L'IA e i robot sono un campo minato di pregiudizi cognitivi.
Noi umani antropomorfizziamo la nostra tecnologia, a volte a nostro danno

La rivista della IEEE, *Spectrum* del 19 luglio 2021 ha pubblicato un'illuminante intervista a Sangbae Kim, direttore del Laboratorio di robotica biomimetica al MIT, in cui descrive i pregiudizi che possono essere incorporati nell'Intelligenza Artificiale e nei robot.
Riportammo alcuni passaggi e ringraziamo Sangbae Kim di averci permesso la pubblicazione.
Il Laboratorio di robotica biomimetica del MIT si occupa della progettazione e controllo dei robot ispirati al mondo naturale e ha realizzato recentememte il robot Mini Cheeta che attua degli incredibili salti all'indietro (https://biomimetics.mit.edu/
Per l'intero articolo: https://spectrum.ieee.org/humans-cognitive-biases-facing-ai (la traduzione è nostra).

Giudichiamo i compiti dei robot o dell'IA paragonandoli a quelli degli umani
di Sangbae Kim

La maggior parte delle persone associa l'intelligenza artificiale ai robot come una coppia inseparabile. Infatti, il termine "intelligenza artificiale" è raramente usato nei laboratori di ricerca. La terminologia specifica per certi tipi di AI e altre tecnologie intelligenti è più accurata e ogni volta che mi viene posta la domanda "Questo robot è gestito dall'IA?", esito a rispondere, chiedendomi se sia appropriato chiamare gli algoritmi che sviluppiamo "intelligenza artificiale".
Sviluppata da John McCarthy e Marvin Minsky negli anni '50, e apparsa in romanzi o film di fantascienza per decenni, Intelligenza Artificiale è oggi utilizzata negli assistenti virtuali degli smartphone e negli algoritmi dei veicoli autonomi. Sia storicamente che oggi, AI può significare molte cose diverse, il che può causare confusione. Infatti, sento spesso esprimere il preconcetto che l'IA sia una versione realizzata artificialmente dell'intelligenza umana. E questo preconcetto potrebbe derivare da un nostro pregiudizio cognitivo.

Credo che molti siano rimasti impressionati nel vedere il Mini Cheetah, sviluppato nel mio laboratorio di robotica biomimetica del MIT, eseguire un salto mortale all'indietro. Mentre il salto all'indietro e l'atterraggio sono molto dinamici, accattivant' e, naturalmente, difficili per gli esseri umani, l'algoritmo per questo particolare movimento è incredibilmente semplice rispetto a quello che consente una camminata bipede stabile che richiede cicli di feedback molto più complessi. Realizzare compiti robotici che sono apparentemente facili per noi è spesso estremamente difficile e complicato. Questo divario si verifica perché tendiamo a pensare alla difficoltà di un compito sulla base di standard umani.
Tendiamo a generalizzare la funzionalità dell'IA dopo aver visto una singola dimostrazione di robot. Quando vediamo qualcuno per strada che fa salti mortali all'indietro, tendiamo a supporre che questa persona sia brava a camminare e a correre, e che sia anche abbastanza flessibile e atletica da

essere brava in altri sport. Molto probabilmente, tale giudizio su questa persona non sarebbe sbagliato.

Tuttavia, possiamo applicare questo giudizio anche ai robot? È facile per noi generalizzare e determinare le prestazioni dell'IA sulla base dell'osservazione di un movimento o di una funzione specifica del robot, proprio come facciamo con gli umani. Guardando un video di un robot che risolve con le mani artificiali il cubo di Rubik all'OpenAI, un laboratorio di ricerca sull'IA, pensiamo che l'IA possa eseguire tutti gli altri compiti più semplici perché può eseguirne uno così complesso. Trascuriamo il fatto che la rete neurale di questa IA è stata addestrata solo per un tipo limitato di compito: risolvere il cubo di Rubik in quella configurazione. Se la situazione cambia - per esempio, tenendo il cubo al contrario mentre lo si manipola – l'algoritmo non funzionerà così bene come ci si potrebbe aspettare.

A differenza dell'IA, gli esseri umani possono combinare abilità individuali e applicarle a più compiti complicati. Una volta che abbiamo imparato a risolvere un cubo di Rubik, possiamo lavorare rapidamente sul cubo anche quando ci viene detto di tenerlo capovolto, anche se all'inizio può sembrare strano. L'intelligenza umana può combinare naturalmente gli obiettivi di non far cadere il cubo e di risolverlo mentre la maggior parte degli algoritmi robotici richiedono nuovi dati o riprogrammazione per farlo. Una persona che può spalmare la marmellata sul pane con un cucchiaio può fare lo stesso con una forchetta. È ovvio. Capiamo il concetto di "spalmare" la marmellata, e possiamo rapidamente abituarci a usare uno strumento completamente diverso. Inoltre, mentre i veicoli autonomi richiedono dati reali per ogni situazione, i conducenti umani possono prendere decisioni razionali basate su concetti preappresi per rispondere a innumerevoli situazioni. Questi esempi mostrano una caratteristica dell'intelligenza umana in netto contrasto con gli algoritmi dei robot, che non possono eseguire compiti con dati insufficienti.

I mammiferi si sono evoluti per più di 65 milioni di anni. E l'arco di tempo che gli esseri umani hanno speso per imparare la matematica, usare le lingue e giocare ammonterebbe a soli 10.000 anni. In altre parole, l'umanità ha speso una quantità enorme di tempo per sviluppare abilità direttamente legate alla sopravvivenza, come camminare, correre e usare le mani. Pertanto, non può essere sorprendente che i computer possano calcolare molto più velocemente degli umani, dato che sono stati sviluppati principalmente per questo scopo. Allo stesso modo, è naturale che i computer non possano ottenere facilmente la capacità di usare liberamente mani e piedi per vari compiti, come fanno gli umani. Queste abilità sono state raggiunte attraverso l'evoluzione per oltre 10 milioni di anni.
Questo è il motivo per cui non è corretto paragonare le prestazioni dei robot o dell'IA a quelle di un animale o di un umano. Così, sarebbe un errore credere che le i robot che vediamo camminare e correre, come Cheeta del MIT, possano fare molto altro così facilmente. Numerose dimostrazioni di robot si basano ancora su algoritmi impostati per compiti specializzati in situazioni limitate e i ricercatori tendono a selezionare per il pubblico quelle dimostrazioni che sembrano difficili, per produrre una forte impressione. Tuttavia, siamo noi umani che valutiamo difficili quelle prestazioni, che dal punto di vista dell'algoritmo non sono così difficili.

Noi umani siamo influenzati più dalle percezioni immediate che dal pensiero logico. Questo bias cognitivo si amplifica quando l'oggetto dell'esperienza è molto complicato e difficile da analizzare logicamente - per esempio, un robot che utilizza il machine learning.
Da dove deriva questo pregiudizio cognitivo? Credo dalla nostra tendenza psicologica ad antropomorfizzare inconsciamente i soggetti che percepiamo. Gli umani si sono evoluti come animali sociali, probabilmente sviluppando la capacità di comprendere ed entrare in empatia con gli altri. La nostra tendenza ad antropomorfizzare i soggetti deriverebbe dallo stesso processo evolutivo. (..) Come disse il filosofo del XVIII secolo David Hume, "Esiste una tendenza universale tra gli uomini a concepire tutti gli esseri come noi stessi. Individuiamo un volto umano nella luna, eserciti nelle nuvole".

Gli umani elaborano le informazioni in modo qualitativo e i computer in modo quantitativo
La nostra vita quotidiana è piena di algoritmi, come si può vedere dalle macchine e dai servizi che funzionano su questi algoritmi. (..) L'obiettivo di compiti come vincere una partita di Go o di scacchi sono relativamente facili da quantificare. Più facile è la quantificazione, meglio funzionano gli algoritmi. Al contrario, gli umani spesso prendono decisioni senza pensare quantitativamente.
Consideriamo, ad esempio, l'azione di pulire una stanza. Il modo in cui puliamo differisce un poco a seconda della situazione, a seconda del proprietario della stessa, di come ci si sente. Possiamo pensare di massimizzare una certa funzione in questo processo? No, l'azione di pulire si basa sull'obiettivo astratto di "pulire abbastanza". Inoltre, lo standard di quanto è "abbastanza" cambia facilmente e questo standard può essere diverso da persona a persona, causando conflitti soprattutto tra membri della famiglia o tra coinquilini.

Ci sono molti altri esempi. Quando vi lavate la faccia ogni giorno, quali indicatori quantitativi intendete massimizzare con i vostri movimenti delle mani? Con quanta forza strofinate? Quando scegliete cosa indossare? Quando scegliete cosa mangiare per cena? Quando si sceglie quale piatto lavare per primo? La lista continua. Siamo abituati a prendere decisioni abbastanza buone mettendo insieme le informazioni che già abbiamo. Tuttavia, spesso non controlliamo se ogni singola decisione è ottimizzata. Il più delle volte è impossibile saperlo, perché dovremmo soddisfare numerosi indicatori contraddittori con dati limitati. Quando si sceglie la spesa con un amico al negozio, non possiamo quantificare per ognuno gli standard della spesa e prendere una decisione in base a questi valori numerici. Di solito, quando uno sceglie qualcosa, l'altro dirà, Va bene, o suggerirà un'altra opzione. Questo è molto diverso dal dire che una data verdura è la "scelta ottimale", ma corrisponderà le a "questa è abbastanza buona".

Questa differenza operativa tra le persone e gli algoritmi può causare problemi quando si progettano lavori o servizi per i robot. Questo perché mentre gli algoritmi eseguono compiti basati su valori quantitativi, la soddisfazione degli umani, il risultato del compito, è difficile da quantificare completamente. Non è facile quantificare l'obiettivo di un compito che deve adattarsi alle preferenze individuali o alle circostanze mutevoli, come i compiti di pulizia della stanza o di lavaggio dei piatti. Questo significa che per coesistere con gli umani, i robot potrebbero evolversi non per ottimizzare

particolari funzioni, ma per raggiungere un “abbastanza buono” umano. Naturalmente, quest’ultimo è molto più difficile da raggiungere in modo preciso, in situazioni di vita reale in cui è necessario gestire così tanti obiettivi contrastanti e con tanti vincoli qualitativi.

In realtà, non sappiamo cosa stiamo facendo

Prova a ricordare il pasto più recente che hai fatto prima di leggere questo articolo. Riuscite a ricordare che cosa avete mangiato? Poi, potete anche ricordare il processo di masticazione e deglutizione del cibo? Sai cosa stava facendo esattamente la tua lingua in quel momento? La nostra lingua fa così tante cose per noi. Ci aiuta a mettere il cibo in bocca, a distribuirlo tra i denti, a ingoiare i pezzi finemente masticati o anche a rimandare pezzi grandi verso i denti, se necessario. Possiamo fare naturalmente tutto questo, anche mentre parliamo con un amico, usando la lingua. Quanto contribuiscono le nostre decisioni coscienti al movimento della nostra lingua che compie così tanti compiti complessi simultaneamente? Può sembrare che stiamo muovendo la lingua come vogliamo, ma in realtà, ci sono più momenti in cui la lingua si muove automaticamente, assumendo comandi di alto livello dalla nostra coscienza. Questo è il motivo per cui non possiamo ricordare i movimenti dettagliati della nostra lingua durante un pasto: perché sappiamo poco del suo movimento.

Possiamo supporre che le nostre mani siano l’organo più controllabile coscientemente, ma anche molti movimenti delle mani avvengono automaticamente e inconsciamente, o al massimo inconsciamente. Per coloro che non sono d’accordo, provate a mettere le chiavi in tasca e tiratele fuori. In quel breve momento, innumerevoli micromanipolazioni si coordinano istantaneamente e senza soluzione di continuità per completare il compito. Spesso non possiamo percepire ogni azione separatamente. Non sappiamo nemmeno in quali unità dovremmo dividerle, quindi le esprimiamo collettivamente come parole astratte come organizzare, lavare, applicare, strofinare, pulire, ecc. Questi verbi sono definiti qualitativamente. Si riferiscono spesso all’aggregato di movimenti e manipolazioni fini, la cui composizione cambia a seconda delle situazioni. Naturalmente, tutto ciò è facile anche per un bambino, ma dal punto di vista dello sviluppo degli algoritmi, la descrizione di queste azioni sono parole infinitamente vaghe e astratte.

Se dobbiamo descrivere come spalmare burro d’arachidi sul pane, lo possiamo fare con poche semplici parole. Supponiamo ora una situazione leggermente diversa. Diciamo che un alieno usa la nostra stessa lingua, ma non sa nulla della civiltà o della cultura umane. (So che questa ipotesi è già contraddittoria..., ma per favore abbiate pazienza). Possiamo spiegare per telefono come si fa un panino al burro d’arachidi? Probabilmente ci bloccheremo già solo cercando di spiegare come tirare fuori il burro d’arachidi dal barattolo. Anche afferrare la fetta di pane non è così semplice. Dobbiamo afferrare il pane abbastanza forte da poter spalmare il burro d’arachidi, ma non così tanto da rovinare il panino che è morbido e nello stesso tempo dobbiamo stare attenti a non far cadere il pane per terra. È facile per noi pensare a come manipolare il panino, ma non sarà facile esprimerlo attraverso la parola o il testo, figuriamoci in una funzione.

Inoltre, possiamo imparare il lavoro di un falegname per telefono? Possiamo correggere con precisione le posture del tennis o del golf per telefono? È difficile discernere fino a che punto i dettagli necessari per questi compiti siano fatti consapevolmente o inconsapevolmente.

Il mio punto è che non tutto quello che facciamo con le mani e i piedi può essere direttamente espresso con il nostro linguaggio. Le cose che accadono tra azioni successive spesso avvengono automaticamente in modo inconscio e quindi, quando descriviamo le nostre azioni, lo facciamo semplificando i processi reali. Questo è il motivo per cui le nostre azioni sembrano molto semplici, e per cui dimentichiamo quanto siano incredibili in realtà. I limiti delle nostre descrizioni portano spesso a sottovalutare la complessità reale. Dovremmo riconoscere il fatto che la difficoltà di rappresentazione mediante il linguaggio può ostacolare il progresso della ricerca in campi in cui le parole non sono ben sviluppate.

Fino a poco tempo fa, l'IA era applicata principalmente nei servizi di informazione relativi all'elaborazione dei dati. Oggi, alcuni esempi importanti includono il riconoscimento vocale e il riconoscimento facciale. E stiamo entrando in una nuova era di IA che può effettivamente eseguire servizi fisici per noi, in mezzo a noi: sta arrivando il momento in cui l'automazione di compiti fisici complessi diventa essenziale.

La nostra società sempre più anziana pone una sfida enorme: la carenza di manodopera non è più un vago problema sociale, ed è urgente discutere su come sviluppare tecnologie che aumentino le capacità degli umani, permettendoci di concentrarci su lavori più preziosi, sviluppare le nostre vite veramente come esseri umani. Questo è il motivo per cui non solo gli ingegneri, tutti dovrebbero migliorare la loro comprensione dell'IA e dei pregiudizi cognitivi inconsci: non è facile capire l'intelligenza artificiale perché è sostanzialmente diversa dall'intelligenza umana.

Concetti, processi che sono molto naturali tra gli umani possono diventare pregiudizi cognitivi nella IA e per i robot. Senza una chiara comprensione dei *nostri* pregiudizi cognitivi, non possiamo impostare le direzioni appropriate per la ricerca tecnologica, le applicazioni e gli indirizzi di policy. Per progredire come comunità scientifica, abbiamo bisogno di una grande attenzione ai nostri processi cognitivi e di un dibattito deliberato che promuova lo sviluppo corretto delle applicazioni della tecnologia.

Per l'intero articolo: https://spectrum.ieee.org/humans-cognitive-biases-facing-ai (la traduzione qui allegata è nostra).

Ringraziamo Sangbae Kim per averci concesso la pubblicazione di passi del suo saggio.