Nella scatola nera

L'idea che i sistemi d'intelligenza artificiale più avanzati siano troppo complessi per spiegarne la logica interna è falsa. Spesso è solo un modo per nascondere errori e distorsioni che incidono sulla vita delle persone

Agata Foryciarz, Daniel Leufer e Kataryna Szymielewicz, Medium, Stati Uniti

Dalle paure sui robot che uccidono al sogno di un futuro completamente automatizzato, l'intelligenza artificiale cattura la nostra immaginazione più di qualsiasi altra tecnologia. Come documentato da numerosi studiosi, l'idea di creare macchine artificiali intelligenti seduce e scandalizza gli esseri umani da millenni. In effetti ciò che rende così affascinante la storia dell'intelligenza artificiale è in parte la combinazione di progresso scientifico e creazione di miti, che qualche volta è puro inganno. In una certa misura la propaganda e il mito sono innocui e possono perfino favorire il progresso scientifico.

Tuttavia, il fatto che oggi i cosiddetti "sistemi d'intelligenza artificiale" siano integrati in una serie di servizi pubblici essenziali e in altri processi delicati deve renderci particolarmente vigili nella lotta contro gli equivoci su questa tecnologia. Nel 2019 si è parlato spesso degli utenti di Google assistant, Alexa (Amazon) e Siri (Apple). Molti sono rimasti sconvolti quando hanno scoperto che le registrazioni di alcune loro conversazioni private erano state ascoltate da persone in carne e ossa. Per chi ha familiarità con i sistemi di addestramento di questi assistenti vocali non è stata certo una sorpresa. Ma per la maggior parte degli utenti, a cui questi sistemi sono presentati come completamente automatizzati, è stato un trauma scoprire che dei lavoratori stranieri sottopagati potevano accedere a conversazioni intime. La notizia che il servizio sanitario nazionale britannico ha firmato con Amazon un accordo che autorizza Alexa a fornire consulenza medica (e, quindi, ad accedere ai dati sanitari dei pazienti) conferma le preoccupazioni sul funzionamento di questi sistemi.

Ci sono molti miti ed equivoci sull'intelligenza artificiale, ma nei casi in cui questi sistemi sono usati in contesti delicati e ad alto rischio, come la sanità e la giustizia, probabilmente l'equivoco più dannoso è che questi sistemi siano "scatole nere" di cui non possiamo sapere niente. Peggio ancora, si sente spesso ripetere che questi sistemi hanno prestazioni superiori (e questo, come sottolinea la ricercatrice Cynthia Rudin, è spesso falso) e che la richiesta di spiegazioni porta automaticamente a una minor precisione ed efficacia. Dovremmo quindi fidarci e addirittura imparare ad amare la scatola nera.

Ma cosa vuol dire che un sistema d'intelligenza artificiale è una scatola nera? Nella guida *Explaining Ai decisions* (di se-

Le foto che illustrano quest'articolo sono state realizzate da Generated photos, un progetto a cui partecipano fotografi ed esperti d'intelligenza artificiale. Partendo da foto di persone comuni, il team di Generated photos crea volti digitali di persone inesistenti. Le immagini sono impiegate per realizzare applicazioni di ricerca o in software aziendali.

GENERATED PHOTOS

guito *Ico-At guidance*), l'Information commissioner's office (Ico, l'agenzia del governo britannico che si occupa della privacy) e l'Alan Turing institute definiscono scatola nera "qualsiasi sistema d'intelligenza artificiale i cui meccanismi e la cui logica interna siano opachi o inaccessibili alla comprensione umana". Questa opacità, tuttavia, ha diverse cause.

In genere, dal punto di vista tecnico si definiscono "scatole nere" alcuni sistemi d'intelligenza artificiale (per esempio le reti neurali) che operano in modo troppo complesso perché un essere umano possa seguirne l'attività nel dettaglio o ricostruirla a posteriori. Ma un sistema può diventare una "scatola nera" anche per questioni legali: ci viene impedito di sapere come funziona per proteggere segreti industriali o per evitare che il sistema sia manipolato (questa possibilità è spesso la spia di un sistema arbitrario e progettato male). Esistono, naturalmente, sistemi di *deep learning* che combinano entrambe le forme di opacità.

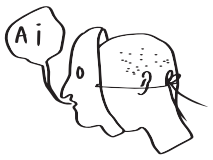
In quest'articolo vogliamo dimostrare che nessuna forma di opacità è del tutto inevitabile, e che anche quando in un sistema resta una certa opacità, ci sono misure di trasparenza che possono e devono essere introdotte per rimuoverla.

Un nuovo nome per la statistica

Prima di addentrarci nelle cause tecniche e a volte anche legali dell'opacità dell'intelligenza artificiale, dobbiamo chiarire un punto fondamentale: è opaca la definizione stessa d'intelligenza artificiale. Oggi questa espressione abbraccia una gamma vastissima di tecnologie, dalle più banali alle più inquietanti. Da una parte, tutto questo parlare d'intelligenza artificiale ha portato le aziende a includere nella categoria anche funzionalità come i suggerimenti su PowerPoint. Dall'altra abbiamo le speculazioni su forme superintelligenti "non allineate" ai valori umani che combattono specie aliene. Cos'hanno in comune tra loro (ammesso che ce l'abbiano) le tecnologie che oggi rientrano nella definizione d'intelligenza artificiale?

Alla base di ogni progetto c'è l'idea di creare una macchina in grado di svolgere un compito complesso. In genere i compiti complessi sono quelli che richiedono l'intervento dell'intelligenza umana, ma non bisogna fare l'errore di descrivere l'intelligenza artificiale solo in termini di imitazione dell'intelligenza umana. Possiamo sviluppare macchine che replicano i meccanismi del cervello umano per capire meglio come funziona la nostra mente, ma l'obiettivo potrebbe essere anche un altro: per esempio creare una macchina in grado di risolvere problemi che per un essere umano sarebbero insormontabili. In questo caso avremmo a che fare con un'intelligenza non umana. Nell'ipotesi puramente teorica in cui l'obiettivo fosse sviluppare un sistema d'intelligenza artificiale con una reale consapevolezza di sé o coscienza, si potrebbe perfino parlare di scopi, obiettivi o intenzioni del sistema stesso. Naturalmente non c'è niente di simile all'orizzonte, resta un'ipotesi fantascientifica.

Anche quando in un sistema resta una certa opacità, ci sono misure di trasparenza che possono e devono essere introdotte per rimuoverla



Da sapere
Distopie europee

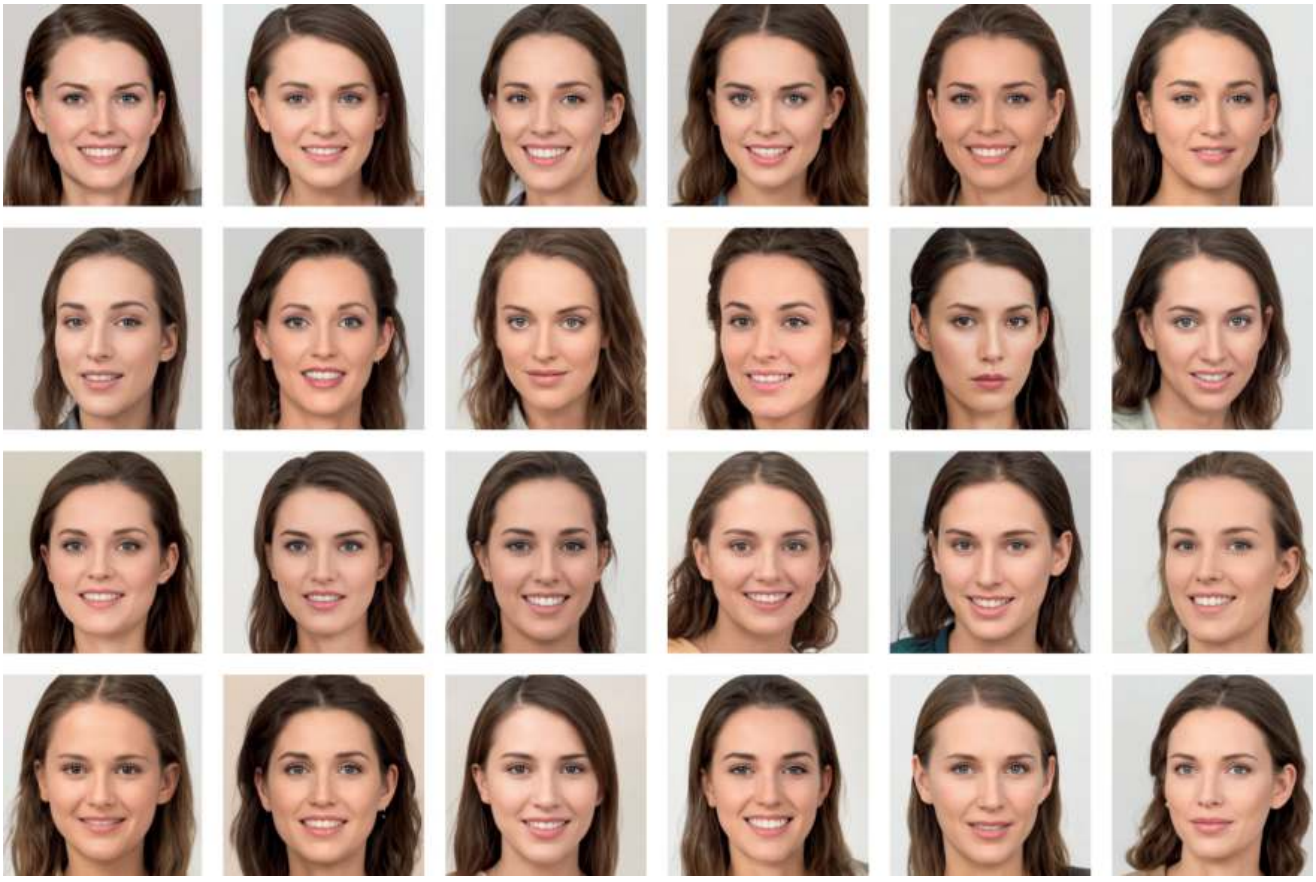
◆ Di solito la discussione sulle conseguenze negative dell'intelligenza artificiale si concentrano sulla Cina o sugli Stati Uniti. In realtà anche in Europa ci sono esempi di sistemi d'intelligenza artificiale discutibili. Queste innovazioni sono destinate ad aumentare soprattutto grazie ad alcuni programmi della Commissione europea. Per esempio l'**Horizon 2020**, che tra il 2014 e il 2020 ha stanziato quasi 80 miliardi di euro e che, secondo indiscrezioni, dovrebbe essere rafforzato e prolungato. Tra i sistemi più contestati promossi dal programma c'è **Sewa** (Automatic sentiment analysis in the wild). Con un finanziamento di 3,6 milioni di euro, questo progetto si propone di sviluppare una tecnologia che legge i sentimenti umani per perfezionare la pubblicità mirata, usando un "motore di annunci" basato sui profili emotivi e dei comportamenti d'acquisto dei consumatori. **Euractiv**

I modelli d'intelligenza artificiale di cui si parla tanto oggi sono semplici modelli statistici (noti anche come modelli di apprendimento automatico), non diversi da quelli usati da anni da sociologi, biologi, statistici e psicologi per attività come la previsione dei valori futuri delle azioni in borsa, la stima degli effetti delle terapie sulla salute o il distinguere testo e immagini.

La differenza principale, in questa nuova "era dell'intelligenza artificiale", è la maggiore efficacia assicurata da alcuni modelli di apprendimento automatico grazie ai progressi tecnologici nella potenza di calcolo e nell'accesso a enormi quantità di dati. Tra questi modelli ci sono le reti neurali, algoritmi studiati fin dagli anni cinquanta ma che solo di recente hanno trovato un'applicazione pratica in attività come il riconoscimento delle immagini e la traduzione automatica. Ma anche altri modelli statistici che esistono da decenni – come la regressione lineare o logistica, gli alberi decisionali e le macchine a vettori di supporto – sono stati riclassificati come intelligenza artificiale, suscitando un rinnovato entusiasmo anche se il loro uso e la loro efficacia restano sostanzialmente identici.

Molte applicazioni accusate di essere palesemente distorte in realtà non erano scatole nere in senso tecnico (cioè non si basavano su reti neurali o altre tecniche di apprendimento automatico molto complesse). In questi casi abbiamo a che fare con tecniche interpretabili e molto meno sofisticate. C'è sicuramente un elemento di mistificazione quando queste tecniche sono inglobate nella definizione generica d'intelligenza artificiale. La prima domanda che dovremmo fare di fronte a un sistema d'intelligenza artificiale è quali tecniche sta impiegando. Un errore comune tra i commentatori e i giornalisti non esperti è quello di raffigurare come una "scatola nera" sia i sistemi semplici sia quelli complessi. La conseguenza è che si giustifica anche l'opacità di chi progetta sistemi semplici. In molti casi l'opinione pubblica è tenuta all'oscuro perché la trasparenza metterebbe a rischio segreti industriali o svelerebbe le scelte discutibili dei proprietari del sistema.

Questa dinamica è già di per sé una buona ragione per mettere in discussione la logica della "scatola nera" ed educare le persone in modo che non tutti i modelli statistici siano messi sullo stesso piano. Tenendo presente quindi che l'intelligenza artificiale attuale – e di fatto l'unica che



GENERATED PHOTOS

abbia un orizzonte di sviluppo realistico – è formata da una serie di modelli statistici avanzati, esaminiamo innanzitutto in che modo i fattori non tecnici possono trasformare in scatole nere sistemi d'intelligenza artificiale potenzialmente interpretabili.

iBorderCtrl

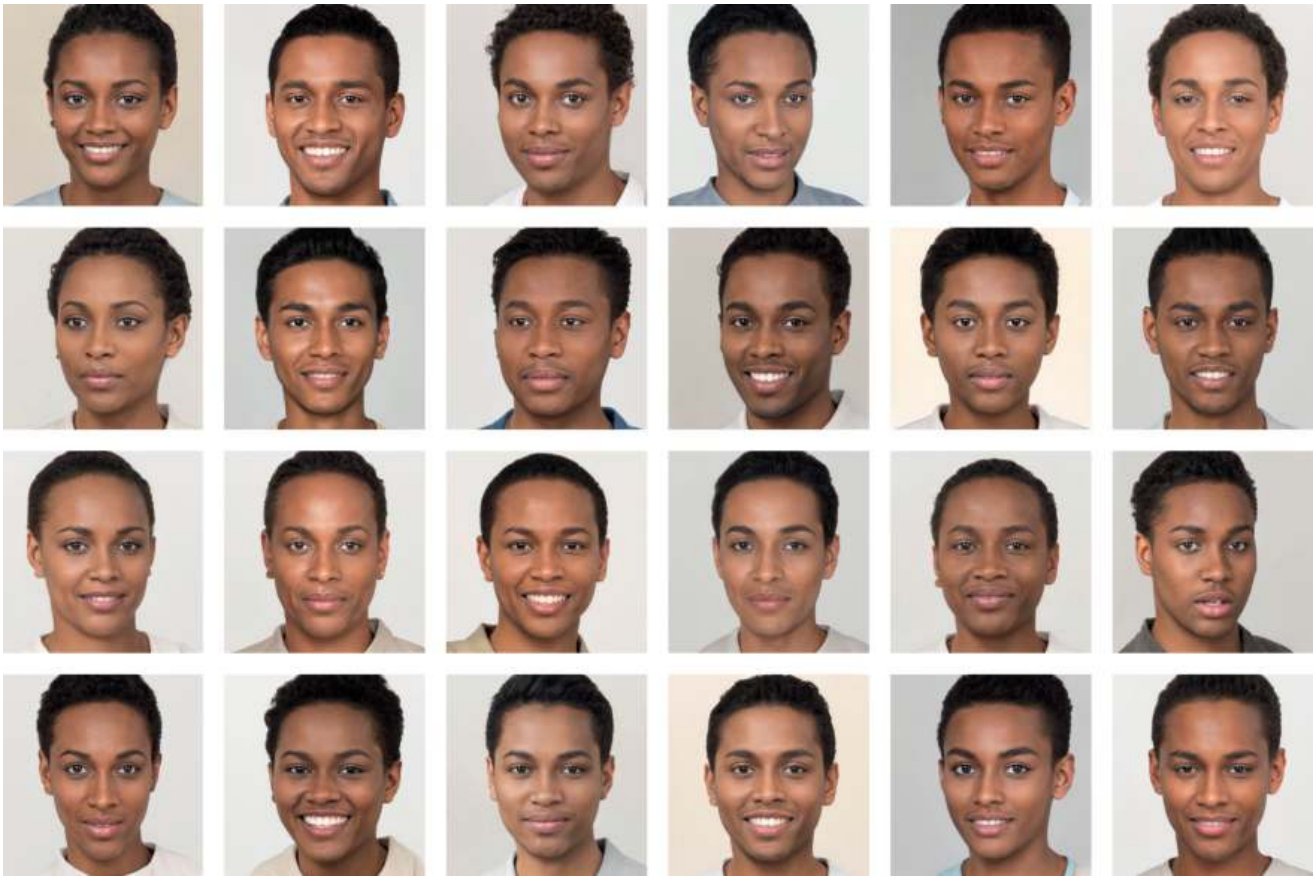
Un esempio perfetto di opacità non tecnica è iBorderCtrl, un progetto finanziato da Horizon 2020, il programma per la ricerca e l'innovazione della Commissione europea. Il sistema iBorderCtrl fornisce un servizio di *lie detecting* (individuazione delle bugie) che ha l'obiettivo di sorvegliare i confini dell'Unione europea. Diversi commentatori hanno sottolineato la mancanza di basi scientifiche per una tecnologia simile. Patrick Breyer, eurodeputato del Partito pirata e attivista per le libertà civili, ha chiesto di rendere pubblici alcuni documenti, tra cui una valutazione etica, ma la sua richiesta è stata rifiutata, ha raccontato lo stesso Breyer, "perché si tratta di 'informazioni commerciali' delle aziende coinvolte che hanno un 'valore commerciale'". Una tecnologia discutibile e scientificamente dubbia, quindi, è stata usata in un contesto molto delicato (il monitoraggio dei flussi migratori) e finanziata con fondi pubblici. Il fatto che un'azienda che fornisce una tecnologia simile possa sfuggire al controllo pubblico sembra contrario al buon senso.

Né la commissione europea né gli sviluppatori di iBorderCtrl hanno fornito elementi concreti per giustificare il fatto che non si possa indagare sul funzionamento del loro software. Il sistema diventa una "scatola nera" per il timore che possano essere rivelati dei segreti industriali. Il caso è particolarmente rilevante, visto che parliamo di una tecnologia molto discutibile che contiene un rischio elevato di discriminazioni e ingiustizie. Tra l'altro, questa evidente mancanza di trasparenza contraddice il concetto di "intelligenza artificiale fidata" che l'Unione europea sembra così ansiosa di promuovere.

In contrasto con la scelta di nascondere le valutazioni etiche al controllo pubblico, il consiglio d'Europa ha raccomandato che le valutazioni d'impatto sui diritti umani (*human rights impact assessment*, Hria) siano fatte da organismi pubblici e rese disponibili a tutti: "Le autorità non dovrebbero acquisire sistemi d'intelligenza artificiale di terze parti quando queste non sono disposte a rinunciare alle restrizioni sulle informazioni (per esempio clausole di riservatezza o di protezione del segreto industriale) che impediscono o vanificano il processo di valutazione dell'impatto sui diritti umani (compreso lo svolgimento di attività di ricerca o revisione esterne) e la divulgazione della valutazione d'impatto al pubblico".

Queste misure sarebbero determinanti per eliminare inutili opacità da parte delle autorità nell'uso dei sistemi d'intelligenza artificiale. La triste realtà, però, è che non solo non abbiamo accesso alle relazioni etiche, alle Hria o alle specifiche tecniche di questi sistemi: non sappiamo neanche se e dove questi sistemi sono usati. Gran parte di quello che sappiamo sui sistemi in uso arriva dal lavoro di giornalisti investigativi e ong come la tedesca Algorithm watch, che ha pubblicato un rapporto in cui c'è una mappa dell'uso dei "processi decisionali automatizzati" in diversi paesi europei.

Rimuovere questi inutili ostacoli alla trasparenza sarebbe un primo passo fondamentale da parte delle autorità. Sapere quali sistemi sono usati nel settore pubblico (soprattutto quando sono finanziati con soldi pubblici) è una condizione ne-



GENERATED PHOTOS

cessaria per un monitoraggio efficace, ma ci sono molti altri modi per affrontare l'opacità dei sistemi d'intelligenza artificiale.

La discussione su quali spiegazioni dovrebbero essere rese obbligatorie e con che grado di dettaglio è in corso. L'*Ico-At guidance* raccomanda che l'intelligibilità e l'interpretabilità di un modello d'intelligenza artificiale siano una priorità fin dall'inizio e che la trasparenza e la responsabilità siano considerate preminenti rispetto ad altri criteri. In altre parole, un'azienda o un'istituzione pubblica che vogliono automatizzare decisioni destinate a influenzare la vita delle persone devono usare, ogni volta che è possibile, un modello che può essere interpretato, non una scatola nera tecnica.

Fare questa scelta fin dall'inizio aiuterebbe a risolvere alcuni problemi, permettendo a chi prende decisioni con il sostegno di sistemi d'intelligenza artificiale di riflettere sui loro limiti, a chi è direttamente interessato di contestare le decisioni, alla società di esercitare un controllo maggiore su come sono usati questi sistemi. È scontato, tuttavia, che ci saranno aziende e istituzioni pubbliche che faranno una scelta diversa: per la discutibile convinzione che un sistema opaco sia automaticamente più preciso, per attirare finanziamenti e interesse grazie al fatto che si usano i modelli più avanzati, per la convenienza di lasciar fare tutto alla rete neurale ed evitare un complicato lavoro sul codice informatico o per la volontà di eludere i controlli scegliendo consapevolmente un modello opaco. Cosa possiamo fare in questi casi? Dobbiamo semplicemente prendere atto che ci saranno sempre sistemi opachi?

Spiegazioni necessarie

In effetti, può essere molto difficile ricostruire il modo esatto in cui un sistema d'intelligenza artificiale ha preso una decisione: per esempio, il modo in cui partendo dalle combinazioni dei pixel stabilisce di aver individuato un gatto. Questa difficoltà è spesso usata come pretesto per bloccare qualsiasi discussione sulla logica o sull'equità di un sistema d'intelligenza artificiale. La difficoltà di fornire spiegazioni caso per caso, tuttavia, non è un motivo sufficiente per rinunciare a una spiegazione. Come raccomandano ancora una volta le linee guida dell'*Ico-At guidance*, dovremmo chiedere ai proprietari dei sistemi d'intelligenza artificiale considerati "scatole nere" di fornire spiegazioni aggiuntive che possano far luce sulla logica alla base dei risultati e del comportamento del loro sistema. Tra queste spiegazioni ci sono quelle interne al modello, per esempio il tipo e l'architettura del modello, i dati usati per addestrarlo, i risultati degli *stress test* (che descrivono come reagisce il modello in una serie di situazioni), ma anche, quand'è possibile, i risultati dei *reverse engineering* interni (ricostruzione a ritroso di un sistema o di un processo), che possono fornire esempi di spiegazioni circoscritte per determinati input.

La *Ico-At guidance* mostra che si può fare molto per spiegare il funzionamento interno di un sistema d'intelligenza artificiale se i soggetti coinvolti sono disposti a fare uno sforzo in questo senso. Il punto che vogliamo sottolineare, ricollegandoci anche al ragionamento delle autorità britanniche, è che ci sono molte decisioni che i proprietari o i progettisti di un sistema d'intelligenza artificiale devono prendere: dalla definizione del problema che si vuole risolvere fino alla scelta del modello da usare e del metodo per valutare la sua efficacia. Queste decisioni dovrebbero essere ben documentate e giustificate,

perché la discussione sui valori che stanno alla base dei sistemi d'intelligenza artificiale parte proprio da una serie di scelte umane su come progettare e ottimizzare il sistema.

Il punto chiave è che le spiegazioni delle decisioni automatizzate dovrebbero essere svincolate dalla capacità dell'opinione pubblica di capire come funzionano gli algoritmi. Per mantenere un livello di controllo su un sistema non è necessario comprendere ogni singola fase del processo di apprendimento automatico. Dobbiamo capire le scelte, le valutazioni e i compromessi fatti dalle persone che hanno progettato il sistema e che influenzano il comportamento dell'algoritmo. Per questo livello di comprensione non c'è bisogno di aprire la scatola nera tecnica.

Ogni algoritmo e ogni sistema decisionale assistito da apprendimento automatico sono progettati dagli esseri umani per raggiungere determinati obiettivi. Questi obiettivi non sono definiti dall'intelligenza artificiale, ma dalle persone che l'hanno progettata, che decidono a monte perché ricorrere a un sistema automatizzato e quali problemi risolvere. Lo scopo può essere aiutare gli esseri umani a trovare dei *pattern* (schemi, ricorrenze) in grandi quantità di dati. Ma i sistemi d'intelligenza artificiale possono essere impiegati anche per aiutare gli esseri umani a formulare giudizi basati su previsioni o per sostituire completamente chi prende una decisione. Esistono esempi di giudizio automatizzato relativamente innocui, come individuare lo spam o fare pubblicità mirate, e altri più discutibili, come l'offerta di annunci immobiliari o di lavoro, o ancora il rilevamento dell'incitamento all'odio, in cui è più probabile che si ripetano forme storiche di discriminazione e in cui eventuali errori possono avere conseguenze gravi.

Infine, ed è l'aspetto più discusso, questi sistemi possono essere usati per prevedere determinati risultati, tra cui i comportamenti umani futuri. Quando questi risultati seguono schemi chiari e disponiamo di dati sufficienti e corretti per interpretarli, l'obiettivo può anche essere realizzabile. Raramente però si presentano condizioni ideali di questo tipo, perciò ci sono ancora dubbi sull'uso dell'intelligenza artificiale per prevedere se una persona commetterà di nuovo un reato o per sostituire gli intervistatori umani nella scelta dei candidati ideali per un lavoro.

Indipendentemente dall'attività o dalla funzione attribuita a un sistema d'intelligenza artificiale, gli ingegneri partono sempre da un problema da risolvere: riconoscere cani e gatti all'interno di un'enorme quantità di dati composta da immagini, nella speranza di ottenere risultati di ricerca migliori; prevedere l'interesse o l'umore di un utente che naviga su internet per massimizzare la possibilità che faccia clic sul contenuto suggerito.

Ci sono ancora dubbi sull'uso dell'intelligenza artificiale per prevedere se una persona commetterà di nuovo un reato

Cosa fa davvero il sistema

In qualsiasi sistema di apprendimento automatico gli obiettivi (quelli per cui il sistema è stato impostato) si possono ricavare da una serie di decisioni tecniche e progettuali. Per capire come funziona un sistema di apprendimento automatico – e in quali circostanze probabilmente sbaglierà – dobbiamo prima capire il compito che gli è stato assegnato o la domanda a cui è stato progettato per rispondere. Possiamo usare metodi sofisticati per "interrogare" il sistema e risalire all'effettivo compito che gli è stato assegnato, ma possiamo anche chiedere a chi lo ha progettato di dircelo. Queste informazioni dovrebbero essere disponibili a tutti.

Supponiamo quindi di sapere perché è stato adottato un sistema assistito da intelligenza artificiale: per esempio, per identificare i clienti che hanno più probabilità di comprare un certo prodotto, oppure per massimizzare la percentuale di clic o il tempo di permanenza degli utenti su un sito web. Siamo soddisfatti? Ancora no. Una volta definito un obiettivo generale, c'è un percorso piuttosto lungo prima che il sistema sia calibrato in modo da produrre i risultati desiderati. Per verificare che il sistema raggiunga il suo obiettivo originario e che non si comporti in modo indesiderato (per esempio discriminatorio) dobbiamo comprendere e verificare le principali scelte tecniche e progettuali.

Di fronte a un compito o una domanda complessi, i progettisti fanno una serie di ipotesi per tradurre (o "operazionalizzare") un obiettivo generale in formule o funzioni matematiche. Una cosa è dire "vogliamo che gli utenti restino coinvolti dai contenuti online", un'altra è definire cosa significa "coinvolgimento", come si misura e quali dati personali possono essere usati per prevedere il comportamento online dell'utente. Una cosa è definire un obiettivo come "indirizzare le prestazioni sociali verso i soggetti bisognosi", un'altra è definire chi è effettivamente "bisognoso", formulare la distribuzione ottimale dei risultati e codificare i risultati inaccettabili, i compromessi accettabili e le procedure di valutazione adeguate per produrre un risultato equo.

L'obiettivo generale di un sistema d'intelligenza artificiale dev'essere tradotto da un linguaggio commerciale o politico in formule matematiche. Le decisioni e le scelte "interne" dei progettisti non sono meno importanti della scelta dell'obiettivo generale, così come il modo in cui è attuato un provvedimento politico non è meno importante dell'obiettivo generale di quel provvedimento.

In realtà la capacità o meno di un sistema d'intelligenza artificiale di raggiungere il suo obiettivo dipende in larga misura da queste piccole decisioni "interne", spesso prese in autonomia dai progettisti e mai comunicate al pubblico. Capire queste decisioni tecniche è particolarmente importante quando si discute dell'equità dei sistemi di apprendimento automatico e delle loro conseguenze sulla vita delle persone.

Quali sono queste scelte? Proviamo a metterci nei panni di un esperto incaricato di costruire un modello che aiuti gli amministratori di un ospedale a selezionare i pazienti di un programma sanitario sulla base dei bisogni. Uno di questi modelli – usato in molti ospedali negli Stati Uniti – è finito al centro delle polemiche dopo che un articolo della rivista Science ha rivelato che l'algoritmo privilegiava sistematicamente i pazienti bianchi più sani rispetto ai pazienti neri più bisognosi di cure mediche. L'effetto, come spiegano gli autori dell'articolo, può essere ricondotto alla scelta di identificare come più bisognosi i pazienti che generano i

costi sanitari più alti, un parametro che notoriamente penalizza i pazienti neri, per i quali gli ospedali statunitensi spendono meno.

Tradurre problemi concreti in formule matematiche comporta necessariamente semplificazioni, ipotesi arbitrarie e compromessi. I parametri di misurazione dei sistemi d'intelligenza artificiale devono corrispondere a fenomeni quantificabili, che non sempre coincidono con la definizione umana di successo in un determinato scenario. Inoltre la loro affidabilità è limitata dalle proprietà matematiche dei modelli usati. Per esempio, è provato che i sistemi calibrati per ottenere il massimo della precisione spesso non funzionano altrettanto bene nel garantire che nessun gruppo sia discriminato. Come ha osservato la matematica statunitense Cathy O'Neil, "non puoi ridurre al minimo i falsi positivi, avere la massima precisione e ridurre al minimo i falsi negativi allo stesso tempo. Ci sono sempre dei compromessi, e bisogna trovare il modo per farli venire alla luce senza dover prendere un dottorato in matematica".

Per quanto i proprietari dei sistemi possano essere in buona fede e possano desiderare un risultato equo, ciò che il sistema restituirà sarà determinato non dal loro pio desiderio, ma da una serie di vincoli matematici e statistici. Non sempre i dati riflettono il mondo reale. La bontà dei sistemi dipende dalla bontà dei dati con cui sono alimentati. Anche il modello migliore non funzionerà se alimentato con dati che sono collegati solo alla lontana con i fenomeni che il sistema dovrebbe prevedere.

Dobbiamo tenere presente che molti fenomeni che potremmo voler prevedere, come il "livello di salute" o il "tasso di recidiva", non hanno definizioni semplici che corrispondono alle misurazioni presenti nei dati. In questi casi, gli esperti si affidano ad approssimazioni, a "surrogati" delle variabili osservate, come il numero di malattie croniche attive o il numero di arresti (è possibile misurare se dei pregiudicati sono stati nuovamente arrestati dopo essere stati scarcerati su cauzione, ma "essere arrestati" non è un sinonimo di "aver commesso un reato", soprattutto tenendo conto degli eccessi di zelo di cui sono vittime le minoranze).

Se c'è un pregiudizio, significa che dietro c'è una decisione umana. Molti studi che prendono in esame gli algoritmi "discriminatori" spiegano il pregiudizio sulla base di fattori esterni come la scarsa qualità dei dati ("gli unici dati disponibili per addestrare il sistema riflettevano il medesimo pregiudizio" o "i progettisti non hanno considerato i dati su un determinato gruppo") oppure un vizio nella scelta dell'obiettivo. Anche se le ragioni possono sembrare estrinseche, ossia derivanti da realtà esterne che i progettisti del sistema in questione non potevano controllare, la verità è che dietro ognuna di queste scelte c'è una decisione umana.

Tradurre problemi concreti in formule matematiche comporta semplificazioni, ipotesi arbitrarie e compromessi

Le possibili conseguenze

I proprietari dei sistemi d'intelligenza artificiale e i progettisti sono responsabili delle scelte in ogni fase dello sviluppo. Sono responsabili della scelta dei dati, anche quando sanno fin dall'inizio che questa scelta è molto limitata. Sono responsabili di aver adottato un sistema pur sapendo di non poterne evitare le distorsioni, di non aver rivisto l'obiettivo principale una volta preso atto che il risultato sperato non era raggiungibile e, infine, della decisione di usare un sistema automatizzato nonostante i suoi limiti e le possibili conseguenze della sua applicazione.

Non si tratta di una decisione isolata, ma di un ciclo di decisioni. Un processo ben gestito di progettazione e addestramento di un sistema d'intelligenza artificiale non può prescindere dal ripensamento degli obiettivi, dall'affinamento della qualità dei dati, dal cambiamento del modello e dalla ricalibrazione del sistema in modo che faccia quello che per cui è stato pensato. Se il modello non raggiunge gli obiettivi desiderati, va semplicemente abbandonato. Non è un processo a senso unico. I progettisti devono costantemente rivedere il processo, tornare sui loro passi e ripensare da capo la progettazione del sistema. Anche se non esiste l'obbligo giuridico di registrare le motivazioni delle scelte progettuali, nella prassi questo processo circolare è abbastanza diffuso e coerente. La decisione di quando interrompere questo processo, tuttavia, non è solo tecnica o economica.

Nei casi in cui il sistema d'intelligenza artificiale ha un'influenza significativa sulla vita delle persone, ogni decisione diventa politica. Per esempio, com'è possibile che un ingegnere del software decida da solo se un modello di previsioni è abbastanza efficace nell'evitare di discriminare le minoranze? Come si fa ad affidare queste decisioni alle intuizioni di persone che non hanno nessuna competenza su questi temi? E, soprattutto, com'è possibile escludere qualsiasi controllo pubblico su decisioni così importanti? Alla fine del suo libro *Automating inequality*, la politologa Virginia Eubanks cita un esperto che sogna di sostituire i dipendenti pubblici con dei sistemi d'intelligenza artificiale: "Le informazioni e le indicazioni saranno immediate, in tempo reale, su misura e facili da confrontare nel tempo. E, idealmente, con il consenso di tutti e perfettamente apolitiche".

Cosa significherebbe sostituire i dipendenti pubblici, responsabili di decisioni molto delicate che hanno una ricaduta enorme sulla vita delle persone, con sistemi d'intelligenza artificiale perfettamente apolitici? Di fatto significherebbe nascondere la politica dentro la scatola nera. Significherebbe oscurare il fatto che in ogni fase del loro sviluppo questi sistemi sono costruiti a partire da una serie di decisioni arbitrarie prese da persone in carne e ossa; e significherebbe mascherare queste decisioni – da quella di concedere o meno la libertà su cauzione a quella di separare o meno un bambino dai suoi genitori – dietro un'obiettività di facciata. Sarebbe semplicemente un modo di spoliticizzare le decisioni, a tutto vantaggio di chi non vuole prendersi la responsabilità delle proprie scelte politiche.

Negli ultimi anni i casi di discriminazioni e problemi di sicurezza derivanti da errori dei sistemi sono stati abbastanza numerosi da farci riflettere sulle loro potenziali implicazioni, che nelle circostanze più gravi possono fare la differenza tra la vita e la morte. Si è discusso molto degli effetti negativi dei fallimenti dei sistemi a cose fatte, ma il problema è che le scelte chiave dei proprietari di quei sistemi o dei

GENERATED PHOTOS

loro esperti – sui dati e le fonti dei dati, sul modello usato – non sono mai state discusse in pubblico. Come osserva la sociologa Ruha Benjamin, tendiamo a vedere le falle e i guasti come "anomalie temporanee di un sistema altrimenti perfettamente innocuo", come semplici problemi tecnologici, mentre dovremmo considerarli "una specie di spia di come funziona il sistema".

Se le decisioni fossero rese trasparenti fin dall'inizio e se le parti interessate fossero coinvolte e consultate, molti di questi effetti negativi potrebbero essere evitati. Oggi i progettisti sono molto più attenti alle conseguenze potenziali del loro lavoro, e spesso si accostano all'intelligenza artificiale perché sperano di poter usare queste tecnologie per il bene comune. Ricercatori e professionisti, scottati da una serie di casi documentati di distorsioni e pregiudizi dei sistemi d'apprendimento automatico, fanno del loro meglio per assicurarsi che i loro algoritmi siano equi. Per quanto ammirevoli, tuttavia, questi sforzi non sono sufficienti: la ricerca del "bene sociale" e dell'"equità" implica una serie di scelte normative e politiche, perché non esiste una definizione condivisa di "bene sociale" e ci sono almeno ventuno definizioni di "equità", che in molti casi si escludono a vicenda. Mentre in scenari a basso rischio queste decisioni possono essere innocue, diventano intrinsecamente politiche quando riguardano sistemi che hanno conseguenze sulla vita umana.

Anche se probabilmente non si arriverà mai a una definizione condivisa di intelligenza artificiale "buona" o "equa", andrebbe almeno condiviso un processo di valutazione e discussione delle scelte politiche di chi progetta questi sistemi. Attualmente, però, non ci sono né il linguaggio né la prospettiva per affrontare tali valutazioni e discussioni. Questo permette agli informatici di fare considerazioni ad ampio raggio sulla soluzione dei problemi sociali senza prendersi però la responsabilità delle conseguenze sociali e politiche delle loro scelte. Spesso l'effetto è quello di complicare ulteriormente o addirittura di aggravare i problemi che si vorrebbero risolvere.

Se concordiamo sul fatto che nell'introduzione dei sistemi d'intelligenza artificiale le questioni tecniche spesso diventano politiche, il passo successivo è pretendere la massima trasparenza e una qualche forma di controllo democratico su queste decisioni. Mentre un controllo completo su ogni singola decisione che riguardi la progettazione dei sistemi non è né ipotizzabile né necessario, possiamo – e dobbiamo – chiedere più trasparenza sulle scelte tecniche alla base dei sistemi addestrati per prendere decisioni che influiscono sulla vita delle persone. Solo così saremo in grado di capire ed eventualmente di mettere in discussione le decisioni politiche alla base delle scelte tecniche, in particolare quando i sistemi sono usati nel settore pubblico. Senza questa trasparenza ci viene negata la possibilità di partecipazione politica, perché le decisioni più discutibili sono nascoste dietro il racconto della "scatola nera". ◆ *fsa*

GLI AUTORI

Agata Foryciarz è una ricercatrice di informatica dell'università di Stanford, negli Stati Uniti.

Daniel Leufer è un filosofo e analista politico che studia le conseguenze dell'intelligenza artificiale sulle decisioni politiche.

Katarzyna Szymielewicz è un'avvocata e atttivista per le libertà civili. Guida la Panoptykon foundation ed è nel consiglio direttivo dell'ong European digital rights.