S&V FOCUS | Intelligenza artificiale. Cosa succederà quando la macchina dovrà compiere una scelta morale, in situazione critiche?

I rapidi progressi nel campo dell'intelligenza artificiale e della robotica hanno consentito lo sviluppo di macchine sempre più sofisticate e intelligenti, in grado di agire anche come agenti morali, prendendo decisioni che riguardano dilemmi e problematiche umane molto complesse, senza la supervisione umana.

Da alcuni anni la ricerca tenta di sviluppare, ad esempio, delle *self driving cars*, veicoli senza conducente che, grazie all''intelligenza artificiale, saranno capaci di compiere spostamenti con una percezione autonoma dell'ambiente che li circonda, quasi senza ausilio dell'essere umano. Cosa

succederà quando la macchina dovrà compiere una scelta morale, in situazione critiche? Il veicolo automatizzato, ad esempio, potrebbe dare priorità alla sicurezza dei propri passeggeri rispetto a quella dei pedoni. Tali veicoli sono tra le prime macchine create dall'uomo in grado di prendere decisioni autonome con potenziali conseguenze sulla vita o sulla morte dell'essere umano. Ciò segna, inevitabilmente, un punto di svolta, per il cambiamento qualitativo delle conseguenze delle scelte progettuali. Stessa problematica può applicarsi anche al campo delle diagnosi mediche assistite dall'intelligenza artificiale o in quello dell'uso di algoritmi predittivi per le attività di polizia o militari.

In questa prospettiva, **la macchina potrà essere considerata un agente morale anche quando la sua programmazione non codifica esplicitamente valori morali, ma quando le conseguenze delle sue azioni ricadranno nell'ambito morale: si parla a tal proposito di agente morale "implicito".**

Il caso tipico è quello di una macchina i cui errori possono creare danni: in campo medico, ad esempio, l'intelligenza artificiale può fornire una diagnosi sbagliata: l'errore ha implicazioni anche in ambito morale. Vi sono, poi, anche macchine che incideranno sempre più "esplicitamente" nel campo morale, come nel caso della macchina che dovrà prendere una decisione nel conflitto tra due principi: ad esempio, una macchina che esegue la moderazione dei contenuti online, a una velocità tale da impedire un controllo umano continuo, potrebbe dover scegliere tra il valore della libertà di manifestazione del pensiero e il dovere di sopprimere contenuti offensivi o dannosi; altro esempio è la decisione in merito all'allocazione di una risorsa scarsa, con conseguenze dannose per gli esseri umani a cui non viene data priorità.

Le macchine saranno, così, implicitamente o esplicitamente, incaricate di collaborare o di prendere decisioni con conseguenze in campo morale, spesso su questioni rilevanti per la vita dell'essere umano (come, ad esempio, allocare un rene

prelevato per finalità di trapianto, o in materia di sicurezza autostradale): come conciliare in questi ambiti così delicati l'autonomia della macchina intelligente e la tutela dell'essere umano e della sua dignità?

Un articolo, dal titolo "*The Moral Psychology of Artificial Intelligence*", pubblicato recentemente su *Annual Review of Psychology*, approfondisce nel dettaglio la tematica: anche la psicologia morale ha iniziato a interessarsi alle macchine intelligenti. Nello scritto si afferma che delegare decisioni morali difficili alle macchine potrebbe, innanzitutto, rimuovere un peso psicologico per gli esseri umani, ma anche influenzare le "aspettative sociali" legate alla decisione. La conseguenza dannosa dovuta a una scelta morale difficile, infatti, avrebbe meno ripercussioni sulla macchina, rispetto alla scelta presa da un essere umano: ad esempio, nel dilemma in cui l'agente deve decidere se salvare diverse vite sacrificandone una, gli esseri umani verrebbero maggiormente ritenuti responsabili per la scelta, una volta presa, e ritenuti "colpevoli" del sacrificio, mentre questo non accadrebbe con la macchina. Si sostiene, inoltre, che per garantire che le macchine risolvano i dilemmi morali in modo compatibile con gli obiettivi e le priorità degli esseri umani, si dovrebbe trovare un accordo su quali siano questi obiettivi e priorità comuni e trovare, poi, un modo per trasmetterlo alle macchine. Tale accordo però non è sempre semplice. Inoltre, le macchine morali esplicite potrebbero dover bilanciare valori o priorità contrastanti. Ad esempio, gli algoritmi per la donazione di rene da vivente – si pensi al caso del trapianto *cross-over* – possono operare il bilanciamento tra diverse priorità.

Inoltre, si segnala il rischio che l'intelligenza artificiale possa essere utilizzata da persone che hanno intenzioni dannose per incrementare comportamenti criminali o non etici. I recenti progressi nel *deep learning* hanno reso semplice la creazione di contenuti falsi che sembrano autentici. Quando le

persone delegano compiti ad agenti di intelligenza artificiale, ciò crea una combinazione di fattori psicologici che possono portare a comportamenti non etici, come l'anonimato, distanza psicologica dalle vittime, non rilevabilità e ambiguità. Lasciare che tali algoritmi eseguano compiti per proprio conto offusca l'attribuzione di responsabilità per il danno causato. Questo offre un nuovo campo di indagine, anche per la psicologia morale.

Più in generale, il rischio è anche la poca chiarezza relativa ai criteri decisionali delle macchine, che possono amplificare pregiudizi e discriminazioni, presenti nei dati di addestramento, pregiudicando la correttezza, la trasparenza e l'equità delle decisioni prese dalle macchine, soprattutto quando hanno conseguenze morali, sociali e legali. Cosa fare quando l'efficacia di un sistema decisionale comporta l'opacità dei suoi criteri di scelta? Le macchine intelligenti sono, poi, spesso "scatole nere": non essendo del tutto chiara l'elaborazione operata degli input per arrivare a una decisione, anche per coloro che le hanno programmate. Inoltre, potrebbero apprendere e modificare costantemente le proprie capacità percettive o processi decisionali.

L'Unione Europea, riconoscendo l'importanza di governare la trasformazione digitale, ha l'obiettivo di individuare regole condivise capaci di garantire uno sviluppo dell'intelligenza artificiale sicura e affidabile, individuando alcuni valori-guida, quali equità, privacy, dignità umana, responsabilità, spiegabilità, non discriminazione, controllo umano significativo. È, però, problematico tradurre tali principi in regole applicative condivise, soprattutto nei casi di potenziale conflitto tra principi, come nel caso di conflitto tra sicurezza pubblica e privacy. L'approccio seguito è basato sulla gestione del rischio, ponendo l'umano al centro del cambiamento tecnologico. È, infatti, la tecnologia che al servizio dell'uomo e non il contrario.

Soprattutto quando la tecnologia si muove nell'ambito dei

dilemmi morali il rischio è la delega dei compiti decisionali alle macchine, dimenticando la responsabilità umana. Si pensi, ad esempio, al caso dell'uso dell'IA in contesti militari: anche solo la distanza rispetto alle conseguenze prodotte sembra rendere l'uomo quasi meno responsabile del male generato.

Il campo della ricerca per la realizzazione di macchine intelligenti, che possano agire nel rispetto e al servizio dell'umano, non può essere solamente quello tecnico – anche in considerazione delle conseguenze sempre più incisive della tecnologia sulla vita umana – e richiede, pertanto, un approccio interdisciplinare, che coniuga discipline scientifiche e umanistiche: anche la bioetica è chiamata a investigare in questo ambito, nel tentativo di comporre conoscenze scientifiche e tecniche con quelle antropologiche, indispensabili per porre l'umano al centro della rivoluzione tecnologica.

Per approfondire:

1. Bonnefon, Jean-Francois and Rahwan, Iyad and Shariff, Azim, The Moral Psychology of Artificial Intelligence (January 2024). Annual Review of Psychology, Vol. 75, pp. 653-675, 2024
2. Awad E, Dsouza S, Bonnefon JF, Shariff A, Rahwan I. *Crowdsourcing Moral Machines*, 2020