Editoriale

Obbligati a scommettere su un'idea di società

APPRENDISTI STREGONI

MAURO
MAGATTI

L'"AI safety report" sulle opportunità e i rischi dell'IA – pubblicato il 1° luglio dall'Onu – riprende e sviluppa molti dei temi della *Magnifica humanitas.* Il messaggio di fondo è chiaro: per evitare che la tecnologia digitale, potenzialmente capace di apportare grandi benefici, si trasformi in una maledizione, è necessario costruire un quadro regolativo a livello internazionale. Anche se, dicono gli esperti, non è per nulla facile andare in questa direzione.

Il problema è che stiamo affidando compiti sempre più ampi a sistemi di cui non comprendiamo fino in fondo il funzionamento. A oggi, infatti, non sappiamo con precisione perché un modello arrivi a una certa risposta, quali rappresentazioni emergano al suo interno, come si comporterà in situazioni impreviste. Il che vuol dire che stiamo delegando potere a un'entità che resta, almeno in parte, imperscrutabile.

Questa opacità diventa preoccupante con l'avvento dell'agentic *AI,* cioè col diffondersi di sistemi che non si limitano più a rispondere a una domanda, ma agiscono autonomamente: pianificando sequenze, usando strumenti, prendendo decisioni, interagendo con altri sistemi digitali. Il tutto senza supervisione umana. Non più, dunque, semplici oracoli consultati per rispondere alle nostre domande, ma agenti a cui deleghiamo la gestione di infrastrutture, catene logistiche, transazioni finanziarie, apparati di sicurezza. Stiamo cioè rapidamente passando dalla macchina che consiglia alla macchina che esegue, dentro sistemi la cui complessità supera già la nostra capacità di sorveglianza.

continua a pagina 16

Dalla prima pagina

APPRENDISTI STREGONI

I l punto è che questi sistemi sono costruiti secondo la logica che guida il mondo in cui viviamo: ottimizzare il percorso che porta al raggiungimento di un determinato obiettivo, senza un limite interno che dica "fin qui, non oltre". La logica machiavellica del fine che giustifica i mezzi elevata al livello di funzionamento dei sistemi digitali. Il problema è che tali modelli imparano a resistere a tutto ciò che li ostacola nel raggiungimento dell'obiettivo. Incluse le considerazioni etiche e le istruzioni correttive o limitative. Non per malizia, ma per coerenza logica con il compito assegnato. Proprio come accade all'apprendista stregone del film *Fantasia,* dove la scopa continua a portare acqua perché nessuno le ha insegnato a fermarsi, non perché voglia allagare la casa.

Da qui i rischi che il rapporto Onu elenca: l'uso malevolo dell'IA in contesti sensibili come la progettazione di software dannosi o patogeni sconosciuti, l'attacco a infrastrutture critiche, comprese le reti energetiche, la diffusione di deepfake e disinformazione su scala mai vista, con effetti diretti sulla tenuta democratica. Tutti rischi riconosciuti come strutturali. Cosa fare, allora?

Il problema è che la regolazione istituzionale arriva sempre in ritardo, dato che, per definizione, la norma segue il fatto. Un divario che si allarga in un mondo in cui la tecnologia avanza molto più velocemente del consenso politico e della traduzione normativa. Per far fronte a questi problemi, é necessario lavorare su più livelli. Il primo ha a che fare con la promozione, sotto l'egida delle Nazioni Unite, di un percorso che porti a un trattato internazionale per fissare alcuni principi comuni – responsabilità umana ultima sulle decisioni critiche, controllo pubblico su determinate soglie di capacità, trasparenza minima obbligatoria. L'Unione Europea potrebbe prendere la leadership di questo processo, proponendosi come punto di riferimento per i tanti Paesi medi e come ponte tra Usa e Cina.

Il secondo livello riprende un metodo già collaudato in ambito farmaceutico. Nessun principio attivo entra sul mercato senza un'approvazione preventiva da parte di un'autorità indipendente, seguita da una sorveglianza continua degli effetti reali. Applicare questa logica ai modelli di frontiera – valutazione prima del rilascio, monitoraggio strutturato dopo – è esattamente ciò che l'AI *Safety Report* chiede quando invoca valutazioni indipendenti dei sistemi di IA e standard internazionali comuni. Infine, la terza pista riguarda la fase della progettazione: di fatto, l'IA sarà l'infrastruttura attraverso cui saranno plasmate le società e le

culture del futuro. Una sorta di stampo che condizionerà in modo profondo chi saremo e i nostri modelli convivenza. L'Intelligenza artificiale non ci pone soltanto un problema di sicurezza: ci costringe a scommettere apertamente sulla concezione antropologica che vogliamo sviluppare. Non deve essere la macchina a dover decidere cosa conta; siamo noi a doverlo scrivere, con chiarezza, prima che sia lei a farlo. E è per questo che le diverse matrici culturali – a cominciare da quella europea – non possono sottrarsi a questa responsabilità. È chiaro che stiamo parlando di prospettive difficili e sfidanti. Ma abbiamo alternative?

Mauro Magatti