cristiani e cittadini / dossier

I robot distinguono tra bene e male?

Aspetti etici dell'intelligenza artificiale

Markus Krienke
Docente di Storia della Filosofia moderna e di Etica sociale presso la Facoltà Teologica di Lugano e di Antropologia filosofica presso la Pontificia Università Lateranense
<krienke@rosmini.de>

Gli scenari aperti dal progresso tecnologico nell'ambito dell'intelligenza artificiale e dall'impatto che esso avrà sulla società sollevano questioni etiche e antropologiche con cui la riflessione filosofica e teologica e la dottrina sociale della Chiesa sono chiamate a misurarsi. Le macchine intelligenti acquisteranno anche la capacità di distinguere il bene dal male? Dovremo quindi considerarle soggetti con una propria responsabilità? O la responsabilità morale resterà una caratteristica peculiare dell'essere umano?

Pontificia Accademia per la Vita, Microsoft, IBM, FAO e Governo italiano sono i primi firmatari del *Rome Call for AI Ethics* (Appello di Roma per un'etica dell'intelligenza artificiale, disponibile in <www.academyforlife.va>), sottoscritto lo scorso 28 febbraio. **L'obiettivo è orientare i progressi dell'intelligenza artificiale (AI) al bene dell'umanità e della casa comune**. Etica, educazione e diritti sono le tre vie per renderlo possibile, evitando derive che asservirebbero l'essere umano alla macchina e ai suoi canoni di funzionamento. Dal principio dell'inviolabile dignità della persona umana emerge l'esigenza che l'AI sia inclusiva e trasparente, in modo da porsi autenticamente al servizio della realizzazione di ogni uomo e ogni donna. È urgente predisporre piani formativi che consentano ai giovani di sviluppare le capacità necessarie per avvalersi dell'AI e adottare norme che regolino gli ambiti della vita

sociale che si realizzano tramite l'AI. A questo scopo il *Rome Call* propone **sei principi di riferimento: trasparenza, inclusione, responsabilità, imparzialità tracciabilità e sicurezza** (privacy). Prende corpo un percorso per concretizzare quanto papa Francesco auspicava nel messaggio indirizzato al World Economic Forum di Davos il 24 gennaio 2018, ossia che l'AI contribuisca «al servizio dell'umanità e alla protezione della nostra casa comune invece che per l'esatto opposto, come purtroppo prevedono alcune stime».

Le dimensioni della trasformazione in atto

Un dato aiuta a mettere a fuoco la dimensione della rivoluzione in atto: **attualmente in un solo anno l'umanità produce tanti dati quanti in tutta la storia precedente**. Tra dieci anni, quando il numero dei dispositivi connessi a Internet sarà di 150 miliardi, il tempo di raddoppiamento si ridurrà a 12 ore (cfr Rasetti 2018, 33). Inoltre, già oggi un terzo delle notizie di *Bloomberg News* è prodotto con l'aiuto di AI (cfr Peiser 2019) e nel 2018 si è registrato un aumento del 18% nell'uso di AI per operazioni chirurgiche. Si stima che in futuro il 49% dei lavori potrà essere svolto da apparati dotati di AI, con una ricaduta anche sui "colletti bianchi". Determinate professioni non esisteranno più, mentre in tutti i lavori l'essere umano dovrà collaborare con macchine intelligenti. Nel 2016 in Arabia Saudita per la prima volta un robot, chiamato Sophia, ha ottenuto la cittadinanza.

Al di là delle sfide tecniche e sociali, **si pone la domanda etica e antropologica se le macchine intelligenti un giorno davvero disporranno di una coscienza simile alla nostra, in grado di autodeterminarsi e di scegliere fra bene e male**. Se così fosse, acquisterebbero lo status di autentiche "persone artificiali", dissolvendo il confine fra l'umano e l'artificiale: è la tesi, decisamente antiumanistica, dell'"AI forte". Invece si definisce "AI debole" l'ipotesi secondo cui la macchina potrebbe essere in grado di simulare il funzionamento di una coscienza umana, ma senza averne le proprietà. Un caposaldo della seconda ipotesi è rappresentato dal "test di Turing". Il matematico britannico Alan Turing (1912-1954) ipotizzò un esperimento in cui un essere umano comunica per iscritto con due soggetti, uno umano e l'altro un'AI. Se non riesce a distinguere tra l'interlocutore umano e la macchina, possiamo affermare che quest'ultima è in grado di pensare (cfr Brand 2018, 39). A questo punto sorgono due domande: questo test vale anche per l'agire morale? E basta questo test per attribuire alle macchine ciò che chiamiamo "moralità" in senso proprio, cioè la capacità di fare scelte sulla base di considerazioni etiche?

Una macchina può agire moralmente?

Nel 1942, nel romanzo *Io, Robot* lo scrittore Isaac Asimov (1920-1992) formulò le tre "leggi della robotica", destinate a disciplinare il comportamento dei robot dotati di AI in un ipotetico futuro:

I. Un robot non può recare danno a un essere umano, né può permettere che, a causa del suo mancato intervento, un essere umano subisca un danno;

II. Un robot deve obbedire agli ordini impartiti dagli esseri umani, purché non siano in contrasto con la prima legge;

III. Un robot deve proteggere la propria esistenza, purché questo non contrasti con la prima o con la seconda legge.

Al di là della finzione narrativa, queste leggi sono un primo tentativo di basare l'agire dei robot intelligenti non solo sul calcolo utilitaristico del miglior risultato, ma su un vero e proprio senso del dovere, che richiama alla mente la filosofia di Immanuel Kant (1724-1804), secondo cui la moralità umana consiste nella capacità di riconoscere la legge morale come vincolante e di obbedirle.

Tuttavia **queste regole basate sul dovere** (e perciò dette deontologiche), hanno un problema in comune con i criteri della morale utilitarista: **non riescono a rispecchiare né a prevedere la complessità della deliberazione morale**, in quanto tendono ad astrarre dal contesto concreto nel quale l'azione si svolge. Proprio questo sembra il punto decisivo da considerare, nel momento in cui l'AI assume tratti sempre più simili all'agire umano.

A differenza delle morali utilitaristiche e deontologiche, per Aristotele e la tradizione che si ispira al suo pensiero l'azione morale va sempre considerata nel suo contesto: solo a partire dalle circostanze concrete è possibile giungere a una scelta ponderata e virtuosa, che si colloca "nel mezzo" fra gli eccessi opposti. La prospettiva suggerita da Aristotele mette al centro il concetto di virtù, considerata come la capacità abituale di individuare i valori in gioco in una situazione e di perseguirli in maniera efficace. Questo approccio potrebbe essere applicato anche all'AI. Poiché, infatti, le virtù sono il risultato di un processo di apprendimento intelligente, **si teorizza anche per l'AI la prospettiva di "imparare" il comportamento morale**. In sintesi, se l'AI fosse in grado di imparare ad agire in modo virtuoso, dovremmo attribuirle una personalità morale al pari dell'essere umano. Tuttavia, questa posizione solleva alcune perplessità.

Come esempio immaginiamo un robot medico che, avendo ricevuto informazioni errate, somministri a un paziente un farmaco sbagliato, uccidendolo, e che, a seguito dell'accaduto, mostri dispiacere per la morte della persona. Possiamo dare una valutazione morale di questo episodio? La nostra risposta è no.

In primo luogo, potremmo "assolvere" il robot perché l'informazione sbagliata non ricadeva nella sua responsabilità. Tuttavia non ha senso affermare che il robot non "volesse" la morte del paziente, perché l'intenzionalità di un robot coincide con il suo agire effettivo.

In secondo luogo, anche il dispiacere non può essere compreso nella valutazione morale, perché è semplicemente la reazione che il robot ha imparato a mostrare dopo tale risultato. Certamente anche per il robot sarebbe stato meglio se il paziente fosse rimasto in vita, perché avrebbe ricevuto un *feedback* positivo, e senz'altro cercherà di fare in modo che questo accada in situazioni analoghe in futuro. Ma questo robot non possiede le caratteristiche di intenzionalità (volere) e autoconsapevolezza (dispiacere) necessarie per giungere a un giudizio morale sul proprio agire (cfr Brand 2018, 119-120). Esso agisce esteriormente in modo conforme al dovere morale, ma ciò che gli manca è il fatto di agire in quanto intimamente convinto che obbedire alla legge morale sia giusto.

In altre parole, **ciò che differenzia il comportamento degli esseri umani da quello delle macchine è che i primi sono consapevoli dei propri stati interiori, riconoscono il proprio agire come libero** ("libero arbitrio") **e sono capaci di scegliere in base a una ponderazione complessa delle circostanze**, e non semplicemente in base a un calcolo di vantaggi e benefici. Queste capacità costituiscono il proprio dell'essere umano e il fondamento di una libertà che chiede di essere rispettata; pertanto già nel 1990, Giovanni Paolo II affermò che «Il tentativo di spiegare il pensare e il volere libero dell'uomo in chiave meccanicistica e materialistica porta inevitabilmente alla negazione della persona e della sua dignità».

Intelligenza senza ragione

Questa riflessione sulla differenza morale fra l'agire umano e l'AI non viene accettata da chi, come l'informatico statunitense Raymond Kurzweil, vede avvicinarsi il momento in cui l'AI raggiungerà il livello umano e avverrà così il «culmine della fusione fra il nostro pensiero e la nostra esistenza biologica con la nostra tecnologia»; verrà così superata ogni «distinzione fra umano e macchina o fra realtà fisica e virtuale» (Kurzweil 2014, 9). Tale tesi trova sostegno in alcune posizioni di autori contemporanei: dall'impressione che la differenza categoriale tra l'AI e quella umana «non può essere sostanziata da argomentazioni filosofiche non controverse» (Boden 2019, 140), all'affermazione che «per quanto riguarda i processi decisionali non è, almeno finora, stata individuata alcuna ragione per credere che esseri umani e macchine obbediscano a principi diversi, naturali o scientifici che siano» (Kaplan 2018, 121).

Tuttavia questi autori omettono di fare distinzione fra due dimensioni della mente umana che possiamo chiamare "intelligenza" e "ragione". Il primo termine riguarda i processi cognitivi e viene applicato, non senza qualche forzatura, alle macchine pensanti. Invece il concetto di ragione ha un maggiore spessore e coinvolge la sfera della deliberazione morale, indispensabile in quei momenti in cui la semplice ottimizzazione intelligente non basta. La macchina intelligente è progettata per ottimizzare i risultati nelle situazioni ordinarie; invece, nelle loro decisioni morali gli «uomini non ottimizzano» (Nida-Rümelin e Weidenfeld 2018, 97, nostra trad.). In altre parole, **la prassi delle decisioni morali non può essere sostituita da algoritmi**. Il fatto che il funzionamento delle macchine possa assomigliare a quello dell'essere umano non significa che sia identico. Il filosofo statunitense John Searle ha distinto su questa base l'intelligenza umana (*intelligence with reason*) da quella artificiale (*intelligence without reason*).

Tuttavia, poiché i sistemi dotati di AI agiscono, è senz'altro necessario dotarli di "ragione etica" per renderli compatibili con il mondo umano (cfr Zamagni 2019, 189): basta pensare ai problemi sollevati dai veicoli a guida autonoma (cfr Cerruti 2018). In questa prospettiva, è necessario attribuire alla macchina, seppure in modo soltanto analogo alla ragione umana, una certa "capacità di agire" (*agency*), così da regolamentarla. In questa linea, **nel 2019 l'Unione Europea si è dotata di un Codice etico sull'AI**, che individua i principi fondamentali in base ai quali il comportamento delle macchine intelligenti deve essere programmato: rispetto dell'autonomia umana, prevenzione dei danni, equità, esplicabilità (cfr AI HLEG 2019). Questa scelta non considera i robot intelligenti «avversari evolutivi dell'*homo sapiens* bensì strumenti (artefatti) che devono essere pensati come cooperativi alla persona» (Benanti 2018, 113). A questo punto si pone la domanda conclusiva sulla differenza antropologica tra persone e macchine intelligenti.

Soggetti e persone

L'aumento della complessità degli strumenti digitali porta con sé un'accresciuta difficoltà di prevederne il funzionamento e di attribuire la responsabilità delle operazioni che essi svolgono. Nascono così i "buchi dell'attribuzione di responsabilità" (*responsibility gap*). In questa situazione, è importante riacquisire un concetto di responsabilità che vada al di là della dimensione, meramente fattuale, del riportare un effetto a una causa. Questa interpretazione riduttiva ha condotto a una crisi della responsabilità, con il risultato di impoverire la nostra esperienza morale.

È questa deriva che papa Francesco denuncia, quando afferma che l'«antropocentrismo moderno, paradossalmente, ha finito per collocare la ragione tecnica al di sopra della realtà, perché questo essere umano "non sente più la natura né come norma valida, né come vivente rifugio"» (LS, n. 115). Si verifica così un paradosso: «la macchina si umanizza non meno di quanto l'uomo si macchinizzi» (Benanti 2016, 63).

Per comprendere correttamente i termini della responsabilità umana, dobbiamo prima chiarire che cosa significa che l'essere umano è persona. È questo, infatti, l'assunto antropologico fondamentale, sotteso anche alla tradizione filosofico-teologica e alla dottrina sociale della Chiesa. Il confronto con l'AI può aiutarci in questo compito. **Le macchine intelligenti e agenti esprimono quella che senz'altro possiamo chiamare soggettività, cioè la capacità di elaborare determinate informazioni e di rispondere autonomamente**. In questo senso, possiamo affermare senza dubbio che l'AI è un soggetto. **Tuttavia un soggetto, descritto in questi termini, è lontano dal realizzare la specificità che la tradizione filosofica e teologica cattolica attribuisce alla nozione di persona**. Essa si realizza nella capacità dell'essere umano di distinguere il bene dal male, detta in termini tecnici sinderesi. Questa facoltà è anteriore alla coscienza morale riflessa e ne costituisce il presupposto. La possibilità di deliberare moralmente, infatti, non potrebbe sussistere se non vi fosse, al di sotto di essa, questa capacità intuitiva dei principi universali dell'ordine morale. È a questo livello che si radica il concetto di persona. Ciò che la persona è non si identifica con l'autoconsapevolezza, la libertà o la capacità di scelta come caratteristiche della soggettività, ma è il presupposto indisponibile di queste capacità.

Tornando al confronto con l'AI, possiamo per la prima volta distinguere fra soggettività e persona, concetti che sono stati spesso assimilati; arriviamo così ad affermare che può darsi soggettività senza persona (la macchina pensante e agente), ma che l'agire di una tale soggettività è svuotato di significato morale, in quanto non c'è vera conoscenza dei principi etici universali. Al contrario, l'essere umano è considerato persona, quindi titolare di un'esperienza morale, anche se è privo delle sue facoltà razionali e di scelta autonoma, come può verificarsi nel caso di un grave handicap cognitivo.

Da questo ragionamento segue anche che difficilmente è possibile immaginare una vita personale senza che l'intelligenza sia realizzata in un corpo biologico dotato di metabolismo e quindi di vita reale (cfr Boden 2019, 141); detto in altri termini: **la macchina può esprimere esteriormente un modello umano di ragionamento e comportamento etico, ma senza realizzarlo veramente in se stessa**.

Possiamo allora affermare che le macchine autonome dotate di AI sono agenti morali solo in senso improprio: possono realizzare istanze di decisione, ma non ha senso attribuire loro una responsabilità morale. **Solo l'essere umano, in quanto moralmente responsabile, è un agente morale in senso proprio.**

Ecco chiarita anche la differenza fra il *problem solving* e il fatto di affrontare un problema morale. Il primo è una procedura di ottimizzazione dei risultati, che può essere svolta da un soggetto artificiale sulla base di parametri previamente assegnati, forse anche in modo più efficiente di quanto un essere umano farebbe. Invece, si dà un problema morale quando una persona mette in gioco la propria intuizione, il proprio libero arbitrio e la propria responsabilità per rispondere a una situazione data. Per questo motivo le macchine non devono acquisire potere decisionale quando si tratta della vita umana: devono essere poste al servizio dell'autodeterminazione umana e non restringerla, per assicurare sempre l'ultima istanza della responsabilità umana.

Marvin Minsky, matematico e informatico statunitense, affermò che «se può farlo una macchina, allora non è una cosa intelligente», sottolineando così la differenza qualitativa fra l'intelligenza umana e quella artificiale. Possiamo riprendere questa frase in chiave etica, dicendo che «se può risolverlo il computer, allora non è più un problema morale» (Brand 2018, 143, nostra trad.).

AI HLEG = Gruppo indipendente di Esperti ad Alto Livello sull'Intelligenza Artificiale (2019), *Orientamenti etici per un'IA affidabile*, in <https://ec.europa.eu/futurium/en/ai-alliance-consultation>.

Benanti P. (2018), *Le macchine sapienti. Intelligenze artificiali e decisioni umane*, Marietti 1820, Bologna.

— (2016), *La condizione tecno-umana. Domande di senso nell'era della tecnologia*, EDB, Bologna.

Boden M.A. (2019), *L'intelligenza artificiale*, il Mulino, Bologna.

Brand L. (2018), *Künstliche Tugend. Roboter als moralische Akteure*, Pustet, Regensburg.

Cerruti M. (2018), «Tra tecnologia ed etica: i veicoli a guida autonoma», in *Aggiornamenti sociali*, 8-9, 570-578.

Francesco (2018), *Messaggio al Presidente esecutivo del World Economic Forum (Davos-Klosters, 23-26 gennaio)*, 24 gennaio.

Giovanni Paolo II (1990), *Discorso in occasione della V Conferenza Internazionale su "La mente umana"*, 17 novembre.

Kaplan J. (2018), *Intelligenza artificiale. Guida al futuro prossimo*, LUISS University Press, Roma.

Kurzweil R. (2014), *La singolarità è vicina*, Maggioli, Santarcangelo di Romagna (RN).

Lenzen M. (2018), *Künstliche Intelligenz. Was sie kann & was uns erwartet*, Beck, München.

LS = Francesco (2015), Lettera enciclica *Laudato si'*.

Nida-Rümelin J. – Weidenfeld N. (2018), *Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz*, Piper, München.

Ohly L. – Wellhöfer C. (2017), *Ethik im Cyberspace*, Peter Lang, Frankfurt a. M.

Peiser J. (2019), «The Rise of the Robot Reporter», in *The New York Times*, 5 febbraio.

Rasetti M. (2018), «Le grandi sfide della nuova cultura tecnologica: scienza ed etica di digitale, Big data, intelligenza artificiale», in *Cultura musicale e nuova cultura tecnologica. Caligara Lectures 2016/2017*, Giappichelli, Torino, 31-68.

Searle J.R. (1980), «Minds, brains, and programs», in *The Behavioral and brain science*, 3, 417-457.

Zamagni S. (2019), *Responsabili. Come civilizzare il mercato*, il Mulino, Bologna.