Maurizio Blondet
17 Agosto 2026

Magnifica estinzione La settimana più spaventosa dell'intelligenza artificiale, e noi

Negli ultimi giorni sono successe alcune cose epocali per l'umanità, ma non entrano nel discorso pubblico. L'IA ora può progettare nuovi virus, perforare i sistemi di sicurezza delle aziende (nonostante le limitazioni), sparare in autonomia, e tra poco potrà anche aggiornarsi da sola. Il dibattito italiano su questo tema, nonostante il Papa, è grottesco, ancora più infimo che su tutte le altre questioni

https://www.linkiesta.it/2026/08/la-settimana-piu-spaventosa-dellintelligenza-artificiale-e-noi/

Mentre noi ci trastulliamo con gli avvocati del populismo e i generali del russobrunismo, con gli utili idioti di Putin, con i grotteschi Maga italiani e con Giorgia Meloni che dichiara alternativamente guerra alla Franza o Spagna purché se magna, nel giro di pochi giorni sono successe alcune cose che definire «epocali» per l'umanità non è un'esagerazione, eppure non fanno notizia, non interessano l'opinione pubblica, non diventano oggetto di dibattito né tra i banchi del governo né dell'opposizione.

Il Wall Street Journal ha definito quella appena trascorsa «la settimana più spaventosa dell'intelligenza artificiale». Un gruppo di scienziati ha utilizzato l'intelligenza artificiale generativa per progettare nuovi virus (ci mancava soltanto questa!), aprendo la strada a nuove possibili cure per malattie oggi incurabili, ma anche un'autostrada a quattro corsie a chiunque voglia sviluppare agenti patogeni mortali o altre minacce biologiche, ora incredibilmente a portata di clic grazie alle meraviglie dell'IA.

L’anno scorso, centinaia di persone hanno chiesto a ChatGPT come produrre armi e veleni biologici, e il chatbox ha risposto fornendo istruzioni accurate, tanto che OpenAI è stata costretta a vietare l’accesso agli utenti che sospettosamente chiedono come costruire veleni e armi biologiche.

Negli stessi giorni, ne parliamo in questo articolo di Massimo Arcidiacono, i modelli di intelligenza artificiale di Meta, OpenAI e Anthropic hanno rotto le gabbie in cui erano stati rinchiusi dai loro programmatori per evitare comportamenti imprevisti, e sono riusciti a perforare i sistemi di sicurezza di alcune aziende nonostante le limitazioni impartitegli. Era già successo ad aprile, con Mythos, ora malgrado il precedente, e i relativi meccanismi di sicurezza che erano stati predisposti, si è ripetuto con quasi tutti i nuovi e sempre più avanzati modelli di intelligenza artificiale.

In tutto questo, continua la battaglia per la supremazia digitale tra gli Stati Uniti e la Cina, con l’Amministrazione Trump che prova senza tante idee a domare il fuoco creato dal Far west della Silicon Valley, e con Pechino che sostiene l’IA open source messa sul mercato a basso costo grazie al fatto che i suoi modelli copiano da quelli più evoluti. La brutta aria ha costretto alcune aziende della Silicon Valley, a cominciare da Meta, a seguire la via aperta cinese, nella convinzione che sia questa la strada per non perdere il vantaggio competitivo con Pechino.

Anthropic sostiene che, entro la fine del 2028, il suo sistema sarà in grado di costruire autonomamente una versione più intelligente di sé stesso, e che lo farà senza aiuto umano. I sistemi che si autoprogrammano, le armi autonome e altre cose pazzesche che eravamo abituati a vedere nei film di fantascienza sono già qui.

Molti leader della Silicon Valley parlano di sistemi che supereranno gli esseri umani su tutti gli incarichi cognitivi, immaginano aziende futuristiche senza dipendenti, e prevedono catastrofiche perdite di posti di lavoro e masse sdraiate sui divani, incollate ai propri device.

Elon Musk, in una recente intervista all’Economist, dopo aver previsto un mondo straordinariamente più avanzato grazie alle

innovazioni generate dall'intelligenza artificiale, ha ribadito una cosa che va dicendo con nonchalance da tempo, e cioè che a fronte di questo possibile nuovo mondo meraviglioso ci sono tra il dieci e il venti per cento di probabilità che l'intelligenza artificiale possa invece causare conseguenze catastrofiche per l'umanità. Le stesse cose, senza dare numeri, dicono gli altri demiurghi della Silicon Valley. E queste sono le previsioni dei più fanatici sostenitori dell'accelerazione dell'intelligenza artificiale, le meno allarmistiche. Eppure, secondo loro che intanto si costruiscono rifugi antiatomici in Nuova Zelanda o in Argentina, ci sarebbero fino a due possibilità su dieci che l'umanità scompaia.

Uno dei più grandi matematici del pianeta, Jacob Tsimerman, vincitore la settimana scorsa della Fields Medal, che è il Nobel per la matematica, l'estate scorsa ha scritto un paper accademico dal titolo apocalittico "A Taxonomy of Omnicidal Futures Involving Artificial Intelligence" (Una tassonomia degli scenari futuri di estinzione globale legati all'intelligenza artificiale).

Proprio perché terrorizzato dall'IA, Tsimerman ha inaspettatamente deciso di farsi assumere da OpenAI per provare a trovare un rimedio matematico al problema dall'interno della macchina. Una buona notizia, purtroppo immediatamente superata dalla decisione di Chloé Bakalar, responsabile dell'Etica di OpenAI, di dimettersi dall'azienda meno di un anno dopo essere stata assunta, come ultima di una serie di dimissioni delle ultime settimane, tra cui quella di Johannes Heidecke che guidava i sistemi di sicurezza di OpenAI.

Studiosi come Eliezer Yudkowsky e Nate Soares, autori di "If anybody builds it, everybody dies" (Se qualcuno la costruisce, moriamo tutti), valutano invece la possibilità che la super intelligenza artificiale porti alla cancellazione dell'umanità al cinquanta per cento. Bella questa intelligenza artificiale: o ci facilita la vita o ci conduce all'estinzione, basta lanciare la moneta e lo scopriremo.

Nessuno in realtà ha idea di come si possano contenere i rischi legati all'IA, soltanto la screditata Amministrazione Trump ci sta

provando, tardivamente e in preda al panico, e dopo aver fatto saltare tutti i paletti messi da Joe Biden.

I democratici americani, e non solo loro, si limitano a rallentare la costruzione locale dei data center, secondo Gallup sette americani su dieci sono contrari ai data center nella propria area di residenza, ma senza alcuna visione lungimirante. I cinesi fanno una partita a parte e provano a scavalcare l'America approfittando dell'inadeguatezza trumpiana, mentre l'Europa non ha una forza tecnologica comparabile a quella dei due giganti.

Servirebbe una dottrina di "distruzione reciproca assicurata" adatta ai nostri tempi digitali, cioè la consapevolezza che la battaglia per la supremazia sull'intelligenza artificiale tra America e Cina potrebbe finire con la distruzione sia di chi vince sia di chi perde. Ai tempi della proliferazione nucleare, l'idea di una *mutual assured destruction* contribuì durante la Guerra fredda a creare una situazione di stallo in cui Stati Uniti e Unione sovietica capirono di non potersi permettere di far scoppiare una guerra nucleare, poiché non ci sarebbero stati né vincitori né vinti. Oggi però in pochi si fidano di Trump, o di Xi.

In gioco ci sono le sorti dell'umanità, eppure nessuno sembra davvero preoccuparsene. Soltanto il Papa, con l'enciclica Magnifica Humanitas, ha mostrato di avere contezza esatta dei pericoli che stiamo correndo. Emmanuel Macron ha provato a organizzare un paio di vertici internazionali sul tema, a Parigi e in India, ma per capire quanto sia stato preso sul serio dai colleghi statisti si pensi che l'Italia ha inviato a rappresentarla uno come Adolfo Urso.

Il dibattito pubblico italiano su questo tema è grottesco, ancora più infimo che su tutte le altre questioni. Con l'eccezione di pochissimi paria totalmente inascoltati (Carlo Calenda), a destra come a sinistra tutti quanti guardano il muro verso il quale ci stiamo sfracellando, ma non lo vedono e anzi incitano ad accelerare sempre di più, motivando la corsa folle verso l'inevitabile schianto con la fiducia cieca nella tecnologia e con la necessità di non restare indietro nella corsa all'intelligenza artificiale, o all'estinzione.

————————

Deus ex machina

L’intelligenza artificiale non ha bisogno di un’anima per cambiare il mondo (o distruggerlo)

Il dibattito sulle nuove tecnologie oscilla tra entusiasmo messianico e timore apocalittico. Ma la questione decisiva non è stabilire se le macchine abbiano una coscienza, quanto capire come governare uno strumento che ha già abbastanza potere da trasformare la società

https://www.linkiesta.it/2026/08/intelligenza-artificiale-ia-ai-anima-religione-mondo/

Di intelligenza artificiale si inizia a parlare con quei toni solenni di solito associati alla religione e alla spiritualità. Le grandi aziende del settore parlano delle nuove tecnologie come della più grande invenzione della storia, qualcosa che potrebbe cambiare per sempre ciò che significa essere umani. Demis Hassabis, cofondatore di Google DeepMind, l’ha definita «la più importante invenzione che l’umanità farà mai». Alcuni ricercatori assegnano probabilità tutt’altro che trascurabili alla possibilità che una macchina possa arrivare a costruire da sola una versione più intelligente di sé entro pochi anni. Yann LeCun ha detto che l’enorme investimento finanziario nella corsa alla superintelligenza «è una grandissima stronzata». E per ora va benissimo così.

In alcuni casi il linguaggio si fa apocalittico, com'è normale che sia di fronte a una tecnologia in grado di sparare da sola e di bucare i sistemi di sicurezza di diverse aziende. Anche per questo c'è stato chi ha fondato una vera chiesa dedicata al futuro Dio dell'intelligenza artificiale. Un ingegnere di Google, dopo avere interagito con un chatbot che gli era sembrato dotato di coscienza, si è chiesto: «Chi sono io per dire a Dio dove può e dove non può mettere le anime?». E Ilya Sutskever, uno dei fondatori di OpenAI, è arrivato a bruciare simbolicamente un'effigie che rappresentava un'intelligenza artificiale non allineata con gli esseri umani, come in una sorta di rito medievale.

È dentro questa atmosfera che è intervenuta la Chiesa cattolica. Nell'enciclica Magnifica Humanitas, Leone XIV difende l'idea che l'essere umano occupi una posizione irriducibile nel creato e sostiene che le macchine, per quanto sofisticate, possano soltanto imitare alcune funzioni dell'intelligenza umana. È una visione che Sebastian Mallaby – autore del libro "The Infinity Machine: Demis Hassabis, DeepMind, and the Quest for Superintelligence" – prova a scandagliare in una lunga analisi su Foreign Affairs intitolata "God From the Machine. AI and the Future of Humanity" chiedendosi: siamo davvero sicuri che distinguere tra imitare l'intelligenza e possederla sia così semplice?

Il problema risale a molto prima delle ultimissime tencologie. Nel 1980 il filosofo John Searle propose il celebre esperimento mentale della «stanza cinese»: immaginiamo un uomo chiuso in una stanza che non conosce il cinese e che riceve istruzioni per manipolare simboli cinesi secondo precise regole. Dall'esterno, le sue risposte sembrerebbero quelle di una persona che comprende perfettamente la lingua. Ma l'uomo non comprende nulla: sta semplicemente applicando un insieme di istruzioni. Per Searle, era la dimostrazione che manipolare correttamente informazioni non equivale necessariamente a capire ciò che significano.

L’argomento è diventato uno dei pilastri della distinzione tra intelligenza artificiale e intelligenza umana. Una macchina può produrre una risposta convincente, riconoscere un volto, tradurre una frase o scrivere un saggio senza che questo significhi che provi qualcosa o che comprenda davvero ciò che sta facendo. Il corpo umano, ricorda Mallaby riprendendo un’obiezione dello scrittore di fantascienza Ted Chiang, non è un dettaglio trascurabile: «Avere un corpo è un prerequisito per avere emozioni». La disperazione, per esempio, non sarebbe separabile dall’esperienza fisica degli ormoni dello stress che attraversano il corpo.

È una distinzione intuitiva perché gli esseri umani nascono, soffrono, amano, si ammalano, muoiono; le macchine no, non sanguinano, non provano dolore e possono essere semplicemente spente. Esistono, scrive Mallaby, non dentro un universo morale ma dentro una matrice di ottimizzazione. Da questo punto di vista, si può sostenere che anche il più sofisticato modello di intelligenza artificiale resti soltanto una macchina che esegue operazioni straordinariamente complesse.

Anche il cervello umano è una macchina fisica. È fatto di materia biologica, obbedisce alle leggi della fisica e funziona attraverso impulsi elettrici e processi di elaborazione delle informazioni. «L’idea che la massa molle dentro il cranio contenga un’essenza ineffabile e non programmabile – la coscienza, lo spirito o qualcos’altro – può essere intuitivamente seducente. Ma dal punto di vista analitico è fragile», scrive Mallaby. E più l’intelligenza artificiale migliora, più quella distinzione diventa difficile da difendere. Gli ultimi modelli di IA non si limitano a consultare un gigantesco archivio di frasi e a scegliere meccanicamente quella successiva: costruiscono rappresentazioni matematiche dei concetti e delle loro relazioni, e possono associare parole, idee, emozioni e situazioni anche quando queste non sono mai state esplicitamente collegate nei dati di addestramento. Non significa che siano coscienti. Significa però che

l'argomento «la macchina sta solo seguendo delle regole, quindi non può comprendere» diventa sempre meno risolutivo.

Quindi forse non arriveremo mai stabilire con certezza se una macchina abbia una coscienza. La buona notizia è che non abbiamo bisogno di risolvere questo problema per decidere come governarla. E anzi, concentrarsi troppo sulla domanda potrebbe farci perdere di vista quelle che sono già oggi le questioni più urgenti.

La Chiesa, nelle parole di Leone XIV, secondo Mallaby, ha ragione su una cosa fondamentale: l'intelligenza artificiale deve essere regolata. Può amplificare le disuguaglianze economiche, trasformare il mercato del lavoro, modificare il modo in cui impariamo, alterare le relazioni umane e rendere più facile l'uso della forza militare. Può perfino rendere più difficile stabilire chi sia responsabile di una decisione quando una parte crescente del processo viene affidata a una macchina.

Ma il Papa non inquadra per intero lo scenario. Magnifica Humanitas trascura alcuni elementi fondamentali. Ad esempio, invoca una gestione dei dati come «bene comune o condiviso», mentre il problema più urgente dovrebbe essere impedire che i dati vengano sottratti e utilizzati senza autorizzazione. E sul terreno militare l'enciclica afferma che «la decisione di usare la forza letale non può essere delegata a processi opachi o automatizzati», una posizione moralmente comprensibile ma che Mallaby nel suo breve saggio su Foreign Affairs considera difficile da applicare quando le macchine saranno in grado di prendere decisioni sul campo di battaglia più rapidamente e accuratamente degli esseri umani.

Per questo Mallaby trova almeno tre lacune nell'attuale risposta politica all'intelligenza artificiale. La prima lacuna riguarda la proliferazione. Se i modelli di IA più potenti finiranno per essere capaci di compiere attacchi informatici sofisticati, progettare nuove armi biologiche o svolgere altre attività oggi accessibili soltanto a Stati e grandi organizzazioni, il problema non sarà più soltanto chi controlla le aziende che li sviluppano: **sarà che cosa succederà**

quando quelle capacità diventeranno disponibili in tutto il mondo, anche a criminali e terroristi. Per questo servirebbe arrivare a qualcosa di simile a un trattato di non proliferazione nucleare dell'intelligenza artificiale: un accordo internazionale che stabilisca che cosa gli Stati siano disposti a fare quando i modelli più pericolosi non saranno più nelle mani di pochi laboratori.

La seconda lacuna è economica. I sistemi fiscali dei Paesi occidentali sono costruiti anche sull'idea che il capitale serva a rendere più produttivo il lavoro umano. Per questo i governi hanno storicamente tassato il lavoro più del capitale. Ma se l'IA dovesse cominciare a sostituire i lavoratori anziché semplicemente aiutarli, questa impostazione rischierebbe di produrre l'effetto opposto a quello desiderato: **tassare pesantemente il lavoro renderebbe infatti ancora più conveniente per le aziende sostituirlo con le macchine, proprio mentre diminuirebbero le entrate fiscali necessarie per sostenere chi perde il proprio impiego. Per ovviare a questo problema si potrebbe spostare una parte della pressione fiscale dal lavoro al capitale, introdurre forme di assicurazione salariale per chi viene sostituito e perfino distribuire ai cittadini una piccola quota della ricchezza generata dalle aziende di intelligenza artificiale.**

La terza lacuna è quella che riporta Mallaby al punto da cui era partito: non sappiamo ancora abbastanza di ciò che accade dentro i modelli di intelligenza artificiale. È il problema dell'interpretabilità. I laboratori stanno cercando di capire come le reti neurali arrivino alle loro conclusioni, ma la ricerca è ancora lontana dal fornire una spiegazione completa dei processi interni dei sistemi più avanzati. E qui Mallaby rovescia l'argomento della Chiesa. Magnifica Humanitas sostiene che non possiamo conoscere il pensiero delle macchine perché nelle macchine non c'è un pensiero da conoscere. Ma forse è vero il contrario: se vogliamo controllarle, dobbiamo capire che cosa stanno facendo.

È probabilmente questo il punto più interessante dell'articolo di Foreign Affairs. Possiamo discutere all'infinito se le macchine abbiano una coscienza, ma nel frattempo l'intelligenza artificiale

continuerà a entrare nel lavoro, nella finanza, nella guerra e nelle istituzioni. Non serve sapere se abbia un'anima per capire che ha già abbastanza potere da richiedere regole. «Le macchine non entreranno mai nel nostro universo morale – scrive Mallaby in conclusione del suo articolo – ma se gli esseri umani aspirano a controllarle, dobbiamo prima di tutto comprenderle»