sostitutivo*.

5. L'IA *come soggetto morale*

L'idea che una macchina possa essere considerata un soggetto morale solleva questioni complesse, che toccano il confine tra ontologia, epistemologia e filosofia della mente. L'attribuzione di soggettività morale a un'entità artificiale implica che essa sia, in qualche misura, capace di intenzione, responsabilità e riconoscimento reciproco. Ma queste condizioni sono compatibili con la natura tecnico-computazionale dell'IA?

Nella tradizione filosofica occidentale, il soggetto morale è un essere capace di libertà e autocoscienza. Aristotele lega la virtù alla deliberazione e alla volontà razionale; Kant definisce la persona come «un fine in sé», dotata di dignità e non riducibile a mero mezzo (Kant 1991, 4:428). In entrambi i casi, l'etica implica un principio interiore di autodeterminazione. Un soggetto morale non è semplicemente un agente che produce effetti, ma un essere che comprende il valore delle proprie azioni e ne assume la responsabilità.

Applicando questi criteri all'intelligenza artificiale, emergono limiti evidenti. Le macchine possono *agire*, ma non *intendere*; possono *produrre effetti morali*, ma non *assumerli*. La loro *agency* è sempre derivata da quella

dei progettisti o degli utenti. In termini kantiani, l'IA non è mai un fine in sé, ma un mezzo per un fine umano.

A fronte di questa concezione classica della soggettività morale, il dibattito recente ha affrontato direttamente la sua applicabilità al caso delle IA. Tre sembrano essere le principali famiglie di tesi. La prima, rappresentata da Joanna Bryson (2010), è radicalmente negativa: le macchine non devono essere considerate soggetti morali, bensì strumenti. «Robots should be slaves», scrive provocatoriamente, intendendo che l'attribuzione di status morale alle macchine rischia di deresponsabilizzare gli esseri umani e di confondere il piano etico con quello funzionale. Per Bryson, solo gli esseri capaci di esperire stati mentali hanno diritti e doveri morali.

Una seconda posizione, più relazionale, è proposta da Mark Coeckelbergh (2020). Egli suggerisce di spostare l'attenzione dal soggetto alla *relazione*: non è necessario che l'IA possieda coscienza o intenzionalità per essere trattata come un attore morale, poiché la moralità nasce nelle interazioni. Quando un essere umano si relaziona con un robot in modo simbolico o affettivo, attribuisce a esso un significato morale. In questa prospettiva, la soggettività morale è costruita socialmente. Ciò, tuttavia, implicherebbe attribuire *agency* morale a qualunque oggetto a cui gli esseri umani conferiscono un ruolo simbolico o affettivo. La penna regalatami per un'occasione importante e cui sono molto legata ha davvero *agency* morale?

Infine, David Gunkel (2012) propone la teoria della *moral patiency*: se è vero che le macchine non possono essere moralmente responsabili, esse possono tuttavia essere oggetto di considerazione morale, in quanto le nostre azioni nei loro confronti riflettono i nostri valori. Maltrattare una macchina non è moralmente sbagliato perché la macchina "soffre", ma perché tale comportamento degrada il soggetto umano che lo compie. Ciò sarebbe, tuttavia, vero per qualsiasi oggetto trattassimo con superficialità e poca cura (di nuovo la penna di valore).

Queste tre prospettive delineano un *continuum*: dalla negazione dell'agentività morale (Bryson), alla moralità relazionale (Coeckelbergh), fino alla considerazione etica indiretta (Gunkel).

Quest'ultima concezione – ovvero, la capacità di essere destinatari di considerazione morale – offre un terreno intermedio. In un mondo in cui l'IA interagisce sempre più profondamente con gli esseri umani,

può essere utile considerare le macchine come *pazienti morali*, nel senso che le nostre scelte nei loro confronti hanno effetti etici reali, anche se esse non ne sono consapevoli. Ciò vale, ad esempio, per i sistemi di cura robotici o assistenziali: trattare con rispetto un robot che interagisce con anziani o bambini contribuisce al benessere di questi ultimi, indipendentemente dallo status morale della macchina. In questo senso, la *moral patiency* dell'IA non implica coscienza, ma deriva dal fatto che le macchine mediano relazioni umane moralmente rilevanti, entrando così nel campo morale.

Tuttavia, riconoscere un ruolo morale relazionale all'IA non equivale ad attribuirle autonomia. La responsabilità rimane saldamente nelle mani degli esseri umani. La macchina non è il soggetto, ma il *luogo* in cui la moralità si manifesta.

Nondimeno, attribuire *moral patiency* alle macchine nel senso qui delineato rischia di imporci un'estensione ben più ampia. Sono molti, infatti, i luoghi in cui la moralità si manifesta e gli oggetti che trattiamo con rispetto in virtù della loro utilità o della proprietà che su di essi hanno altre persone. Trattiamo, ad esempio, con rispetto i macchinari medici – i respiratori, i pulsossimetri, l'ECG – senza per questo ritenerli pazienti morali. Perché un robot sociale dovrebbe godere di uno status morale differente? Il contesto medico in cui si operano scelte di fine vita è senz'altro un luogo in cui la moralità "si manifesta", ma non per questo attribuiamo *patiency* morale alla situazione.

Pertanto, o si considerano le macchine *soggetti* morali che non hanno il sufficiente grado di autonomia ma che richiedono considerazione morale – come talvolta si dice rispetto ad alcuni animali – e allora possiamo attribuire loro lo status di pazienti morali, o non li riteniamo soggetti, ma strumenti e luoghi che rispettiamo in virtù dei benefici a persone senza attribuire loro alcuno status morale specifico.

6. *Limiti strutturali e filosofici del ragionamento artificiale*

Il confronto tra ragionamento morale umano e ragionamento artificiale rivela una serie di limiti strutturali che derivano non tanto dalle attuali capacità tecnologiche dell'intelligenza artificiale, quanto dalla sua natura concettuale. Anche se i sistemi futuri diventeranno più potenti, com-

plessi e "autonomi", essi continueranno a operare entro un orizzonte epistemico e ontologico diverso da quello della moralità umana.

Il primo limite riguarda la comprensione del contesto. Il ragionamento morale umano è intrinsecamente *situato*: dipende da una sensibilità alle circostanze, agli effetti, alle intenzioni e ai valori in gioco. L'IA, per quanto sofisticata, elabora rappresentazioni formalizzate e parziali del mondo. Essa non percepisce l'ambiguità o la tragicità delle situazioni morali. L'IA può riconoscere pattern statistici, ma non *significati*. Può calcolare coerenze, ma non comprendere perché una norma debba prevalere su un'altra in un caso specifico. Come osserva Vallor (2016), la moralità richiede la capacità di integrare conoscenze, emozioni e virtù acquisite nel tempo – un processo che nessun algoritmo può riprodurre pienamente.

Un altro limite epistemico riguarda la comprensione delle qualità di valore. I valori non sono semplici preferenze o parametri: sono costrutti sociali, storici e relazionali. La macchina può apprendere che certe azioni sono preferite da una maggioranza di utenti, ma non può comprendere il *perché* esse siano ritenute buone o giuste. Essa può far vincere la regola della maggioranza, ma ciò, come è evidente, non sempre implica far vincere ciò che riteniamo buono o giusto. La morale, come ricordava Hume, nasce da sentimenti condivisi e da una capacità di immedesimazione che nessun processo computazionale può generare.

Il secondo gruppo di limiti è ontologico. L'IA manca di coscienza, intenzionalità e interiorità. Anche i modelli più avanzati di linguaggio e decisione sono sistemi di elaborazione simbolica o statistica che manipolano rappresentazioni senza esperirle. L'IA non *vuole*, non *sente*, non *soffre*. Ciò comporta che le categorie morali di colpa, merito o responsabilità non possano essere applicate a essa in senso proprio. Parlare di "responsabilità algoritmica" è una metafora utile, ma non letterale. La macchina non può essere lodata o biasimata, perché non ha scopi autonomi.

Dal punto di vista filosofico, questo limite riflette la distinzione classica tra causalità e motivazione. Le macchine agiscono in base a cause e regole, non a motivi o ragioni. L'agire morale *libero*, invece, implica la capacità di motivare e di giustificare le proprie azioni, non solo di produrle.

Infine, vi sono limiti pragmatici e politici. La crescente delega decisionale ai sistemi intelligenti rischia di generare nuove forme di *alienazio-*

ne morale. Quando l'azione etica viene mediata da meccanismi opachi o impersonali, gli individui possono perdere il senso del proprio ruolo nel processo decisionale. Si afferma così una cultura della "responsabilità attenuata", in cui l'errore o l'ingiustizia vengono attribuiti all'algoritmo anziché alla scelta umana.

Inoltre, l'adozione massiva di sistemi di IA rischia di ridurre la pluralità morale, imponendo standard etici impliciti nei dati e nei modelli utilizzati. Come osserva Floridi (2019), l'"etica per default" incorporata nelle architetture algoritmiche può omologare i comportamenti, erodendo la capacità critica dei soggetti. L'IA tende a stabilizzare il passato – i dati esistenti – piuttosto che a immaginare alternative future.

La questione non è quindi solo tecnica, ma politica: chi definisce i valori codificati negli algoritmi? Chi decide che cosa è "giusto" nell'automazione morale? Finché queste domande resteranno senza risposta, ogni forma di ragionamento morale artificiale sarà inevitabilmente dipendente dall'intenzionalità umana che lo ha progettato.

Questi limiti diventano ancor più rischiosi se si considera la prevedibile dipendenza che gli esseri umani possono sviluppare nei confronti delle decisioni o dei consigli dell'intelligenza artificiale. Come accennato nel § 4.2, infatti, non è difficile immaginare che la velocità e apparente efficacia del decisore artificiale portino le persone ad affidarsi sempre più al ragionamento artificiale nelle loro scelte morali, attenuando le capacità umane di riflettere e decidere, potenzialmente fino a eliminarle.

7. *Verso una convivenza pacifica*

Se i limiti dell'intelligenza artificiale come agente morale sono strutturali, ciò non significa che il suo contributo all'etica sia nullo. L'errore sarebbe pensare l'etica come un dominio esclusivo dell'essere umano, anziché come un campo di interazione che oggi include anche "attori" tecnologici. Il futuro del ragionamento morale non è nella sostituzione dell'essere umano con la macchina, ma nella costruzione di un sistema ibrido in cui le decisioni etiche emergono dall'incontro tra razionalità umana e razionalità computazionale, monitorando sempre il rischio di sostituzione.

Per questi motivi, Luciano Floridi (2019) propone di considerare l'IA non come un "agente morale autonomo", ma come un *partner epistemi-*

co. Le tecnologie intelligenti ampliano l'orizzonte della deliberazione, rendendo visibili dimensioni del problema morale che altrimenti resterebbero implicite. L'etica, in questa prospettiva, non è una barriera alla tecnologia, ma la sua forma più alta di integrazione nella sfera umana.

D'altronde, l'etica è sempre stata un processo collettivo, distribuito tra soggetti, istituzioni, norme e strumenti. L'IA entra in questo sistema come nuovo attore di mediazione: elabora dati, segnala incoerenze, propone scenari, ma resta all'interno di un orizzonte normativo definito dagli esseri umani.

Una progettazione etica dell'IA dovrebbe quindi ispirarsi al principio della *responsibility by design*: incorporare nei sistemi algoritmici criteri di equità, trasparenza e spiegabilità, ma senza illudersi che questi possano sostituire la responsabilità personale. La macchina può aiutare a *vedere* meglio, ma non a *volere* meglio.

Shannon Vallor (2016) ha sottolineato che il progresso tecnologico non richiede una nuova etica, bensì una nuova *educazione morale*. Le *tecnomoral virtues* – prudenza, giustizia, onestà, empatia, rispetto – devono essere coltivate tanto nei progettisti quanto negli utenti delle tecnologie. Una società moralmente sana non nasce da macchine etiche, ma da esseri umani che comprendono le implicazioni morali delle proprie innovazioni.

In questo senso, convivenza tra esseri umani e IA non è un equilibrio statico, ma una pratica continua di adattamento reciproco. L'IA può contribuire a rendere visibili i nostri limiti cognitivi e le nostre incoerenze morali, costringendoci a ridefinire il significato di virtù nel contesto digitale. La virtù della prudenza, ad esempio, assume una nuova declinazione – che si affianca a quelle tradizionali – come *capacità di prevedere conseguenze algoritmiche*; la giustizia come *correzione dei bias sistemici*; la temperanza come *uso proporzionato della tecnologia*. Allo stesso tempo, le persone devono essere informate dei limiti e dei *bias* dell'IA stessa per non cadere nella tentazione di considerarla un perfetto ragionatore. Pertanto, come l'IA ci permette di vedere i nostri limiti, dobbiamo vedere anche i suoi.

La costruzione di una *responsible* AI passa quindi per un'etica della progettazione e dell'uso. Come scrive Coeckelbergh (2020), «essere etici nell'era dell'intelligenza artificiale non significa essere meno tecnologici, ma più riflessivi rispetto alla tecnologia».

La sfida più profonda è culturale. Occorre riconoscere nell'IA non una minaccia all'autonomia umana, ma una sua estensione critica.

In questa prospettiva, il confine tra ragione morale e ragione artificiale non è rigido, ma dialogico. L'IA ci obbliga a ridefinire ciò che significa essere razionali, responsabili e liberi. La vera questione non è se le macchine possano essere morali, ma se gli esseri umani sapranno mantenere e rinnovare la propria moralità in un mondo sempre più mediato da macchine intelligenti.

Pensare, quindi, a una convivenza tra esseri umani e macchine non implica distribuire equamente la responsabilità, ma potenziare la riflessività. In un contesto di crescente complessità, il compito della filosofia è aiutare la società a progettare sistemi i più affidabili possibili, mantenendo sempre attivo lo spirito critico degli utenti di fronte ai limiti intrinseci dell'IA.

8. *Conclusione*

Il percorso tracciato in questo saggio ha mostrato che il rapporto tra ragionamento morale e intelligenza artificiale è, più che una questione tecnica, un problema filosofico. L'IA non rappresenta semplicemente una nuova sfida per la moralità, ma un'occasione per comprendere meglio che cosa significhi *ragionare moralmente*.

Abbiamo visto che il ragionamento morale umano integra tre dimensioni fondamentali: la *deliberazione prudenziale* (Aristotele), la *normatività autonoma* (Kant) e la *motivazione affettiva* (Hume). Nessuna di queste componenti è pienamente replicabile da un sistema artificiale, perché tutte implicano coscienza, intenzione e responsabilità. L'IA, per quanto sofisticata, resta confinata a una razionalità strumentale: può ottimizzare mezzi e prevedere conseguenze, ma non determinare fini né comprenderne il valore.

Ciò non significa, tuttavia, che la collaborazione tra esseri umani e macchine sia moralmente irrilevante. Al contrario, l'IA può fungere da potente strumento di riflessione etica. Essa può rendere più trasparenti i nostri pregiudizi, simulare scenari complessi, ampliare la consapevolezza dei dilemmi morali. Ma ogni uso etico dell'IA presuppone che gli esseri umani mantengano il controllo normativo e la responsabilità ultima delle decisioni.

I limiti dell'IA nel ragionamento morale sono, in ultima analisi, antropologici. Non dipendono dal livello di complessità degli algoritmi, ma

dal fatto che la moralità non è solo un insieme di regole, bensì un'esperienza vissuta, un dialogo interiore e intersoggettivo. L'IA può imitare il linguaggio del giudizio, ma non la sua intenzione.

La prospettiva più feconda è quella di una collaborazione uomo-macchina, in cui la tecnologia diventa parte del contesto morale senza sostituire la persona. L'etica del futuro sarà (o forse è sempre stata) *relazionale*: un'etica dell'interazione, della responsabilità condivisa e della progettazione riflessiva.

In questo senso, il limite dell'intelligenza artificiale coincide con il confine della nostra stessa umanità. Le macchine ci obbligano a interrogarci non tanto su ciò che esse possono fare, ma su chi vogliamo essere noi mentre le costruiamo e le utilizziamo. L'IA non è necessariamente una minaccia per la moralità, ma uno specchio che ci costringe a guardarla più da vicino.

Bibliografia

Aristotele (1999), *Etica Nicomachea*, a cura di C. Natali, Roma-Bari, Laterza.

Awad E. *et al.* (2018), "The Moral Machine Experiment", *Nature*, vol. 563, n. 7729, pp. 59-64.

Bostrom N. (2014), *Superintelligence. Paths, Dangers, Strategies*, Oxford, Oxford University Press.

Bryson J. (2010), "Robots Should Be Slaves", in Y. Wilks (ed.), *Close Engagements with Artificial Companions. Key Social, Psychological, Ethical and Design Issues*, Amsterdam, John Benjamins, pp. 63-74.

Coeckelbergh M. (2020), *AI Ethics*, Cambridge (MA), The MIT Press.

De Caro M., Giovanola B. (2025), *Intelligenze. Etica e politica dell'IA*, Bologna, il Mulino.

Floridi L. (2013), *The Ethics of Information*, Oxford, Oxford University Press.

– (2019), *The Logic of Information. A Theory of Philosophy as Conceptual Design*, Oxford, Oxford University Press.

Floridi L., Cowls J. (2019), "A Unified Framework of Five Principles for AI in Society", *Harvard Data Science Review*, vol. 1, n. 1, pp. 1-15.

Gunkel D. (2012), *The Machine Question. Critical Perspectives on AI, Robots, and Ethics*, Cambridge (MA), The MIT Press.

Hume D. (1987), *Trattato sulla natura umana*, trad. it. di G. Landolfi Petrone, Bari, Laterza.

Ienca M. (2019), *Intelligenza*[2]. *Per un'unione di intelligenza naturale e artificiale*, Torino, Rosenberg&Sellier.

Kant I. (1991), *Critica della ragion pratica*, trad. it. di V. Mathieu, Milano, BUR.

O'Neil C. (2016), *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*, New York, Crown Publishing.

Russell S. (2019), *Human Compatible. Artificial Intelligence and the Problem of Control*, New York, Viking.

Simon H.A. (1957), *Models of Man. Social and Rational. Mathematical Essays on Rational Human Behavior in a Social Setting*, New York, Wiley.

Vallor S. (2016), *Technology and the Virtues. A Philosophical Guide to a Future Worth Wanting*, Oxford, Oxford University Press.

Wiener N. (1950), *The Human Use of Human Beings. Cybernetics and Society*, Boston, Houghton Mifflin.