Se all'Intelligenza artificiale diamo il potere di vita o di morte

di Andrea Granelli

In un celebre incidente stradale durante un test, una macchina con guida autonoma Uber ha rilevato la presenza di un pedone – Elaine Herzberg – sei secondi prima di colpirla e di ucciderla, ma l'ha identificata come una bicicletta (la Herzberg stava infatti spingendo la sua bici).
Come spiega il rapporto preliminare pubblicato dalla National Transportation Safety Board, Uber non aveva impostato il suo sistema per "agire in autonomia" su quel tipo di decisione. Gli ingegneri hanno impedito alla loro auto di frenare improvvisamente da sola in situazioni analoghe, afferma il rapporto, "per ridurre il potenziale di comportamento erratico del veicolo".
L'azienda ha deciso di affidarsi all'operatore umano dell'auto in tali situazioni per evitare incidenti. E questa scelta apre a un sistema complesso di interazioni possibili uomo-macchina.
La fallibilità degli algoritmi di guida non sono però l'unico problema legato al corretto funzionamento delle macchine a guida autonoma. Ve ne sono altri due potenzialmente problematici e dalle molteplici implicazioni.
Il primo è che questi sistemi di guida usano un algoritmo "di morte" – peraltro protetto dal segreto industriale – che decide chi uccidere in una situazione critica: il/i passanti o il/i passeggeri. Un'indagine fatta dalla rivista del MIT – Technology Review – sul criterio che tali algoritmi debbano adottare dava una risposta unanime: minimizzare il costo delle vite umane.
Ma la complessità del tema non finisce qui: e se le persone che stanno attraversando hanno la fedina penale sporca? Se sono antagonisti di chi è al potere? Se hanno semplicemente un colore diverso della pelle? Nell'analisi "numerica" che vuole minimizzare il costo in vite umane valgono sempre uno? Oggi la capacità di riconoscimento facciale, unita alla possibilità di costruire algoritmi parametrici – che in questo caso non si limiterebbero a "contare" tre persone, ma ne valuterebbero le caratteristiche per scontare quel numero – sono un fatto consolidato, non una ipotesi futuribile.

> **I SISTEMI DI GUIDA AUTONOMA CHE DECIDONO CHI SALVARE SONO ALGORITMI PROTETTI DAL SEGRETO INDUSTRIALE**

Il secondo aspetto problematico è che è relativamente semplice impossessarsi del sistema di guida e controllarlo direttamente da remoto per guidare il veicolo dove si vuole. I recenti casi del terrorismo ci ricordano che questa opportunità non è un semplice caso di scuola: guidare auto, camion sulla folla inerme è una delle strategie adottate in tempi recenti dal terrorismo islamico.
Per questi motivi il dibattito sugli algoritmi che guidano la "macchina senza guidatore" ad evitare gli incidenti "scegliendo il male minore" incominciano a destare preoccupazione. E vengono recuperate le riflessioni dei filosofi etici. Una per tutte il cosiddetto "Trolley Problem" formulato per la prima volta dalla filosofa Philippa Foot nel 1967: «un tram ferroviario ha perduto il controllo. Il guidatore non può frenare, ma può solo azionare lo scambio tra i binari. A un certo punto si trova di fronte a un bivio: seguendo il percorso previsto, ci sono cinque persone sul binario; mentre sull'altro binario – che può percorrere solo decidendo di azionare lo scambio – ce n'è solo una. In entrambi i casi, le persone moriranno nell'impatto. Cosa deve fare il guidatore? Subire passivamente quanto programmato e assistere alla morte di cinque persone o azionare deliberatamente lo scambio e ucciderne "solo" una?». Questo esercizio del pensiero vuole porre l'attenzione sul fatto che non basta la comparazione numerica del male minore; vi è anche la differenza tra assistere e determinare una morte.

È venuto il momento di contrastare l'adorazione magica delle tecnologie

di Andrea Granelli

Una recente copertina dell'Economist riporta la nostra attenzione a un tema spesso dimenticato ma molto importante: la presenza del pensiero magico che tende ad attribuire a persone e tecnologie un potere molto maggiore di quello che sono in grado di fornire.

L'immagine – come spesso accade su quelle copertine – è potente ed efficacissima: si vede infatti il classico coniglio che appare nel cappello a cilindro del prestigiatore; ma è un coniglio ferito e ammaccato, evidentemente non più in grado di far credere che con un colpo di bacchetta magica si possano risolvere i problemi che ci attanagliano. Questa propensione umana – nei fatti una vera e propria bias cognitiva – non vale solo nella politica, con la disperata ricerca dell'uomo forte, ma si applica anche al settore delle tecnologie, nella ossessiva ricerca della tecnologia magica che risolve tutti i problemi, oltretutto senza effetti collaterali.

Fu una grande intuizione di Gartner Group comprendere che dietro ogni innovazione c'è sempre un hype, un'illusione potente, tenace e spesso autocostruita (ma comunque rinforzata dai fornitori di tecnologie e dai finanziatori di tali aziende che sanno come influire sui media mainstream) che fa sì che creiamo aspettative tecnologiche che vanno molto oltre quanto la nuova tecnologia è in grado di dare. Questo pensiero magico è pericoloso non solo per la delusione che necessariamente comporta – tanto più cocente quanto più alte e irrealistiche sono le illusioni – una volta che la tecnologia si manifesta nelle sue capacità e limiti. Ma lo è anche perché inquina la razionalità dei meccanismi decisionali nel nostro modo di valutare le cose. E, a ben vedere, questo modo di ragionare non è così lontano dal movimento "no-X" e dal suo parente stretto: il complottismo. Dai commenti sui media mainstream che stanno accompagnando il lancio di chatGPT mi sembra che stiamo ricadendo nello stesso errore. Il tema è dunque più generale. Stiamo partendo da considerazioni su un prodotto specifico – la chatGPT – ma di fatto ciò che serve è un approccio più sistematico e corretto sulla valutazione tecnologica che eviti di cadere nei due estremi su cui spesso il dibattito si polarizza: i tecno-fan adoratori dell'innovazione in tutte le sue manifestazioni e i tecno-fobici, retrogradi, conservatori dello status quo e avversatori del futuro e del progresso. Questo semplificare questioni complesse in schemi binari – tecnica adorata da alcuni politici ("o con me o contro di me") – è chiamata dai retori fallacia del falso dilemma ed un virus che contamina i ragionamenti e annichilisce le possibilità di convergenza tra diversi punti di vista. Si tratta piuttosto di sviluppare e diffondere un approccio che consenta a chi ama l'innovazione ed è curioso, di imparare a non farsi abbindolare dalle false chimere, spesso alimentate dai molteplici interessi dietro l'innovazione tecnologica. Ci vuole cautela nel buttarsi su cose di cui non abbiamo capito tutto.

Ci sono allora due filoni di pensiero – molto diversi fra di loro – che possono però aiutarci in questo percorso. Il primo è riconducibile alla cultura "slow", che non si applica solo al cibo ma sta diventando una sorta di fermiamoci-e-riflettiamo-prima-di-decidere per contrastare la frettolosità tipica delle persone imprudenti e superficiali. È giusto ricordare che la prudenza era una delle doti più importanti dei leader (oltre a essere una delle quattro virtù cardinali). Come non ricordare il potente libretto "L'arte della prudenza" scritto a metà del Seicento dal gesuita Baltasar Gracián. Essere prudenti non vuol dire essere timorosi e procrastinare sine die possibile decisioni e azioni. Vuol dire, invece, decidere nel momento opportuno – cogliendo quello che i Greci chiamavano kairos – e agire poi senza tentennamenti e con assoluta determinazione. È la frettolosità, l'impazienza, dunque, il vero male.

Il secondo filone è debitore di grandi pensatori come ad esempio Popper e deve al filosofo Hans Jonas la sua formulazione più efficace, detta Principio di precauzione. Si tratta di riapplicare in modo sistematico il pensiero critico anche all'innovazione tecnologica. Pensiero critico non per creare alibi al non fare, ma per costruire su fondamenta solide. Lo diceva Cartesio nelle sue Meditationes de prima philosophia – "Il dubbio è l'origine della saggezza" – e lo riprende in modo icastico Bertrand Russell (Storia della filosofia occidentale): "Il problema dell'umanità è che gli sciocchi e i fanatici sono estremamente sicuri di loro stessi, mentre le persone più sagge sono piene di dubbi".

Su ChatGpt il confronto ideologico arriva prima dei benefici (e delle criticità)

di Andrea Granelli

Barbara Carfagna esordisce in una delle sue riflessioni su ChatGtp con la seguente affermazione: «Probabilmente Adriano Olivetti solo oggi vedrebbe il suo sogno realizzarsi pienamente; quando, all'inizio degli anni Sessanta prese la decisione di sviluppare un "computer da tavolo" voleva che ogni individuo potesse creare quello che solo una grande azienda poteva progettare. ChatGpt di OpenAI...».

Parlare di ChatGpt in modo obiettivo, equilibrato e definitivo è molto difficile non solo per la complessità e articolazione della materia, ma anche perché la stiamo osservando in opera da poco; abbiamo appena incominciato a capirne e prefigurarne possibilità e implicazioni. Ma su una cosa sono certo: è davvero molto pericoloso azzardare valutazioni precise e giudizi tranchant. Oltre ad essere imprecisi, se non in errore, si rischia di cadere in un confronto ideologico che alla fine si riduce nelle due polarità dei tecnofan e dei tecnofobici, degli amanti del futuro e dei retrogradi.

Fatte queste doverose premesse, non si può non rimanere affascinati e colpiti dalla potenza ed efficacia di ChatGtp. La vera questione, io credo, è però identificare il tipo di utilizzo che è giusto fare, i benefici che può ragionevolmente portarci e le criticità ad oggi prevedibili che una tecnologia così potente può causare.

Partiamo dalla fine: nonostante la sua relativamente breve vita operativa, ci sono già manifestazioni evidenti di abusi di questa tecnologia. Per esempio la fotocopia digitale: sono già apparsi siti e applicazioni creati a perfetta immagine e somiglianza di siti reali, che consentono ai criminali di carpire informazioni personali anche a utenti esperti, che non si rendono conto di essere su un sito fake. Ma questo meccanismo va oltre e sfrutta le potenzialità dell'intelligenza artificiale, arrivando addirittura a costruire landing page e contenuti aggiornati nel tempo. E ciò grazie alla capacità di ChatGpt di creare testi non solo grammaticalmente corretti ma anche sofisticati e in linea con lo stile editoriale del sito. Anche la creazione seriale di mail personalizzate è un compito semplice per ChatGpt. Quello che colpisce è la estrema facilità con cui queste azioni vengono condotte, aspetto che fa presagire una futura invasione di siti fake.

Vediamo ora le possibili applicazioni che si possono auspicabilmente prefigurare, tenendo ovviamente in mente queste possibili derive negative. Ne vedo in particolare tre.

Innanzitutto, come strumento di apprendimento per imparare a considerare possibili opzioni rispetto a quelle che ci vengono in mente e a estendere, quindi, la nostra capacità di inquadrare un problema in tutti i suoi aspetti. Sarebbe interessante poter costruire dei veri e propri dibattiti con ChatGpt a fini educativi (molti già si esercitano con i programmi di scacchi). Come ci ricorda infatti un retore medioevale, «nessuna verità può essere veramente capita e predicata con ardore se prima non sia stata masticata dai denti della disputa».

In secondo luogo, come strumento per simulare possibili alternative a un'ipotesi progettuale a cui si sta lavorando. Anche se queste alternative sono teoriche o perfino impossibili, la loro analisi aumenta la nostra consapevolezza sul tema, ci forza a entrare in profondità e capire meglio il contesto che stiamo analizzando.

Infine, per aumentare la nostra efficienza, producendo semilavorati che però devono essere integrati da un attento lavoro sia sui contenuti che sullo stile, soprattutto integrando elementi mancanti o togliendo affermazioni o argomentazioni che sono irrealistiche o lontane dal nostro punto di vista.

Se usato in maniera opportuna, dunque, questo strumento può rafforzare le nostre competenze, capacità cognitive e qualità argomentative. Non dobbiamo mai cadere, però, in affermazioni generiche o irrealistiche né in uno stile anonimo e universalizzante, perdendo le nostre specificità culturali. L'ultima parola deve essere sempre la nostra.

È giusto, dunque, buttarsi nella mischia di ChatGpt: come ci ricorda Friedrich Hölderlin in uno dei suo Inni: «Dove c'è il pericolo cresce anche ciò che salva». Non dobbiamo, però, mai staccare il pensiero critico, facendo nostro un suggerimento di un intellettuale di rango, il grande scrittore Hemingway, che, intervistato nel lontano 1954, affermò come era solito fare in modo caustico: «Every man should have a built–in automatic crap detector operating inside him» ("Ogni uomo dovrebbe avere un rilevatore automatico di schifezze incorporato che opera al suo interno").

Deep Fake, quando l'AI gioca sporco

di Andrea Granelli

Al rientro dalle ferie, prima Beppe Grillo fa un intervento celebrativo sul ruolo e le capacità progressive della Cina e lo fa in cinese; a stretto giro Matteo Salvini gli fa da controcanto e si rivolge ai francesi quasi per giustificare l'invito di Marine Le Pen a Pontida. È anche curioso notare gli scherzetti dei traduttori automatici. Il leader della lega utilizza una simpatica excusatio non petita "non è un raduno" ma poi afferma "è un raduno" – la parola francese usata due volte è "un rassemblement" – in qualche modo smentendosi... o forse ricordandosi improvvisamente e inconsciamente che la sua ospite è il presidente del Rassemblement National.

Nei fatti queste comunicazioni non hanno ingannato nessuno e hanno raggiunto un'audience più ampia. Inoltre, gli autori non hanno smentito né voluto nascondere il fatto di aver utilizzato tecnologie di AI... forse anche per continuare a dare il senso di essere al passo con i tempi, di essere anticipatori dei cambiamenti.

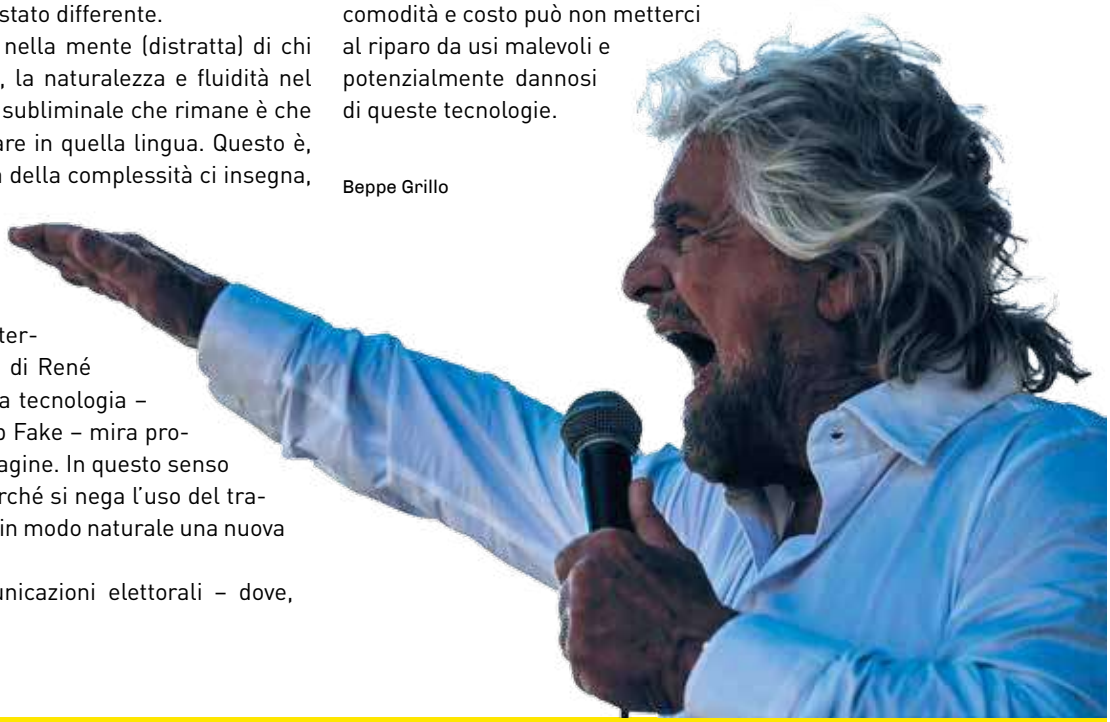
Il punto naturalmente non è se questi programmi potranno funzionare meglio o se è scorretto usarli in politica. La questione è più fine. Se per esempio Salvini avesse diffuso un podcast – anche legato alla sua foto per firmarne i contenuti – il messaggio subliminale sarebbe stato differente.

Non si sarebbe, cioè, impressa nella mente (distratta) di chi normalmente ascolta dai social, la naturalezza e fluidità nel parlare quella lingua. La traccia subliminale che rimane è che quella persona "sa" anche parlare in quella lingua. Questo è, io credo, il punto. Come la teoria della complessità ci insegna, piccoli comportamenti (il battito d'ali di una farfalla) possono avere grandi impatti (un uragano, per continuare l'esempio usato nel film The Butterfly Effect che richiama le teorie di René Thom). La scelta di usare questa tecnologia – nel gergo tecnico chiamata Deep Fake – mira proprio a potenziare la propria immagine. In questo senso la comunicazione è fake: non perché si nega l'uso del traduttore ma perché si suggerisce in modo naturale una nuova abilità non posseduta.

In questo solco vanno le comunicazioni elettorali – dove, come noto, aumenta molto la distanza tra l'autenticità e verità e ciò che si comunica. Tra i primi utilizzi di un traduttore automatico in questo tipo di comunicazione, si ricorda quello di Manoj Tiwar, un leader indiano che nel 2020 ha parlato in Haryanvi (un dialetto hindi che ovviamente non conosceva) per accreditarsi con quegli elettori, grazie al semplice mostrare di conoscere la loro lingua e quindi di rispettare e apprezzare la loro cultura. È stato naturalmente smascherato e la cosa ha suscitato polemiche e perplessità.

Le frontiere del Deep Fake possono essere anche molto subdole. Tra i casi più "sofisticati" vi è un filmato fatto girare in Rete a maggio 2019 dai (suppongono gli osservatori politici) supporter di Donald Trump, e relativo a un intervento di Nancy Pelosi. Il video è stato "semplicemente" rallentato di un poco. L'effetto, però, è sorprendente: l'ex Speaker della Camera statunitense sembra infatti ubriaca; le sue parole risultano impastate e incerte, quasi zoppicanti. Tecnicamente è sempre il suo discorso, nessuna parola è stata infatti cambiata. Il possibile travisamento del suo messaggio si origina dunque da una semplice variazione della velocità del video, che di per sé non è un atto manipolativo (basti pensare all'importanza della moviola nel calcio).

È quindi chiara la complessità del fenomeno e la difficoltà di un suo semplice inquadramento schematico. Per questo motivo, come in altri casi riconducibili ai temi dell'intelligenza artificiale, la pura valutazione basata su utilità, comodità e costo può non metterci al riparo da usi malevoli e potenzialmente dannosi di queste tecnologie.



Beppe Grillo

La tragedia della guerra ha spiazzato i tecnofan

di Andrea Granelli

Vedendo il materiale video relativo alla drammatica incursione di Hamas sul territorio israeliano del 7 ottobre – sia il materiale di propaganda diffuso da Hamas che quello raccolto in maniera fortuita sul campo – e dopo naturalmente aver con molta difficoltà superato il dramma, il dolore e il disorientamento che le immagini suscitano, la memoria ritorna su una serie cinematografica di grande successo – Mad Max – e soprattutto sul sequel del 2015: Fury Road.

Il contesto della serie è noto: il nostro pianeta è ormai ridotto ad una landa desolata dopo una tragica guerra nucleare autodistruttiva e acqua e benzina sono diventati gli elementi più importanti per la sopravvivenza dei sopravvissuti e il loro controllo diventa la causa permanente di continue guerriglie e scontri frontali.

Dato lo status post-atomico, le armi da guerra usate sono tutte basate su vecchi reperti civili (moto, camion, tubi, aste...) riadattati in modo creativo ma sempre primitivo, per diventare offensivi. Una sorta di armata Brancaleone del futuro.

Il fatto è che gli strumenti di guerra usati da Hamas in questo attacco ricordano in qualche modo questa narrativa: deltaplani, droni commerciali, moto di piccola cilindrata, auto civili, missili terra-aria fai-da-te, mini-sommergibili, tunnel sotterranei...

Come è stato possibile che questo esercito artigianale – quasi post-atomico – abbia potuto agire indisturbato e superare le difese elettroniche – probabilmente tra le più sofisticate esistenti – mettendo in ginocchio uno degli eserciti meglio armati e più esperti al mondo. Alcuni spunti per la risposta potrebbero venirci da un altro film, questa volta del 2008 e diretto dal grande Ridley Scott – "Nessuna verità" –, per bocca di Ed Hoffman alias Russel Crowe. Nel film Hoffmann è uno dei responsabili della Cia per le azioni in Medio Oriente e deve coordinare la cattura di uno dei più pericolosi terroristi islamici. In un passaggio clou del film esprime la sua visione e la sua preoccupazione: «Il nostro nemico ha capito che sta combattendo contro gente del futuro. Se tu vivi come nel passato e ti comporti come nel passato, allora per la gente del futuro diventa difficile vederti. Se getti via il tuo telefonino, se smetti di usare la tua email, passi tutte le informazioni di bocca in bocca, volti le spalle alle tecnologia... semplicemente sparisci nella folla... Quindi la novità è che il nostro presunto rozzo e grossolano nemico si è ormai reso conto di quella che è una nuda e cruda verità: che siamo un bersaglio facile».

Forse è proprio questo che è capitato. Gli occhi e gli orecchi digitali israeliani – quelli che controllavano il confine senza le incertezze e le stanchezze tipiche dell'umano – sono una summa delle tecnologie più avanzate e per questo gestite senza personale. Vedere allora neutralizzare questi sofisticati sistemi di difesa tecnologica da semplici droni commerciali che sganciavano piccole bombe artigianali dalla potenza contenuta... ma sufficiente a distruggere le fragili antenne e batterie di queste reti di sensori ha certamente spiazzato l'immaginario collettivo tecnofan. Ma comprendere che non vi erano neanche presenti esseri umani in grado di dare l'allarme a valle di questi piccoli ma chirurgici attacchi ha spiazzato ancora di più.

Ed è una situazione che si incomincia a vedere anche nelle aziende. Pensare che la tecnologia possa totalmente sostituire l'essere umano è proprio un ragionamento fallace e pure rischioso: certo la tecnologia può essere più efficiente, talvolta anche più efficace. Ma la sicurezza, la prevenzione di ogni rischio è tutt'altra partita. Come direbbe Taleb, nonostante l'abilità predittiva degli algoritmi, c'è sempre un possibile cigno nero, e il 7 ottobre ne abbiamo avuto conferma.

L'IA generativa deve imparare a disimparare

di Andrea Granelli

Una delle caratteristiche delle piattaforme generative che viene tenuta in gran conto e considerata con riverenza soprattutto dai neofiti è la mole di dati utilizzati: si ipotizza infatti che più dati ci sono più aumenta la qualità delle risposte. Anche nella scelta di adozione di queste piattaforme, la quantità viene usata come elemento convincente. È come se potessimo attingere all'intero patrimonio dell'umanità. Ma siamo sicuri che questa proprietà sia efficace?

I nostri nonni ci ricordano che "il troppo stroppia". Il motore, per avere risposte pertinenti e non superficiali, è infatti la qualità più che la quantità; e soprattutto il poter discriminare le fonti informative. Impedire l'utilizzo da parte della piattaforma di sezioni della Rete con dati superficiali o manipolati equivale a insegnare loro a disimparare, a non considerare come validi alcune parti del materiale consultato. Il tema è ovviamente soggettivo, ma qui sta l'efficacia e anche la democraticità di questi sistemi.

La statistica si è abituata a gestire grandi moli di dati dove ci sono dati sporchi o rumore di fondo; e lo fa con tecniche matematiche che in qualche modo li isolano e ne riducono il potere predittivo. Molto più complicato fare ciò sulla Rete e con la massa testuale di informazioni presenti. Il fenomeno diventa ancora più complesso in quanto la produzione di queste informazioni non è accidentale; inoltre chi le introduce fa di tutto per moltiplicarle con il fine di aumentare la loro cattura da parte dei motori di ricerca e rafforzare la loro sedicente credibilità.

La questione è dunque se accontentarsi di risposte medie che rappresentano il punto di vista generale di quello che c'è sulla Rete (una sorta di "media del pollo" di Trilussa) oppure cercare risposte insightful, illuminanti, anche imprevedibili, che non ci danno conferma di quanto la gente già sostiene ma ci aprono a nuove interpretazioni e punti di vista ...anche insinuando nuovi dubbi.

Il problema diventa dunque poter pesare il contributo delle informazioni a cui il motore di ricerca attinge.

In un'intervista a Umberto Eco sul funzionamento di tag e keyword nel classificare gli articoli scientifici, lo studioso sottolineava che questo meccanismo funziona bene per i lettori poco ferrati nella materia. L'esperto, invece, trova spesso il valore di un articolo non necessariamente nella parte più rappresentativa e centrale – e cioè nella tesi sostenuta e nelle ipotesi portate a dimostrazione – ma in affermazioni a latere o addirittura in una nota a pié di pagina.

Il punto è allora poter differenziare in modo obiettivo e consapevole le fonti utilizzate e quindi definire modalità per predisporre le piattaforme di IA generativa in modo che non solo riducano l'impatto del rumore di fondo e delle fonti manipolate, ma consentano a chi le consulta di scegliere determinati domini di conoscenza e dare loro un "peso" rilevante nella costruzione della risposta.

Ad esempio, io ho costruito nel corso degli anni un contenitore su web (che chiamo "zaino digitale") che contiene tutto ciò che ho letto (ovviamente ciò che mi ha colpito e che ho deciso di conservare) e i lavori che ho fatto. Rappresenta il mio punto di vista ...ma non è una fonte poi così soggettiva: gli estratti dei libri letti presenti nel sito sono più di 2.000. È quindi un corpus conoscitivo non solo "obiettivo" ma per me particolarmente rilevante e ispirativo; infatti il mio modo di pensare e argomentare è in qualche modo dipendente da questa base di conoscenza; il suo utilizzo "tradizionale" sconta, però, tutti i limiti della memoria biologica. L'idea è allora di fare in modo che la "mia" piattaforma generativa acceda con particolare attenzione a questa base di conoscenza; non perché ritengo che sia la più valida di tutte, ma perché rappresenta il mio punto di vista, la mia sensibilità culturale, il mio stile cognitivo costruito negli anni. Nulla toglie che io poi possa interrogare la piattaforma generativa su altri domini creando una sorta di confronto fra differenti tipi di risposta generati dalla stessa domanda.

Qui sta, secondo me, il valore più importante che le piattaforme di IA possono darci: non un'unica risposta – ipse dixit – ma gruppi di risposte che ci forzano a cambiare il nostro punto di vista e a guardare lo stesso tema o problema da diverse angolature.