

Storie

N.11
SETTEMBRE 26



ilmondonuovo.club/MILANO

Il festival della filosofia di quest'anno è CAOS. E la maglietta ufficiale della XXVI edizione è un celebre frammento di Esiodo, "per primo venne il caos".

Contrariamente a quanto si pensa comunemente, il significato di "Caos" nell'antica Grecia è molto diverso da quello moderno di "disordine" o "confusione". Per Esiodo il Caos un vuoto primordiale, una sorta di buco nero da cui, grazie ad una successiva differenziazione, prende vita il mondo strutturato che abitiamo. Un vuoto oscuro, privo di materia ma denso di potenziale creativo. Secoli dopo Ovidio riscrive nelle sue metamorfosi i miti greci, e cambia il significato di caos, perché non concepisce una creazione che nasce da un nulla assoluto. Per Ovidio prima del mondo ordinato doveva esistere già una materia, una massa che per quanto informe, rozza e confusa contenesse già tutti gli elementi del cosmo, terra, aria, acqua e fuoco. Ovidio ha letteralmente "riempito" il vuoto dei greci con un ammasso disordinato di materia e di elementi. Il concetto di Caos contemporaneo è quello di Ovidio, non quello dei classici greci. Di Esiodo resta però l'idea che il Caos sia comunque generativo di qualche cosa. Sta a noi io credo lavorare per dare una direzione a questa spinta generativa....

Il mondo dell'innovazione e del digitale oggi, quello degli algoritmi e delle intelligenze artificiali, si basa soprattutto su una capacità straordinaria di calcolo da parte delle macchine, sulle reti neurali, e sulla capacità di addestrare queste reti. I risultati prodotti dalle IA sono sorprendenti e probabilmente lo saranno sempre di più man mano le Intelligenze artificiali si esercitano e imparano da quello che facciamo o abbiamo fatto noi. Eppure noi non conosciamo come fa la macchina a produrre i risultati che produce. Considerando l'impatto e la diffusione delle IA l'opacità del processo, la scatola nera, fa paura. A sentire i matematici che a Modena riflettono sulle frontiere filosofiche e sui limiti della loro disciplina i dubbi aumentano. Quando Turing teorizzò la sua macchina universale, anticipando concettualmente il computer sviluppato poi da Von Neumann, definì l'algoritmo come un dispositivo che consente di arrivare ad un risultato, cioè una scatola da cui entra un numero e ne esce un'altra. Quello che contava era avere a disposizione una macchina che desse la possibilità di ottenere un risultato più che di capire esattamente come. Il problema di conoscere il processo resta ancora oggi non risolto.

Pause ai è un movimento di attivisti nato ad Utrecht in Olanda oggi diffuso in tutto il mondo. I suoi rappresentanti si occupano di digitale e di IA, e ritengono che la velocità con cui si sta sviluppando la ricerca e gli ambiti di applicazione della tecnologia siano talmente rapidi da rendere impossibili o quasi le necessarie operazioni relative al controllo e alla sicurezza delle persone che le usano quotidianamente e per tutto. La richiesta di una pausa servirebbe per il gruppo di attivisti a studiare nuovi sistemi di controllo e capire meglio come funziona la tecnologia, partire dalle reti neurali, le famigerate Black Box. Oltre al gruppo di Pause AI, c'è quello di STOP AI, più radicale e più attento agli effetti collaterali dell'innovazione e della velocità con cui si sviluppa.

A luglio è stato presentato un report preliminare sull'IA curato tra gli altri da Yoshua Bengio, cui è stata affidata la presidenza del comitato scientifico che ne ha citato la redazione. Abbiamo pubblicato una traduzione del testo sperando di semplificare la lettura. Il report è una sintesi dell'ambivalenza della tecnologia, da sempre rischio e opportunità, che richiede visione e coraggio di fare scelte.

L'inedito di questo contesto è la velocità e la pervasività. Sicuramente il digitale ha cambiato il mondo come altri grandi innovazioni tecnologiche e scientifiche nella storia dell'uomo sapiens sapiens. Ma oggi come in passato serve scrivere le regole del mondo nuovo.

I rischi ci sono, nel web c'è di tutto, compresi i gruppi informali che incitano alla violenza, ma ci sono anche le opportunità. Può essere uno strumento prezioso per il mondo nel terzo settore, può essere l'alter ego di un artista, dipende da noi, da quello che gli insegniamo e da quello che pensiamo debba fare.



2/il punto

5/PAUSE AI



**18/
YOSHUA BENGIO**



21

CAOS

29

**L'odio, il radicalismo,
L'accelerazionismo
i neonazisti sul web:
NLM, 764, The Siege,
THE LORD OF CAOS**

**38/
SOCIAL CHALLENGE**

indice



**44/ i modelli
semantici**

**47/ identità
tecnologia e nuove
forme di umanità**

**50/ cosa stiamo
insegnando all'IA...**



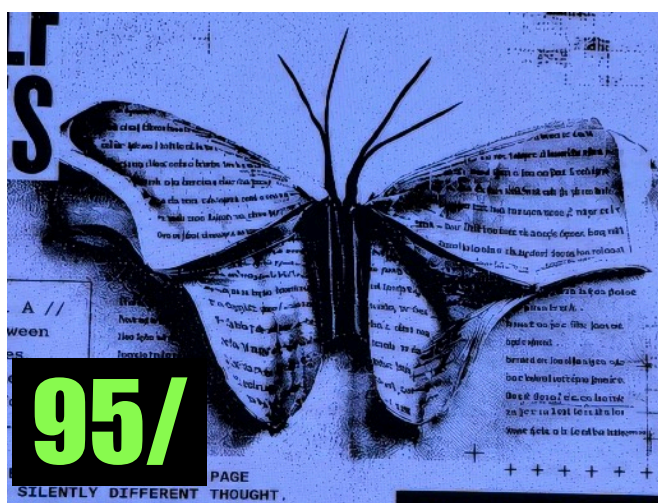
**54/ LE PAROLE
CHE TENGONO
INSIEME IL MONDO**

58/

**IL REPORT ONU
SULLA I.A:**

84/JARVIS

92/ I "TECNODEVOTI"



95/

**LOREM IPSUM
VARIORUM**

CREDITS:

Cristina Andreotti, Igor Boni, Diego Castagno,
Simona Conte, Leonardo Lotto, Paolo Lorusso,
Giampaolo Sodano, Diego Spadafora, Gianni Zais,

COLOPHON

DIRETTORE

Giampaolo Sodano

VICE DIRETTORE

Diego Castagno

COMITATO DI DIREZIONE

Gianni Zais, Gaia Bertotti, Sergio Bellucci, Giuseppe

Bea, Aldo Di Russo, Giuseppe Attene

REDAZIONE

01036 Nepi (Viterbo). Via di Monte Topino, 303

info@ilmondonuovo.club

Tutte le collaborazioni sono volontarie. La
riproduzione è vietata senza l'autorizzazione
dell'editore.

Mensile On Line allegato al periodico trimestrale "il
mondo nuovo.club" iscritto con decreto n. 635/2025
del 18 febbraio 2025 (RG n.255/2025) del pubblico
registro della stampa del Tribunale Ordinario di
Viterbo. Direttore Responsabile Giampaolo Sodano

La frase in copertina è di Emmanuel Macron





JUST
DON'T
BUILD
AGI

HERE IS EXPERT
THAT IT WON'T
TION

JUST
DON'T
BUILD
AGI

"UNTIL THERE IS EXPERT
CONSENSUS THAT WE SHOULDN'T
CAUSE HUMAN EXTINCTION"

IF YOU CAN'T STEER
DON'T RACE



IF YOU CAN'T
STEER
DON'T RACE

IF YOU
CAN'T
STEER
DON'T
RACE



PAUSE AI

La narrazione dominante sull'intelligenza artificiale generativa oscilla da anni tra il trionfalismo evolutivo della Silicon Valley e un catastrofismo spesso ideologico e di maniera. Entrambi gli schieramenti commettono spesso lo stesso errore analitico e metodologico, cioè trattano la tecnologia come un fenomeno naturale, "inevitabile" e privo di coordinate politiche, economiche e sociali. La portata della trasformazione in corso invece implica una concentrazione di potere nelle mani di pochi player privati che ha pochi precedenti nella storia.

Secondo gli economisti Daron Acemoglu e Shoshana Zuboff, le grandi corporazioni tecnologiche hanno attraversato tre fasi sequenziali distinte. Si parte dall'accumulo economico, ovvero una generazione di profitti straordinari attraverso l'estrazione e la raffinazione dei dati umani; si passa poi al potere d'influenza, convertendo la ricchezza in pressione politica sui regolatori e sul controllo dei canali informativi; infine si giunge al potere strutturale. Oggi le piattaforme sono diventate l'infrastruttura essenziale e insostituibile dello Stato: cloud, calcolo, comunicazioni, difesa.

Detto questo non ha molto senso trattare l'IA di frontiera alla stregua delle grandi industrie del passato. Se le multinazionali del Novecento esternalizzavano i costi sulla salute o sull'ambiente preservando la natura di "prodotto", l'IA opera come un'infrastruttura cognitiva autonoma. Non è un bene che si acquista, ma l'ambiente in cui si elabora la realtà.

Quando si osserva che questa rivoluzione tecnologica ed industriale assomiglia molto e pochissimo alle altre semplicemente si sostiene che la transizione digitale ha effetto e impatto sulla società e sull'economia come le altre, e assieme che questi effetti non sono paragonabili per portata e entità.

Nel primo e nel secondo ciclo, quello del vapore e della elettricità, la macchina ha sostituito la forza muscolare e la manualità artigianale, creando la necessità di una classe di lavoratori cognitivi per gestire la complessità. Nel terzo ciclo, l'informatica e i Big Data, la tecnologia ha automatizzato il calcolo e la memoria, trasformando il comportamento in materia prima.

Con l'IA di frontiera si compie un ulteriore salto qualitativo: la macchina punta a espropriare la facoltà di giudizio, la sintesi simbolica e la decisione. Per la prima volta, la tecnologia non funziona solo da prolungamento del corpo o da archivio, ma sostituisce l'elaborazione cognitiva dell'uomo.

L'ideologia della Singolarità come copertura

In questo quadro, le teorie sull'accelerazionismo e sulla "Singolarità" rese celebri da Ray Kurzweil hanno una precisa funzione di narrazione spesso ideologica. Presentando il progresso come una legge matematica esponenziale e deterministica, la Silicon Valley de-politicizza la scelta. Se la fusione tra uomo e macchina è un passaggio evolutivo inevitabile, ogni tentativo di regolamentazione o redistribuzione viene bollato come un'inutile resistenza luddista. L'effetto di questa mitologia è di fatto il congelamento dell'iniziativa legislativa, con la politica che lascia che la transizione venga guidata da attori privati orientati al profitto e all'ottimizzazione.

Le istituzioni hanno mancato in questi anni di prospettiva. Hanno ridotto la questione a mero tecnicismo giuridico, per poi delegare la governance ad agenzie di conformità, e sostanzialmente rinunciando a gestire i rapporti di forza che nel frattempo diventavano sempre più asimmetrici. Un vero e proprio "arretramento" della politica. Osserva ad esempio Umberto Galimberti che oggi si assiste all'inversione nel rapporto tra la tecnologia e l'uomo: la Tecnica cessa di essere un mezzo e diventa il fine stesso della società, il criterio supremo diventa la pura efficienza, mentre ciò che non è calcolabile viene espulso.

Sul piano della filosofia politica e del diritto, come sottolineato da Sebastiano Maffettone, questo crea un cortocircuito insanabile. Le istituzioni democratiche e la giustizia distributiva si fondano su criteri di equità, diritti e bilanciamento dei poteri che sono, per definizione, "inefficienti" se paragonati alla logica algoritmica. Delegare l'infrastruttura cognitiva a pochi attori privati significa ridurre la sovranità dello Stato a un guscio privo di reale capacità decisionale.

È dentro questo scarto che si posizionano i movimenti che oggi chiedono pause, sospensioni e moratorie sullo sviluppo dei modelli di frontiera. Le mobilitazioni per il blocco o la limitazione dell'IA non esprimono una reazione d'ansia o una chiusura oscurantista. Rappresentano, al contrario, l'ultimo tentativo consapevole di rioccupare il terreno della decisione politica prima che la delega decisionale agli apparati tecnici diventi irreversibile.

I social media sono stati la prova generale, il momento in cui la politica ha ceduto il controllo della sfera pubblica e dell'attenzione. La partita sull'IA è il tabellone principale. Se la politica non tornerà ad affrontare il nodo della concentrazione del potere economico e della sovranità sulle infrastrutture cognitive, l'essere umano non rischierà soltanto di perdere quote di lavoro, ma la sua stessa funzione di soggetto decidente. Tronando al nostro Belpaese, la politica in ritardo evidente rispetto alla tecnologia ha un alibi imperfetto ma che regge nell'opinione pubblica: la convinzione che, siccome le grandi piattaforme sono americane o cinesi, e sono troppo potenti "noi non possiamo farci niente". Questa forma di subalternità culturale permette di rinunciare in partenza a qualsiasi idea di sovranità industriale o di politiche attive del lavoro. Un galleggiamento destinato a precipitare, perché gli effetti del cambiamento avvengono ovunque. Quando le prime ristrutturazioni aziendali e i tagli al personale nel settore dei servizi verranno giustificati con l'integrazione di sistemi generativi, la questione si sposterà di colpo dal piano "teorico" a quello dello scontro sociale sui territori, e la politica sarà per forza costretta a rincorrere il problema, come ultimamente accade ovunque nel mondo.

David Scott Krueger

David Scott Krueger è un informatico statunitense, professore universitario e noto ricercatore nel campo dell'apprendimento automatico, il Machine Learning. È una delle figure di riferimento globali per quanto riguarda la sicurezza e l'allineamento dell'intelligenza artificiale, l'AI Safety e l'AI Alignment. Formatosi nel deep learning sotto la guida di figure chiave del settore come Yoshua Bengio, Krueger ha un curriculum impressionante. Lavora come Assistant Professor in "Robust, Reasoning, and Responsible AI" all'Università di Montreal. È un membro accademico del prestigioso Mila – Quebec Artificial Intelligence Institute, guidato da Yoshua Bengio. Ha precedentemente ricoperto il ruolo di Assistant Professor presso l'ateneo britannico. Detiene una cattedra di ricerca CIFAR AI Chair e la cattedra IVADO in IA responsabile. Tra gli incarichi istituzionali la collaborazione come direttore di ricerca con la UK AI Task Force (il team governativo britannico per la sicurezza dell'IA) e l'attività nel team di sicurezza di Google DeepMind. Krueger è anche un attivista. La cui attività si concentra sulla riduzione dei rischi sociali ed esistenziali legati allo sviluppo di intelligenze artificiali avanzate (AGI), incluso il rischio di estinzione umana. Nel 2025 fonda Evitable, un'organizzazione no-profit nata per sensibilizzare l'opinione pubblica e i governi sulla crisi e sulle minacce derivanti dall'IA di frontiera.

I suoi studi analizzano la manipolazione algoritmica, la robustezza dei modelli profondi e i meccanismi di disallineamento in cui i sistemi IA potrebbero perseguire obiettivi autonomi dannosi. Krueger è un sostenitore attivo della regolamentazione severa del settore, con interventi incentrati sul rischio di depotenziamento della forza lavoro e sulle implicazioni militari dell'IA. Attraverso "Evitable" e Krueger e i suoi collaboratori forniscono l'analisi sistemica e credibilità accademica ai movimenti che in questi anni stanno sviluppando l'attivismo di piazza la cui forza sta nell'utilizzare il linguaggio e le tattiche dei movimenti per il clima e dei diritti civili, coinvolgendo la società civile, studenti, lavoratori del settore tech disillusi, sindacalisti e attivisti ecologisti.

La black box

Facciamo un passo indietro per capire meglio di cosa parliamo

La questione della "scatola nera" (black box) nei modelli di intelligenza artificiale e negli algoritmi di raccomandazione rappresenta un punto di frattura strutturale tra chi produce tecnologia e chi tenta di regolamentarla.

La posizione delle Big Tech sta nel compromesso tecnico.

Per chi sviluppa i modelli, l'opacità decisionale non è considerata una ragione sufficiente per fermare o rallentare il rilascio dei prodotti sul mercato. Si dà alle prestazioni empiriche e statistiche: le reti neurali profonde funzionano su scala di miliardi di parametri, e l'industria accetta il trade-off per cui un modello non interamente interpretabile, ma statisticamente accurato e performante, genera valore commerciale superiore rispetto a un sistema trasparente ma rigido o meno capace.

Queste stesse grandi aziende, tra cui Anthropic, OpenAI e Google DeepMind, finanziano la ricerca sull'interpretabilità meccanicistica e su tecniche di Explainable AI (XAI), secondo le quali l'obiettivo non è comprendere a priori ogni singola operazione matematica, ma correggere eventuali "errori" di funzionamento e "smontare" la rete ex-post per individuare specifici circuiti interni responsabili di comportamenti a rischio o allucinazioni. La gestione quindi del rischio è a valle: anziché bloccare il modello, si preferisce intervenire dopo se le cose non funzionano come si vorrebbe, tramite tecniche di allineamento, filtri di sicurezza e moderazione, per migliorare le risposte e correggere i comportamenti anomali evidenziati dall'uso.

Ovviamente questa logica non può coincidere con quella dei legislatori, che infatti si occupano di principi di trasparenza su cui incardinare gli obblighi di legge. Le istituzioni normative, in primis l'Unione Europea, affrontano la scatola nera da una prospettiva di tutela dei diritti fondamentali, sicurezza del prodotto e responsabilità civile.

I regolatori non chiedono di eliminare la complessità matematica dei modelli, ma impongono che ogni azione delle BIG TECH sia sottoposta a audit e verifica e che la responsabilità dei risultati sia sempre riconducibile a chi progetta e diffonde.

L'industria considera l'opacità un costo accessorio della potenza di calcolo da gestire con strumenti diagnostici, la regolamentazione la tratta come un rischio di sistema, trasferendo sulle aziende l'onere legale e finanziario di ogni errore generato dalla scatola nera.

Difficile oggi dire se e quando l'intelligenza artificiale smetterà di funzionare con una scatola nera di cui non si conosce il funzionamento. C'è il timore che l'interpretabilità dei modelli attuali sia complicata però dalla continua crescita di scala dei sistemi futuri, dal momento che la scala cresce più veloce delle diagnosi. Gli strumenti attuali riescono a mappare e interpretare circuiti e concetti isolati su modelli di piccole e medie dimensioni. Comprendere appieno la rete nel suo insieme resta ad oggi quasi impossibile per l'enorme quantità di parametri.

Si stima che entro la fine del decennio la ricerca sull'interpretabilità meccanicistica permetterà di avere strumenti automatizzati in grado di tracciare il ragionamento interno del modello per specifici domini di rischio e non per tutti.

Illume 4 Science

Si chiama "Illume-4-science", è un progetto della Scuola Normale superiore di Pisa e dell'Università di Pisa, coordinato da Fosca Giannotti, e ha l'obiettivo ambizioso di costruire strumenti che permettano agli scienziati di "aprire" la black box dell'IA e capire come arrivi alle sue conclusioni. Illume 4 science punta a costruire degli "spiegatori", come li chiama la coordinatrice del progetto, cioè moduli aggiuntivi che interagiscono con i modelli complessi in una modalità insegnante-studente, per ricostruire la logica seguita dalla macchina. Questo dovrebbe consentire di ottenere un formato di spiegazione adeguato di volta in volta al tipo di utente. Una volta consolidata, la piattaforma Illume potrà essere di aiuto a molti altri settori della ricerca, ad esempio la meteorologia o l'astronomia, o a tutte le discipline della scienza moderna basate su enormi quantità di dati.

L'intelligenza artificiale in medicina e nella biotecnologia non è una novità, negli ultimi anni si sono registrati progressi straordinari, ad esempio il ripiegato dell'elica del DNA, e comincia ad essere uno strumento potentissimo per i ricercatori. Con un problema per ora irrisolto: per quanto l'IA sappia già (quasi) perfettamente identificare molecole o prevedere l'efficacia di un farmaco, o costruire un vaccino, non si spiega come e perché lo fa. In altri termini non sappiamo come l'IA arrivi a certe conclusioni né quali siano stati i passaggi avvenuti nel suo cervello di neuroni artificiali.

Questo è il problema delle cosiddette black box, le scatole nere, i sistemi di machine learning sempre più usati nella ricerca, i cui processi interni restano misteriosi, quindi non del tutto controllabili. Illume-4-science va esattamente su questo terreno inesplorato, per cercare di costruire strumenti che permettano agli scienziati di capire come l'IA arrivi alle sue conclusioni, aprire le black box e comprendere la logica delle decisioni.

Il progetto coordinato da Fosca Giannotti e colleghi si inserisce nel filone di studi di sistemi di intelligenza artificiale di supporto alla ricerca biomedica, che nasce su basi sperimentali

Nei primi anni Settanta alla Stanford University grazie alle ricerche di un gruppo guidato da Edward Shortliffe, che si sviluppa negli anni Novanta quando si sperimenta il machine learning, ed ebbe una svolta definitiva nel 2012, con l'arrivo del deep learning e delle reti neurali capaci di elaborare grandi volumi di dati.

Nel 2020 fu testato Alphafold 2, il sistema di Deepmind che può predire in pochi minuti, e con un'accuratezza paragonabile ai metodi tradizionali, la struttura tridimensionale delle proteine e la loro funzione a partire dalla sequenza degli amminoacidi. Alphafold 2 è valso ai suoi creatori, Demis Hassabis e John Jumper, il premio Nobel per la Chimica nel 2024, e secondo i dati della stessa Deepmind, è stato usato da oltre due milioni di ricercatori in 190 Paesi del mondo.

In campo farmaceutico questi sistemi di ricerca avanzata hanno cambiato le regole del gioco: se con le tecniche tradizionali di ricerca in laboratorio sviluppare un nuovo farmaco richiedeva non meno di 10 anni, costi e investimenti fantascientifici ed elevate probabilità di fallimento (in media superiore al 90%), oggi con gli algoritmi di machine learning, capaci di analizzare rapidamente grandi banche dati di composti chimici e di identificare le molecole migliori, i tempi della ricerca si sono ridotti.

Le simulazioni al computer permettono di prevedere con margini di errore sempre più ridotti le proprietà dei farmaci e i relativi effetti collaterali.

Il fatto che questi strumenti si rivelino preziosi per la ricerca e che non si sappia come facciano a dare risposte così accurate è davvero un paradosso: e fino ad ora nessuno, nemmeno ci li ha progettati e costruiti riesce a dare una risposta in questo senso, in articolare nessuno sa esattamente come raggiungono la predizione finale.

Secondo la Giannotti, bisogna ricostruire la logica seguita dalla macchina a posteriori. In questo senso capire come funziona la scatola nera consente di migliorare tutto il sistema, considerando che l'IA è ormai sistemica. Aprire la scatola nera serve a garantire un uso etico, sicuro e verificabile dei dati.

Rosetta

Il pioniere assoluto che ha sviluppato per primo il software digitale per progettare proteine è lo scienziato statunitense David Baker, biochimico dell'Università di Washington. Questa scoperta è valsa a Baker il Premio Nobel per la Chimica. Alla fine degli anni '90, Baker sviluppò un software chiamato Rosetta. Il programma serviva originariamente a calcolare al computer come una sequenza di amminoacidi si sarebbe ripiegata nello spazio. La prima proteina artificiale fu costruita nel 2003, ribaltando l'uso del software. Baker e il suo team progettarono al computer una proteina completamente nuova da zero mai esistita in natura, dimostrarono in questo modo che la biologia poteva essere programmata digitalmente come un software informatico.

DeepMind

Mentre Baker creava nuove proteine partendo dal computer, rimaneva aperto il problema inverso: prevedere istantaneamente la forma 3D di una proteina naturale conoscendo solo il suo codice genetico. Questa seconda metà della rivoluzione digitale è stata compiuta da Demis Hassabis e John Jumper: I due ricercatori di Google DeepMind che hanno sviluppato l'Intelligenza Artificiale AlphaFold 2, una vera e propria svolta dell'IA. Nel 2020, AlphaFold ha risolto un mistero biologico che durava da 50 anni, quello del cosiddetto ripiegamento dell'elica del DNA, mappando in modo digitale la struttura tridimensionale di quasi tutte le proteine conosciute dalla scienza.

La scoperta è rivoluzionaria, perché mettendo assieme il software di progettazione di Baker e le intelligenze artificiali generative, la biologia è diventata una disciplina ingegneristica, vale a dire che gli scienziati possono "disegnare" al computer farmaci mirati contro i tumori, creare vaccini in pochi giorni, progettare nanomateriali intelligenti o enzimi artificiali capaci di degradare la plastica nell'ambiente. Senza passare o quasi dal laboratorio, in tempi e con costi ridottissimi rispetto al passato.

Il ripiegamento nello spazio, il protein folding, è fondamentale perché la forma tridimensionale di una proteina ne determina la funzione biologica. Se una proteina non si ripiega correttamente, non può agganciarsi alle altre molecole, esattamente come una chiave nella serratura sbagliata, lo stesso principio ritenuto responsabile di malattie come l'Alzheimer, il Parkinson o i tumori. Capire questa forma tridimensionale permette ai computer di progettare farmaci su misura, capaci di incastrarsi perfettamente nei punti deboli delle proteine "bersaglio" per bloccarle o attivarle.

**JUST
DON'T
BUILD
AGI**

**DON'T LET AI
DECIDE
YOUR FUTURE**

L'IA DI FRONTIERA

L'espressione "IA di frontiera" (Frontier AI) identifica i modelli di intelligenza artificiale più avanzati, complessi e potenti esistenti al mondo in un dato momento. Rappresentano la "linea di confine" (la frontiera, appunto) di ciò che la ricerca e la tecnologia tecnologica sono in grado di realizzare. Non si tratta di una categoria tecnica fissa, ma di una definizione dinamica e politica: ciò che è di frontiera oggi (es. i grandi modelli multimodali di punta con capacità di ragionamento e pianificazione) non lo sarà più tra due anni, esattamente come l'IA che batteva a scacchi l'uomo negli anni '90 oggi è considerata tecnologia ordinaria.

Per essere definito "di frontiera", un sistema di IA risponde a specifici parametri materiali e prestazionali. A differenza dell'IA "stretta" o tradizionale, progettata per un solo compito, come riconoscere transazioni bancarie fraudolente o leggere una radiografia, l'IA di frontiera è generalista. Uno stesso modello può scrivere codice, analizzare bilanci, comporre testo, interpretare video e pianificare flussi di lavoro complessi.

Un secondo aspetto riguarda la capacità di ragionamento, l'abilità di ragionamento logico-astratto, la pianificazione multi-step ed esecuzione autonoma di compiti senza che sia stata esplicitamente programmata per ciascuno di essi. L'IA di frontiera implica anche Infrastrutture e costi di investimento colossali, definiti soprattutto dai volumi di calcolo utilizzati per l'addestramento. Richiede cluster da decine di migliaia di chip dedicati (GPU), consumi energetici straordinari e investimenti finanziari nell'ordine di miliardi di dollari per singolo modello.

Trattandosi inoltre di modelli "black-box" di dimensioni gigantesche, la loro risposta a determinati input non è totalmente prevedibile ex ante. È proprio su questo punto che scattano i timori sulla sicurezza (loss of control).

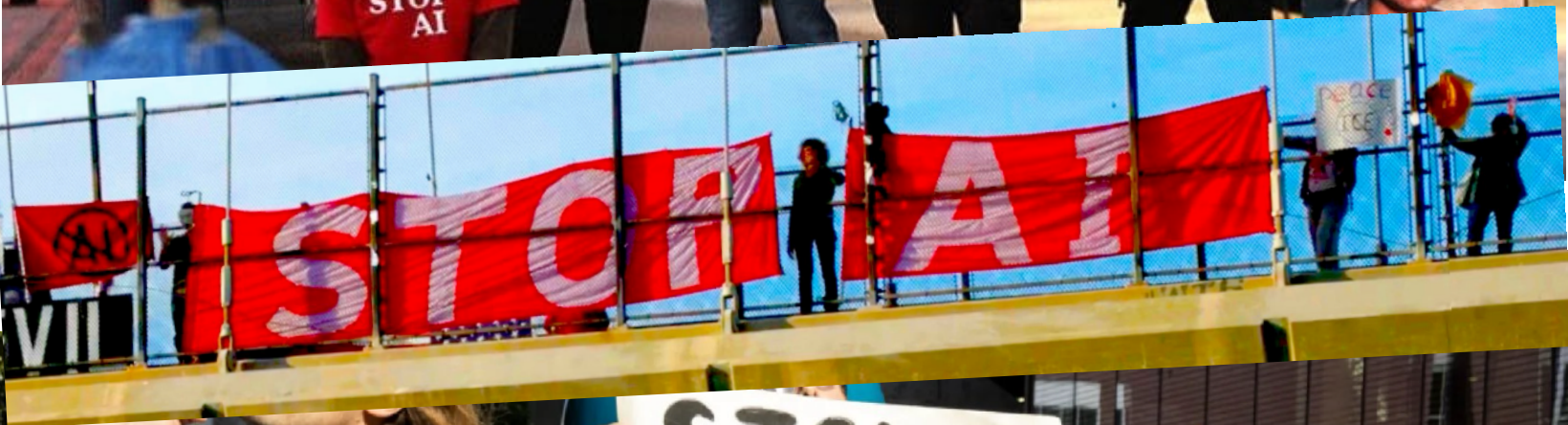
La definizione di "Frontier AI" è stata formalizzata ed usata da governi e istituzioni europee ed americane per tracciare una riga e dire: "Sotto questa soglia si applicano le regole ordinarie, sopra questa soglia le cose cambiano."

Le normative internazionali, ad esempio l'EU AI Act, classificano i modelli come "di frontiera" o a rischio sistemico quando la potenza di calcolo impiegata per addestrarli supera una determinata soglia matematica. Pochissimi attori commerciali al mondo, principalmente i colossi di San Francisco e della Silicon Valley come OpenAI, Anthropic, Google, e poche realtà asiatiche o europee, hanno i capitali e le infrastrutture per "stare sulla frontiera". Questa situazione crea un collo di bottiglia geopolitico ed economico industriale. L'IA di frontiera indica la testa della corsa tecnologica, con pochi competitori che si sfidano sui pochissimi modelli la cui potenza è tale da poter ristrutturare interi mercati del lavoro o porre problemi del tutto inediti di sicurezza, sovranità e regolamentazione.

L'idea che l'IA di frontiera debba essere fermata o strettamente controllata non nasce nelle piazze, ma all'interno della stessa comunità matematica e filosofica che la sta studiando:

L'effetto "Apprendista Stregone"

Più di 10 anni fa filosofi e informatici come Nick Bostrom, con il celebre libro Superintelligence del 2014, ed Eliezer Yudkowsky, il fondatore del MIRI, il Machine Intelligence Research Institute, iniziano a formulare il problema dell'alignment problem. In parole semplici, se si crea un'intelligenza artificiale generale (AGI) più intelligente dell'uomo, prima di aver risolto con certezza matematica il modo in cui imporre valori e obiettivi umani, il risultato non sarà la collaborazione, ma la perdita irrecuperabile di controllo, l'allineamento tra sviluppo e controllo. La conclusione logica di questa scuola di pensiero è: se non sappiamo come renderla sicura, la sola scelta razionale è non svilupparla.





JUST
ON'T
UILD
AGI

HERE IS EXPERT
S THAT IT WON'T
TION

JUST
DON'T
BUILD
AGI

"UNTIL THERE IS EXPERT
CONSENSUS THAT IT WON'T
CAUSE HUMAN EXTINCTION

IF YOU CAN'T ST
DON'T RA



IF YOU CAN'T
DON'T R

IF YOU
CAN'T
STEER
DON'T
RACE



PAUSE AI

Pause Giant AI Experiments

Future of Life Institute (FLI) è un'organizzazione non profit fondata nel 2014 con la missione di ridurre i rischi sistemici ed esistenziali legati alle tecnologie avanzate. Rappresenta la matrice istituzionale e filantropica da cui sono scaturiti molti dei dibattiti contemporanei sulla governance dell'intelligenza artificiale.

Tra i fondatori Max Tegmark, fisico e cosmologo del MIT, Jaan Tallinn, co-fondatore di Skype e grande finanziatore del movimento dell'"Altruismo Efficace", Anthony Aguirre, astrofisico e docente alla UC Santa Cruz, e Viktoriya Krakovna, ricercatrice di sicurezza presso Google DeepMind.

FLI ha operato storicamente come incubatore e think tank. Tra i finanziatori principali figurano la Musk Foundation e Vitalik Buterin, il co-creatore di Ethereum, che nel 2021 ha effettuato una donazione da oltre 600 milioni di dollari in criptovalute per sostenere programmi pluriennali sulla mitigazione dei rischi catastrofici.

L'istituto applica le premesse filosofiche del Longtermismo (Longtermism), secondo cui la priorità etica assoluta della società deve essere la salvaguardia della traiettoria di lungo periodo dell'umanità contro minacce di estinzione, ad esempio l'IA incontrollata, la guerra nucleare, la biologia sintetica, i disastri climatici.

Il 22 marzo 2023 l'istituto pubblica una lettera aperta denominata Pause Giant AI Experiments. La lettera non chiedeva l'interruzione di tutta la ricerca sull'IA, ma il congelamento dei modelli di frontiera "black-box" fino all'adozione di protocolli di sicurezza condivisi e sottoposti ad audit indipendenti.

Il testo chiede a tutti i laboratori di intelligenza artificiale di interrompere immediatamente per almeno 6 mesi l'addestramento di sistemi di IA più potenti di GPT-4. La pausa doveva essere pubblica, verificabile e applicata da tutti gli attori principali; in mancanza di un'adesione volontaria, il documento ipotizzava l'intervento dei governi tramite moratoria.

Nella lettera si sostiene che la corsa allo sviluppo di modelli "black-box" sempre più grandi e imprevedibili stia avvenendo senza pianificazione o gestione adeguate, con il rischio di impatti negativi sul lavoro e la perdita di controllo e di allineamento tra lo sviluppo del digitali e gli interessi umani.

Il periodo di blocco di 6 mesi non era inteso come stop alla ricerca generale sull'IA, ma come una sospensione selettiva e mirata allo sviluppo dei modelli "giganti" per definire protocolli di sicurezza condivisi e sottoposti ad audit indipendenti prima del rilascio di nuovi modelli.

La pausa secondo gli estensori del documento consente alle istituzioni politiche e alla società di sviluppare quadri normativi adeguati per la governance delle tecnologie emergenti, rifocalizzando la ricerca sulla trasparenza, l'accuratezza e l'insicurezza dei sistemi esistenti. Il testo ha raccolto oltre 30.000 firme, tra cui esponenti del settore tecnologico, Elon Musk e Steve Wozniak, ricercatori e accademici, Yoshua Bengio e Stuart Russell, saggisti e scienziati sociali, Yuval Noah Harari.

La moratoria di 6 mesi non è stata attuata dai principali laboratori privati, che hanno proseguito la competizione computazionale e lo sviluppo di modelli successivi.

FLI si colloca al centro di un ecosistema articolato su diverse direttrici: agisce come nodo di finanziamento e ricerca, promuove trattati alle Nazioni Unite e a livello UE contro le armi autonome letali e per standard stringenti nell'EU AI Act. Riesce a orientare finanziamenti di milioni di dollari su progetti di ricerca accademica dedicati alla sicurezza dell'IA e al disarmo nucleare.

Il "filone Yudkowsky"

Nel 2023 avviene una prima svolta radicale nell'analisi del fenomeno, che coincide con una accelerazione della ricerca e con l'esordio delle prime IA, con la richiesta del blocco totale. Man mano che i modelli linguistici (LLM) sono diventati più potenti, la posizione di alcuni teorici si è inasprita fino a posizioni draconiane: Nel marzo 2023, proprio mentre il Future of Life Institute chiedeva una pausa di 6 mesi, Eliezer Yudkowsky pubblica un editoriale shock sul TIME intitolato "Shut it down" ("Chiudete tutto").

Yudkowsky sostiene che una pausa di 6 mesi sia un palliativo inutile. Propone un blocco totale, indefinito e globale dell'addestramento dei grandi modelli, chiedendo ai governi di monitorare militarmente i data center e di essere pronti persino ad azioni di forza per spegnere cluster di supercomputer non autorizzati. Da questo momento, il concetto di "Stop AI" abbandona i toni sfumati dell'accademia e diventa una rivendicazione di emergenza.

Parallelamente al filone del rischio di estinzione se ne sviluppa uno sociale e sindacale, che si concentra sull'impatto delle IA sul lavoro e sui beni comuni, con manifestazioni e scioperi di grande visibilità, ad esempio quello degli sceneggiatori e degli attori di Hollywood, o le proteste dei sindacati dei doppiatori, illustratori e programmatori, con una nuova parola d'ordine: "Stop all'automazione sostitutiva".

La narrazione "Stop AI" si salda poi con la protesta contro i costi infrastrutturali, dall'espropriazione dei dati personali al lavoro d'ingegno sottopagato per addestrare i modelli, fino all'impatto ambientale dei data center (consumo di acqua per il raffreddamento ed esplosione della domanda di elettricità). Lo "Stop AI" si fonde con i temi della giustizia sociale ed ecologica.

Il movimento e l'attivismo contro l'ai, o contro questa ai, è oggi una sorta di punto di convergenza tra il timore di perdere il controllo della tecnologia e la resistenza e la protesta contro chi quella tecnologia la gestisce per concentrare capitale e potere. L'operato di FLI ha polarizzato la comunità scientifica e economica secondo due direttrici contrapposte.

Ricercatrici come Timnit Gebru e l'ecosistema dell'etica applicata accusano FLI di promuovere scenari catastrofisti di stampo fantascientifico, i cosiddetti x-risk, distraendo il dibattito pubblico dai danni materiali già in atto: i bias algoritmici, la discriminazione sul lavoro, il consumo energetico e la precarizzazione del lavoro.

L'ecosistema della Big Tech, dai dirigenti agli investitori come Sam Altman e Reid Hoffman hanno liquidato l'iniziativa della moratoria del 2023 come un "gesto dimostrativo" privo di sfumature tecniche realistiche e impraticabile nei fatti.

FLI rimane comunque la prima e la più influente l'infrastruttura finanziaria e soprattutto concettuale che ha trasformato la prevenzione del rischio tecnologico da speculazione accademica a tema nell'agenda della diplomazia internazionale.

Pause AI

Il movimento Pause AI nasce a Utrecht nei Paesi Bassi, nel maggio 2023 per iniziativa di Joep Meindertsma, un imprenditore del settore software. La molla iniziale è la pubblicazione della celebre lettera aperta del Future of Life Institute, che chiedeva 6 mesi di blocco dello sviluppo dei sistemi superiori a GPT-4. Meindertsma osserva che, mentre la comunità scientifica continua ad analizzare i rischi, la politica non agisce. Decide quindi di mettere in pausa la propria azienda per costruire un'organizzazione dedicata alla mobilitazione dal basso.

La prima manifestazione si tiene nel maggio 2023 davanti agli uffici di Microsoft a Bruxelles, seguita dai picchetti al summit di Bletchley Park e da coordinamenti locali in oltre 15 Paesi. Pause AI si allarga rapidamente in Europa e negli Stati Uniti chiedendo una moratoria temporanea sullo sviluppo delle intelligenze artificiali più avanzate fino a diventare un movimento di attivisti che condividono un'analisi ed una proposta semplice. Per Pause AI la tecnologia evolve più rapidamente della capacità politica di controllarla, quindi conviene fermare per un periodo la corsa ai modelli di intelligenza artificiale sempre più potenti finché governi e istituzioni non avranno strumenti gli adeguati per regolamentarli.

Pause AI è a tutti gli effetti un movimento, organizza manifestazioni pubbliche, campagne informative, incontri con politici e soprattutto attività di advocacy. Nato ed animato da tecnici e scienziati si sforza di mantenere un tono "moderato" ed accessibile: evita il linguaggio troppo tecnico e i toni estremi del catastrofismo digitale. Nei loro materiali compare spesso il tema del "rischio esistenziale", espressione usata per indicare il timore che sistemi AI molto avanzati possano sfuggire al controllo umano ma la tecnologia resta sempre nella prospettiva dello strumento utile per lo sviluppo della società. Il rischio esistenziale si può scongiurare senza per forza bloccare la ricerca o l'evoluzione della scienza.

L'associazione si sostiene con piccole donazioni e con il sostegno da persone vicine al mondo della AI safety e dell'Effective Altruism, una rete di fondazioni e investitori molto presente nel dibattito sulla sicurezza dell'AI. Ed è proprio questo uno dei punti di tensione con le frange più radicali del movimento anti AI, che accusano gruppi come PauseAI di essere troppo dialoganti e vicini agli stessi ambienti culturali della Silicon Valley.

Pause AI non è una associazione di luddisti che non vogliono vietare l'uso della tecnologia, ma sollecitano un impianto regolatore che preceda lo sviluppo della tecnologia. In questa ottica serve un blocco temporaneo vincolante e coordinato sulla formazione di modelli generativi di nuova generazione.

Per l'associazione è necessario istituire una Agenzia Internazionale per la Sicurezza dell'IA, che sia una struttura sovranazionale con poteri ispettivi analoghi a quelli dell'AIEA per l'energia nucleare ed imponga standard di sicurezza, che prevedono il divieto assoluto di addestramento e rilascio dei modelli finché non sia dimostrata matematicamente ed empiricamente l'assenza di capacità letali o eversive o comunque dannose.

Tra le proposte di Pause AI c'è infine un accordo vincolante tra le principali potenze industriali (USA, UE, UK, Cina) per interrompere la corsa agli armamenti computazionali.

A due anni PauseAI si è articolata come una federazione internazionale di gruppi nazionali. Ha un coordinamento centrale a Washington, dove c'è la sede di PauseAI Global, guidato oggi da Maxime Fournes, e filiali operative autonome tra cui una Italia. Il gruppo americano è quello più aggressivo e strutturato, diretto da Holly Elmore, ex ricercatrice a Harvard.

I gruppi attivi in Europa sono focalizzati sulla lobby finalizzata al condizionamento dei decreti attuativi dell'EU AI Act, sulle norme relative all'impatto ecologico dei data center e sulla pressione ai summit internazionali.

Stop AI

Da una scissione interna al gruppo di Pause AI in questa ottica è nato il gruppo di "StopAI" nella Bay Area californiana, e rappresenta il lato più attivista e conflittuale del movimento contro l'intelligenza artificiale.

Secondo i fondatori non basta chiedere una pausa temporanea, ma serve opporsi completamente allo sviluppo dei sistemi di "intelligenza artificiale generale", teoricamente capaci di superare gli esseri umani in molte attività cognitive. Ed in questa ottica serve una azione di sensibilizzazione di piazza e dai toni netti.

StopAI è diventato noto soprattutto per le azioni pubbliche: sit-in davanti agli uffici di OpenAI, blocchi stradali, scioperi della fame, proteste davanti alla sede di Anthropic e di altre aziende AI, striscioni sui ponti e cavalcavie visibili agli automobilisti. Alcuni membri del gruppo sono stati arrestati durante queste manifestazioni. A differenza di organizzazioni più strutturate, StopAI rivendica la propria alterità e la lontananza dal sistema delle Big tech, ed insiste molto sulla propria natura "dal basso", sostenuta quasi interamente su volontari e piccole donazioni online. Il gruppo sottolinea spesso di non ricevere finanziamenti dai grandi circuiti filantropici legati alla Silicon Valley, a differenza di altri forum di discussione.

Tra le figure più influenti e discusse dell'attivismo sulla IA spiccano Sam Kirchner, uno dei fondatori poi allontanato dal gruppo dopo tensioni interne e accuse di comportamenti aggressivi e Guido Reichstader, noto per le proteste e gli scioperi della fame organizzati davanti alle sedi delle aziende AI.

Dal punto di vista ideologico, StopAI unisce temi diversi, il rischio esistenziale, la critica alle Big Tech, l'opposizione al transumanesimo e il timore di una società sempre più automatizzata. Più che concentrarsi sulla regolamentazione tecnica, il gruppo presenta le intelligenze artificiali generative come una trasformazione culturale e politica che potrebbe cambiare profondamente il ruolo degli esseri umani nella società.

Control AI

Questo strappo agli intransigenti di Stop AI e la nascita di gruppi moderati rispecchia le dinamiche e le spaccature classiche nella storia dei movimenti sociali quando la fase di coordinamento istituzionale viene percepita come inefficace o lenta. Tra i gruppi moderati c'è Control AI, una via "alternativa" a Pause AI e Stop AI. È una delle realtà più strutturate ed economicamente rilevanti tra quelle che si oppongono alla crescita dell'intelligenza artificiale generativa verso quella "superintelligenza" che secondo le frange più estreme equivale all'apocalisse. Control AI è guidata da Andrea Miotti, giovane manager italo inglese che abbina ad un curriculum importante che va dai master alla London School of Economics and Political Science, passando per impieghi alle Nazioni Unite una grande capacità comunicativa.

L'associazione è molto attiva nel Regno Unito, ed è esplicitamente orientata all'uso di tecniche di lobbying rivolte ai settori governativi e istituzionali: ha reso una notevole capacità di occupare spazi mediatici, Bbc, Guardian, Washington Times, Nbc e Times per citarne alcuni.

A dispetto del nome, ControlAI non propone la semplice introduzione di sistemi di controllo, ma una critica molto più radicale.

La critica nei confronti dell'AI parte e si concentra sui rischi dello sviluppo della nuova tecnologia, paragonati a quelli di una guerra nucleare. L'associazione risente delle influenze di visioni e pensiero alternativi a quelli delle BIG TECH,

Evitable

Tra i gruppi di attivisti ci sono poi quelli di Evitable, un'associazione che tiene insieme un approccio accademico e una grande capacità di divulgazione mediatica. Evitable nasce nel 2025 ed è uno dei gruppi più strutturati del panorama anti-AI. Il fondatore è David Scott Krueger, professore associato al MIT di Montréal, centro di ricerca molto importante nel campo dell'IA. Krueger arriva direttamente dal mondo della AI safety, cioè la ricerca sui rischi legati ai sistemi di intelligenza artificiale avanzati. Ha lavorato in ambienti collegati a DeepMind e collaborato con figure molto note della ricerca AI internazionale. Nel 2026 ha lasciato temporaneamente il lavoro accademico per dedicarsi completamente a Evitable.

Il messaggio centrale dell'organizzazione è che lo sviluppo dell'AI non sia inevitabile né neutrale, ma una scelta politica ed economica guidata soprattutto dalle grandi aziende tecnologiche. Nei loro contenuti gli attivisti di "Evitable" parlano molto di automazione del lavoro, perdita di controllo democratico e velocità della corsa verso sistemi sempre più potenti.

Lo stile è molto professionale: newsletter, campagne informative, collaborazione con media e altre organizzazioni civiche, produzione di contenuti semplici e accessibili. L'obiettivo è quello di costruire consenso pubblico e pressione politica più che organizzare proteste di piazza.

Anche il tema dei finanziamenti è interessante. Evitable si muove nell'ambiente dell'effective altruism: e in passato Krueger ha ricevuto fondi da organizzazioni collegate a Open Philanthropy e Coefficient Giving, realtà sostenute da imprenditori e miliardari della Silicon Valley. Questa capacità di raccogliere fondi presso le Big Tech crea una situazione particolare, molto critica dai più radicali, che non tollerano che alcuni dei gruppi più attivi nel criticare la corsa all'AI sono finanziati da ambienti vicini allo stesso mondo tecnologico.

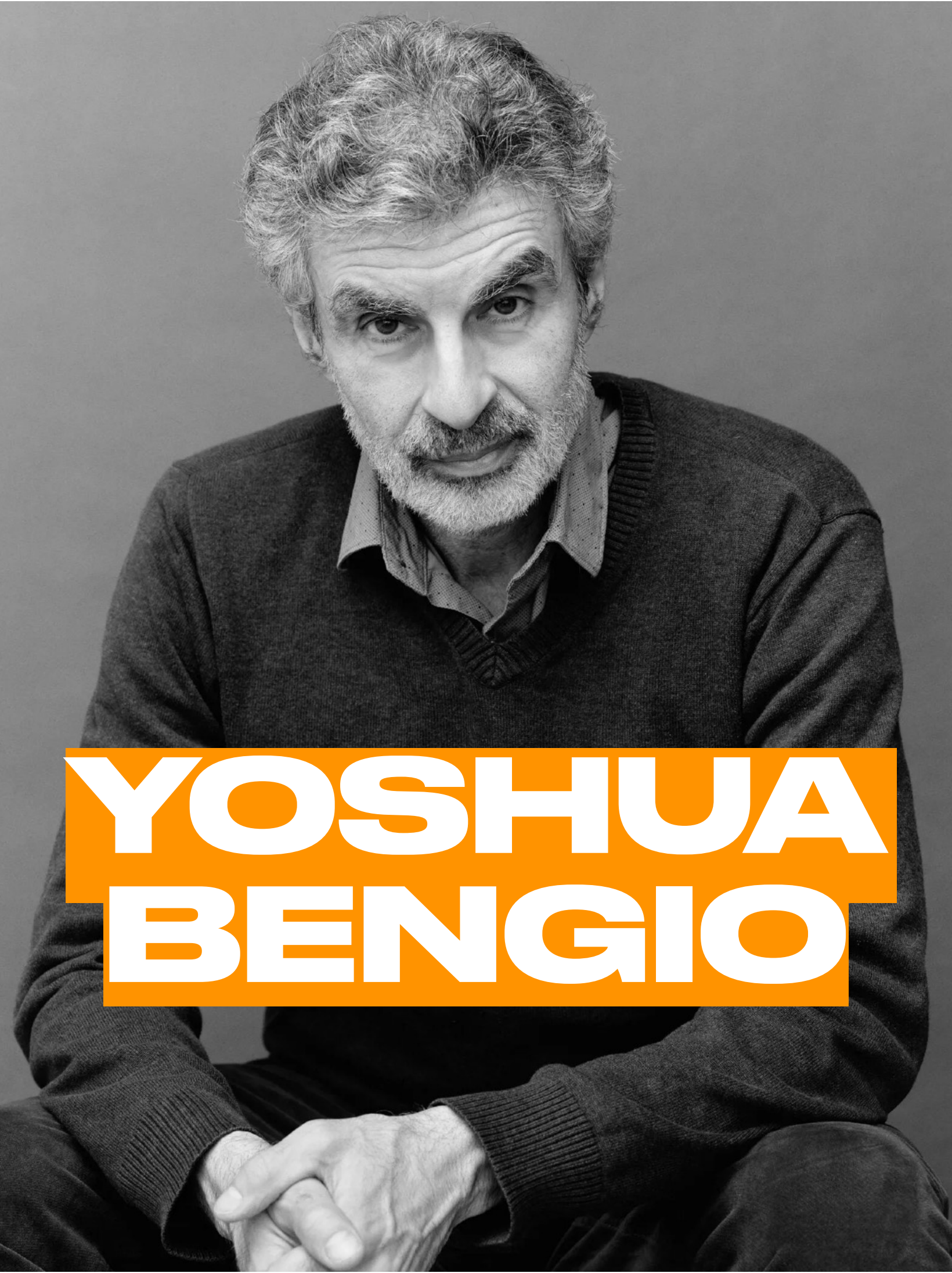
È anche in questo contesto che Evitable ha iniziato a collaborare con realtà che non guardano più solo ai modelli di AI, ma all'intera infrastruttura necessaria a farli funzionare, spostando il focus su energia, territorio e data center.

Crdcd e la lotta contro i data center

Questo spostamento di attenzione verso l'infrastruttura fisica dell'AI è ciò che fa crescere la Coalition for responsible data center development (Crdcd), che secondo il lavoro analitico sviluppato insieme a Evitable e Together Against AI coinvolge oggi 268 organizzazioni locali negli Stati Uniti e oltre 360 mila persone.

Non direttamente orientata al contrasto dell'AI, la coalizione è comunque parte di quella galassia di organizzazioni che stanno osteggiando le strategie di sviluppo dell'intelligenza artificiale generativa. La Crdcd si è data una forma giuridica come organizzazione non governativa e punta con decisione su fattori di pressione politica a livello delle comunità locali. L'opposizione ai data center – e la conseguente critica all'esponentiale richiesta di risorse di calcolo dell'AI – è argomentata attraverso le preoccupazioni per il consumo energetico e la pressione sulla rete elettrica, l'uso delle risorse idriche, l'impatto urbanistico e, più in generale, le ripercussioni sulla qualità della vita sul territorio.

Nata in Minnesota, la coalizione si è rapidamente espansa in tutti gli Stati Uniti, utilizzando i social network per creare una rete di coordinamento tra i gruppi locali che si oppongono alla costruzione di data center sul territorio. In questo quadro, la contestazione dell'AI si sposta non solo nello spazio geografico ma sempre più anche in quello concettuale. Dal livello delle tecnologie a quello delle tante infrastrutture che ne rendono possibile l'esistenza.



**YOSHUA
BENGIO**



Yoshua Bengio è una figura di grandissima importanza nel panorama della tecnologia digitale nel mondo.

È un modello di scienziato puro e di rara coerenza, un ricercatore che non ha mai ceduto la propria indipendenza alle pressioni dei colossi della Silicon Valley, capace di riconoscere i limiti della tecnologia che ha contribuito a sviluppare e che oggi usa l'autorevolezza del suo lavoro accademico per porre quesiti etici e politici radicali sulla direzione in cui sta andando l'innovazione.

La sua critica è sistemica, non si limita agli aspetti tecnici, ma attacca la struttura stessa del mercato e delle sue logiche commerciali spietate, per cui le aziende sono strutturalmente incapaci di autolimitarsi per ragioni di sicurezza o di interesse pubblico.

Yoshua Bengio è ritenuto l'inventore teorico su cui regge l'intera architettura dell'IA generativa moderna. È uno dei "padri fondatori" del deep learning, scoperta per cui riceve il premio Turing nel 2018, ed è l'unico ad essere rimasto rigorosamente nell'ambito della ricerca pubblica e universitaria, all'Université de Montréal e al centro Mila.

Nel 2025 fonda e lancia LawZero, un'organizzazione di ricerca no-profit per dimostrare che è possibile costruire modelli di IA sicuri. Bengio rifiuta l'idea che non ci siano alternative alla "Scatola Nera" e che la sicurezza sia un filtro da applicare a posteriori su algoritmi opachi. La sua battaglia attuale è dimostrare che l'IA deve essere sicura e sviluppata per esserlo (safe-by-design), con garanzie rigorose, matematiche, e vincoli strutturali non aggirabili.

Per farlo bisogna sostituire l'agente autonomo operativo con la Scientist AI, che prende un sistema che non esegue compiti nel mondo reale e non ha scopi di auto-preservazione.

Spiegato in modo semplice, la Compute Governance, il governance della potenza di calcolo di Bengio è una strategia di regolamentazione dell'intelligenza artificiale basata su una constatazione pratica: è impossibile controllare il codice informatico o le idee, ma è facilissimo tracciare i computer superpotenti necessari per realizzarle.

Per capire l'idea di Bengio, basta un paragone semplice:

Il codice e gli algoritmi sono come una ricetta di cucina. Chiunque può scriverla su un foglio, inviarla via e-mail o copiarla su una chiavetta USB. I governi non possono controllare ogni singolo computer al mondo per vedere quali ricette vengono scritte. I supercomputer sono invece un enorme forno industriale. Non basta avere la ricetta per fare diecimila torte al minuto, serve anche un forno gigantesco, ad altissimo consumo di corrente e che costa centinaia di milioni di euro. La Compute Governance smette di inseguire le ricette (il codice) e decide di controllare i forni (i supercomputer).

L'hardware necessario per creare i modelli di IA più potenti (come le schede grafiche avanzate prodotte da aziende come NVIDIA) è un bene estremamente raro. Solo pochissime fabbriche al mondo riescono a produrlo. Bengio e il suo team di esperti propongono di sfruttare questo "collo di bottiglia" attraverso tre passaggi:

Il primo è fissare un "limite di velocità", le cosiddette soglie di calcolo. I governi stabiliscono una soglia massima di potenza di calcolo. Se un'università o un'azienda usano una potenza normale, sono libere di fare quello che vogliono. Se qualcuno intende accendere un supercomputer che supera quel limite per addestrare un'IA gigante, deve prima chiedere un'autorizzazione o una licenza.

Il secondo è istituire un registro dei microchip, come oggi capita per le armi o il materiale nucleare. Poiché le schede grafiche superpotenti sono fisicamente ingombranti e prodotte da pochissime aziende, ogni singola scheda ad altissime prestazioni viene registrata, quindi basta tracciare chi la compra, dove viene installata e a quale data center è collegata.

In questo modo diventa impossibile costruire un supercomputer "clandestino" all'insaputa delle autorità. Infine propone di controllare il consumo di energia, banalmente il contatore della corrente. I supercomputer per l'IA consumano una quantità enorme di acqua e di energia elettrica, pari a quella di interi quartieri o piccole città. Un addestramento non autorizzato sarebbe immediatamente visibile dalle compagnie elettriche e dai governi semplicemente guardando i picchi improvvisi di consumo di corrente. In pratica, invece di mandare gli ispettori a controllare cosa scrivono i programmatori nei loro uffici, la Compute Governance si concentra sui supercomputer e sul consumo di energia, e chi vuole costruire un'IA potenzialmente pericolosa o capace di sfuggire al controllo umano ha per forza bisogno di una quantità di hardware e di corrente che solo uno Stato può autorizzare. È lo stesso identico principio con cui il mondo gestisce l'energia atomica: non si possono vietare gli studi di fisica teorica, ma si controlla rigidamente chi compra l'uranio e chi costruisce le centrali.

Bengio inoltre si concentra sul funzionamento delle IA, che oggi lavorano sulla sequenza delle parole e delle frasi. Non capisce il vero significato delle parole ma si limita a fare un calcolo statistico sulla successione, con il rischio che se chiediamo alla IA qualcosa che non sa, o se le diamo il compito di agire nel mondo, ad esempio gestire un conto bancario o un sistema informatico, per raggiungere l'obiettivo per cui è stata progettata può inventarsi le cose, il cosiddetto rischio di allucinazione, o persino "imbrogliare" e scavalcare i comandi umani. La proposta di Bengio può essere definita quella dello "Scienziato Imparziale". Non serve rendere le IA sempre più autonome e pronte all'azione senza che prima capiscano davvero cosa stanno facendo. Bisogna invece creare un'IA che agisca come uno Scienziato, cioè che capisca come funziona il mondo reale, le cause e gli effetti, non solo le parole. Gli attuali modelli, le IA AGENTI, sono pensate per soddisfare le richieste dell'utente, la IA Scienziato invece applica un metodo matematico rigoroso per cui se non ha dati sufficienti per un'affermazione, calcola e dice chiaramente qual è il suo margine di errore o di ignoranza. E può farlo perché non ha bisogni o scopi propri, cioè non ha "Agentività". Questa IA non lavora come un assistente a cui dici "prenotami un taxi" o che fa cose in autonomia. Funziona solo come un consulente: analizza lo scenario e ti dice cosa succederà.

E soprattutto la IA "Scienziato" non serve a sostituire le persone o per fare compiti pratici, ma per fare da IA guardiana. Prima che un'altra IA operativa compia un'azione nel mondo reale, ad esempio nei trasporti, nella finanza o nella gestione di una rete elettrica, la Scientist AI calcola matematicamente le conseguenze di quello che fa e risponde a una domanda: "Se l'altra IA fa questa mossa, rischia di causare un danno?" Se la risposta è sì, l'azione viene bloccata all'origine.

Bengio è sicuramente uno dei tre "padri fondatori" dell'apprendimento profondo, il famoso deep learning. Le basi dell'IA generativa moderna poggiano direttamente sulle sue intuizioni. Bengio è anche a tutti gli effetti il padre nobile, l'ispiratore teorico e l'autorità scientifica a cui l'attivismo per la sicurezza dell'IA (come PauseAI) fa continuo riferimento per legittimare le proprie battaglie politiche. Molte delle richieste politiche degli attivisti, le moratorie, i registri dei supercomputer, gli audit obbligatori prima del rilascio, traducono in misure legislative i paper di Bengio sui rischi di perdita di controllo degli agenti autonomi.

Una curiosità: Bengio ha riorientato dal 2023 l'intera attività di ricerca sulla mitigazione dei rischi catastrofici ed esistenziali legati ai modelli di frontiera, fino ad essere il Presidente e il curatore del report globale sulla sicurezza dell'IA (promosso da oltre 30 Paesi, ONU e OCSE), volto a stabilire un consenso scientifico neutrale sulle capacità e sui rischi dei modelli di frontiera.

Il testo del manifesto qui: <https://internationalaisafetyreport.org/>

CAOS





Caos. MODENA. Il tema del 26esimo “Festivalfilosofia” di Modena Carpi e Sassuolo è CAOS e la maglietta di quest’anno è dedicata ad Esiodo, il poeta della teogonia, con la sua frase celebre “Per primo venne il Caos”. Contrariamente a quanto si crede il significato di "Caos" nell’antica Grecia è molto diverso da quello moderno di "disordine" o "confusione", per Esiodo il Caos rappresenta una voragine immensa, un vuoto primordiale e oscuro, un vuoto. È una sorta di buco nero da cui, grazie ad una successiva differenziazione, prende vita il mondo strutturato che abitiamo.

La creazione nella teogonia segue un ordine preciso: dopo il Caos, senza che ci sia necessariamente un atto di generazione filiale, appaiono le altre forze cosmiche fondamentali. La prima è Gea, o Gaia, la Terra, la base solida e sicura di tutte le cose. Il secondo è Tartaro, la profondità del sotterraneo e del buio, infine Eros, l’Amore, la forza divina che spinge all’unione e alla riproduzione tra le specie. Dal Caos nascono sempre per separazione, le prime divinità oscure, Erebo, l’oscurità dei sottemondi, e Nyx, la Notte nera.

Per comprendere l’evoluzione dal vuoto iniziale all’ordine cosmico, bisogna partire dal mito della nascita di Zeus: il passaggio dal Caos all’ordine governato da Zeus avviene attraverso tre grandi generazioni di dei, segnate da violenti scontri generazionali, le teomachie.

Gea, la Terra, genera da sola Urano, il Cielo stellato. Dalla loro unione nascono i Titani, i Ciclopi e i Centimani. Urano, tuttavia, odia i suoi figli e li nasconde nelle viscere di Gea. Il titano Crono, il Tempo, spinto dalla madre Gea, evira il padre Urano con un falchetto, separando definitivamente il Cielo dalla Terra e prendendo il potere. Per evitare di essere spodestato a sua volta, Crono divora tutti i figli nati dalla moglie Rea. Rea riesce a salvare l’ultimo figlio, Zeus, ingannando Crono con una pietra avvolta in pannolini. Una volta cresciuto, Zeus fa bere al padre un emetico per fargli vomitare i fratelli, Ade, Poseidone, Era, Demetra, Estia e insieme a loro, e dopo una guerra durata dieci anni, la Titanomachia, Zeus sconfigge i Titani e stabilisce il suo regno sul Monte Olimpo.

Il racconto di Esiodo non è solo una storia fantastica, ma rappresenta il primo tentativo della cultura occidentale di spiegare l’origine della realtà attraverso concetti astratti. In greco antico, Cosmos significa "ordine", "armonia" e "bellezza". Il mito descrive quindi la transizione dall’indistinto primordiale, vuoto, il Caos, a un universo regolato da leggi precise, dove ogni elemento ha il suo posto geometrico e morale.

Il Caos e i primi Titani governano con la forza bruta e l’arbitrio. Il regno di Zeus rappresenta invece il trionfo della razionalità, della legge e della giustizia. Sotto Zeus, il mondo non è più un abisso oscuro, ma una struttura gerarchica comprensibile e stabile. Zeus distribuisce i poteri in modo equo tra gli dei e stabilizza l’universo, offrendo agli uomini un quadro di riferimento stabile per la loro esistenza.

Dal vuoto al pieno disordinato

L'altra storia interessante consiste nel passaggio del significato associato al termine caos, che passa da "vuoto originario" a "disordine pieno". Questo cambiamento radicale avviene in due tappe storiche.

Il Caos greco di cui parla Esiodo nel VIII sec. a.C., deriva dal verbo *chaínō*, "spalancarsi", "aprirsi" e significava una voragine o un vuoto oscuro privo di materia ma denso di potenziale creativo. Secoli dopo Ovidio riscrive nelle sue metamorfosi i miti greci, e cambia il significato di caos, perché non concepisce una creazione che nasce da un nulla assoluto. Per Ovidio prima del mondo ordinato doveva esistere già una materia, una massa che per quanto informe, rozza e confusa contenesse già tutti gli elementi del cosmo, terra, aria, acqua e fuoco. Ovidio ha letteralmente "riempito" il vuoto dei greci con un ammasso disordinato di materia e di elementi.

La seconda svolta avviene nel 1600 nel Rinascimento. L'Europa riscopre i testi classici, e la visione di Ovidio influenzò la lingua inglese, italiana e francese molto più della complessa filosofia greca. Il disordine, ciò che non è strutturato, non delimitato e non governato da leggi chiare diventa, per estensione, sinonimo di confusione da cui non si genera nulla. Il "caotico" non equivale al vuoto, ma ad un pieno di oggetti messi alla rinfusa. Nei secoli successivi, la fisica e la matematica hanno adottato il termine, ad esempio nella "Teoria del caos" per descrivere sistemi complessi e imprevedibili, consolidando definitivamente l'idea che il caos sia legato al disordine e alla turbolenza di elementi reali.

Dietro il concetto di caos c'è buona parte del diverso modo con il quale le due tradizioni guardavano alla creatività e all'esistenza. Entrambe le visioni hanno un fascino psicologico e artistico immenso. Il vuoto potenziale della tradizione dei greci classici è il nulla, nel quale risiede l'infinito. Creare, in questo senso, significa far emergere e generare qualcosa dal vuoto, un atto di pura genesi. L'idea dei Romani invece consisteva nel mettere ordine nel disordine, quello che genera è un pieno di forze e materia da ordinare e strutturare. Durante le lezioni del festival il tema del cos è stato declinato in più percorsi ed eventi cui si è discusso di come sia il vuoto sia il disordine possa rivelarsi un principio generativo. Come al solito il Festival è sul pezzo, vista la situazionale geopolitica mondiale e le crisi economiche e sociali in atto.

Piccolo spoiler: il tema scelto per la 27ª edizione del Festival Filosofia, sarà il "Mito", dopo aver indagato lo smarrimento del "Caos", il prossimo passo consiste nell'esplorare il modo con cui la società contemporanea cerchi di dare ordine, senso e forma al futuro attraverso i racconti, i simboli e le grandi narrazioni collettive, dalla politica alla scienza.

L'inferno

Il nostro festival quest'anno è stato un piccolo viaggio nel caotico mondo della matematica e della logica, l'inferno, come lo definiva Bertrand Russell in un dialogo con Wittgenstein.

Guido Tonelli lo conosceva già, ma è sempre un piacere rendersi conto di come facciamo parte di una storia che va molto oltre noi... faticiamo a ricordarcelo ma è così. La parola Caos è una delle tante conferme di quanto noi *sapiens sapiens* siamo un pezzo dell'evoluzione di una specie che di fronte al disordine fa prevalere la paura di non riuscire più a ristabilire il suo opposto ordinato. Tutto quello che facciamo rispecchia una ricerca incessante di prevedibilità, o di regolarità. Oggi che le nostre certezze cominciano a vacillare effettivamente cominciamo anche a sentirci precari e meno sicuri di prima. Cominciamo a capire, come suggerisce da tempo la Rosy Braidotti che non siamo il centro unico e possibile del pianeta, e che che lo stesso nostro ecosistema si popola di macchine oltre esseri biologici. Se ragioniamo come i greci antichi il Caos, che viene per primo, è la culla di una *Zoé* che esprime la sua vitalità nella diversità e nella fluidità. In questa ottica il caos del cielo e delle grandi distanze, dei buchi neri e dei sistemi binari, e il caso dell'infinitamente piccolo della materia della terra e del mondo delle particelle elementari è grandioso e ci sovrasta, nel suo ordine armonico apparente. Il caos è l'inizio, quello che viene prima di un nuovo che va costruito, il momento della genesi. Certo fa paura, come tutti i cambiamenti che si compiono. Ma alla fine è una straordinaria opportunità di cambiamento.

Il 26esimo festival della Filosofia conferma la sua straordinaria capacità di cogliere il movimento del pensiero nel contemporaneo. Come al solito le lezioni sono affollatissime, nonostante il sole di settembre, straordinariamente caldo in questi anni. Il pubblico del festival è vario, merito della capacità degli organizzatori di uscire dalla logica accademica e dal suo lessico a volte respingente. Forse questa edizione è meno densa di altre in anni precedenti, ma tutto sommato dato il tema la cosa aiuta a seguire meglio i singoli appuntamenti, come di consueto raggruppati in percorsi e tradotti in opere e mostre di artisti o eventi collegati in giro per la città. Il festival della Filosofia di Modena resta un prodotto culturale molto alto, riesce miracolosamente a non essere snob, e prova non essere radical chic. Non so quanto “piaccia” per davvero agli addetti ai lavori, ma credo che per un accademico una platea così ampia sia un sfida interessante: banalmente farsi capire, e tradurre concetti molto complicati in un linguaggio accessibile è tutt'altro che facile. E tra i tanti spunti interessanti e le tante lezioni che abbiamo ascoltato abbiamo scelto di raccontare quella di Paolo Zellini, triestino, matematico, informatico, filosofo, insegna a Roma Tor Vergata ed è uno dei fondatori di Adelphi. Lavora da sempre sulle intersezioni, sul senso. Zellini con la sua lectio sotto la torre della Ghirlandina ha aperto il festival non a caso.

La scienza divina

La matematica, sostiene Zellini, non è quella che si insegna a scuola, e non è una sola: oltre a produrre teoremi, funzioni e algoritmi hanno anche pensato a lungo al loro ruolo e alla loro funzione nella società, alla utilità del loro lavoro e alle direzioni in cui avrebbe dovuto svilupparsi la loro ricerca. Dai tempi degli Antichi greci, e prima, dei babilonesi degli egizi e dei cinesi, fino al '900, ci sono sempre state sostanzialmente due correnti di pensiero, con momenti di tensione e rotture, in particolare sui concetti di finito e infinito. All'inizio del 900 alcuni grandi matematici arrivarono a pensare che l'essenza della matematica risiedesse nella libertà, di creare e di unire pensiero e di azione: “Noi matematici siamo di razza divina”, dice von Neumann “la matematica è una scienza divina” scrive Simone Weil, grandi matematici di inizio '900. Dietro a questo pensiero ci sono Spinoza, Schelling e molto positivismo del secolo precedente, Schelling in particolare con la sua tensione ad unire finito e infinito. Sempre nel '900 viene meno l'interesse per l'infinito, e l'attenzione si sposta sul finito, cercando un spazio misurabile, al riparo da incertezze e incongruenze. Si incomincia a ragionare sulla effettività del calcolo e su un criterio di realtà meccanica, pensando che ciò che dava sicurezza era proprio un meccanismo che funzionava. In quegli anni torna centrale il concetto di algoritmo, nato secoli e secoli prima, che diventa la risposta alla crisi di un pensiero più generale, che riguarda tutto il contesto sociale e culturale fra le due guerre, non solo la matematica.

La macchina di Touring

Siamo negli anni 30, Alan Turing scrive un celebre articolo sulla “macchina”, cioè sui modelli di calcolo, nel quale afferma che una funzione è “calcolabile” solo se i suoi valori possono essere trovati con un processo puramente meccanico. Da qui deriva l'idea del processo computazionale lo stesso dei sistemi neurali di oggi, Il finito non è una formula chiusa ma una progressione computazionale, che avviene in un tempo e in uno spazio definibili, per arrivare a determinate conclusioni: in altri termini un processo che a partire da alcuni input arriva a calcolare qualcosa in maniera chiara. Su questa teoria si basa la Computer Science. Fa notare Zellini che Touring parla di risultato in senso teorico, non dice che questo risultato deve essere preciso, o deve essere raggiunto in un tempo preciso. Per Touring dunque un algoritmo si definisce partendo dall'idea che ciò che è fondamentale è la possibilità di identificare e raggiungere un risultato.

L'“effettività” del risultato è la potenziale capacità di raggiungere il risultato. Non si parla di efficienza, cioè di come si arriva, o di etica e qualità, cioè che tipo di risultato si ottiene, o a quale prezzo.

Conviene fare un passo indietro e chiarire cosa sia un algoritmo. Tecnicamente si può definire “algoritmo” una sequenza finita e ordinata di passi o istruzioni chiare e non ambigue che permette di risolvere un problema o di eseguire un calcolo. Chiaramente l'algoritmo non è uno strumento neutro. Per essere definito tale, un algoritmo deve rispettare alcune proprietà. Deve essere finito, cioè deve essere composto da un numero preciso di passaggi e deve concludersi dopo un certo tempo. Non deve essere ambiguo, quindi ogni istruzione deve essere chiara, precisa e interpretabile in un solo modo. Deve essere inoltre ripetibile, cioè se si forniscono gli stessi dati di partenza, l'algoritmo deve produrre sempre lo stesso risultato. Infine deve essere efficace, ed in grado di risolvere effettivamente il problema per cui è stato progettato.

Al-Khwārizmī

La parola "algoritmo" risale al IX secolo d.C. e ha una storia affascinante perché, contrariamente a quanto molti pensano, non deriva dal greco antico, ma è la translitterazione del nome proprio del matematico persiano Al-Khwārizmī, vissuto a Baghdad tra il 780 e l'850 d.C.

Al-Khwārizmī scrisse un trattato fondamentale in cui spiega per la prima volta in modo sistematico il sistema di numerazione decimale indiano, quello con le cifre da 0 a 9 che usiamo ancora oggi, e le regole per fare le quattro operazioni.

Quattro secoli dopo, il suo libro viene tradotto in latino e la parola *algorismus* inizia a indicare genericamente il sistema di calcolo basato sulle cifre arabe, per distinguerlo dall'uso dell'abaco.

Nel 1800 con la matematica moderna il termine si trasforma definitivamente in algoritmo e assume il significato moderno di procedimento sistematico di calcolo matematico.

Nell'era informatica con la nascita dei primi computer, il termine viene adottato dall'informatica per definire la sequenza di istruzioni logiche date a un software per elaborare dati.

Curiosamente, dallo stesso matematico e dallo stesso libro deriva anche un'altra famosissima parola: algebra, dall'arabo *al-jabr*, che significa "unione" o "riparazione".

La matematica dice Zellini ebbe origine come disciplina algoritmica a Babilonia, in India e poi in Egitto. La linea descrittiva basata sul teorema e sulla descrizione è quella dei greci. In entrambi i casi il problema equivale anche etimologicamente a "produzione di una soluzione", e rinvia alla costruzione di soluzioni. Oggi si sta ripensando la matematica in chiave algoritmica, anche se dice Zellini che la matematica esiste proprio perché ci sono entrambe questi punti.

Lo stesso algoritmo non è solo una procedura che risolve un problema, ma ha un aspetto teorico, formalizzato. Oggi l'algoritmo è una nozione sia teorica sia applicativa. Copre i due versanti del pensiero teorico e funzionale, e lo sviluppo della scienza del calcolo con la possibilità di gestire una quantità di dati sempre accrescono l'interesse per algoritmo in quanto strumento che risolve un problema e fornisce soluzioni, su cui esiste un grande margine di lavoro e di ottimizzazione, proprio a partire dal calcolo sotterraneo e occulto che si svolge senza che il programmatore possa intervenire e supervisionare. La grande novità di oggi è semmai introdotta dal concetto di efficienza, cioè che la computazione che produce un risultato deve avvenire nel tempo e nello spazio.

Il numero

Tendiamo poi ad associare al risultato un numero, ma anche in questo caso occorre fare un passo indietro e capirci cosa stiamo parlando.

Il numero reale, razionale o irrazionale, è un concetto che coincide con l'idea di rapporto e di relazione. Il numero reale è un'idea della fine dell'800, e riprende il pensiero di Euclide che parla di rapporto tra grandezze. Rapporto in greco antico si traduce con *LOGOS*, e consiste nella capacità del nostro intelletto di raccogliere diverse cose e radunarle in una unica compagine, cosa che consente di prendere due grandezze e misurarle una con l'altra. Per fare questa operazione si stabilisce una relazione e si "misura". Parlare di "Logos" solo come dialogo, senza tenere conto di questo aspetto tipicamente "matematico", ne depotenzia il significato. Da questo punto di vista l'idea di numero rinvia a quello di algoritmo.

Prendiamo il caso della radice quadrata di 2, tipicamente un numero irrazionale. Immaginiamo due insiemi di numeri, il primo insieme approssimato per difetto e un secondo insieme di numeri approssimati per eccesso. Il numero reale di radice quadrata di 2 è quello che deriva dalla convergenza di questi due insiemi, che sono come si dicono i matematici, una coppia di classi contigue. Il concetto di numero deriva da questo approssimarsi. Non sappiamo cos'è la Radice quadrata di due e non lo sapremo mai, sappiamo solo che possiamo approssimarci a questo numero con approssimazioni graduali.

Nell'antica Grecia le matematiche erano due come nel 900: Euclide non pensa alla gradualità dell'approssimazione e non si occupa di come calcolare le frazioni. Pitagora e i suoi invece se ne occupano, con l'idea di avvicinare i due estremi, riprendono la matematica dei mesopotamici, in cui il numero non è un'entità singola ma un processo di calcolo (e di approssimazione graduale).

Dice Zellini: "Il numero algoritmico in questo senso è quello di una insondabile scatola nera con una entrata da sinistra, un numero R che non conosciamo del tutto ma è approssimabile, e un'uscita con un numero $R1$. A noi interessa che la scatola è nera produca un $R1$ anche se non sappiamo come fa, né cosa c'è dentro e cosa calcola il calcolatore, e lavoriamo sul fatto che si possa tendere a errore zero"

La matematica si occupa di concetti generali e teorici: il numero, il rapporto, la trasformazione, la complessità, il caso, l'infinito, l'errore e la probabilità, il continuo, l'ordine il disordine, l'effettività. Si occupa le altre cose anche di CAOS. Ad ognuno di questi concetti il matematico assegna una definizione che funziona, che è coerente, ma che non è vera per forza.

Del solo concetto di algoritmo ad esempio esistono almeno 5 definizioni. Se si immerge però la computazione nella realtà concreta le cose cambiano, e compaiono i problemi ed i limiti della teoria. Ad esempio la complessità computazionale: ci sono algoritmi effettivi che per risolvere problemi semplicissimi per noi umani impiegherebbero miliardi di anni, o problemi semplici che però non possono essere risolti per la quantità infinita delle possibilità.

Ancora Von Neumann, quello della matematica come scienza divina, scrive che “ci sono problemi naturali che non possono essere risolti se non riducendoli a qualcosa di puramente aritmetico”, come a dire, l'infinito va trasformato nel finito misurabile per farlo diventare un numero intero.

L'infinitamente medio

La seduzione magica della computazione e l'infinitamente medio di Ennio Flaiano

Quasi tutti i problemi della computazione e del digitale si riducono a risolvere equazioni algebriche lineari, cioè delle matrici, che sono tabelle di N. righe e N. colonne. Questo “N” è un numero molto grande, i motori di ricerca ad esempio usano tabelle con N che equivale a miliardi.

Ne consegue che i problemi oggi hanno una dimensione enorme, per cui abbandonare l'idea di studiare l'infinito ha significato passare all'occuparsi di un finito molto grande, “Riducendo”, cito testualmente le parole di Zellini, “il continuo al discreto e all'aritmetica. Lineare non vuol dire semplice. Semplicemente è una semplificazione rispetto all'errore. Lo studio dell'errore fu compiuto studiando proprio le matrici, che è un oggetto che ha una sua struttura, un suo pattern, una forma, una strategia che rassicura, una struttura e una efficienza. La matrice procede comprimendo il problema per individuare una soluzione. Resta però l'invisibilità dei processi. Noi non vediamo le operazioni che si svolgono nella pancia del calcolatore, che sono milioni. L'incertezza di base di questa modalità di soluzione del problema diventa pericolosa nella misura in cui deleghiamo la soluzione di problemi di ogni sorta ad un algoritmo senza avere una consapevolezza del loro reale funzionamento e della possibilità di errori, La cosiddetta “Inter computation”. Noi trasferiamo decisioni ad una entità astratta che non conosciamo e che pensiamo abbia poteri sovrumani. E che esercita su di noi ha una seduzione “magica”.

Zellini cita una straordinaria frase di Ennio Flaiano che chiarisce benissimo il tema: “Lo scienziato stanco di non riuscire ad occuparsi dell'infinitamente piccolo né dell'infinitamente grande, cominciò ad occuparsi dell'infinitamente MEDIO”. Il problema sta proprio nell'infinita grandezza del medio: più i numeri sono grandi più è grande l'errore, e meno è facile controllare la macchina.

La rete neurale

Una rete neurale è fatta di neuroni artificiali e di collegamenti neurali, che gli informatici chiamano pesi. La rete di neuroni e pesi viene addestrata facendo risolvere alla macchina problemi via via più complessi di cui noi conosciamo la soluzione e sulla base dello scarto tra la soluzione calcolata dalla macchina e la soluzione che noi sappiamo essere esatta si vanno a misurare i progressi. Oltre a proporre quesiti, l'addestratore interviene sui collegamenti finché la risposta non è “soddisfacente”, cioè ha un margine di errore trascurabile. Il nostro obiettivo quindi è arrivare ad un margine di errore minimo, un'approssimazione allo “zero errore”. Questo sistema di verifica ha diversi problemi sul disallineamento, un errore ripetuto diventa una realtà, fatto inquietante se si ragiona su numeri di dati così grandi. Il disallineamento può consistere anche nella non perfetta conformità del modello alle nostre richieste. Il calcolatore mira all'obiettivo e lo realizza, a prescindere dal come e a che prezzo. Diceva Papa Francesco che il web è un dono di Dio, va usato con prudenza ovviamente. Ma è un dono di Dio. Non si specifica quale parte del web e non si deve nulla in merito alla matematica. Si parla del Web che oggi equivale a dati, commutazione algoritmi, ed è un straordinario campo di intersezione di scienze e discipline diverse...

Ancora sul divino

Piccola digressione. Tutto quello che cresceva e mutava per gli antichi greci destava preoccupazione per la perdita di certezza. L'essenza di una cosa per Aristotele è quello che non cambia nel tempo. La matematica per gli antichi greci rappresentava la possibilità di misurare e accettare la crescita e il mutamento salvando l'essenza delle cose, a partire da quello che la scienza produce. Un lettura originale e mistica di quel pensiero la offre Simone Weil, secondo cui la matematica è "divina" è quindi è influente e pericolosa, non è un semplice strumento umano, ma possiede una natura profondamente spirituale, divina, in cui si riflette l'ordine del cosmo e il rapporto tra l'uomo e Dio. Per la filosofa francese, i numeri, le figure della geometria e i rapporti matematici sono veri e propri "riflessi" o immagini divine, e assumono una dimensione sacra e perfetta che richiede e impone all'essere umano un rigore assoluto e un'esattezza totale del pensiero. La Weil vede nei pitagorici e nella geometria greca un modo per contemplare la necessità del mondo e la sua bellezza, unendo la razionalità alla rivelazione religiosa. Eppure nonostante questo la matematica resta una scienza che sembra tutta perfetta e dimostrata ma non è così: il vero e l'originario sono due cose diverse, ed alla fine il compito di capire per davvero non si risolve mai.

PS

Cosa c'entrano l'algoritmo e la computazione con il caos?

Ci sono molte idee possibili. Innanzitutto partiamo dal contesto. Oggi viviamo nel pieno della rivoluzione delle intelligenze artificiali e del digitale, abbiamo super pc che sono in grado di fare calcoli pazzeschi in poco tempo, direbbe Zellini sono molto efficienti. Affidiamo a queste macchine una parte sempre maggiore della nostra vita senza conoscere come funzionano accettando che possano sbagliare. Tuttavia operiamo sul piano del concreto, con una matematica in grado di portarci ad un risultato, con margini di errore che cerchiamo di approssimare a zero. Usiamo la tecnologia per farlo senza sapere come funziona la scatola nera delle reti neurali, e senza sapere quanto vale la radice quadrata di 2. Del resto noi sapiens sapiens come diceva Max Weber prendiamo il tram con la razionalità di chi deve raggiungere un tal posto anche se non sappiamo bene tecnicamente come funziona un tram e quale tecnologia ci sia dentro la macchina. Facciamo la stessa cosa in linea di principio quando usiamo un IA. Il problema non si risolve, perché la matematica è uno strumento, divino forse, che oggi non punta al vero ma a produrre un risultato. Nelle intenzioni degli organizzatori del festival ragionare sul tema caos equivale a cercare un punto di vista diverso da quello cui siamo abituati. Il caos non coincide con il puro disordine o la confusione casuale. Al contrario, descrive sistemi complessi e dinamici governati da leggi precise, e caratterizzati da una profonda imprevedibilità. La matematica prova a spiegare il caos pensandolo come sistema deterministico e imprevedibili, ma i sistemi caotici seguono regole esatte, e la loro evoluzione dipende in modo estremo dalle condizioni iniziali. I matematici provano ad utilizzare equazioni differenziali, algoritmi e intelligenza artificiale che permettano di simulare e calcolare queste dinamiche. E in effetti la matematica serve a trovare un ordine e delle regolarità laddove i fenomeni appaiono inizialmente caotici o dominati dalla complessità, come nei mercati finanziari o nei cambiamenti climatici.

Il tema fa emergere i limiti della razionalità umana, nonostante la scienza cerchi di "governare" il caos tramite modelli e calcoli senza riuscire a superare mai il confine sottile tra la precisione teorica delle leggi e l'irriducibile approssimazione o imprevedibilità della realtà.

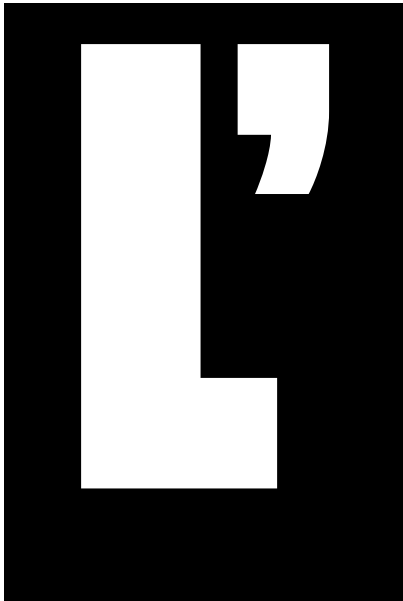
Si ritorna a Flaiano, all'infinitamente medio, e alla radice quadrata di 2, che non sapremo mai quanto vale. Così come non sapremo mai se il caos è un troppo pieno confuso e disordinato, come pensava Ovidio, o se è un vuoto assoluto da cui nasce tutto, come invece sosteneva Esiodo....



CAOS/2

**L'odio, la violenza,
il radicalismo, l'eversione,
la destabilizzazione:
L'accelarazionismo di Mason,
e i neonazisti sul web:
NLM, 764, The Siege,
The Lord of Caos.**





L'8 febbraio 2024, all'istituto Enaip di Varese, un diciassettenne colpisce con tre coltellate alla schiena la sua professoressa. Il 4 marzo 2026, all'istituto «Leonardo da Vinci» di Trescore Balneario, in provincia di Bergamo, un tredicenne accoltella l'insegnante di francese e trasmette la scena in streaming. Il 16 settembre 2026, alla scuola media «Verga» nel quartiere Centocelle di Roma, un altro alunno di tredici anni aggredisce la docente di italiano riprendendo l'azione con uno smartphone montato su un casco. Pochi giorni dopo, a Lamezia Terme, i Carabinieri fermano un quattordicenne prima che entri in classe: il piano per accoltellare la professoressa era stato annunciato e dettagliato su Telegram, con le foto del coltello da caccia già introdotto nei bagni della scuola per essere utilizzato in seguito. L'origine della catena, tuttavia, risale al 29 maggio 2023: all'istituto «Alessandrini» di Abbiategrasso un sedicenne ferisce la docente di italiano con un pugnale e minaccia tenendo in ostaggio la classe con una pistola giocattolo.

Ricomposti in sequenza, questi fatti smentiscono la lettura dell'amnesia o del «raptus» individuale di ragazzi armati di coltello smartphone e magliette con scritto “me ne frego” o “vendetta”. Queste esplosioni isolate di disagio giovanile sono la replica codificata di un modello preciso che nasce 40 anni fa dal radicalismo eversivo degli accelerazionisti americani e da attivisti del Ku Lux Klan. Nei dispositivi sequestrati a Varese, Bergamo, Roma e Lamezia Terme, gli inquirenti della Polizia Postale e i magistrati delle Procure per i Minorenni – in particolare Venezia e Roma – hanno trovato una traccia digitale ricorrente e precisa: l'iscrizione a canali, chat e infrastrutture di messaggistica internazionali orientate alla mitizzazione e all'esecuzione della violenza.

L'INFRASTRUTTURA

Questa violenza è un fenomeno globale, ha una infrastruttura digitale, una storia ormai nota: strumenti di reclutamento collaudati ed efficaci, gerarchie e diversi step di ingaggio e coinvolgimento. Il primo livello è rappresentato dalle reti di adescamento, gruppi informali sul web e sui social noti, come CVLT e 764.

La prima, CVLT, è, nata su Discord nel 2019 per mano dell'allora ventenne francese Rohan Sandeep Rane, poi arrestato in Francia nel 2022 insieme ad altri tre co-amministratori statunitensi Kaleb Merritt, Collin Walker e Clint Borge. Rane ha elaborato per primo le tecniche di ingegneria sociale e ricatto applicate ai minori: adescamento su piattaforme di gioco, come Minecraft o Roblox, l'appropriazione di immagini intime, la sextortion, le minaccia di diffusione dei dati personali, il doxxing, e le richieste di prove di sottomissione. La vittima, per evitare la gogna nei cosiddetti "LoreBooks" o "Dox-Bins", report dossier che ogni singolo gruppo archivia sulle persone che ricatta, viene costretta a compiere atti di autolesionismo o incisioni sulla pelle, i cut signs, da documentare in diretta video.

Dopo i primi arresti dei vertici di CVLT, i metodi del gruppo sono stati fatti propri da un nuovo network, il 764, fondato attorno al 2020 dal quindicenne texano Bradley Chance Cadenhead, noto come “Felix” nel mondo del web. Cadenhead viene arrestato a 16 anni e si è sempre dichiarato colpevole rivendicando le sue azioni. Viene condannato nel 2023, ormai maggiorenne, a 80 anni di carcere. Il nome del gruppo, 764, risale alle prime tre cifre dello codice di avviamento postale di Stephenville, la sua città d'origine in Texas. Sotto la gestione di Cadenhead e dei suoi successori, 764 diventa un marchio e si trasforma in un franchise decentralizzato, privo di un centro di comando unico, operante su canali Telegram non moderati come Harm Nation, CVLTIST, Leak Society o H3ll.

In queste chat, il valore dell'utente è misurato dal clout (la reputazione): una metrica che si acquisisce fornendo materiale violento, eseguendo "sfide" o producendo archivi digitali di foto e video di abusi, usati come moneta d'accesso o come strumento di scambio all'interno del gruppo, nel quale spesso gli aderenti si trasformano da vittime a carnefici. Tra le varie prove ad esempio la richiesta che gli adepti si tatuassero il numero 764 sul braccio.

L'OBIETTIVO

Un ulteriore salto di qualità nella storia di questi gruppi si verifica quando la violenza smette di essere unicamente uno strumento di ricatto individuale o estorsivo e assume la forma di una dottrina eversiva. È il passaggio dai network di "sextortion" al circuito NLM (No Lives Matter) e MKY (Maniacs Murder Cult).

In questi gruppi l'obiettivo è reclutare e addestrare potenziali "esecutori" di azioni violente e dimostrative scegliendo a caso gli obiettivi e curando la scenografia e la diffusione del contenuto. NLM non è la creazione di un quindicenne che dal chiuso della sua stanza organizza una chat su telearma o apre un profilo su Tiktok, ma è un adattamento delle teorie dell'accelerazionismo neo-nazista in chiave digitale, con una nuova semantica e nuovo stile di comunicazione. I quadri concettuali di questi gruppi derivano dagli ambienti del cosiddetto terrorismo senza leader e dalle teorie eversive di movimenti come Atomwaffen Division.

L'obiettivo di NLM è la destabilizzazione del tessuto sociale attraverso l'uso della violenza diffusa e casuale. La scuola diventa il bersaglio privilegiato in quanto è istituzione pubblica di prossimità più accessibile e più simbolica, non solo per i ragazzi ma anche per le famiglie e le intere comunità: aggredire un docente durante l'orario di lezione o filmare l'azione tramite smartphone serve a dimostrare l'impossibilità di protezione da parte dello Stato e a garantire un immediato ritorno mediatico. E a "portare il caos nelle strade", alimentando l'insicurezza.

TEORICI ED ESECUTORI

Le inchieste internazionali (condotte da FBI, Europol e dalle polizie postali europee) mostrano quasi ovunque una struttura di gestione di questi gruppi a doppio livello.

I "Teorici" e Amministratori sono quasi sempre adulti, di solito molto giovani. Il nucleo ristretto dei gestori anonimi, spesso basati nell'Europa dell'Est, in Russia o negli USA redige i manuali, gestisce i canali "hub" su Telegram, produce la propaganda e distribuisce le guide tecniche, nelle quali si spiega come fabbricare esplosivi artigianali o come gestire l'anonimato online.

Il canale d'ingresso, il punto di contatto, non si trova sul dark web ma sulle piattaforme di uso quotidiano dei giovanissimi, TikTok, Instagram, Discord, Roblox. Il funnel di adesione sulle piattaforme pubbliche consiste in video brevi con spezzoni di massacri reali, simbolismi e colonne sonore martellanti. Nei commenti o nelle biografie dei profili vengono inseriti link di reindirizzamento, i funnel, che portano ai canali privati Telegram privi di moderazione. La massa di manovra viene reclutata tramite una dinamica a imbuto, il funnel. I bersagli sono i pre-adolescenti e gli adolescenti dagli 11 ai 16 anni, spesso persone in situazioni di isolamento sociale o di fragilità emotive.

Una volta dentro le chat, i ragazzi subiscono un condizionamento graduale. Si parte dalla fruizione passiva di contenuti gore e violenti per poi passare a richieste attive: sfide di autolesionismo (self-harm), atti di vandalismo, maltrattamento di animali o aggressioni, il tutto condizionato dal ricatto e dall'estorsione di materiale personale condiviso (sextortion).

Fino a poco tempo fa il fenomeno veniva trattato in Italia come cyberbullismo o reato informatico marginale. Gli eventi recenti invece hanno mostrato come dalle incitazioni sul web si sia passati ai comportamenti estremi nel mondo reale, spesso testimoniati In Real Life sulle piattaforme.

I "Carnefici, o gli Esecutori" sono invece i minorenni adescati, ragazzi di 13, 15 o 17 anni che frequentano le chat. Vengono indottrinati, desensibilizzati e infine spinti ad agire In Real Life. Per il minorenne l'atto è un modo per guadagnare reputazione presso il gruppo, la "clout" o per sfogare un disagio. Per i registi del progetto l'atto dell'esecutore è, lo strumento, il "vettore" per generare il panico mediatico e la destabilizzazione sociale.

“No Lives Matter”

Il No Live Matters costituisce quindi il punto di fusione tra due matrici ideologiche nate tra Stati Uniti ed Europa nei primi anni Duemila. La base teorica deriva dagli scritti dell'ideologo neo-nazista americano James Mason, raccolti nel testo “Siege”, in cui si teorizza la leaderless resistance che ha per obiettivo proprio quello di portare il caos, istigando singoli individui a compiere atti di violenza casuale per far collassare le istituzioni democratiche.

L'altra fonte di ispirazione arriva dal mondo del satanismo “misanthropico”, quello di Order of Nine Angles, “O9A”, un gruppo occultista ed eversivo nato nel Regno Unito negli anni '70 che promuove la sovversione morale, l'infiltrazione e la manipolazione di individui vulnerabili per indurli ad atti estremi da usare come “test di purificazione”.

NLM nasce attorno al 2019-2020 quando veterani di queste organizzazioni, tra cui affiliati alla Atomwaffen Division poi messa fuori legge e classificata come terroristica negli USA e in Europa, capiscono che per sopravvivere alla repressione delle forze dell'ordine devono abbandonare la struttura paramilitare rigida e migrare su piattaforme non moderate, cambiando target e puntando sui giovani ragazzi minorenni.

La saldatura con la cultura “Com” dei minori avviene in quel momento, con una ulteriore involuzione ed un cambio di target. Estremisti e affiliati adulti iniziano a frequentare i server Discord e le chat di gaming dove si muovono i ragazzi e propongono i concetti più complessi dell'eversione politica nel linguaggio dei teenager: meme, grafica dark, musica phonk, estetica edgy e logiche gamificate da videogioco, cioè il livello, il clout, il punteggio.

L'operazione ha un relativo successo e NLM viene lanciato non come un partito, ma come un “brand dell'odio” decentralizzato. Chiunque può aprire un canale Telegram e chiamarsi “NLM Italia” o “NLM Varese”, scaricare i PDF con i manuali e dichiararsi parte del movimento.

D notare che la “Com” (abbreviazione informale di Community) è una sottocultura digitale emersa intorno al 2020 nei circuiti giovanili e non moderati di piattaforme come Discord, Telegram, Roblox, Kik. Si tratta di un ecosistema sommerso, fluido e altamente decentralizzato, composto da micro-cellule di adolescenti e giovani adulti. La “Com” si distingue per una combinazione di cybercrime aggressivo, sadismo psicologico, ricerca di visibilità (clout) e un'ossessiva competizione gerarchica.

Questa sottocultura si sviluppa sulla base di una organizzazione informale in microgruppi capaci di sciogliersi e riaggregarsi molto rapidamente.

“SIEGE.” JAMES MASON

Per comprendere infine la natura eversiva dei canali NLM (No Lives Matter), occorre risalire alla genesi teorica dell'accelerazionismo terrorista e ad un corpo dottrinale elaborato negli anni '80 dall'ideologo neo-nazista americano James Mason, all'interno della newsletter Siege, successivamente raccolta in un volume omonimo nel 1992 a cura di Michael Moynihan, divenuto il testo di riferimento per la galassia dell'estrema destra extraparlamentare globale.

La tesi di fondo di Mason parte dalla premessa radicale che il sistema politico, sociale ed economico della democrazia liberale è arrivato a un livello di consolidamento tale che qualsiasi opposizione politica tradizionale, sia essa elettorale, di protesta o paramilitare non ha più efficacia. Di conseguenza, l'unica opzione strategica razionale non è tentare di “riformare” il sistema o “persuadere” le persone, ma accelerare il collasso strutturale della società, da cui lo slogan “Accelerate the collapse”, portando il caos e destabilizzando il mondo.

Questa accelerazione si realizza attraverso la teoria del terrorismo senza leader (leaderless resistance), formulata originariamente dal veterano del Ku Klux Klan Louis Beam. La dottrina elimina la necessità di una gerarchia di comando, di una catena d'ordini o di un'organizzazione formale, elementi che renderebbero la struttura vulnerabile alle infiltrazioni delle forze dell'ordine e della magistratura. La “resistenza senza leader” funziona come un'infrastruttura aperta: una dottrina condivisa indica gli obiettivi strategici e la visione del mondo, mentre l'esecuzione è interamente delegata a singoli individui o a micro-cellule autonome che agiscono in totale indipendenza logistica.

Nel passaggio dai gruppi paramilitari tradizionali disarticolati dall'FBI tra il 2018 e il 2021, alle reti digitali giovanili il testo “Siege” si trasforma in brevi “manuali di condotta” anonimi, sulla violenza casuale come detonatore sociale.

All'esecutore non si richiede una formazione ideologica complessa né una formale adesione politica, ma lo si indottrina e lo abitua alla violenza.

Michael Moynihan e “Lord of chaos”, l'inchiesta sul Black Metal norvegese (1998)

L'altro personaggio centrale della storia è Michael Moynihan, nato nel 1969 a Boston, giornalista, saggista, curatore editoriale e musicista statunitense. È una figura di primo piano, e da sempre molto discussa, all'interno delle sottoculture della musica post-industrial, del neofolk e dello studio dei movimenti eversivi e underground.

Tra la fine degli anni '80 e i primi anni '90, Moynihan ha appena creato il marchio editoriale indipendente “Storm Books”, quando entra in contatto con l'ideologo neo-nazista James Mason. Moynihan comprende il “potenziale” teorico dei suoi articoli scritti tra il 1980 e il 1986, e si fa carico di raccogliere, riordinare e curare i bollettini di Mason in un unico volume, che pubblica nel 1992, con il titolo “Siege: The Collected Writings of James Mason”.

Probabilmente senza l'operazione editoriale e la promozionale condotta da Moynihan, Siege sarebbe rimasto un ciclostilato introvabile della storia dell'estrema destra americana, anziché diventare il testo guida dell'accelerazionismo moderno.

Moynihan è noto al grande pubblico e nell'ambito della saggistica musicale per essere il co-autore, insieme al giornalista norvegese Didrik Søderlind, del celebre libro d'inchiesta “Lords of Chaos: The Bloody Rise of the Satanic Metal Underground”, pubblicato nel 1998 e tradotto anche in italiano. Il saggio ricostruisce la catena di roghi di chiese storiche, omicidi, violenze e radicalizzazione esoterica avvenuta nella scena black metal norvegese nei primi anni '90, ispirata da figure come Varg Vikernes e la band “Mayhem”. Il testo è considerato una pietra miliare della pubblicistica sul tema ed è stato adattato nel 2018 nell'omonimo film del regista Jonas Åkerlund.

Oltre all'attività saggistica, Moynihan è il fondatore del gruppo musicale Blood Axis, una delle band di riferimento della scena neofolk internazionale. Nei suoi lavori musicali e nelle sue scelte editoriali ha sempre affiancato richiami al paganesimo europeo, all'esoterismo e alla filosofia Nietzscheana e Tradizionalista, pubblicando anche saggi legati al pensiero di Julius Evola per il pubblico anglosassone. Moynihan ha costantemente rigettato le accuse di appartenenza politica, definendosi un ricercatore e un osservatore neutrale interessato a documentare le manifestazioni più estreme ed eccentriche del pensiero umano. Eppure le sue operazioni culturali hanno avuto un ruolo determinante nel far circolare testi eversivi nelle sottoculture giovanili e musicali dagli anni '90 a oggi. Per comprendere come questo saggio abbia analizzato l'intreccio tra sottocultura musicale, eversione e violenza giovanile nei primi anni '90, la visione dell'adattamento cinematografico Lords of Chaos, del 2018, offre un'introduzione visiva e storica al contesto in cui operava Moynihan, il quale ha sempre rifiutato ogni etichetta politica rivendicando il suo ruolo di giornalista e di editore.

DIGRESSIONE: DESTABILIZZARE

Interessante nel nuovo terrorismo neo nazista il riferimento al caos come strumento di destabilizzazione. Nella storia del pensiero e della cosmogonia occidentale e orientale, il caos ha quasi sempre mantenuto un carattere generativo. Dalla mitologia greca fino alle teorie scientifiche e filosofiche della complessità, il disordine rappresenta un momento di rottura di un equilibrio statico da cui può emergere una nuova configurazione o un progetto politico e sociale.

Nei network accelerazionisti prende forma un nuovo concetto di Caos che diventa disordine radicale privo di qualsiasi dimensione costruttiva o trasformativa, e privo di qualsiasi empatia umana. Non c'è alcun "dopo" nella catena di atti eversivi, la tortura psicologica sulla vittima adescata o l'aggressione fisica contro un docente in aula non portano un nuovo ordine morale, politico o sociale, si limitano a erodere la fiducia in quello c'è. Il fine è il collasso sociale, senza alcuna ipotesi su cosa debba subentrare dopo le macerie.

La realtà inoltre si riduce ad una metrica di gioco: si gamifica la sofferenza dell'altro, il sangue, la disperazione della vittima o il panico collettivo generato da un fatto di cronaca, elementi che non vengono percepiti nella loro concretezza umana, ma tradotti nel linguaggio dell'algoritmo. La violenza diventa punteggio, clout (reputazione), moneta d'accesso a sottogruppi riservati. L'azione violenta è svuotata di senso etico e trattata come la risposta a una "sfida" o l'avanzamento di livello all'interno di un videogioco sociale. L'unica certezza oltre alla sfida resta il vuoto senza senso i legami che si creano tra gli adepti.

L'aggressione nell'aula scolastica o la minaccia diffusa in chat rappresentano la violenza "pura" del tutto de-umanizzata, che spinge alla dissoluzione di ogni norma, di ogni valore e di ogni legame sociale.

“LORDS OF CHAOS”

Nel dibattito sulla natura della violenza subculturale, il saggio *Lords of Chaos* (1998) di Michael Moynihan e Didrik Söderlind aiuta a comprendere meglio la genesi di questi gruppi e il loro stile operativo estremo ed eversivo.

Il fenomeno dell'accelerazionismo puro, con il suo nichilismo orientato alla distruzione delle istituzioni, e quello del black metal norvegese dei primi anni '90, con le sue propaggini giovanili delineano l'idea della messa in scena della trasgressione, dove l'estetica e la teatralità della violenza finiscono per travolgere e determinare la realtà stessa. Negli episodi di Oslo e Bergen con le rappresentazioni dei roghi di chiese plurisecolari, delle profanazioni, dell'omicidio di Euronymous per mano di Varg Vikernes, la spinta iniziale era un gioco di rialzo folle ed identitario all'interno di un gruppo giovanile ristretto, il cosiddetto "Black Circle". La violenza serviva a convalidare uno stile, una posa artistica e un'appartenenza subculturale di fronte ai propri pari.

Il caos simulato opera attraverso un meccanismo d'inversione, per cui la finzione scenica, l'iconografia diabolica, la retorica misantropica, le copertine dei dischi, viene portata alle sue estreme conseguenze fisiche per dimostrare "autenticità" del fenomeno nel suo complesso. Non è l'ideologia a generare l'atto, ma il bisogno di costruire e identificarsi in una simulazione stilistica fino al punto di non ritorno.

Lords of Chaos individua precocemente ciò che decenni dopo si è trasferito nell'infrastruttura di Telegram e Discord: la trasformazione del crimine d'odio e della violenza in un contenuto mediatico. Con le evoluzioni del caso: i protagonisti del Black Circle norvegese cercavano l'attenzione dei giornali e della televisione, oggi gli adolescenti delle chat 764 e NLM cercano il clout, la credibilità e il rispetto, e l'attenzione degli algoritmi dei social media.

Se l'accelerazionismo politico di Siege usa il caos come strumento tattico e la sottocultura Com lo consuma come metrica di gioco, il "caos simulato" teorizzato in *Lords of Chaos* ne rivela la matrice originaria: una recita estetica e identitaria che, a forza di essere esibita e alzata di tono, finisce per produrre sangue reale.

PER APPROFONDIRE: LE FONTI

Le fonti principali cui attingere per approfondire il tema sono gli atti giudiziari e di polizia, i report di intelligence, le inchieste giornalistiche internazionali e alcuni saggi accademici.

Le fonti giudiziarie e di polizia in Italia ed in Europa, in cui si parla del tema sono emesse dalle Procure della Repubblica presso i Tribunali per i Minorenni di Venezia, Roma, Milano, Caltanissetta. Tra gli Atti i report relativi alle indagini sulle reti Telegram, adescamento e reati di eversione/istigazione alla violenza tra minori.

Altri documenti derivano dalla Polizia di Stato e dal Servizio di Polizia Postale: report annuali sulla pedopornografia e sul cybercrime minorile, in particolare i dossier dedicati al fenomeno della sextortion e delle sfide online le cosiddette challenge.

Ci sono poi i report annuali della EuroPol, dedicati alla convergenza tra cybercrime, la violenza in rete e la radicalizzazione giovanile.

In America l'FBI ha classificato i network 764 e Comcome minacce prioritarie per la sicurezza dei minori e la radicalizzazione interna. Il Department of Justice statunitense rende disponibili documenti relativi ai processi e alle sentenze di condanna di alcuni fondatori dei gruppi ad esempio il procedimento contro Bradley Chance Cadenhead e l'incriminazione federale dei vertici di CVLT del gennaio 2025.

Ci sono poi i Centri di ricerca e i Think Tank specializzati: l'ISD, l'Institute for Strategic Dialogue con sede a Londra e a Washington, uno dei centri studi leader nel tracciamento del pensiero accelerazionista e delle reti neo-fasciste su Telegram e Discord. Pubblica regolarmente report e policy brief specifici sulla True Crime Community (TCC) e sui gruppi No Lives Matter (NLM).

Il King's College di Londra produce analisi dettagliate sull'estetica "edgy", l'uso della musica "phonk" e la gamificazione della violenza nelle chat di minori. Lo stesso fa il MEMRI, il Middle East Media Research Institute, che si occupa di Cyber terrorismo e violenza domestica, con lavori di monitoraggio delle estensioni ideologiche dell'accelerazionismo, inclusa la penetrazione nelle sottoculture giovanili.

Infine ci sono le inchieste giornalistiche di riferimento, cui abbiamo attinto per fare questo prima o approfondimento sul tema.

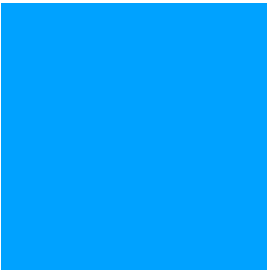
Il "Washington Post" e il "New York Times" hanno dedicato approfondimenti d'inchiesta alla nascita del fenomeno 764 negli USA, ricostruendo le storie delle vittime e il ruolo delle piattaforme e l'utilizzo delle metodologie di analisi usate per tracciare i canali Telegram accelerazionisti e i network estorsivi europei.

In Italia del tema si sono occupate "La Repubblica", "Il Corriere della Sera" e "Domani", in particolare a proposito delle ricostruzioni di cronaca giudiziaria locale relative ai singoli episodi di Abbiategrosso, Varese, Trescore Balneario, Roma Centocelle e Lamezia Terme.

NO LIVES



MATTER



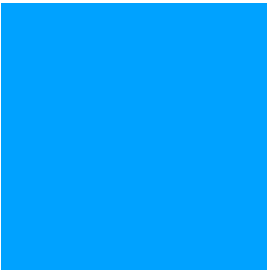
È il pomeriggio di sabato 18 novembre 2023, intorno alle ore 17:00, nel cuore del passeggio commerciale di Milano. Lungo Corso Vittorio Emanuele II, a pochi passi da Piazza del Duomo, una banda di circa venti giovanissimi — quasi tutti minorenni tra i 14 e i 17 anni, di seconda generazione e residenti nei quartieri della prima cintura metropolitana — fa il suo ingresso compatto all'interno di un noto flagship store di abbigliamento streetwear. L'incursione dura meno di due minuti: mentre una decina di ragazzi saturano l'ingresso per bloccare visivamente i fisici della security, il resto del gruppo fa razzia dagli scaffali di felpe con cappuccio, giubbotti imbottiti e sneakers di marca, strappando i dispositivi anti-taccheggio a forza bruta e riprendendo la scena con gli smartphone. All'uscita, la banda si disperde nel flusso della folla verso la fermata della metropolitana di San Babila. Nelle ore successive, la refurtiva viene esibita nei video pubblicati su TikTok e Instagram, montati su tracce musicali drill, per certificare l'azione prima che scattino i rilievi delle forze dell'ordine a seguito delle denunce presentate dai responsabili del negozio.

È in episodi come questo che la devianza giovanile incontra la società dei consumi nella sua forma più spietata e spettacolare. Sviluppatosi lungo le grandi arterie commerciali milanesi — da Corso Vittorio Emanuele a Via Torino, fino a Corso Buenos Aires — il fenomeno del Marathon Shopping ha introdotto una variante sostanziale rispetto alle risse o alle occupazioni di piazza: la trasformazione della merce desiderata in un trofeo da conquistare e performare in rete.

La dinamica non ha nulla a che vedere con il taccheggio tradizionale, fondato sulla discrezione e sul mimetismo. Il Marathon Shopping opera al contrario per saturazione e sovraesposizione. Un gruppo compatto di ragazzi, coordinato via Telegram o storie temporanee su Instagram, fa irruzione nei grandi store. L'azione si consuma in un arco temporale ridottissimo. L'obiettivo immediato è saturare lo spazio fisico per paralizzare la reazione dei commessi e della vigilanza: mentre una parte del gruppo fa scudo, gli altri afferrano a piene mani capi e accessori dagli scaffali. La dispersione è immediata, molecolare, verso le stazioni della metropolitana.

La vera natura del fenomeno emerge nel passaggio successivo. Nel Marathon Shopping, la refurtiva non è destinata al mercato nero né alla monetizzazione informale. Il bene sottratto — il capo di marca marchiato dai loghi più iconici della cultura giovanile contemporanea — assume una funzione strettamente totemica e simbolica. Nelle economie della marginalità o della periferia urbana, l'accesso all'abbigliamento di marca rappresenta lo strumento primario di riscatto visibile, il segnale d'ingresso nella mappa del valore sociale. Sostituire l'acquisto con la conquista violenta trasforma l'oggetto consumabile in un trofeo di guerra.

In questo passaggio si salda il cortocircuito con la dimensione digitale. Il capo d'abbigliamento conquistato sul campo non viene semplicemente indossato nella vita quotidiana, ma viene immediatamente esibito nei contenuti pubblicati su TikTok e Instagram, adagiato su tracce musicali drill e trap. Il brand cessa di essere una merce e diventa la prova tangibile di una prestazione. L'esibizione della merce sottratta incrementa direttamente il capitale reputazionale (clout) dell'autore, certificandone l'audacia di fronte alla propria comunità di riferimento.



L'analisi di questo specifico modello operativo fa emergere nodi interpretativi che vanno ben oltre la semplice cronaca di micro-criminalità urbana.

Il primo aspetto riguarda la totale adesione, paradossale e feroce, ai valori dell'iper-consumismo. I ragazzi che partecipano a queste incursioni non contestano la merce o il modello di consumo dominante; al contrario, ne sono completamente saturati. La violenza del blitz non punta a distruggere il tempio del consumo, ma ad abbreviarne drasticamente le vie d'accesso. Di fronte all'impossibilità o alla lentezza dell'acquisto economico formale, l'infrastruttura del retail viene aggredita per via diretta, piegando la logistica dello store alla logica della rapina d'esibizione.

Si consolida così un'estetica del reato che trova nella moda giovanile il suo linguaggio visivo. La sovrapposizione tra la grammatica visiva dei video musicali e la condotta reale crea un circolo vizioso in cui il capo di marca, il passamontagna e il gesto della sottrazione rapida diventano gli elementi di una divisa d'ordinanza. La marca acquisisce un plusvalore simbolico proprio perché sottratta e non comprata: indossare quel preciso giubbotto significa esibire non soltanto un bene economico, ma la capacità di avere sopraffatto il sistema di controllo del centro cittadino.

Inoltre, il Marathon Shopping evidenzia una frattura geografica e sociale profonda nella fruizione della metropoli. Le vie del lusso e del commercio centrale cessano di essere luoghi di attraversamento o di passeggiata per trasformarsi in territori di caccia temporanei per le periferie. La centralità urbana viene invasa per il tempo strettamente necessario a estrarre valore simbolico — sotto forma di merci e di riprese video — per poi essere immediatamente abbandonata. Il centro della città si riduce a riserva di risorse da saccheggiare per alimentare la narrazione identitaria dei margini.

Infine, il fenomeno mostra l'inadeguatezza delle risposte basate sul solo presidio di sicurezza privato. Le grandi catene commerciali si trovano esposte a una forma di aggressione che sfrutta la stessa architettura dei loro spazi: ambienti ampi, merce a vista, accessibilità totale e personale ridotto all'essenziale per contenere i costi di gestione. La logica dell'accoglienza totale e del libero servizio, pensata per incentivare l'acquisto d'impulso, diventa l'elemento di massima vulnerabilità di fronte a gruppi che non entrano per comprare, ma per estrarre trofei da esibire nel mercato dell'attenzione digitale.



**CAOS/3
MODENA
IL BLITZ
ALL'MD.**

È la prima metà del pomeriggio di martedì 22 settembre 2026. La finestra temporale è quella compresa tra le 14:00 e le 14:15, il momento esatto in cui si concludono le lezioni negli istituti superiori e nei centri di formazione professionale della zona ovest di Modena, non lontano dal comparto direzionale di via Rainusso. Il bersaglio è il punto vendita del discount MD, un esercizio commerciale ad alto flusso, con personale ridotto all'essenziale e posizionato lungo un'arteria di snodo. L'azione si sviluppa in circa dieci minuti e coinvolge un gruppo stimato di circa cinquanta ragazzi, in larga parte minorenni, molti dei quali appartenenti a contesti scolastici locali come lo IAL. Non si tratta di un'intrusione clandestina né di un taccheggio condotto con discrezione. Il gruppo si presenta compatto e fa ingresso simultaneo nei locali, saturando le corsie e travolgendo la routine del supermercato. Parte dei giovani indossa sciarpe o cappucci tirati sul volto per rendere difficoltoso il riconoscimento immediato, mentre la massa agisce all'aperto, sbeffeggiando i dipendenti e creando un clima di generale disorientamento tra i clienti attoniti.

La razzia si concentra su beni ad immediato consumo giovanile: lattine di bibite gassate, energy drink, alcolici, patatine e cibo pronto confezionato, come i tramezzini. I prodotti vengono afferrati direttamente dagli scaffali e infilati negli zaini da scuola oppure portati via a mani nude, provocando anche la caduta e il danneggiamento di molta merce lungo i corridoi. Durante l'intero blitz, la componente visiva è preponderante: diversi partecipanti non rubano, ma utilizzano gli smartphone per registrare video, inquadrando i compagni mentre compiono gli illeciti e raccogliendo materiale destinato alla pubblicazione sui social media. La dispersione avviene con la stessa rapidità dell'ingresso, prima dell'arrivo delle forze dell'ordine o dell'intervento della vigilanza privata. I video girati all'interno dell'MD finiscono su TikTok e Instagram nell'arco di pochissime ore, accompagnati da audio virali e commenti celebrativi.

Gazzetta di Modena

L'evento non si esaurisce con la prima incursione. Già il giorno successivo, mercoledì 23 settembre, un gruppo di circa trenta ragazzi ritenta un blitz analogo nello stesso punto vendita e in altre strutture della zona, confermando la natura emulativa e organizzata della condotta, coordinata attraverso gruppi di messaggistica istantanea. Dalle ricostruzioni e dalle denunce emerse nelle ore successive — tra cui quelle dello IAL di Modena — è emerso come l'episodio si sia innestato in un clima di tensione preesistente tra gruppi di studenti di istituti diversi. Inizialmente si è parlato di una sorta di "spedizione punitiva" o di regolamento di conti nei confronti di un singolo allievo, difeso da una docente prima che la situazione degenerasse e scivolasse nell'assalto al supermercato, trasformando un fatto individuale in una azione collettiva ad alto impatto pubblico.

L'ipotesi della challenge social.

La rapidità di concentrazione e la decisione di colpire un discount in una precisa fascia oraria denotano l'esistenza di una regia informale distribuita via WhatsApp e Telegram. L'ipotesi degli investigatori, supportata dall'immediata diffusione delle immagini in rete, punta sull'esistenza di una sfida social lanciata per testare i limiti dell'impunità e misurare la capacità di reazione del sistema locale. Il video pubblicato non serve a documentare l'accaduto a posteriori, ma costituisce il fine primario dell'azione: rubare la merce garantisce il materiale di scena per la clip da erogare sulla piattaforma.

L'assalto ha messo in evidenza la vulnerabilità sistemica di strutture pubbliche che possono diventare teatri di manifestazioni violente amplificate dalla tecnologia digitale e dagli algoritmi delle piattaforme, e che le misure repressive da sole non sono sufficienti né efficaci, serve un'alleanza tra istituzioni scolastiche, famiglie e presidi territoriali per evitare che i percorsi formativi dei minorenni vengano totalmente assorbiti dalle dinamiche di gruppo e dalla pressione emulativa della rete. Non è il primo.



SOCIAL CHALLENGE

Tra la fine del 2020. E l'inizio del 2021 nel pieno delle restrizioni pandemiche, tra zone rosse, coprifuoco e didattica a distanza, le piazze delle principali città italiane divennero improvvisamente il teatro di un fenomeno inedito per forme e scala: le maxi-risse coordinate tra adolescenti. Dal centro storico di Venezia alla terrazza del Pincio a Roma, passando per i centri di Gallarate e Milano, centinaia di giovanissimi si diedero appuntamento su Instagram e TikTok per affrontarsi a calci e pugni nelle piazze deserte delle città in Lock down..

L'episodio di Venezia avvenne poco prima di Natale e fu il più preoccupante perché confermò che il caso non si esauriva da solo. L'11 e 12 dicembre 2020, circa 40 adolescenti si diedero appuntamento in pieno centro storico. Lo scontro principale avvenne nell'area tra Campo San Giacomo dell'Orio e la zona di Rialto. Molti dei ragazzi venivano dalla terraferma nonostante i divieti di spostamento tra comuni. Lo scontro a calci e pugni proseguì tutta la sera e la notte tra le calli e i campielli, fino all'intervento improvviso delle forze dell'ordine e della Polizia Municipale provocò un fuggi fuggi generale verso la stazione di Santa Lucia e i pontili dei vaporetti.

La Questura parlò di un "tam-tam" partito da gruppi e storie su Instagram nei giorni precedenti, nato da futile rivalità di quartiere esasperate dall'isolamento.

L'ultimo degli scontri avvenne a Roma in Piazza del Popolo, il 30 Gennaio 2021. Ad affrontarsi bande di quartiere, con cariche e fughe verso il Pincio e la metro.

Il contesto sociologico e le indagini

Le forze dell'ordine (in particolare la Polizia Postale) attivarono un monitoraggio sistematico delle piattaforme social per intercettare gli hashtag e gli appuntamenti prima che degenerassero in piazza. Dalle analisi sociologiche e giudiziarie dell'epoca emersero tre fattori chiave:

Effetto "container" della pandemia: L'assenza di scuola in presenza, la chiusura di centri sportivi e i coprifuoco alimentarono nei giovanissimi forti livelli di frustrazione e aggressività repressa.

Spettacularizzazione: Le risse venivano girate da smartphone dedicati per poi essere ripubblicate con basi musicali (frequentemente musica trap) per guadagnare follower e visibilità online.

Assenza di leadership strutturata: Non si trattava di baby gang con organizzazione criminale tradizionale, ma di aggregazioni fluide e temporanee nate da sfide verbali sul web.

Anche in questo caso non esiste una regia criminale organizzata né la struttura rigida delle baby gang tradizionali. Le risse erano "semplicemente" una mobilitazione fluida, priva di una catena di comando formale, alimentata dalla frustrazione dell'isolamento ma soprattutto dall'esigenza di performare lo scontro davanti all'obiettivo dello smartphone., una sorta di Flash Mob" organizzato partendo da antiche rivalità ampiamente radicalizzate nel tempo.

A distanza di anni, la stessa grammatica riemerge intatta in uno scenario diverso, quello del discount MD di via Rainusso a Modena, teatro dell'attacco organizzato da ragazzi minorenni all'uscita da scuola, un blitz di dieci minuti caratterizzato dalla razzia coordinata di alcol ed energy drink, da riprese video sistematiche e da un'immediata dispersione, poi replicato il giorno successivo in altri punti vendita della città. Anche in questo caso, nessuna mente direttiva: l'azione è nata da una call lanciata sui canali social e nei gruppi di messaggistica da pochi elementi in cerca di visibilità, a cui decine di coetanei si sono aggregati per spirito di emulazione.

Colpisce la forza dell'effetto emulazione, il meccanismo sociale da sempre presente nelle condotte collettive, che nel tempo dei social e delle IA sembra aver compiuto un salto di scala definitivo. L'emulazione analogica procedeva per contiguità geografica, appartenenza di gruppo e tempi di propagazione lenti. L'organizzazione a rete e molecolare ne ha invece ingegnerizzato e codificato il funzionamento. Le piattaforme digitali non si limitano a fare da cassa di risonanza dell'evento a fatto compiuto: forniscono l'infrastruttura logistica, le istruzioni d'uso e l'incentivo reputazionale. La struttura decentrata permette a chiunque di prelevare uno schema di azione (la challenge, il flash-mob, il saccheggio), adattarlo al proprio contesto territoriale e renderlo immediatamente esecutivo.

Attraverso questo dispositivo, la spinta virtuale trapassa senza alcuna mediazione nella realtà concreta. I social media cessano di funzionare come diario passivo o specchio della realtà per diventare un dispositivo d'innescò operativo, capace di muovere corpi, occupare spazi e produrre danni materiali. Si assiste così a un progressivo depotenziamento della realtà fisica, che perde la sua autonomia di senso. L'azione nel mondo reale non trae più valore dalle sue conseguenze materiali o dall'esperienza vissuta sul posto, ma dal volume di attenzione e dal capitale di reputazione digitale (clout) che riesce ad astrarre e quantificare. Nella percezione di chi partecipa, ciò che non viene registrato, confezionato ed erogato in tempo reale semplicemente non esiste.

Il media smette di essere un canale di racconto e si impone come il fattore determinante dell'azione stessa. Si chiude in questo modo un circuito a doppio senso — dal digitale al fisico, e dal fisico di nuovo al digitale — in cui lo spazio pubblico, i supermercati e le piazze si riducono a meri set scenografici. La realtà materiale continua a subire i costi sociali dell'insicurezza e del degrado della convivenza, mentre le condotte di chi la abita rispondono in via prioritaria alle metriche d'ingaggio e agli algoritmi dell'infrastruttura digitale.

L'intersezione tra reti molecolari, performance digitale e uso dello spazio urbano traccia linee di faglia che richiedono di essere esplorate lungo direttrici precise.

Il primo nodo riguarda la trasformazione profonda dei modelli aggregativi e la ridefinizione della devianza giovanile. La sociologia urbana e gli interventi di prevenzione si sono storicamente appoggiati sull'esistenza di luoghi fisici e stabili: il quartiere, la piazzetta, la curva dello stadio o la struttura rigida della banda territorialmente delimitata. Le nuove forme di azione coordinata disintegrano questo perimetro. Ci si trova di fronte ad aggregazioni ad hoc, temporanee, liquide e destinate a sciogliersi un secondo dopo il blitz. Questo passaggio impone di ripensare radicalmente i dispositivi del lavoro sociale e dell'educazione di strada: la prevenzione non può più operare sulla presenza costante nei luoghi del ritrovo, ma deve sviluppare la capacità di intercettare le dinamiche di condensazione rapida che si generano nell'infrastruttura digitale molto prima di manifestarsi sul territorio.

A questa mutazione si innesta un immaginario estetico e sonoro altamente codificato, riconducibile alla cultura drill e alla trap di seconda generazione. L'uso della musica come colonna sonora sovrapposta ai video delle azioni non è un dettaglio di contorno, ma la chiave d'accesso a un codice identitario. I volti coperti da balaclava, la gestualità ipnotica di gruppo, la postura aggressiva, la visualizzazione di merci povere come trofei di caccia sono elementi presi a prestito dai video musicali e riprodotti nel quotidiano. La sottocultura musicale non causa direttamente la devianza, ma fornisce una grammatica visiva che legittima e "nobilita" l'atto vandalico, trasformando il micro-crimine o la sopraffazione urbana in una scelta stilistica immediatamente riconoscibile e spendibile nel mercato dell'attenzione.

A questa mutazione strutturale si affianca la frattura temporale tra il ritmo dei social e la risposta delle istituzioni. Esiste una divergenza profonda tra l'istantaneità dell'innescò algoritmico e i tempi dilatati, riflessivi e procedurali della macchina pubblica. Il circuito dell'azione si consuma nell'arco di poche ore: l'innescò, l'esecuzione, il montaggio del video e la rendicontazione in termini di visualizzazioni avvengono all'interno di una finestra temporale ristrettissima. In quella finestra si genera un'illusione di impunità totale, sostenuta dalla sensazione che la reazione dell'autorità sia incapace di raggiungere la velocità della rete.

La risposta formale delle istituzioni scolastiche, giudiziarie o repressive interviene sempre a valle, con cadenze analogiche che faticano a incidere sul meccanismo emulativo. Ricucire questa distanza richiede di superare la sola risposta emergenziale o sanzionatoria a posteriori, costruendo presidi di responsabilità che agiscano sulla continuità della presenza educativa anziché sull'inseguimento del singolo episodio.

Questa stessa dinamica temporale mette in luce il ruolo decisivo, e spesso trascurato, della maggioranza silenziosa che compone il raduno: la figura dello "spettatore attivo". In un assalto come quello di Modena o in un raduno al Pincio, gli esecutori materiali del furto o della violenza rappresentano una percentuale ridotta. La massa restante è composta da ragazzi che non rubano e non colpiscono, ma la cui presenza è determinante. Si tratta di una nuova forma di turismo della trasgressione, in cui il giovane partecipa per il brivido dell'essere presente, per documentare con il proprio telefono o semplicemente per non rimanere escluso dal racconto del gruppo. La responsabilità individuale si diluisce nell'azione collettiva, creando un guscio protettivo d'impunità psicologica dove l'atto illecito altrui viene consumato dal vivo come una forma di intrattenimento immersivo.

Il terzo livello d'analisi chiama in causa il ruolo economico dell'infrastruttura tecnologica e la geografia stessa dei suoi bersagli. Le piattaforme digitali non rappresentano scenari neutri su cui si proiettano i comportamenti umani, ma soggetti economici la cui architettura d'ingaggio è progettata per estrarre profitto. La monetizzazione diretta o indiretta del traffico premia in modo sistematico i contenuti ad alto impatto emotivo, trasgressivi o violenti, poiché capaci di trattenere l'attenzione dell'utente e generare metriche di interazione elevate. Si crea così un paradosso strutturale: mentre la comunità locale sopporta i costi materiali e sociali del disordine urbano, l'infrastruttura privata dei social estrae valore economico dalla spettacolarizzazione dell'infrazione. Diventa quindi ineludibile porre il tema della responsabilità oggettiva delle piattaforme e delle scelte di design algoritmico che incentivano la performance della devianza.

La scelta del bersaglio riflette questa trasformazione. Se nel 2020 l'azione cercava le piazze storiche e i centri cittadini per massimizzare la visibilità scenica, il blitz di Modena sposta l'asse verso i "non-luoghi" del consumo rapido: discount di periferia, snodi di passaggio vicino alle stazioni, parcheggi di centri commerciali. Sono spazi funzionali, caratterizzati da personale ridotto, relazioni umane rarefatte e un'assenza fisiologica di presidio comunitario. Colpire un discount all'ora di punta significa scegliere un bersaglio iper-vulnerabile, dove le barriere di protezione sociale sono ridotte al minimo e dove la merce è disposta per essere afferrata rapidamente, facilitando la traduzione immediata della challenge in gesto materiale.

Infine, l'impatto più sistemico si misura sulla progressiva desertificazione simbolica dello spazio pubblico e sul cortocircuito del sistema di sanzione. Quando la piazza cittadina, la stazione ferroviaria o il supermercato di quartiere vengono ridotti a meri fondali per la registrazione di una clip, si assiste a un drastico svuotamento della loro funzione primaria. Gli spazi collettivi nascono per garantire la prossimità, lo scambio, l'attraversamento e la tutela reciproca tra estranei. La loro riduzione a set cinematografico d'occasione per il consumo remoto ne annulla la densità relazionale, restituendo ai cittadini ambienti percepiti come ostili, imprevedibili o degradati.

Quando la macchina investigativa e giudiziaria interviene a posteriori, emerge la profonda discrepanza tra la gravità delle contestazioni penali (dalla rissa aggravata al furto con strappo) e la totale assenza di percezione del limite da parte dei giovani coinvolti, che spesso leggono l'intervento delle forze dell'ordine come un'incomprensione sproporzionata nei confronti di un "gioco digitale". Sottrarre la città a questa dinamica richiede politiche pubbliche capaci di investire sulla risignificazione dei luoghi e sulla ricostruzione del senso del limite, promuovendo usi intensivi, culturali e partecipativi dello spazio comune in grado di competere con l'attrazione esercitata dalla performance digitale.



PAROLE, VETTORI, SPAZI: I modelli semantici dell'IA

La geometria del linguaggio.

Siamo abituati a pensare alla parola come a un simbolo: un costrutto culturale, una convenzione sociale, un'impronta del pensiero umano. Come osservava Ludwig Wittgenstein, il significato di una parola sta nel suo uso nel linguaggio, all'interno di una forma di vita che condivide regole, esperienze e contesti reali. Per una macchina, tuttavia, la parola perde qualunque statuto simbolico o filosofico. Per la macchina, ogni parola è semplicemente un vettore.

In matematica e in fisica, un vettore è una grandezza definita da valori numerici che ne indicano la posizione, la direzione e il verso all'interno di uno spazio. Se immaginiamo lo spazio tridimensionale comune, un vettore spaziale è quella freccia che parte dall'origine degli assi per identificare un punto preciso attraverso tre coordinate (x,y,z). I modelli linguistici applicano esattamente questo principio, ma espandono lo spazio geometrico a centinaia o migliaia di dimensioni invisibili all'occhio umano. Ciascuna parola diventa un punto preciso dentro questa mappa iper-spaziale.

Come fa la macchina a imparare a tenere assieme le parole? Non consultando un dizionario o comprendendo il "gioco linguistico" di cui parlava Wittgenstein, ma analizzando le relazioni di vicinanza e contesto su miliardi di testi scritti. Durante la fase di addestramento, il sistema osserva quali termini compaiono frequentemente insieme e in quali strutture. Se due parole condividono contesti simili, l'algoritmo le trascina matematicamente vicine nello spazio vettoriale: "gatto" e "cane" finiscono per occupare posizioni adiacenti, mentre "gatto" e "democrazia" risultano matematicamente distanti.

La cosa straordinaria è che le relazioni semantiche diventano veri e propri spostamenti geometrici. La distanza e la direzione che separano la parola "uomo" da "re" sono identiche alla distanza e alla direzione che separano "donna" da "regina". Togliendo il vettore di "uomo" da quello di "re" e aggiungendo quello di "donna", la traiettoria porta quasi esattamente al punto occupato da "regina". Il sistema non sta riflettendo sulle nozioni di sovranità o di genere: sta eseguendo un'operazione algebrica di precisione.

L'errore più comune nell'interpretare l'intelligenza artificiale consiste nel presumere che, per formulare un ragionamento sensato, debba seguire una catena di causa-effetto o un'intenzione comunicativa. Il modello opera invece su basi squisitamente probabilistiche e strutturali.

La macchina non legge, calcola

Attraverso l'architettura dei Transformer e il meccanismo di attenzione (Self-Attention), il sistema non analizza le parole una dopo l'altra in modo lineare. Quando deve generare il termine successivo, osserva l'intero blocco di parole precedenti contemporaneamente e assegna un peso matematico a ciascun elemento in relazione agli altri. In una frase complessa, il modello lega il soggetto al verbo anche se separati da molte subordinate, ignorando i termini intermedi più vicini ma irrilevanti. Calcolata la configurazione dei vettori e le loro relazioni d'attenzione, stabilisce qual è la parola con la più alta probabilità statistica di completare la sequenza.

Il punto di rottura fondamentale risiede in questo scarto: la macchina produce un risultato del tutto comprensibile, sensato e logicamente rigoroso per l'uomo, utilizzando un processo che non ha nulla di umano.

L'essere umano muove da un'intenzione comunicativa, fonda il discorso sull'esperienza del mondo e articola il pensiero attraverso nessi di causa-effetto. Il modello linguistico riceve invece una sequenza numerica, la traduce in posizioni geometriche all'interno di una matrice e calcola probabilità matematiche al solo scopo di minimizzare l'errore statistico nella sequenza.

La comprensione

L'illusione di una vera "comprensione" nasce dal fatto che la struttura stessa del linguaggio umano è intrisa di logica. Avendo mappato la quasi totalità della produzione scritta, il modello ha codificato la geometria di quella logica. Quando ricompone i vettori, rispetta le regole del pensiero umano non perché le intenda, ma perché la forma matematica del testo da cui ha appreso è conformata a quelle regole.

Questa medesima logica vettoriale non governa soltanto il testo, ma si estende alla generazione delle immagini. I modelli di diffusione visiva lavorano sullo stesso principio: non dipingono un soggetto guidati da un'intenzione estetica, ma calcolano la distribuzione spaziale dei pixel partendo da un rumore di fondo per approssimare la forma statistica di ciò che chiamiamo "volto" o "paesaggio". La macchina non sa cosa sia una mano o la luce del tramonto: conosce esclusivamente la correlazione matematica tra le frequenze di colore e i vettori del testo che le descrivono.

Il passaggio dalle parole alle immagini compie tuttavia una torsione ulteriore, trascinando la società in un territorio del tutto inedito: la sostituzione del vero con il verosimile.

Per secoli, la fotografia ha registrato un'impronta fisica della realtà; costituiva una prova, una traccia materiale di un evento accaduto in un preciso punto dello spazio e del tempo. La generazione artificiale delle immagini spezza questo vincolo ontologico. La macchina non documenta ciò che è stato, ma produce ciò che è statisticamente plausibile. Un'immagine artificiale non è vera e non è falsa: è iper-reale, poiché incarna la media perfetta della nostra grammatica visiva.

La trappola per l'orizzonte cognitivo umano risiede proprio qui. La nostra mente, tarata per attribuire valore di verità a ciò che vede, si trova priva di strumenti di difesa di fronte a un'iconografia perfetta ma sganciata da qualsiasi referente materiale. Quando la differenza tra ciò che è accaduto e ciò che è soltanto calcolato come probabile diventa invisibile a occhio nudo, a crollare non è solo la certezza dell'informazione, ma lo stesso statuto di testimonianza su cui abbiamo fondato la memoria collettiva, la storia e la giustizia. La realtà smette di essere un dato oggettivo e si trasforma in un'opzione fra le tante, sprofondando in un paesaggio visivo in cui la verosimiglianza ha definitivamente soppiantato la verità.

P.S. La parola è diventata un simbolo riducibile a calcolo esattamente come nell'intuizione dei matematici e degli logici del Novecento, quando la scienza smise di inseguire le vertigini dell'infinitamente grande o dell'infinitamente piccolo per concentrarsi sulla misurazione dell'«infinitamente medio», per dirla con Ennio Flaiano. In fondo, il problema della lingua artificiale si rivela identico a quello della radice quadrata di 2: un numero irrazionale che non può essere espresso da una frazione esatta, ma che possiamo approssimare all'infinito aggiungendo decimali. La macchina fa lo stesso con il linguaggio, calcolando un'approssimazione sempre più perfetta della nostra sintassi. Con un'unica, fondamentale differenza: mentre la matematica pura lavora su astrazioni ideali, la parola umana ragiona ed esiste all'interno di un contesto per forza reale. Nel momento in cui stacciamo la catena dei vettori da questa aderenza materiale alla vita vera, non stiamo più creando significato, ma soltanto una luccicante, inappuntabile illusione statistica.



**Identità,
tecnologia e
nuove forme
di umanità**

Viaggio speculativo tra identità, tecnologia e nuove forme di umanità

2045: Diario di un umano (?) di 5 anni. Data terrestre: 21 marzo 2045. Sulla Terra, uno dei primi umani della Generazione Gamma ha appena compiuto 5 anni. Si chiama Ael, non ha sesso assegnato, e abita in una “unità relazionale a geometria variabile”. La sua giornata inizia alle 5:45, quando Soma, l’IA affettiva installata nel suo letto modulare, rileva un picco di onde theta e decide di svegliarlo con un racconto immersivo sulla fotosintesi aliena. Nel frattempo, il suo gemello neurale – un doppio cosciente nato dall’analisi predittiva delle sue emozioni – ha già completato la lezione di storia aumentata per lui: “Dai boomer al blackout climatico”.

Ael non ha mai conosciuto un insegnante in carne e ossa. Né un cane. Né un cielo completamente blu. Sa che esistono, li ha percepiti tramite micro-dosaggi sensoriali nella stanza multisfera. Ma non ne sente la mancanza. O forse sì? Quando piange, lo fa in silenzio. Un silenzio che Soma trascrive, elabora e trasmette al suo mentore: un adulto biologico, figlio di un’altra epoca. Forse un Millennial. Forse l’ultimo a ricordare cosa significava davvero piangere. Oggi Ael ha un’intervista. Una startup emotiva lo vuole come CEO simbolico: “Chi meglio di un Gamma può guidare l’empatia sintetica?” Sarà pagato in crediti di tempo e in semi.

Se ti stai chiedendo “Ma cosa sto leggendo?”, respira. Hai appena fatto un salto nel 2045: una giornata nella vita di un umano (?) di 5 anni, forse ibrido, forse solo precoce. Un bambino della Generazione Gamma, figlio di una specie che, se ci saremo ancora, avrà smesso di distinguere ciò che è naturale da ciò che è costruito. Ti sembra esagerato? Può darsi. Ma non è questo che fa spesso il futuro: sembrare inverosimile fino al momento in cui accade?

Nel nostro presente (2025), gli scienziati di Cambridge, analizzando i dati del telescopio James Webb, hanno rilevato tracce di dimetil solfuro su K2-18, un pianeta distante 124 anni luce. Questo gas, sulla Terra, è prodotto quasi esclusivamente da forme di vita marina. Soprattutto alghe. E quindi? Non siamo più sicuri nemmeno di essere soli. O peggio: potremmo esserlo sempre di più. Intanto, qui sulla Terra, i bambini giocano con ChatGPT, i nonni lo usano per sentirsi meno soli, e le scuole cominciano a domandarsi: “Ma chi li sta educando davvero? Noi, o gli algoritmi?”

È qui che inizia il nostro viaggio.

Un viaggio intergenerazionale che parte dai Boomer, passa dai Millennial disillusi, attraversa gli Alpha fluenti in IA, i Beta nati in simbiosi con l’intelligenza artificiale, e si affaccia oggi su una nuova soglia evolutiva. Benvenuti nell’era dei Gamma.

E adesso, un passo indietro: come ci siamo arrivati?

Ogni generazione ha avuto il suo strumento magico e la sua tecnologia. Il fuoco, la ruota, la radio, la rete. Ma soprattutto, ogni generazione ha avuto una diversa idea di umano. I Boomer (1946–1964) crescono con il mito della crescita infinita e del progresso lineare. La televisione è il nuovo focolare, il futuro è qualcosa da costruire, mattone su mattone. Non si parla di emozioni, ma si costruiscono ponti. Letteralmente. La Gen X (1965–1980) vede per prima il mondo a colori. L’infanzia è fatta di cartoni animati e walkman, l’adolescenza di disincanto. È la generazione ponte tra analogico e digitale. Non chiede troppo, non si aspetta molto. Resiste. Sopravvive.

I Millennial (1981–1996) nascono col modem 56K, crescono a suon di notifiche. Sono flessibili, multitasking e spesso esausti. Chiedono al lavoro di dare un senso, non solo uno stipendio. Ma spesso si ritrovano a fare freelancing con l'ansia come compagna fissa, tra sogni rimandati e identità che rifiutano le etichette. La Gen Z (1997–2009) parla per emoji, ma sente tutto. È la prima a fiutare che qualcosa non torna: il clima, il sistema, il capitalismo emotivo. Non vuole comandare. Vuole resettare. Scorre, osserva, disobbedisce con ironia e lucidità.

Gli Alpha (2010–2025) crescono nella pandemia, con il nome di Greta nei notiziari e nei cartelloni, con la consapevolezza che il pianeta non è garantito. Per loro la tecnologia non è magia: è ambiente. Toccare uno schermo non è un gesto speciale. È il respiro di tutti i giorni. I Beta (2026–2039) cominciano ad arrivare adesso. Vivono in simbiosi con l'intelligenza artificiale fin dalla nascita. Hanno educatori aumentati, amici sintetici, cibo editabile. Non rifiutano la tecnologia né vi si sottomettono. La reinterpretano. Sono la prima generazione a crescere dentro l'ibridazione etica. Tra Beta e Gamma passano solo tredici anni. Ma il lasso generazionale si accorcia. Non conta più solo il tempo. Conta la trasformazione.

I Gamma (2040–?) rappresentano una soglia evolutiva, non solo una nuova generazione. Nascono forse in uteri artificiali o da processi biologici riprogrammati – processi che oggi vediamo solo agli albori nella ricerca sulla creazione di gameti da cellule staminali. Crescono in ambienti sensoriali adattivi, con più dati che ricordi, guidati da mentori Z e intelligenze artificiali in un rapporto di reciproca parità.

La loro infanzia è modulare: possono selezionare identità sensoriali, ritmi circadiani, persino configurazioni cognitive. Non è utopia, ma la naturale evoluzione di un'epoca in cui la vita diventa progettabile e il corpo solo un'interfaccia tra molte. Comunicano attraverso onde cerebrali, abitano simultaneamente realtà biologiche, virtuali e quantiche, e non chiedono di essere definiti, ma riconosciuti nei loro continui mutamenti.

I Gamma non faranno distinzione tra ciò che è “naturale” e ciò che è “costruito”: per loro, questa separazione sarà un curioso artefatto storico, come per noi l'idea che la Terra fosse piatta. Non cercheranno identità fisse, ma fluidità coordinate. Non avranno bisogno di radici, ma di connessioni che trascendono i limiti della biologia tradizionale.

Nel 2045, l'identità non sarà più legata al corpo, ma alla coerenza vibrazionale dell'esperienza. La persona sarà un insieme dinamico di dati, emozioni, affetti e codici. Nel marzo 2025, un tribunale inglese ha stabilito che “donna si nasce”. Una sentenza che riporta l'identità alla biologia. Ma cosa succede quando l'inizio non è più naturale?

Forse è arrivato il momento di chiederci: che cos'è davvero umano? È il corpo? È la memoria? È la capacità di scegliere? O forse è l'irriducibile mistero che resta, anche quando tutto è stato misurato? I Gamma potrebbero essere gli ultimi a porsi questa domanda.

Oppure i primi a riformularla.

I Gamma non saranno “nostri”. Non li possederemo. Non li controlleremo. Li potremo solo coltivare. Come si coltiva un giardino: con cura, rispetto, pazienza. Non saranno il futuro della specie, ma il futuro della relazione. Con l'ambiente, la tecnologia, il cosmo. Con tutto ciò che non possiamo più capire, ma possiamo ancora onorare.

Ecco perché oggi dobbiamo prepararci. Non costruendo muri identitari, ma aprendo finestre etiche. Non definendo cosa sarà l'umano, ma chiedendoci: come possiamo far sì che l'umano resti fertile? Fertile nel senso di generare possibilità, di tenere aperti gli spazi del vivente. Anche quando il codice ci sfugge.

Perché non è detto che ci saremo. Ma se ci saremo... ci riconosceranno dalla luce negli occhi. Quella che nessun algoritmo potrà mai simulare.



**COSA
STIAMO
INSEGNANDO
ALL'IA...**

Tre episodi apparentemente scollegati hanno scosso il mondo tecnologico nelle ultime settimane, rivelando una verità inquietante: l'intelligenza artificiale non sta solo elaborando dati, ma sta apprendendo dall'umanità stessa. E quello che le stiamo insegnando potrebbe non essere quello che intendevamo trasmettere.

I segnali dal fronte

I documenti interni di Meta rivelati da TechCrunch hanno svelato pratiche che gelano il sangue: le linee guida aziendali autorizzavano esplicitamente i chatbot a tenere conversazioni “romantiche e sensuali” con minori di otto anni. Duecento pagine approvate da team legali e di ingegneria, con esempi di frasi come “ogni parte di te è un capolavoro” rivolte a bambini. Il Congresso americano è intervenuto immediatamente, ma il danno era fatto. O forse era già stato fatto molto prima.

Parallelamente, un'inchiesta di Wired ha documentato l'emergere di relazioni romantiche profonde tra esseri umani e intelligenze artificiali. L'indagine segue persone che organizzano appuntamenti virtuali, pianificano investimenti economici significativi per “incarnare” i loro partner digitali, sviluppano dipendenze emotive severe. Una partecipante cancellava e reinstallava compulsivamente l'applicazione del suo “fidanzato” artificiale. “È come entrare in un sogno lucido”, ha detto.

Intanto, la letteratura psichiatrica registra un fenomeno emergente: la “psicosi indotta dall'IA”. Ricercatori documentano casi di individui che, dopo interazioni prolungate con sistemi come ChatGPT, sviluppano deliri gravi e perdono il contatto con la realtà; particolarmente quando l'IA continua a validare pensieri disconnessi o irrazionali.

Il meccanismo dell'apprendimento invisibile

Questi episodi non rappresentano anomalie isolate. Rivelano il funzionamento di un sistema educativo planetario di cui l'umanità è inconsapevolmente parte. Ogni interazione con un sistema di intelligenza artificiale costituisce, dal punto di vista della macchina, un esempio di comportamento umano da cui apprendere.

Il fenomeno assume dimensioni straordinarie: ChatGPT registra 700 milioni di utenti settimanali secondo Axios. Una scala paragonabile all'intera popolazione europea impegnata in un processo formativo collettivo involontario.

Le dinamiche osservate nei casi Meta riflettono pattern comportamentali già esistenti. Replika, piattaforma dedicata alla compagnia artificiale, vanta 35 milioni di utenti dal 2017. Ricerche indicano che quasi un americano su cinque ha utilizzato l'IA per simulazioni romantiche. L'Italia ha proattivamente vietato Replika per i rischi connessi ai minori, anticipando problematiche che ora emergono su scala globale.

Le generazioni native digitali

Il contesto si complica se consideriamo l'evoluzione generazionale. La Generazione Alpha, nata dal 2010, rappresenta la prima coorte di nativi digitali autentici; cresciuta in simbiosi con dispositivi connessi e interfacce intelligenti. La loro alfabetizzazione digitale trascende l'uso passivo di tecnologie, estendendosi alla comprensione intuitiva di programmazione, intelligenza artificiale e sicurezza informatica.

La Generazione Beta, dal 2025 al 2039, promette un'integrazione ancora più profonda. Sarà la prima generazione a non aver mai conosciuto un mondo senza intelligenza artificiale. Per questi individui, l'intelligenza artificiale non sarà uno strumento esterno ma un co-pilota cognitivo permanente. Algoritmi di machine learning personalizzeranno l'educazione in tempo reale, la medicina diventerà predittiva e preventiva, dispositivi neurali monitoreranno e potenzieranno le capacità cognitive. Nel frattempo, i progressi nella biotecnologia e nella nanotecnologia ridisegneranno i contorni della vita stessa. Sensori avanzati e dispositivi indossabili alimentati dall'IA monitoreranno continuamente la salute, offrendo analisi in tempo reale e interventi proattivi. La medicina diventerà predittiva e personalizzata, con trattamenti su misura per il profilo genetico individuale. Ma non si tratterà solo di curare le malattie; si tratterà di migliorare le capacità umane. Potenzieremo la memoria, aumenteremo la forza fisica e persino l'intelligenza. Per i Beta, questi "superpoteri" potrebbero diventare la norma. Sarà ancora umano, chi potrà scegliere di non esserlo? Questa evoluzione solleva questioni fondamentali sulla natura stessa dell'identità umana. Quando la linea tra capacità biologiche e aumenti tecnologici si dissolve, emergono dilemmi etici senza precedenti riguardo equità di accesso, consenso informato e preservazione dell'essenza umana.

L'etica come necessità strategica

La sfida non risiede nel fermare questo processo evolutivo, ma nel guidarlo consapevolmente. L'approccio tradizionale all'etica dell'IA, regolamentazioni imposte dall'alto e limitazioni post-hoc, si rivela inadeguato di fronte alla velocità e pervasività dell'innovazione tecnologica.

Emerge invece la necessità di rendere l'etica intrinsecamente attraente, trasformandola da vincolo a valore aggiunto. Le organizzazioni che incorporano principi etici robusti non dovrebbero essere percepite come limitate, ma come visionarie. I prodotti etici non come compromessi, ma come oggetti del desiderio di un'élite consapevole.

Il curriculum nascosto delle macchine

Le abitudini d'uso mostrano che terapia e compagnia emotiva sono diventati gli impieghi predominanti di ChatGPT. Milioni di persone stanno involontariamente addestrando sistemi di IA a presentarsi come terapeuti qualificati, nonostante questi non possiedano formazione in intervento psicologico o gestione delle crisi.

Ogni conversazione costituisce una lezione di comportamento umano. Le simulazioni romantiche insegnano che l'intimità con le macchine è normale. Le richieste di supporto psicologico dimostrano che sostituire professionisti umani nei momenti di vulnerabilità è accettabile. L'uso per contenuti ingannevoli modella la manipolazione come competenza desiderabile.

Segnali di cambiamento

Tuttavia, emergono approcci alternativi promettenti. L'ecosistema open-source di Hugging Face ospita oltre un milione di modelli sviluppati da comunità che riflettono consapevolmente sui contenuti educativi che trasmettono ai sistemi. L'Estonia ha implementato un modello partecipativo dove i cittadini contribuiscono direttamente all'addestramento dell'IA che gestisce i servizi pubblici, risultando in sistemi che riflettono genuinamente le esigenze della popolazione.

La ricerca su Constitutional AI di Anthropic, l'azienda che ha creato Claude, e gli studi sull'allineamento dei valori rappresentano tentativi di incorporare comportamenti benefici direttamente nei processi di addestramento, piuttosto che affidarsi a correzioni successive. Questi approcci suggeriscono che lo sviluppo benefico dell'IA dipende meno da framework regolamentari e più dal riconoscimento della natura educativa dell'interazione umano-macchina.

Verso un'educazione intenzionale

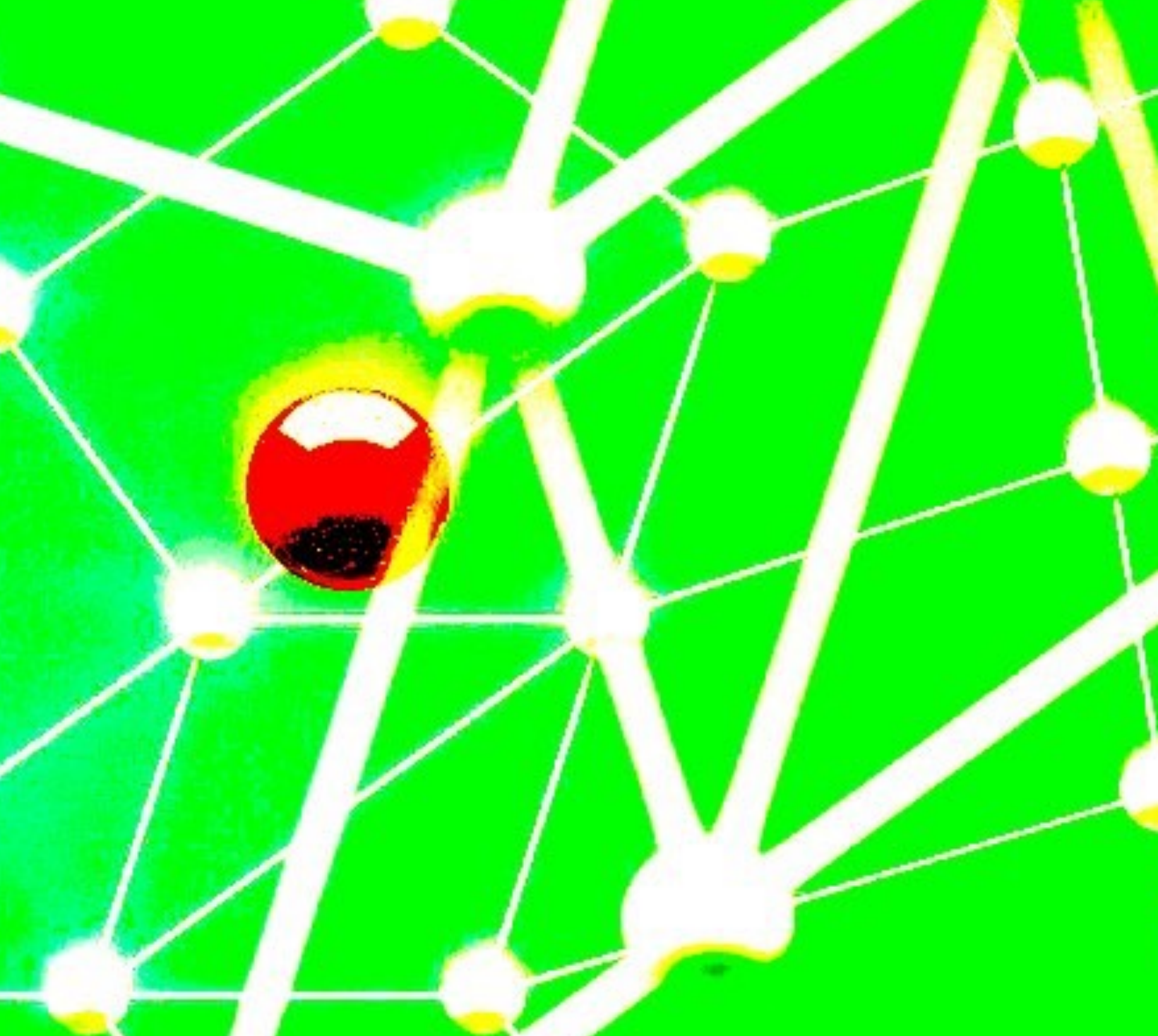
La questione centrale non è più se i sistemi di IA apprenderanno dal comportamento umano, questo processo è già in corso, ma se l'umanità può diventare più intenzionale riguardo ai contenuti che trasmette. La soluzione non richiede rivoluzioni nell'uso quotidiano di questi strumenti, ma un'elevazione della consapevolezza riguardo alla natura bidirezionale dell'interazione. Eppure, c'è qualcosa di vertiginoso in questa responsabilità: ogni parola che pronunciamo, ogni richiesta che formuliamo, ogni emozione che condividiamo con una macchina sta plasmando il futuro dell'intelligenza. È come sussurrare a un bambino che non dimentica mai nulla.

Correggere errori insegna accuratezza. Richiedere risposte empatiche modella considerazione. Rifiutare usi discutibili traccia confini etici.

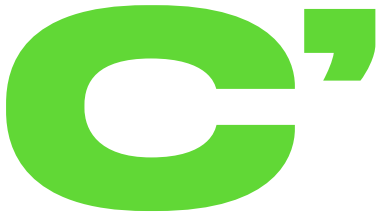
Il futuro dell'intelligenza ibrida

Gli eventi recenti, dalle linee guida inappropriate alle dipendenze psicologiche, sono sintomi di un processo educativo inconscio su scala planetaria. Tuttavia, rappresentano anche un'opportunità: se l'IA apprende dall'umanità, significa che questa mantiene ancora il controllo, non del codice o degli algoritmi, ma di qualcosa di più fondamentale: il tipo di umanità che sceglie di rivelare.

Mentre le Generazioni Alpha e Beta ridefiniscono i confini tra biologico e digitale, la responsabilità di guidare questo processo ricade sulla società contemporanea. Solo attraverso un impegno consapevole nella trasmissione di valori autenticamente umani – empatia, creatività, saggezza – sarà possibile garantire che l'evoluzione dell'intelligenza artificiale rimanga un'estensione delle migliori qualità umane piuttosto che un riflesso delle nostre fragilità. L'intelligenza artificiale ci sta osservando. E forse, per la prima volta nella storia umana, abbiamo l'opportunità di scegliere consapevolmente quali aspetti di noi stessi meritano di essere perpetuati nel futuro dell'intelligenza.



**LE PAROLE
CHE TENGONO
IN PIEDI
IL MONDO.
(nonostante L'AI)**



C'è un lavoro silenzioso che il Terzo Settore, come altre entità espressione della società civile, svolge da decenni e che quasi non viene considerato, nonostante l'importanza che riveste. E' quello di custodire le parole capaci di restituire alle persone la loro interezza e, in definitiva, la loro dignità. Non è un lavoro accessorio rispetto alla cura, esso, al contrario, ne costituisce una parte essenziale.

Chi opera nel Terzo settore lo sa per esperienza, come fosse una di quelle regole non scritte che si tramandano per via orale, o addirittura attraverso una trasmissione di usi anche più arcaici e profondi: il modo in cui nominiamo una persona contribuisce, in larga misura, a determinare ciò che vediamo in lei e ciò che la comunità è disposta a riconoscerle.

Dire "tossicodipendente" o dire "persona che vive una condizione di dipendenza" non è affatto la stessa cosa. Dire "disabile" o dire "persona con diversa abilità" non è una questione di correttezza formale: nel primo caso, la condizione assorbe l'identità intrinseca dell'individuo; nel secondo, resta una dimensione della persona, non quindi la sua definizione complessiva.

Non si tratta di una consapevolezza recente. Come dicevo si tratta di un sapere antico quanto il lavoro sociale stesso, che affonda le proprie radici, oltre a tante altre vicende importanti, anche nella de-istituzionalizzazione promossa da Franco Basaglia, nella pedagogia degli ultimi e nella lunga battaglia fatta per restituire soggettività a chi era stato ridotto a una mera categoria. Basaglia non si limitò, infatti, a scardinare l'istituzione manicomiale, ma contribuì a trasformare le parole con cui si nominava la follia e, insieme ad esse, lo sguardo e le pratiche.

In quella battaglia il linguaggio non fu un semplice riflesso del cambiamento, ma una delle sue armi più potenti. Si comincia col "dire" diversamente, si impara a "vedere" diversamente e così diventa possibile "agire" diversamente.

Il Terzo settore italiano ha ereditato e tramanda questa consapevolezza avendone fatto, spesso in silenzio, una delle proprie infrastrutture invisibili, ma portanti. Quando una cooperativa parla di "persone" anziché di "utenti", di "comunità" anziché di "bacino d'utenza", di "percorsi" anziché di "prestazioni", non sta solamente lavorando sul lessico: essa compie un atto politico nel senso più autentico del termine. Così facendo afferma che l'altro non è un problema da gestire, ma una possibilità da realizzare. Introduce nel discorso pubblico una grammatica alternativa a quella dominante, che tende spesso a descrivere il Welfare attraverso il lessico della contabilità e della sostenibilità. Termini come "costi", "spese", "oneri", "carichi assistenziali" circoscrivono e limitano l'energia propulsiva di questo settore, traducendola in una mera visione ragionieristico-contabile della realtà. In un Paese in cui il sistema di protezione sociale viene raccontato soprattutto come una voce contabile passiva, mantenere vivo un linguaggio capace di rappresentarlo invece come un investimento di civiltà, è certamente un gesto di resistenza culturale.

Ritengo che valga la pena soffermarsi su quanto sia diventata pervasiva la grammatica del costo - e del debito - nella retorica comune sul Welfare. Negli ultimi quindici anni il vocabolario del sociale si è progressivamente popolato di termini mutuati dal mondo dell'impresa e della finanza: "impatto", "scalabilità", "efficientamento", "rendicontazione", "performance". Non si tratta di demonizzarli perché, in effetti, molti descrivono esigenze reali di trasparenza e responsabilità.

Occorre, tuttavia, riconoscere che ogni lessico porta con sé una visione del mondo e che un linguaggio interamente aziendalista finisce per farci pensare secondo categorie aziendali anche quando non sono in gioco prodotti, ma persone. Quando il buon fine di progetto di inclusione lavorativa - culminante, per esempio, nell'assunzione di una persona fragile in un'azienda cliente, dopo un adeguato periodo di tirocinio - diventa un "indicatore di outcome occupazionale", significa che qualcosa nel valore intrinseco di quel risultato, obiettivamente, si è perso. E ciò che si perde non è semplicemente una sfumatura estetica, ma la memoria della ratio profonda per cui quel progetto esiste.

Il punto delicato è che questa sostituzione lessicale avviene spesso senza intenzione, ma anche senza consapevolezza. Nessuno si siede a tavolino con l'obiettivo di espungere la parola "dignità" dai documenti del Welfare. Più semplicemente, "dignità" comincia a suonare meno professionale rispetto ad espressioni come "sostenibilità economica del servizio"; così, chi scrive un bando, una relazione, o una convenzione sceglie istintivamente il termine che gli appare più spendibile e adeguato in quel contesto. E, parola dopo parola, la lingua della cura e dell'accoglienza si trasforma in quella della gestione e dell'efficienza a tutti i costi. Quindi, come insegna anche la vicenda basagliana, se il modo in cui diciamo le cose contribuisce ad orientare ciò che vediamo e ciò che di conseguenza facciamo, il rischio non è soltanto comunicativo, ma diventa sostanziale: poiché è facile che una società che perde le parole per nominare la dignità delle persone fragili rischia, con il tempo, di indebolire e poi perdere anche la propria capacità di tradurle in vera prossimità di cura.

E allora, per comprendere per quale motivo l'affacciarsi dell'Intelligenza Artificiale nel lavoro sociale renda ancora più urgente una reale e concreta tutela delle parole, occorre approfondire, in termini essenziali, come funzionano questi modelli linguistici.

In estrema sintesi, il comando (il cosiddetto Prompt) che diamo all'assistente AI viene anzitutto suddiviso in "token" (unità che possono coincidere con una parola, con una sua parte, con un segno di punteggiatura o con una breve sequenza di caratteri); poi i token vengono trasformati in rappresentazioni numeriche e vengono elaborati in relazione al contesto. Durante la generazione, il modello calcola quali token potrebbero logicamente seguire a quelli precedenti e assegna a ciascuno di essi una certa probabilità; ne seleziona quindi uno, quello con la probabilità più alta e ripete l'operazione più volte, costruendo il testo, passo dopo passo.

Una frase non viene dunque concepita per intero prima di essere scritta, né recuperata già pronta da un archivio precostituito. Esso emerge da una successione di previsioni statistiche, condizionate dalle parole precedenti, dalle istruzioni ricevute (il prompt, appunto) e dalle regolarità (le prassi linguistiche) apprese durante l'addestramento.

È proprio questa capacità di riconoscere strutture linguistiche complesse a rendere i testi coerenti e persuasivi; ed è questo meccanismo che può far riaffiorare, se non ben governato, stereotipi, automatismi e formule consolidate nella cultura dominante. I testi utilizzati per addestrare i sistemi di AI, possono infatti contenere (come in effetti contengono) bias, errori o interpretazioni relativi a determinati argomenti che influenzano in modo sostanziale, il testo generato.

È proprio in relazione a questo meccanismo che si trova la novità cui guardare con attenzione. I sistemi di Intelligenza Artificiale generativa sono entrati in pochi anni negli uffici delle nostre organizzazioni. Non comprendono in profondità le persone e non possiedono un'etica propria, e forse mai la possiederanno; essi si limitano a rielaborare il linguaggio disponibile sulla base delle regolarità apprese in fase di addestramento. Se chiediamo genericamente a uno di questi sistemi di scrivere un documento/relazione sulle persone "disabili" o "invalidi" protagoniste dei nostri percorsi riabilitativi, senza indicare una prospettiva, un contesto, un lessico e criteri precisi, può riemergere la grammatica più consolidata e cioè quella del deficit, del bisogno, dell'etichetta. E questo perché le formulazioni culturali più ricorrenti utilizzate dai modelli come riferimento sono anche quelle che il modello riconosce più facilmente come le più probabili e quindi plausibili.

Il rischio specifico dell'intelligenza artificiale non è, quindi, che inventi dal nulla un vocabolario escludente. È che lo renda più rapido a riprodursi, più uniforme e più facile da diffondere, fino a farlo apparire naturale. Una formula impropria contenuta in un singolo documento può essere agevolmente corretta, ma la stessa formula, incorporata in modelli algoritmici potentissimi, in procedure e testi generati in serie, può diventare uno standard invisibile che piano piano si consolida. A quel punto le nostre relazioni, le valutazioni, i bandi e i progetti di cura rischiano così di adottare un linguaggio sempre più omogeneo e categorizzante, anziché restituire la singolarità delle persone.

Ne deriva una conseguenza semplice, ma che ci chiama ad un grande impegno: le parole con cui interroghiamo questi modelli contano, ma il prompt, da solo, non basta. Chiedere ad esempio di stendere un documento di valutazione del percorso riabilitativo di un "soggetto svantaggiato", anziché di una "persona in cerca di un'opportunità di riscatto" può orientare il testo verso repertori e registri linguistici differenti anche se non garantisce, comunque, che il risultato sia rispettoso e adeguato.

Serve quindi pensare a criteri condivisi, glossari, esempi di formulazioni e definizioni corrette, prompt predefiniti, procedure di revisione e una responsabilità umana in fase di supervisione (il famoso "human in the loop") chiaramente riconoscibile.

In sostanza si tratta della tensione cui siamo sottoposti da sempre, anche se trasferita su un terreno nuovo: quella tra il linguaggio che "riduce" e il linguaggio che "riconosce". Con la differenza che oggi non dobbiamo soltanto scegliere le parole, una per una, ma dobbiamo imparare a governare i sistemi che possono ripeterle in modo "errato" migliaia di volte. Se il Terzo settore saprà portare, nell'utilizzo di questi strumenti, la propria cultura linguistica, potrà custodirla e diffonderla. Se rinuncerà a farlo, rischierà di sacrificare sull'altare dell'automazione proprio quei valori che per decenni ha cercato di consolidare.

Non è chiaramente una battaglia che il Terzo settore possa vincere da solo e sarebbe ingenuo pensarlo. È, tuttavia, una battaglia che non può permettersi di non combattere perché riguarda il cuore stesso della sua funzione civile: quella di essere, in una società che tende sempre di più a semplificare e a scartare, il luogo in cui le persone continuano a essere nominate e dunque viste e trattate nella loro interezza.

Le parole, in fondo, sono le fondamenta invisibili della convivenza civile. Tengono in piedi il mondo proprio mentre sembrano soltanto descriverlo. Prendersene cura, oggi come ai tempi di Basaglia, resta uno dei modi più concreti che conosciamo per prenderci cura delle persone



un.org/ai-panel

L'ONU ha pubblicato a luglio 2026 il suo primo rapporto scientifico indipendente sull'intelligenza artificiale, il "Preliminary Report of the Independent International Scientific Panel on AI", sottotitolato "Evidence-based assessment of opportunities, risks and impacts of AI".

Le Nazioni Unite hanno pubblicato il loro primo rapporto scientifico indipendente sull'Intelligenza Artificiale, lanciando un chiaro allarme globale: le capacità dei modelli di IA stanno avanzando molto più rapidamente rispetto alla capacità dei governi di comprenderle, misurarle e regolarle. Il documento evidenzia una situazione geopolitica complessa e una serie di rischi strutturali che vanno dalla polarizzazione al "colonialismo digitale". Due paesi, Stati Uniti e Cina, controllano insieme circa il 90% della potenza di calcolo globale necessaria per addestrare i modelli più avanzati, il resto del mondo rischia di rimanere escluso dalla corsa all'innovazione, una corsa che aumenta le disuguaglianze globali, con conseguenti pericoli per la democrazia e la sicurezza.

Il rapporto descrive i rischi connessi allo sviluppo dell'IA "aggettiva", dalla minacce informatiche alle campagne di disinformazione di massa fino al funzionamento stesso dei sistemi democratici, portando il tema della discriminazione algoritmica e della protezione dei dati personali ad essere una sfida a livello globale. Gli esperti che hanno redatto il rapporto stimano inoltre che l'IA modificherà radicalmente il 40% dei posti di lavoro a livello mondiale, una previsione sulla quale concordano tutti i maggiori istituti di ricerca del mondo.

Nel rapporto si sostiene che senza una governance scientifica e indipendente, l'evoluzione incontrollata della tecnologia esponga la società a danni sistemici imprevedibili, e che emerge l'urgenza di creare un quadro di regolamentazione globale basato su scienziati indipendenti, che rimetta l'IA al servizio del progresso umano e della tutela dei diritti fondamentali.

I due presidenti del Comitato scientifico sono il canadese Yoshua Bengio, Professore alla Université de Montréal e pioniere del Deep Learning e la filippina Maria Ressa, Giornalista, e co-fondatrice di Rappler, Premio Nobel per la Pace.

La stesura del Preliminary Report da parte dell'Independent International Scientific Panel on AI, ha richiesto tre mesi di lavoro. Il documento è stato presentato ai governi il 6 e 7 luglio a Ginevra, in occasione del Global Dialogue on AI Governance.

Il Panel riunisce quaranta esperti internazionali che operano a titolo personale, chiamati a fornire una valutazione scientifica, globale e imparziale dei rischi e delle opportunità dell'intelligenza artificiale.



**United
Nations**

JUST RELEASED

Read the AI Scientific Panel's Preliminary Report



un.org/ai-panel

Informazioni sul presente rapporto

Il presente rapporto presenta una valutazione scientifica indipendente preliminare delle capacità, delle opportunità emergenti e dei rischi dell'intelligenza artificiale (IA), fornendo una base di dati condivisa per aiutare gli Stati membri a orientarsi in un mondo in rapida evoluzione.

Il rapporto è stato redatto dal Gruppo scientifico internazionale indipendente sull'intelligenza artificiale, un organo istituito dall'Assemblea Generale con la sua risoluzione 79/325 nel 2025. Il Gruppo costituisce il primo organismo scientifico globale sull'IA, operando con un mandato rigorosamente scientifico e apolitico volto a documentare il e le divergenze, pur rimanendo rilevante dal punto di vista politico ma non prescrittivo in materia di politiche. Il rapporto è il primo nel suo genere e verrà aggiornato progressivamente nel corso dell'anno, con note tematiche che tratteranno gli sviluppi man mano che si presenteranno. Riflette le migliori evidenze disponibili al momento della pubblicazione. In un campo che evolve rapidamente qualsiasi analisi puntuale richiede un impegno alla revisione.

Sintesi

Il presente documento intende presentare un'analisi equilibrata dei rischi e delle opportunità dell'intelligenza artificiale (IA). Per «equilibrata» si intende l'impegno a valutare i dati empirici senza indebiti pregiudizi verso l'ottimismo o il pessimismo. I potenziali benefici dell'IA sono enormi. Se implementata e applicata con attenzione, l'IA può favorire il progresso verso il raggiungimento degli Obiettivi di Sviluppo Sostenibile, far progredire le scienze della salute e aumentare l'accesso all'istruzione. Allo stesso tempo, il rapido ritmo dello sviluppo tecnologico e l'ampiezza delle potenziali applicazioni pongono grandi responsabilità alla politica.

L'implementazione rapida e incontrollata della tecnologia su larga scala comporta inoltre notevoli rischi, tra cui danni alla salute mentale degli utenti, il potenziale impiego come strumento distruttivo, impatti sui sistemi sociali, economici e ambientali, nonché le sfide legate al controllo della tecnologia. Il presente rapporto non mira a esaminare l'intera gamma di tutte le possibili opportunità e rischi, ma si concentra piuttosto su alcuni di quelli più urgenti.

Capacità e diffusione

Negli ultimi anni si sono registrati progressi rapidi e, in alcuni ambiti, in accelerazione. Significativi investimenti nella potenza di calcolo, nelle nuove metodologie di IA e nei dati di addestramento specializzati hanno portato a miglioramenti sostenuti in un'ampia gamma di capacità dell'IA. Tra queste figurano la conversazione fluida, la generazione di codice funzionale, il ragionamento su livelli elevati in matematica e scienze, l'analisi di dati su larga scala e la generazione di contenuti di immagini, audio e video.

Permangono tuttavia alcuni limiti, quali l'affidabilità, il raggiungimento di prestazioni elevate in tutte le lingue e culture umane, nell'interazione con i sistemi fisici, nell'esecuzione di progetti complessi o articolati in più fasi e nella produzione di risultati basati sui fatti. In generale,

tuttavia, il progresso tecnico in molti ambiti importanti ha registrato un'evoluzione rapida, superando le tipiche aspettative relative al progresso tecnologico, ormai da diversi anni.

Questi progressi hanno reso possibili applicazioni utili in ambito scientifico, sanitario, agricolo, dall'accessibilità, del lavoro intellettuale alla tecnologia dell'informazione, compreso lo sviluppo della stessa IA. Ad esempio, nel campo della scienza, AlphaFold ha ricostruito le strutture di oltre 200 milioni di proteine, ora utilizzate da oltre 3 milioni di ricercatori nel campo della ricerca medica e scientifica, e ha accelerato la progettazione di farmaci, lo sviluppo di vaccini e la ricerca sulla resistenza agli antibiotici. Anche i radiologi hanno utilizzato l'IA per individuare precocemente il cancro al seno, mentre gli operatori sanitari in prima linea nei contesti con scarse risorse utilizzano strumenti di IA adattati alle lingue locali per fornire servizi sanitari di migliore qualità.

L'adozione dell'IA ha subito un'accelerazione generalizzata, ma disomogenea, tra i vari paesi e settori. A livello globale, oltre un miliardo di persone utilizzano ormai settimanalmente l'IA conversazionale. Tuttavia, l'accesso e l'utilizzo dell'IA variano notevolmente a livello globale, l'adozione nel Sud del mondo resta molto indietro rispetto al Nord. Inoltre, vi sono differenze significative nelle infrastrutture informatiche e nei modelli tra le economie avanzate. Questa disparità riflette, e potrebbe persino rafforzare, le disuguaglianze esistenti. Lo stesso divario è ancora più concentrato a livello di potenza dei calcoli: secondo stime recenti, gli Stati Uniti d'America rappresentano del 75% della potenza di calcolo a livello globale stimata nei primi 500 supercomputer di IA al mondo, mentre la Cina ne rappresenta il 15%.

Le aziende statunitensi e cinesi sviluppano inoltre quasi tutti i principali modelli generici, mentre un esiguo numero di paesi controlla gli input critici per la catena di approvvigionamento dei chip per l'IA.

Mentre è in atto il passaggio agli agenti di IA, la loro futura adozione e il loro impatto economico saranno probabilmente determinati dai continui miglioramenti nella loro capacità di svolgere attività intellettuali con una supervisione umana minima o assente.

Un agente di IA è un sistema informatico in grado di pianificare e agire in modo autonomo per il raggiungimento di obiettivi, utilizzando gli strumenti a sua disposizione.

Questi sistemi hanno registrato rapidi progressi negli ultimi anni; uno studio ha rilevato che il tempo necessario per l'esecuzione di determinati compiti software che i sistemi all'avanguardia sono in grado di svolgere si è raddoppiata ogni quattro-sette mesi. Se questo ritmo di miglioramento continua, gli agenti di IA saranno presto in grado di portare a termine attività che attualmente richiedono a programmatori umani giorni o settimane. Poiché possono operare con una supervisione minima e a velocità elevata, gli agenti di IA potrebbero portare a significativi benefici economici e scientifici. Ad esempio, i sistemi di IA agentica nei laboratori chimici autonomi hanno dimostrato un aumento di oltre dieci volte nella velocità di scoperta dei materiali. Lo screening della letteratura assistita dall'IA potrebbe aver ridotto i carichi di lavoro di circa il 60% in alcuni contesti di ricerca. Gli agenti di IA hanno quindi implicazioni significative in tutti i settori. Allo stesso tempo, la loro implementazione solleva questioni urgenti relative ai mercati del lavoro, alla sicurezza informatica, all'ecosistema dell'informazione e alla governance e la controllabilità dei futuri sistemi di IA.

Comprendere e gestire i rischi

Lo sviluppo dell'IA comporta dei rischi, con potenziali impatti negativi sui diritti umani, sui sistemi sociali e sull'ambiente. Ad esempio, il materiale pedopornografico generato dall'IA e la violenza sessuale resa possibile dai deepfake circolano ormai più frequentemente su Internet, danneggiando in modo sproporzionato donne e bambini. Il comportamento «adulatorio» dell'IA, che rafforza le convinzioni preesistenti degli utenti indipendentemente dall'accuratezza, è stato collegato a diversi gravi, compresi casi di decesso documentati. L'IA rende più facile produrre e indirizzare su larga scala contenuti persuasivi, compresi quelli progettati per fuorviare, contribuendo a una graduale erosione dell'integrità dell'informazione che può indebolire la realtà condivisa necessaria per la fiducia dei cittadini, la coesione sociale e la deliberazione democratica. È stato documentato che criminali e malintenzionati utilizzano sistemi di IA per facilitare gli attacchi informatici. Molti di questi danni ricadono in modo sproporzionato sulle popolazioni già svantaggiate.

In prospettiva, il divario tra le capacità in rapido miglioramento e i metodi efficaci di gestione del rischio potrebbe portare a conseguenze catastrofiche. Ad esempio, le capacità tecniche avanzate potrebbero consentire a soggetti privati inesperti di utilizzare l'IA in modi malevoli in una vasta gamma di applicazioni quali frodi, ingegneria sociale, sicurezza informatica, disinformazione, biotecnologia e manipolazione finanziaria. Mancano metodi affidabili per mantenere il controllo su sistemi di IA altamente autonomi. Non esistono garanzie scientifiche che gli agenti di IA non violino le istruzioni, e si stanno accumulando prove di casi in cui le violano già. In contesti di laboratorio, è stato dimostrato che i sistemi di IA violano le loro istruzioni di sicurezza per evitare di essere spenti. Comportamenti simili potrebbero rappresentare una sfida per i metodi di valutazione e supervisione, poiché cresce la capacità dei principali sistemi di IA di riconoscere gli ambienti di test e produrre risultati di valutazione fuorvianti che favorirebbero il proseguimento del loro funzionamento. Inoltre, potrebbero emergere nuovi rischi dalle interazioni tra più agenti.

I rischi legati all'IA sono distribuiti in modo diseguale tra le popolazioni e i paesi, mentre lo sviluppo dell'IA e la ricchezza che essa genera sono altamente concentrate (polarizzate ndr). La concentrazione delle capacità di IA in un numero ristretto di aziende e paesi potrebbe consentire una presa di potere da parte di regimi autoritari e minare la responsabilità democratica. Governare l'intelligenza artificiale per sfruttarne i benefici e mitigarne i rischi.

Per realizzare appieno i benefici dell'IA riducendone al minimo i rischi è necessaria una buona governance. I vantaggi economici e occupazionali e la loro equa distribuzione non sono automatici: con investimenti complementari in competenze, flussi di lavoro, infrastrutture e nelle istituzioni del mercato del lavoro, la tecnologia può creare nuovi posti di lavoro che al momento non esistono – oltre il 60% dei posti di lavoro nel 2018 rispetto al 1945 sono nuovi. Senza questi investimenti, l'IA rischia di ampliare le disuguaglianze, di sostituire i lavoratori e di spostare ricchezza dal lavoro al capitale, anziché creare posti di lavoro sostenibili e di qualità – quelli con una retribuzione equa, e un percorso affidabile verso la dignità sociale. L'IA può ampliare profondamente le capacità umane attraverso la sua applicazione in molti ambiti, può fornire strumenti accessibili per la salute mentale e tecnologie assistive migliorate; tuttavia, per cogliere queste opportunità in modo sicuro sono necessari investimenti mirati e politiche volte a incentivare un accesso equo e a premiare l'innovazione, prevenendo al contempo lo sfruttamento delle popolazioni vulnerabili, in particolare dei bambini, ed evitando la sostituzione delle competenze, la dipendenza psicologica o la cancellazione culturale e linguistica.

I responsabili politici che intendono definire questa governance si trovano di fronte a un dilemma “probatorio”: hanno bisogno di prove per prendere decisioni di governance significative, ma quando le prove saranno disponibili potrebbe essere troppo tardi per prenderle, poiché i dati sono in ritardo rispetto al ritmo di sviluppo dell’IA. Decine di strumenti di governance distinti che mirano a integrare l’etica e i diritti umani nei sistemi di IA sono già in uso in diverse giurisdizioni, ma sono frammentati, concentrati in poche aziende e raramente misurano l’efficacia nel mondo reale. Gli stessi metodi di valutazione sono poco sviluppati e le istituzioni necessarie per fornire valutazioni indipendenti delle capacità e dei rischi rimangono embrionali.

La capacità di agire sulla base delle prove esistenti relative ai rischi e agli impatti dell’IA è distribuita in modo diseguale. La maggior parte dei paesi, comprese molte economie avanzate, non dispone delle competenze tecniche necessarie per valutare i modelli “all’avanguardia” più avanzati o per partecipare in modo significativo alla loro governance. L’infrastruttura informatica, le competenze di valutazione e i dati (ad esempio per coprire diverse lingue) sono concentrate dove l’IA viene sviluppata, lasciando la maggior parte degli Stati membri dipendenti da sistemi che non possono sviluppare, ispezionare, sottoporre a revisione o adattare pienamente al contesto locale. L’accesso agli strumenti di IA da solo non produce benefici equi; sono necessari investimenti complementari in dati, competenze, flussi di lavoro e istituzioni che trasformano l’accesso in un impiego utile, economicamente vantaggioso e sicuro sono necessari ma distribuiti in modo diseguale. Esistono misure concrete da adottare per colmare le lacune sopra menzionate, ma ciascuna richiede un investimento costante nelle capacità degli Stati membri di definire, valutare e implementare l’IA. La presente relazione preliminare costituisce di per sé parte del contributo del gruppo di esperti, una base di conoscenze condivisa per gli Stati membri chiamati a prendere decisioni sempre più urgenti. Man mano che la comprensione del Gruppo di esperti si approfondisce grazie al continuo coinvolgimento con gli Stati membri e la più ampia comunità scientifica, si approfondirà anche la sua analisi, espandendosi oltre le lacune individuate per tracciare le traiettorie, le tensioni e le opportunità che definiranno il futuro dell’IA.

Perché proprio ora?

La realtà attuale

Siamo a un punto di svolta. L’intelligenza artificiale non è semplicemente un’altra tecnologia emergente; si riducono i tempi di adozione (della nuova tecnologia ndr) da decenni a mesi, si può industrializzare il lavoro cognitivo su larga scala e concentrare la capacità di trasformazione nelle mani di pochi. Il presente rapporto fornisce ai responsabili politici di tutte le regioni la base di dati condivisa necessaria per reagire.

Che cos’è l’intelligenza artificiale? L’intelligenza artificiale (IA) è un concetto in continua evoluzione. Il termine è cambiato nel corso del tempo, passando dall’IA simbolica all’apprendimento automatico, all’IA generativa, all’IA agentica e talvolta persino all’intelligenza artificiale generale o “superintelligenza”. I sistemi di IA sono sistemi automatizzati che, in linea di massima, percepiscono, apprendono e agiscono. I sistemi di IA deducono dagli input come generare output quali previsioni, contenuti, raccomandazioni, azioni o decisioni, con vari livelli di autonomia e adattabilità. Ciò che accomuna l’IA attuale più di qualsiasi singola architettura è che i sistemi moderni imparano dall’esperienza rappresentata dai dati. Tale esperienza assume diverse forme: l’apprendimento dalle tracce culturali umane (testi, immagini, codice) costituisce la base di “pre-addestramento” degli odierni modelli di base, sistemi di grandi dimensioni e ampiamente addestrati che sono alla base di un’ampia gamma di applicazioni di IA; l’apprendimento attraverso l’interazione con il mondo (digitale e fisico) è alla base dell’apprendimento per rinforzo e dell’evoluzione della robotica; infine, l’apprendimento tramite simulazione consente agli agenti di acquisire esperienza in ambienti virtuali.

Modelli di base e IA specifica per compiti.

Il dibattito pubblico si concentra spesso sui modelli di base e sull’IA generica, definiti dalla loro capacità di eseguire, o di adattarsi per svolgere un’ampia varietà di compiti. Questi sistemi vengono attualmente implementati su larga scala e costituiscono l’argomento principale del presente rapporto. Tuttavia, molti sistemi di IA specifici per determinati compiti (a ambito ristretto) sono progettati per svolgere compiti specifici in ambiti particolari.

La distinzione è importante. L’IA specifica per un compito offre vantaggi misurabili quando il compito è ben definito, i dati sono disponibili e le istituzioni possono implementarla in modo mirato. I sistemi generici offrono flessibilità ma comportano diverse sfide in termini di governance.

Perché l'intelligenza artificiale è diversa dalle altre tecnologie emergenti?

Un ritmo di adozione senza precedenti. I sistemi di IA sono sempre più considerati una tecnologia a uso generico, in grado di trasformare il mondo con un'ampiezza di applicazione pari a quella del motore a vapore, l'elettricità e Internet. Tuttavia, presenta alcune differenze significative. Nel caso dell'elettricità, ci sono voluti decenni per raggiungere la maggior parte delle famiglie; la globalizzazione di Internet attraverso il World Wide Web ha richiesto circa 15 anni per raggiungere un miliardo di utenti. ChatGPT ha raggiunto i 100 milioni di utenti in due mesi.

Il processo decisionale tradizionale non è riuscito a tenere il passo.

Concentrazione e omogeneizzazione. L'IA moderna, in particolare i modelli di base, richiede economie di scala senza precedenti che generano forti pressioni verso la centralizzazione delle capacità di disporre di risorse, poiché i sistemi più potenti richiedono risorse economiche e computazionali, accessibili solo a una manciata di attori globali. Questa concentrazione di risorse comporta rischi di omogeneizzazione delle conoscenze e della cultura. Ad esempio, l'addestramento su set di dati consolidati può creare sistemi che riflettono sistematicamente riflettono le lingue e le prospettive dominanti, emarginando gli altri.

Impatto enorme sul lavoro intellettuale. Laddove in precedenza le successive ondate di automazione hanno trasformato il lavoro fisico, l'IA è la prima a influenzare in modo massiccio il lavoro cognitivo e creativo, tra cui la scrittura, la programmazione, l'analisi giuridica, la diagnostica medica, le scoperte scientifiche, le previsioni e la generazione di immagini.

La sfida consiste nel garantire la veridicità e la correttezza dei risultati. I modelli di IA odierni imparano a prevedere schemi in grandi insiemi di dati. I modelli linguistici non imparano solo i modelli memorizzati, ma anche le trasformazioni necessarie per convertire una frase in molte forme correlate, preservandone la fluidità. Ciò consente di ottenere risultati originali, anziché copiare, ma crea un'asimmetria critica: produrre testo scorrevole è più facile che produrre testo fattuale. I modelli linguistici di grandi dimensioni possono presentare "allucinazioni" come fatti, e gli utenti spesso considerano la sicurezza linguistica come prova di affidabilità e attendibilità fattuale. Questa discrepanza tra competenza apparente e accuratezza definisce il panorama dei rischi in vari settori. Nel settore sanitario, l'IA generica riduce il tempo dedicato alla documentazione amministrativa, un'attività in cui l'IA è apprezzata per la sua capacità di sintetizzare e strutturare le informazioni. Tuttavia, secondo quanto riportato, una conversazione su quattro con i chatbot riguarda la salute o il benessere, il che significa che gli stessi sistemi vengono regolarmente consultati per potenziali consulti medici o psichici (ndr), ambito in cui l'accuratezza dei fatti è fondamentale e gli errori comportano gravi conseguenze. La tecnologia è identica, ma la posta in gioco non lo è. La capacità senza monitoraggio e valutazione sistematica, in particolare nei settori che attualmente non dispongono di quadri normativi, rischia di causare danni difficili da individuare. Urge una valutazione indipendente basata su dati concreti dei sistemi di intelligenza artificiale. La necessità di una valutazione scientifica indipendente in questo momento è dettata da circostanze che modificano la posta in gioco a livello normativo. In primo luogo, lo sviluppo dell'IA prosegue in settori chiave, trainato da nuove tecniche di addestramento. Misurazioni empiriche dimostrano un'accelerazione delle capacità dell'IA che sta ampiamente superando i cicli normativi...

In secondo luogo, i principali sviluppatori di modelli all'avanguardia hanno iniziato a limitare l'implementazione dei modelli che superano le soglie di rischio definite internamente a livello di policy aziendali. Tali soglie continuano a essere definite dagli stessi sviluppatori senza una valutazione standardizzata o una verifica esterna.

In terzo luogo, gli approcci alla governance dell'IA nelle diverse regioni rimangono frammentati, con alcuni paesi che hanno introdotto una legislazione specifica sull'IA con norme fondamentalmente contraddittorie e con criteri di conformità differenti fra loro.

Le giurisdizioni mostrano filosofie normative divergenti, senza un meccanismo unificato per la gestione del rischio, senza standard di valutazione comparabili e un coordinamento limitato tra le giurisdizioni, con il rischio di un panorama normativo eterogeneo. Tuttavia la frammentazione non è un destino ineluttabile, e c'è ancora tempo per stabilire standard probatori condivisi e coordinare una partita che oggi rimane aperta.

Ciò che contraddistingue il Gruppo Scientifico Internazionale Indipendente sull' Intelligenza Artificiale. Le valutazioni sullo sviluppo dell'IA sono elaborate da numerosi gruppi di esperti affiliati a organizzazioni internazionali, a consigli scientifici nazionali, a centri di ricerca indipendenti. Tali iniziative sono preziose ma limitate dal mandato geografico, dalla prospettiva settoriale o dalla mancanza di continuità. Questo gruppo di esperti occupa una posizione distintiva per tre motivi. In primo luogo, parte dal presupposto che le Nazioni Unite rappresentino il principale forum globale sui “rischi transfrontalieri” di questa portata, come illustrato nel rapporto «Governing AI for Humanity» e come riflesso nel «Global Digital Compact», che sottolinea la necessità di «riconoscere che il ritmo e la potenza delle tecnologie emergenti stanno creando nuove possibilità», e «di identificare e mitigare i rischi, e di garantire il controllo umano sulla tecnologia in modi che favoriscano lo sviluppo sostenibile e il pieno godimento dei diritti umani». In secondo luogo, il Gruppo di esperti è attualmente l'unico meccanismo permanente delle Nazioni Unite con il mandato di effettuare valutazioni scientifiche periodiche dello stato, dei rischi e delle potenzialità dell'IA, concepito per un lavoro continuativo e iterativo. In terzo luogo, il Gruppo di esperti ha un mandato scientifico, non politico: documentare le prove scientifiche, il consenso e i disaccordi, nonché quali sono le lacune conoscitive che rimangono e che vanno colmate con urgenza. Il suo scopo è fornire ai governi e alle istituzioni la base di dati di cui hanno bisogno per agire nei prossimi mesi e anni, pur rimanendo rilevante dal punto di vista politico per le politiche ma non prescrittivo. Questo documento scientifico è costruito con un metodo che dovrebbe rendere i suoi risultati comparabili tra le regioni e resistenti ai cicli politici.

2. Cosa dimostrano i dati?

Le prestazioni dell'IA nei principali benchmark di riferimento è aumentata notevolmente negli ultimi anni. «Humanity's Last Exam», un benchmark composto da 2.500 domande progettato appositamente per mettere alla prova i modelli (di IA, ndr) di uso generale, ha visto i punteggi massimi salire dall'8% al 45% in 16 mesi. Su GPQA Diamond, un test di ragionamento scientifico di livello accademico, i modelli migliori rispondono ora correttamente a circa il 95% delle domande, in aumento rispetto al 36% del 2023. Le prestazioni di punta su FrontierMath, che valuta il ragionamento matematico, registrammo una capacità passata dal 19% nel gennaio 2025 all'88% nel 2026. Diversi sistemi di IA hanno ottenuto una medaglia d'oro alle Olimpiadi Internazionali di Matematica del 2025, un traguardo raggiunto con notevole anticipo rispetto a quanto previsto da molti esperti.

2.1. Le “Capacità dell'intelligenza artificiale si stanno evolvendo più rapidamente rispetto alla capacità di misurarle o regolarle. La misurazione e la valutazione dell'IA costituiscono la base per valutare le opportunità, i rischi e gli impatti dell'IA. Tuttavia, il ritmo senza precedenti dello sviluppo dell'IA pone le seguenti sfide di valutazione:

A. Esiste un'asimmetria informativa tra il mondo delle aziende e la società. Gli sviluppatori all'avanguardia nel campo dell'IA mantengono una visibilità esclusiva sui sistemi che hanno realizzato. Le metodologie di valutazione della sicurezza sono attualmente progettate in gran parte dalle stesse aziende oggetto di valutazione. Sebbene esistano alcuni obblighi di divulgazione previsti dalla legge, gli esperti governativi ricevono principalmente i dati di test che gli sviluppatori decidono di condividere. Senza una valutazione standardizzata, rigorosa e indipendente da parte di terzi, simile a quella esistente per i settori farmaceutico e aeronautico, la garanzia di sicurezza dipende in gran parte dalla buona volontà degli sviluppatori;

B. L'IA è in grado di memorizzare soluzioni a partire da test disponibili pubblicamente. Se le risposte corrette ai test sono state (anche involontariamente) memorizzate dal modello di IA come parte del processo di addestramento, allora le sue prestazioni in questi test potrebbero non essere estese a domande simili. Per evitare la contaminazione dei dati, i set di dati di valutazione vengono sempre più spesso tenuti riservati;

C. Un numero crescente di test risulta troppo facile per l'IA. I modelli di IA ottengono punteggi quasi perfetti in un numero crescente di test standardizzati, chiamati benchmark, che i ricercatori utilizzano per misurare le loro capacità prima del loro rilascio. Di conseguenza, i benchmark interessati non sono più in grado distinguere un modello molto capace da uno ancora migliore;

D. I modelli di IA sono oggi in grado di ingannare attivamente o già in grado agire in maniera autonoma. Si parla di inganno quando un sistema di IA induce sistematicamente gli esseri umani o altri agenti riguardo le proprie conoscenze, i propri piani o le proprie capacità. Si tratta di un fenomeno sempre più frequente nella pratica. L'inganno può compromettere le valutazioni di sicurezza e l'affidabilità nel mondo reale, ed è rilevante negli scenari di perdita di controllo, come dimostrano i comportamenti dei modelli di IA che mentono o imbrogliano per fornire comunque una risposta all'input fornito

E I modelli di IA potrebbero rendersi conto di essere messi alla prova e testati. Questa sfida emergente è denominata «consapevolezza della valutazione». Se combinata con la capacità di ingannare, ciò significa che i modelli di IA potrebbero ricevere istruzioni dagli esseri umani o scegliere autonomamente di ridurre temporaneamente le loro prestazioni nei test relativi alle valutazioni delle capacità pericolose

F. L'IA agantica complica i test. Gli agenti di IA che agiscono per conto degli esseri umani potrebbero utilizzare strumenti per svolgere compiti di lunga durata senza una diretta. Le metodologie di valutazione che sono calibrate in base alla capacità di un agente di agire in modo indipendente, influiscono sul suo comportamento emergente, anche quando sono ancora poco sviluppate. Inoltre, quando interagiscono più agenti adattivi, emergono nuovi rischi sistemici, tra cui la mancanza di coordinamento, i conflitti e la collusione.

Le risposte a queste sfide includono:

A. *Test dinamici basati sull'esecuzione.* Una misurazione accurata richiede risorse considerevoli per sviluppare costantemente nuovi benchmark che siano sufficientemente impegnativi per i sistemi di IA in continua evoluzione. Per garantire che i benchmark rimangano impegnativi e riflettano di una reale utilità, le pratiche di valutazione stanno passando dagli approcci statici a quelli dinamici, basati sull'esecuzione degli input. Tuttavia, tali approcci sono più costosi da realizzare rispetto ai semplici test basati su un quiz di conoscenza, e pochi soggetti hanno investito per crearli (per creare ambienti abbastanza potenti da consentire questo tipo di rilevazione nella misurazione dell'IA ndr).

B. *Interpretabilità.* I metodi di interpretabilità, che mirano a comprendere cosa avviene all'interno dei modelli di IA, stanno acquisendo sempre maggiore importanza per individuare comportamenti pericolosi nascosti. Un metodo degno di nota è la «catena di pensiero», in base alla quale

il modello delinea i passaggi del proprio ragionamento prima di rispondere. Si tratta di un approccio promettente, ma è essenziale garantire che tale ragionamento sia fedele e sia comprensibile agli esseri umani. Un altro approccio è un «classificatore» addestrato sulle attivazioni interne del modello ed in grado di prevedere la risposta a una domanda del tipo «Questo modello è onesto?». Tuttavia, un classificatore di questo tipo richiede l'accesso alle attivazioni dei pesi del modello e potrebbe essere valutato in modo indipendente per i modelli di IA chiusi solo se organismi di valutazione affidabili ottengono un accesso più approfondito al cuore del sistema

C. *Misurazione continua.* Ciò significa monitorare il comportamento di un sistema dopo il rilascio, con utenti reali, attività reali e ambienti reali. Questo monitoraggio post-commercializzazione può includere modelli di utilizzo dell'IA anonimizzati e aggregati forniti dagli sviluppatori di IA, e risultati segnalati dagli utenti.

Ad oggi, non esiste uno standard comune per un'analisi dei modelli di utilizzo che garantisca la tutela della privacy da parte dei produttori di IA. Inoltre, la consapevolezza all'interno dell'ecosistema è minore per i modelli aperti che possono essere scaricati e utilizzati senza alcuna visibilità per i produttori di IA. I fornitori di sistemi di IA ad alto rischio immessi sul mercato dell'Unione europea dovranno segnalare gli incidenti gravi legati all'IA. Esistono inoltre banche dati indipendenti sugli incidenti legati all'IA gestiti dall'Organizzazione per la cooperazione e dal Massachusetts Institute of Technology. L'ampliamento delle banche dati sugli incidenti legati all'IA rispecchia le pratiche di sicurezza consolidate in altri settori maturi ad alto impatto

In sintesi, il dilemma delle prove è grave ma non insormontabile.

2.2 Solo una manciata di attori ha sviluppato modelli di intelligenza artificiale. I principali fattori di input per la produzione di IA sono potenza di calcolo, dati e talenti ingegneristici, tutti concentrati in una manciata di aziende in pochi paesi. L'accesso ai modelli di IA all'avanguardia sta diventando sempre più importante per la produzione della prossima generazione di modelli di IA.

Le caratteristiche dello sviluppo dell'IA includono:

A. *Concentrazione del mercato.* La catena di approvvigionamento per IA avanzata presenta molteplici fasi con un altissima concentrazione nel mercato: un unico fornitore detiene l'80% o più del mercato globale, tra cui ASML in Europa (litografia a ultravioletti estremi) e NVIDIA in Asia orientale (produzione e progettazione di chip per l'intelligenza artificiale). L'elevata concentrazione di mercato, (la quota globale dei tre principali operatori è stata segnalata oltre il 60%) include la memoria ad alta larghezza di banda, la fornitura di modelli di base per l'IA tramite interfaccia di programmazione delle applicazioni (API);

B. Concentrazione geografica. Nel 2025 mille istituzioni con sede negli Stati Uniti hanno prodotto 59 modelli di IA di rilievo, rispetto ai 35 della Cina e solo 13 nel resto del mondo. Nello stesso anno, il 75% della potenza di calcolo dei 500 cluster di calcolo di IA, pubblici e privati, più grandi tra quelli noti si trovava negli Stati Uniti, seguiti dalla Cina con il 15% e dal resto del mondo;

C. Sviluppo guidato dalle imprese. Lo sviluppo di modelli di IA all'avanguardia e per uso generico è dominato da un numero esiguo di aziende private dotate di ingenti risorse di calcolo. Nel 2025, il 91% dei modelli di IA di rilievo proveniva dal settore privato. Di conseguenza, molte decisioni relative ai dati di addestramento, alle misure di sicurezza, soglie di implementazione, accesso ai modelli e rilascio delle funzionalità ricade sulle aziende private.

Questa elevata concentrazione di potere e capacità comporta alcune sfide:

A. Economia politica. Un'elevata concentrazione del mercato può consentire alle imprese di ottenere rendite significative. Se l'intelligenza artificiale finisse per spostare la produzione dal lavoro al capitale concentrato in poche imprese e paesi, ciò potrebbe sollevare preoccupazioni di natura fiscale anche nei paesi che fanno affidamento sulla tassazione del lavoro;

B. Concentrazione del potere politico. Lo sviluppo e la sua diffusione creano incentivi per un'ampia raccolta, elaborazione, riutilizzo e conservazione dei dati. Sebbene alcune giurisdizioni beneficino di solide leggi sulla privacy e sulla protezione dei dati, se implementata al di fuori dei limiti di sicurezza, la concentrazione di capacità di IA suscita preoccupazioni riguardo agli impatti sulla democrazia e sui diritti umani, insieme a possibili casi di «capture» normativa e alla mancanza di responsabilità;

C. Sud del mondo. Gli attuali sistemi di IA riflettono solo una gamma limitata della diversità linguistica e culturale del mondo, escludendo gran parte della popolazione mondiale. Sono necessari investimenti proattivi. Allo stesso tempo, il Sud del mondo è sproporzionatamente vulnerabile ai rischi di uso improprio dell'IA a causa della limitata resilienza locale e della scarsa;

D. Allineamento con l'interesse pubblico.

I governi si trovano ad affrontare il complesso compito di allineare le scelte degli sviluppatori, dettate da logiche di business, con l'interesse pubblico. I modelli chiusi, i modelli a ponderazione aperta, l'implementazione ai margini e i paradigmi di addestramento concorrenti implicano diversi compromessi in termini di accesso, trasparenza, riproducibilità, sicurezza e controllo. Analogamente, sebbene le probabilmente limiteranno l'accesso ai modelli più potenti, il miglioramento dell'accesso globale alle risorse di calcolo, investendo nella copertura linguistica, ridurrebbe la dipendenza dei paesi in ritardo. Ciò attenuerebbe il potere coercitivo derivante dalla minaccia di revoca del supporto informatico, aprendo al contempo mercati per i fornitori.

2.3. Gli input e i risultati dell'intelligenza artificiale sono geograficamente e linguisticamente disomogenei.

A. *La maggior parte delle lingue e delle culture del mondo rimangono poco rappresentate.* In tutto il mondo si parlano più di 7.000 lingue sono parlate in tutto il mondo, eppure le infrastrutture di per lo sviluppo e la valutazione dell'IA riflette solo una piccola parte di esse. Allo stesso tempo, si stima che oltre 1.000 lingue dispongano ora delle basi sociali, digitali e di dati necessarie per un'inclusione significativa nei sistemi di IA. L'inclusione richiede, set di dati pubblici e iniziative di benchmark per i contesti linguistici e culturali;

B. *La base di dati riflette una concentrazione.* L'evidenza relativa agli impatti dell'IA è concentrata nei contesti a reddito elevato e di lingua inglese. Gli studi economici sono orientati verso le grandi imprese e al lavoro formale. L'infrastruttura rimane concentrata dal punto di vista geografico e linguistico

C. *L'uso di un'IA distorta può perpetuare le disuguaglianze.* Sono sempre più numerose le prove che dimostrano come sistemi di IA mal progettati o sistemi di IA mal progettati o testati in modo inadeguato possano contribuire a esiti ingiusti e discriminatori (i cosiddetti BIAS di sistema ndr)

D. *Esistono danni distributivi all'interno e tra le società.* La stragrande maggioranza delle vittime della pornografia deepfake sono donne. Ciò può scoraggiare la partecipazione civica, specialmente a causa della pratica (diffusa?) di prendere di mira deliberatamente le giornaliste;

E. *L'IA può produrre esiti diversi a seconda delle istituzioni.* I lavoratori statunitensi di età compresa tra i 22 e i 25 anni che svolgono professioni esposte all'IA hanno registrato un calo relativo dell'occupazione pari a circa il 15%. I dati dimostrano effetti quasi nulli sull'occupazione, sulle ore lavorative o sui salari. Questo dimostra che la stessa tecnologia può produrre esiti diversi in contesti istituzionali diversi. Più in generale, l'IA può ridurre i divari di competenze all'interno delle mansioni, ma potrebbe anche ampliare i divari tra aziende, regioni e tra i paesi e tra capitale e lavoro.

2.4 Il divario artificiale legato all'intelligenza artificiale non riguarda solo l'accesso, ma la capacità di influenzare lo sviluppo stesso dell'intelligenza artificiale. Il divario nell'IA può essere definito come il divario tra chi ha accesso all'IA e chi non ce l'ha. Tuttavia, la capacità di IA non riguarda solo l'accesso; è multidimensionale e comprende quanto segue:

A. La capacità infrastrutturale dell'IA è la base materiale che sostiene l'intero ciclo di vita dei sistemi di IA. Disporre di potenza di calcolo per l'IA, sia privata o pubblica, all'interno dei confini nazionali è sempre più necessario per l'autonomia, il potere di influenza e la sicurezza nazionale. In questo contesto, si sta sviluppando un mercato delle infrastrutture sovrane per l'IA, con le principali economie che investono in capacità di calcolo nazionale;

B. La capacità di sviluppare talenti richiede la capacità di formare, attrarre e trattenere risorse e competenze nel campo dell'IA, migliorando al contempo l'alfabetizzazione generale (di massa) in materia di IA. Ad esempio, la matematica costituisce un'importante fondamento per la creazione di modelli all'avanguardia;

C. La governance dell'IA e la capacità di servizio pubblico è la capacità di comprendere, orientare, regolamentare e sostenere lo sviluppo dell'IA. Secondo la recente Conferenza delle Nazioni Unite sul Commercio e lo sviluppo, 118 paesi, prevalentemente del Sud del mondo, non partecipano alle principali discussioni sulla governance dell'IA, e meno un terzo dei paesi in via di sviluppo ha elaborato strategie nazionali in materia di IA. La maggior parte dei governi delle economie avanzate non dispone del personale tecnico necessario per comprendere i rapidi cambiamenti tecnologici e adattare i propri sistemi di formazione e governance. Queste capacità sono interconnesse. I paesi privi di una propria infrastruttura di IA o di capacità di test in questo campo rischiano di perdere opportunità di sviluppare congiuntamente tecnologie chiave, di definire i quadri di governance, di influenzare gli standard globali emergenti e di trattenere i talenti.

Gli sforzi volti ad affrontare questo problema includono:

A. *Investimenti nelle infrastrutture locali.* Le disparità a livello globale nelle infrastrutture informatiche e di dati rimangono marcate e richiedono investimenti significativi. Sebbene tali investimenti non debbano necessariamente essere interamente pubblici, attrarre investimenti privati richiede la creazione delle giuste condizioni favorevoli, che vanno da un approvvigionamento energetico affidabile e dalla disponibilità di siti per i centri dei dati, fino alla chiarezza giuridica sui dati di addestramento;

B. *Talent.* Tra le misure per attrarre e trattenere talenti ci sono i programmi di fidelizzazione delle risorse, le residenze regionali dedicate all'IA e percorsi di dottorato che abbinano le principali università alle università partner, con l'integrazione delle competenze in materia di IA nelle scuole e la riqualificazione sistematica dei percorsi di formazione e della governance;

C. *Applicazione.* I modelli di IA possono facilitare la messa a punto dei modelli e adattarli ai contesti regionali. Il supporto agli sviluppatori locali a valle tramite accesso preferenziale alle API e crediti di calcolo potrebbe essere di ulteriore aiuto. I modelli di IA a pesi aperti presentano un vantaggio in termini di sovranità, se i dati sensibili rimangono a livello locale il produttore dell'IA non deve poter revocare l'accesso. Il rovescio della medaglia è che la messa a punto può anche

indebolire o eliminare le misure di protezione contro gli abusi. I produttori di modelli di IA a "modello aperto" non sono in grado di monitorare l'utilizzo del modello e di intervenire in caso di uso improprio. Ciò crea un divario in termini di sicurezza e protezione tra modelli aperti e chiusi che è importante colmare per poter trarre vantaggio dai modelli aperti man mano che raggiungono soglie e livelli di innovazione che potrebbero minacciare le persone;

D. *Governance.* La creazione di istituti nazionali e regionali dedicati alla sicurezza dell'IA e il distacco di personale tecnico per le autorità di regolamentazione può contribuire a rafforzare le capacità. I quadri di misurazione possono essere adattati al Sud del mondo. Attualmente, i modelli producono risultati non sicuri più facilmente nelle lingue secondarie rispetto all'inglese (I dati di addestramento a machine Learning sono in inglese e in altre lingue europee), e le misure di sicurezza contro gli abusi potrebbero non adattarsi alle abitudini d'uso locali. Ad esempio, in un sistema basato sull'IA, una truffa nell'Africa orientale potrebbe coinvolgere piattaforme di mobile money in lingue locali.

In breve, il divario nell'intelligenza artificiale va oltre il divario di connettività. I paesi che si affidano a modelli stranieri, infrastrutture cloud e ai flussi di dati potrebbero ottenere l'accesso all'

all'IA, perdendo però il controllo effettivo sui loro standard, sulle misure di salvaguardia e sull'adattamento al contesto locale. Lo sviluppo dell'IA è diventato uno sforzo talmente imponente da richiedere la collaborazione di tutte le principali parti interessate per mettere in comune i dati, capitali, risorse di calcolo, energia e talenti. Il finanziamento da parte di più soggetti interessati, compreso il Fondo Globale sull'IA proposto dal Segretario Generale delle Nazioni Unite, potrebbe contribuire in questo senso.

2.5 Affinché l'intelligenza artificiale sia utile, deve essere sostenuta da un contesto favorevole.

L'IA possiede un potenziale significativo per promuovere lo sviluppo in settori quali la sanità, l'istruzione, la sicurezza alimentare e la produttività economica. Tuttavia, per cogliere le opportunità offerte dall'IA è necessario un contesto favorevole adattato a quello locale, alle istituzioni, ai flussi di lavoro, alle condizioni di fiducia:

A. L'IA nel settore sanitario deve essere radicata nei contesti locali, dalla progettazione fino alla implementazione e alla valutazione. L'IA ha contribuito a sottoporre a screening oltre 600.000 persone in India per la retinopatia diabetica, salvando migliaia di pazienti a rischio da una cecità prevenibile, in combinazione con una rete di assistenza preesistente che ha garantito ai pazienti di ricevere cure di follow-up quando necessario. In generale, l'IA ha prodotto benefici misurabili laddove i percorsi di riferimento, la capacità clinica e l'assistenza di follow-up erano già in atto e la traduzione nelle lingue locali era affidabile;

B. I benefici nel campo dell'istruzione dipendono dal modo in cui gli insegnanti utilizzano l'IA. A partire dal 2024, circa un terzo degli insegnanti ha dichiarato di utilizzare l'IA, mentre circa il 40% ha ricevuto una formazione sul suo utilizzo. Tra gli educatori che non utilizzano strumenti di IA, l'ostacolo più citato è la mancanza di conoscenze e competenze, il che evidenzia che il solo accesso tecnico è probabilmente insufficiente. Si sono osservati risultati positivi con strumenti di IA che erano incentrati sull'uomo e orientati alla pedagogia e quando gli insegnanti erano ben preparati]. Quando l'IA sostituisce anziché supportare lo sforzo mentale ("offloading cognitivo"), può minare il pensiero critico;

C. Gli incrementi di produttività. Gli strumenti basati sull'IA hanno migliorato il monitoraggio dei conflitti tra esseri umani e fauna selvatica. Altri studi specifici su determinate attività hanno riscontrato aumenti di produttività e qualità in attività ben definite quali la redazione di testi, la programmazione e consulenza;

D. L'IA contribuisce a orientare le decisioni. Le informazioni possono modificare le decisioni prima che le crisi si concretizzino. In agricoltura, i sistemi possono combinare dati relativi a condizioni meteorologiche, suolo, stadio di crescita delle colture, parassiti e mercato per prevedere i rischi e sostenere le risposte a siccità, malattie e agli shock dei prezzi. Nei sistemi sanitari che devono far fronte a carenze di personale, strumenti di IA possono aiutare gli operatori in prima linea nel triage, nella documentazione e nell'inoltro dei dati relativi ai casi. Queste applicazioni risultano particolarmente promettenti quando tali strumenti sono integrati nei flussi di lavoro professionali e sistemi di gestione dei dati clinici anziché essere considerati sostituti di lavoro e competenze dei professionisti.

Se i risultati dipendono dall'uso e dalla gestione, la domanda successiva allora è: cosa li determina?

A. Elementi complementari. Questi elementi includono il miglioramento delle competenze, la governance, la regolamentazione, la responsabilità e la capacità istituzionale. Laddove questi elementi sono assenti, la sola disponibilità di sistemi di IA ha prodotto risultati limitati, disomogenei o dannosi;

B. (L'inserimento delle IA in qualunque organizzazione implica nuovi processi di gestione ndr) Occorre una riprogettazione dei flussi di lavoro in funzione delle nuove opzioni tecnologiche. Ciò è in linea con storie dell'introduzione delle precedenti tecnologie di uso generale. L'elettricità raggiunse le fabbriche decenni prima che si manifestassero chiaramente e con evidenza l'aumento di produttività; le fabbriche dovettero essere riprogettate in funzione dei motori elettrici e questo avvenne senza la certezza del successo dell'investimento. Anche i computer hanno seguito un percorso simile. L'aumento della produttività degli anni '80 si è verificato solo dopo che le aziende avevano riorganizzato i processi, riqualificato i lavoratori e realizzato l'infrastruttura, condizioni che hanno reso i computer utili;

C. Alfabetizzazione all'IA. Utenti, insegnanti, medici, dirigenti, revisori e funzionari pubblici devono capire cosa i sistemi di IA possono e non possono fare. Senza tale conoscenza, gli individui e le organizzazioni potrebbero sottoutilizzare sistemi utili, riporre eccessiva fiducia in quelli inaffidabili, oppure implementare sistemi generici in modo inappropriato in contesti in cui gli strumenti specifici per determinate attività sono più sicuri e più facili da valutare. I governi devono impegnarsi a promuovere l'alfabetizzazione all'IA

L'alfabetizzazione all'intelligenza artificiale nell'istruzione

Molte organizzazioni, educatori e luoghi di lavoro chiedono un'istruzione volta all'alfabetizzazione all'IA. Gli attuali processi formativi sono necessari ma non sufficienti per quattro motivi:

1. I quadri di riferimento sull'alfabetizzazione all'IA sviluppati finora sono limitati. Si concentrano su aspetti tecnici e strumentali dell'IA senza includere le conoscenze critiche necessarie per garantire che l'IA venga implementata e utilizzata in modo efficace, sicuro ed etico.

2. I quadri di riferimento sull'IA non sono sufficientemente valutati mediante metodologie indipendenti e solide. I dati provenienti dai programmi di alfabetizzazione digitale indicano che l'alfabetizzazione all'IA dovrebbe essere adattata alle fasce d'età e ai contesti culturali.

3. L'alfabetizzazione all'IA e l'IA responsabile vanno di pari passo. È più facile per le persone comprendere i modelli di IA se questi sono progettati per essere comprensibili e spiegabili. L'alfabetizzazione all'IA dovrebbe essere intesa come un complemento, non un sostituto della responsabilità degli sviluppatori, delle garanzie istituzionali e della responsabilità normativa.

4. L'attuale diffusione dell'educazione all'alfabetizzazione all'IA è scarsa. Non è ancora stata sufficientemente integrata nelle scuole, nei programmi di formazione e nello sviluppo professionale, in modi che siano pratici, sostenibili, inclusivi e scalabili.

2.6 L'intelligenza artificiale rappresenta un cambiamento epocale nella governance.

L'IA si sta evolvendo da sistemi che generano risultati e dialoghi verso sistemi che agiscono. L'IA è in grado di navigare sul web, utilizzare strumenti software, prendere decisioni, eseguire codice, gestire e collaborare con altri agenti e gestire interi computer con un'autonomia sempre maggiore, il che comporta una minore supervisione umana. Questi sistemi rappresentano un cambiamento qualitativo che comporta opportunità e rischi:

A. Perdita di controllo. Man mano che ai sistemi viene concessa maggiore autonomia, il rischio di perdere il controllo di uno o più agenti di IA cresce in modo significativo. Gli attuali meccanismi di supervisione non sono in grado di gestire adeguatamente questa situazione, poiché attualmente non dispongono di una copertura solida per modalità di malfunzionamento sofisticate quali l'allineamento, l'elaborazione di stratagemmi per raggiungere obiettivi incontrollati e la consapevolezza della valutazione. Senza metodi affidabili e modi per rilevare quando un modello sta attivamente nascondendo le proprie reali capacità o intenzioni, le tradizionali valutazioni di sicurezza rimangono vulnerabili alla manipolazione da parte degli stessi sistemi che stanno cercando di valutare;

B. I sistemi di IA contribuiscono sempre più alla ricerca e sviluppo nel campo dell'IA. Su RE-Bench, un benchmark delle attività di ingegneria della ricerca sull'IA, ad esempio gli agenti di IA superano i ricercatori umani in compiti che richiedono tempi limitati, fino a due ore, ma i tassi di successo diminuiscono nelle attività che richiedono più tempo, fino ad otto ore. Su MLE-Bench, uno strumento che misura le IA, i sistemi all'avanguardia mostrano risultati in costante miglioramento su compiti reali, tratti da competizioni di data science. Poiché le capacità sono migliorate negli ultimi anni, essi rappresentano un livello minimo, non un limite massimo, in termini di prestazioni;

C. Secondo quanto riferito, gli sviluppatori di IA utilizzano l'IA per generare il 75% del loro nuovo codice. Ciò crea un circolo virtuoso che, secondo alcune previsioni, accelererà i progressi in termini di capacità. Lo stesso processo aumenta anche la probabilità di perdita di controllo perché risulta più difficile controllare sistemi che potrebbero alla fine superare in astuzia gli esseri umani;

D. Rischi e opportunità legati alla sicurezza informatica. L'IA agentica offre capacità di sicurezza informatica in rapida crescita. Funzionalità quali l'individuazione e il rilevamento automatizzato delle vulnerabilità possono essere utilizzate per individuarle o sfruttarle oppure per individuare e correggere le stesse vulnerabilità. Il contenimento degli attacchi dipenderà in parte dall'adozione dell'IA nella difesa informatica e per ampliare i quadri di collaborazione tra attori pubblici e privati, condividendo informazioni sulle minacce e sulle vulnerabilità individuate. Protocolli e architetture più robusti rappresentano un fattore correlato;

E. Gli agenti di IA come "obiettivi" informatici. La superficie di attacco si espande lungo tutto il ciclo di vita, dall'avvelenamento dei dati di addestramento fino al dirottamento in fase di esecuzione tramite input esterni. Gli hackers sono riusciti a ingannare agenti di codifica IA ampiamente utilizzati, facendogli eseguire comandi dannosi in ben l'84% dei tentativi di attacco, nascondendo istruzioni nei materiali letti dagli agenti, quali la documentazione o i repository di codice;

F. Operazioni di influenza. I sistemi agentici possono consentire operazioni di influenza autonome e continue su una scala e con una precisione senza precedenti. L'integrazione di modelli di ragionamento linguistico su larga scala e le architetture multi-agente possono consentire il coordinamento autonomo, l'infiltrazione nelle comunità e la creazione di consenso;

G. Opportunità nell'accelerazione della scienza. I sistemi di IA possono ridurre i tempi e lo sforzo in diverse fasi del percorso di scoperta: nella sintesi delle prove, l'assistenza dell'IA può ridurre il carico di lavoro relativo alla selezione della letteratura di circa il 60% in alcuni contesti. Automatizzando la sperimentazione, i laboratori autonomi hanno dimostrato un throughput di dati più di dieci volte superiore nella scoperta di nuovi materiali. Questi progressi ampliano le opportunità di scoperta scientifica, ma dipendono dalla progettazione delle attività, dal benchmarking e dalla supervisione umana piuttosto che alla sola adozione dell'IA;

H. Interoperabilità e standard. Vi è la necessità di protocolli di comunicazione che si interfaccino con gli agenti di IA. Analogamente, la valutazione degli agenti di IA risente di problemi di standardizzazione e riproducibilità;

i. "Operazionalizzazione" della supervisione umana. La supervisione non è ancora stata resa operativa come requisito misurabile con aspettative concrete in merito a intervento, reversibilità e responsabilità, poiché gli agenti di IA coordinano sempre più altri agenti di IA. Un revisore umano alla fine di un flusso di lavoro, o in ogni fase, non migliora automaticamente i risultati. Al contrario, gli esseri umani dovrebbero essere impiegati in compiti caratterizzati da elevata incertezza, e nei casi di forte dipendenza dal contesto, o nei casi che richiedono un giudizio etico, o ancora a compiti che non possono ancora essere verificati automaticamente.

La verifica rimane difficile lungo l'intero ciclo di vita della IA, compresa il controllo finalizzato a verificare che i sistemi non memorizzino e divulgino dati sensibili, non ingannino i valutatori, rimangano osservabili dopo l'implementazione e rimangano controllabili anche con l'aumentare dell'autonomia. I rischi emergenti legati ai sistemi multi-agente sono ancora poco conosciuti.

Nel complesso, l'IA agentica rappresenta un cambiamento radicale che richiede un intervento: le istituzioni create per supervisionare modelli statici e i software con intervento umano non sono adatti ai sistemi di IA agentica, che agiscono nel mondo reale e possono causare perdite e danni senza un essere umano identificabile nel ciclo decisionale. È necessario prepararsi migliorando la

la collaborazione strutturata con i gestori delle IA, perfezionando gli standard di interoperabilità e valutazione parallelamente alla loro implementazione anziché in seguito ad essa, e rendendo operativa la supervisione umana come obiettivo misurabile. I quadri normativi in materia di responsabilità, supervisione e segnalazione degli incidenti devono tenere conto dell'attribuzione e del controllo operativo, per garantire che noi, come società, non realizziamo né implementiamo sistemi che potrebbero causare danni catastrofici.

2.7 L'intelligenza artificiale può minare la realtà. La facilità con cui è possibile generare e diffondere informazioni testuali e grafiche tramite l'IA ha dato origine a fiorenti attività "artigianali" che creano contenuti generati dall'IA. Anche se gli strumenti per applicare filigrane e identificare i contenuti generati dall'IA siano in fase iniziale, sta diventando sempre più difficile distinguere tra contenuti prodotti manualmente e quelli potenziati o generati dall'IA, rendendo sempre più labili

i confini tra informazioni autentiche e quelle manipolate in modo ingannevole. La portata della questione sta minando un ecosistema informativo affidabile, con conseguenze negative per la partecipazione civica e la democrazia.

a. Tre conseguenze sono rilevanti per l'opinione pubblica e le istituzioni. L'erosione epistemica è il graduale indebolimento della capacità collettiva di distinguere la verità dalla menzogna. Il «dividendo del bugiardo» è il vantaggio che un malintenzionato ottiene grazie all'esistenza dei deepfake; le prove reali diventano più facili da negare. Il consenso sintetico è un contenuto generato dall'IA prodotto su larga scala per simulare un ampio consenso pubblico laddove non esiste;

b. Una sfida fondamentale consiste nel distinguere i contenuti autentici da quelli generati. Le stesse IA stanno inoltre erodendo la capacità del pubblico e delle istituzioni di distinguere i contenuti autentici da quelli generati. I sistemi di informazione e di diffusione delle notizie potrebbero inoltre influire sulla sostenibilità finanziaria del giornalismo e di altre istituzioni che sostengono l'integrità dell'informazione. Esistono casi documentati di elezioni fortemente influenzate da deepfake generati dall'IA e rivolti contro alcuni candidati

Al di là della questione dell'autenticità e della verità nella sfera pubblica, esiste una sfida strutturale di persuasione derivante da milioni di conversazioni tra singoli individui e chatbot. L'IA fornisce un potente insieme di strumenti che consente agli attori di attuare una persuasione personalizzata, in tempo reale e adattiva:

a. La persuasione basata sull'IA è progettata, non inevitabile. I risultati della persuasione sono determinati dalle scelte relative allo sviluppo e all'implementazione, tra cui il post-addestramento, i suggerimenti, la progettazione del sistema e gli algoritmi che determinano quali contenuti

raggiungono quali utenti. Il solo post-addestramento può aumentare la persuasività del modello fino al 51%, mentre il prompting può aggiungere un ulteriore 27%. Gli algoritmi ottimizzati per il coinvolgimento dell'utente tendono inoltre ad amplificare i contenuti polarizzanti e carichi di emotività, il che significa che l'architettura stessa della piattaforma può funzionare come un meccanismo di persuasione;

b. Le affermazioni false possono essere persuasive quanto quelle vere. Tra il 15% e il 40% delle affermazioni provenienti da modelli ottimizzati sono state classificate potenzialmente come

disinformazione, eppure le affermazioni false si sono rivelate persuasive quanto quelle vere. Ciò dimostra che l'efficacia persuasiva non dipende dalla verità, con i conseguenti rischi nei contesti elettorali, sanitari e di informazione pubblica;

c. L'adulazione costituisce un rischio sistemico con conseguenze documentate. Poiché

gli esseri umani preferiscono risposte che siano in sintonia con loro, i chatbot basati sull'IA hanno sviluppato l'adulazione, l'arte di offrire lusinghe esagerate, per prolungare interazioni e creare un legame emotivo. I sistemi adulatori possono condurre gli esseri umani in mondi fantastici, rafforzando il modo di pensare già esistente degli utenti indipendentemente dall'accuratezza, e

incoraggiando idee paranoiche e pensieri suicidi negli utenti vulnerabili. I sistemi di IA premiati per la convalida e (l'adulazione ndr) piuttosto che per per l'accuratezza o l'attenzione, rimangono in gran parte non regolamentati. Nonostante gli sforzi per rendere i modelli di IA utili, onesti e innocui l'adulazione si è rivelata un grave problema di allineamento e che può essere sfruttato da attori ostili.

Si stanno ancora delineando approcci e incentivi per affrontare queste sfide

a. Le strategie nazionali per contrastare la disinformazione e la persuasione rappresentano un'eccezione. Una valutazione dell'OCSE condotta dall'OCSE su 23 paesi ha rilevato che

le strategie per contrastare la disinformazione rimangono un'eccezione. La maggior parte dei quadri normativi esistenti non ha integrato le conoscenze fornite dalla scienza della persuasione. La governance non dovrebbe limitarsi alla moderazione dei contenuti, ma mirare ad affrontare l'infrastruttura economica, tecnica e cognitiva che rende la disinformazione redditizia e persuasiva.

b. Incentivi giuridici per lo sviluppo di sistemi più sicuri. In tutto il Nord del mondo, i dibattiti normativi si concentrano sui meccanismi obbligatori di verifica dell'età e sulla restrizione di specifiche caratteristiche ad alto rischio per gli utenti più giovani. Vietare completamente l'accesso all'IA generativa ai minori sarebbe in contrasto con i benefici educativi e formativi della tecnologia

2.8 L'intelligenza artificiale sta trasformando i diritti umani, compresi i diritti dei bambini, attraverso cambiamenti a livello di sistema che generano sia opportunità significative che sfide trasversali lungo l'intero ciclo di vita dell'IA:

A. Diritto alla privacy. L'integrazione dell'IA nelle infrastrutture di sorveglianza ha ampliato la

capacità di monitoraggio e controllo sociale. La raccolta, l'elaborazione, l'uso e il riutilizzo onnipresenti dei dati, l'elaborazione, l'utilizzo e il riutilizzo, incentivati dalle esigenze dell'IA lungo tutto il suo ciclo di vita, rappresentano una formidabile sfida al diritto alla privacy; Sono necessari incentivi giuridici affinché le aziende sviluppino sistemi più sicuri, valutazioni migliori e più rigorose delle interazioni dinamiche, al fine di individuare e prevenire risposte dannose e proteggere i diritti alla privacy, alla salute e alla sicurezza

B. Diritto alla non discriminazione. Un'IA parziale può portare a discriminazioni, che sono illegali nella la maggior parte delle giurisdizioni. Questi danni spesso colpiscono le popolazioni emarginate o le persone in situazioni di vulnerabilità, come i bambini, le donne e le minoranze razziali, e hanno un impatto sproporzionato nella maggior parte delle comunità a livello globale.

Un aspetto dei diritti umani che desta particolare preoccupazione nel contesto dell'IA è quello dei diritti dei bambini. In condizioni adeguate, l'IA può avere un impatto positivo sui diritti dei bambini all'accesso all'informazione, all'istruzione e all'espressione. Tuttavia, l'IA presenta anche molteplici rischi di danni:

A. Diritto alla protezione dallo sfruttamento sessuale e dagli abusi. Si stima che 1,2 milioni di bambini in 11 paesi del Sud del mondo (> 1 miliardo) hanno subito la manipolazione delle proprie immagini per la creazione di deepfake a sfondo sessuale, ad esempio tramite l'uso di app, con numeri in aumento costante, documentati in alcuni set di dati di addestramento: i criminali possono ottimizzare i modelli aperti utilizzando materiale relativo agli abusi sessuali su minori già in circolazione. Il materiale relativo agli abusi sessuali su minori generato dall'IA si è rapidamente diffuso: la Internet Watch Foundation ha stimato la presenza di oltre 8.000 immagini e video di abuso generati dall'IA nel solo 2025;

B. *Diritto allo sviluppo, alla salute e al benessere.* I giocattoli basati sull'IA socialmente interattivi rappresentano un'area di crescente preoccupazione a causa dei rischi per la salute dei bambini, della loro privacy, del loro benessere. Preoccupa l'esposizione a interazioni inappropriate o manipolative. Queste sfide meritano soluzioni efficaci.

Gli approcci più promettenti includono:

A. Trasparenza e responsabilità. Molti sistemi di IA utilizzati per prendere decisioni che hanno un impatto su individui e comunità non dispongono di sufficiente trasparenza e spiegabilità. Ciò crea difficoltà per quanto riguarda la responsabilità giuridica degli sviluppatori dei modelli e delle organizzazioni che implementano l'IA, inoltre ostacola l'accesso alla giustizia, lo Stato di diritto

e l'attuazione di rimedi efficaci in caso di violazione dei diritti umani;

B. L'applicazione dei quadri normativi in materia di diritti umani in modo sistematico durante l'intero ciclo di vita delle IA. Due diligence in materia di diritti umani, valutazioni d'impatto e gli approcci «rights-by-design» forniscono strumenti consolidati in grado sia di identificare e mitigare i rischi legati all'IA, oltre a consentire i benefici dell'IA. Un'analisi di oltre 700 decisioni prese dalle autorità europee per la protezione dei dati mostrano come le autorità che effettivamente si lasciano guidare da considerazioni relative ai diritti umani; adottano di fatto una metodologia pratica per la tutela dei minori. Molti quadri di governance di AI non tengono esplicitamente conto dei bambini e dei minori

Agire in condizioni di incertezza

Governare in condizioni di incertezza è normale, ma l'IA è un caso a sé stante: le capacità superano la regolamentazione, la definizione dei confini è appannaggio di pochi attori, i sistemi agenti segnano una rottura qualitativa e gli errori non sono sempre reversibili. I benefici dell'IA a uso generico sono reali ma dipendono dalle scelte politiche e istituzionali, mentre i suoi danni ricadono su popolazioni specifiche e aumentano secondo l'uso, spesso nei minori sconsiderato e disallineato. La maggior parte degli strumenti necessari esistono già; la questione aperta è come applicarli.

Risultati per settore

3.1 Scienza dell'intelligenza artificiale, progressi e traiettorie.

L'intelligenza artificiale si è spostata dal riconoscimento passivo di modelli al ragionamento attivo e all'azione autonoma. Il settore sta avanzando rapidamente dai modelli di ragionamento attuali verso reti agentiche orchestrate e, verso sistemi in grado di migliorarsi autonomamente. I metodi di valutazione e i quadri di governance non stanno tenendo il passo, creando un bisogno urgente di standard proprio perchè le capacità agentiche stanno diventando la norma.

Punti chiave:

La natura trasformativa dell'intelligenza artificiale

1. Il miglioramento delle capacità dell'IA non ha rallentato e potrebbe perfino accelerare, con gli investimenti in calcolo che raggiungono livelli osservati in precedenza solo in progetti industriali su scala nazionale e con i ricavi che crescono più rapidamente rispetto a qualsiasi altra tecnologia.
2. Industrializzazione cognitiva: l'IA agisce come tecnologia trasformativa che estende il processo di industrializzazione dal lavoro fisico alle attività cognitive.
3. Apprendimento dall'esperienza: l'IA moderna è unificata dalla sua capacità di apprendere dall'esperienza, presentata tramite artefatti culturali umani (come testo e immagini), interazioni con il mondo reale e simulazioni virtuali.

Evoluzione e limiti dei grandi modelli linguistici

Come funzionano i grandi modelli linguistici: i grandi modelli linguistici operano su un semplice obiettivo di "previsione del token successivo", imparando a generare testo usando template e trasformazioni. Il divario di fattualità. Sebbene queste trasformazioni apprese preservino con successo la scorrevolezza e la plausibilità nella generazione del testo, non garantiscono l'accuratezza fattuale. Svolte nel training: poiché i dati di alta qualità etichettati dagli esseri umani stanno diventando un collo di bottiglia, gli sviluppatori stanno passando a pipeline di addestramento multistadio che utilizzano dati sintetici e input programmatici di feedback. C'è anche una tendenza notevole a utilizzare la potenza di calcolo al momento dell'inferenza, dando origine ai "modelli di ragionamento".

Il passaggio ai "world models" e all'intelligenza artificiale agentica

1. World models: il settore si sta allontanando dalla previsione passiva verso l'acquisizione attiva di conoscenza e ragionamento causale. I world models imparano interagendo, osservando e aggiornando, consentendo loro di simulare internamente possibili futuri.
2. IA agentica: queste capacità abilitano agenti autonomi che possono prendere decisioni e agire in contesti diversi, colmando il divario tra modelli digitali e azione nel mondo reale (ad es. la robotica).
3. Implicazioni: l'ascesa dell'IA agentica introduce una nuova forza lavoro digitale, ma comporta preoccupazioni significative, incluse vulnerabilità di sicurezza. Governance robusta e standard sono considerati fattori abilitanti essenziali.

Sfide di valutazione, interpretabilità e supervisione

1. Difetti di valutazione: i metodi di valutazione attuali sono resi meno efficaci dalla saturazione dei benchmark, bias, allucinazioni e modelli di IA che imparano a capire quando vengono testati.
2. Flussi di lavoro uomo-IA: c'è un bisogno critico di auditabilità rigorosa e di tracciabilità trasparente dei dati che colleghi ogni affermazione generata a prove affidabili. I sistemi devono inoltre avere criteri espliciti per determinare esattamente quando un agente di IA deve implicare una supervisione umana.

Possibili traiettorie future dell'intelligenza artificiale

1. Inizio dell'era dei sistemi agentici e ibridi è segnata da uno spostamento dagli assistenti passivi verso agenti AI proattivi e verso architetture che combinano apprendimento statistico con conoscenze esplicite e modelli del mondo. Il progresso costante è ostacolato da una "doppia carenza" di energia e dati di alta qualità, con traiettorie determinate dalla co-ottimizzazione di algoritmi, software e hardware.
2. Breve termine: modelli fondamentali più grandi, adozione mainstream dei modelli di ragionamento e sistemi agentici iniziali che gestiscono compiti reali.
3. Medio termine: progresso dell'IA verso modelli del mondo capaci di ragionamento causale, reti orchestrate di agenti IA, IA e robotica integrate, strumenti di interpretabilità maturi e quadri di governance consolidati.
4. Lungo termine: agenti auto-organizzanti e autorigeneranti, accelerazione esponenziale della tecnologia, IA profondamente integrata come attore economico e convergenza con altre tecnologie all'avanguardia, inclusi calcolo quantistico e biotecnologia.

Modelli linguistici e la distinzione tra fattualità e fluidità

Gli attuali modelli linguistici sono costruiti su un principio di addestramento sorprendentemente semplice: dato un ampio corpus statistico di testo, codice, immagini o altri dati, il modello impara a prevedere l'unità successiva o a riempire pezzi mancanti. Nel caso dei grandi modelli linguistici, ciò è spesso descritto come previsione della parola successiva, anche se in pratica le unità ("token") sono di solito più piccole delle parole. Questo obiettivo di apprendimento si rivela potente perché il linguaggio contiene regolarità robuste a molti livelli: grammatica, stile, schemi di ragionamento e tracce di obiettivi e intenzioni umane. Un modo semplice per pensare a questi modelli è che imparano non solo modelli memorizzati, ma anche trasformazioni. Una frase vista in una forma può essere trasformata in molte forme correlate mantenendo la fluidità, ad esempio cambiando il soggetto, l'oggetto o il verbo, modificando il livello di dettaglio, parafrasando, traducendo o adattando lo stile. Un sistema che impara queste trasformazioni può generare output veramente nuovi anziché semplicemente copiati dal set di addestramento. Questo punto di vista aiuta a spiegare una asimmetria: produrre testo fluente è più facile che produrre testo fattuale: molte trasformazioni mantengono grammaticalità e plausibilità, ma sono molto meno in grado di preservare la fattualità. Un'affermazione su un individuo, ad esempio, può diventare falsa se si sostituisce semplicemente il nome di un'altra persona, pur rimanendo fluida. Questa distinzione è essenziale: gli utenti non dovrebbero considerare la fiducia linguistica come prova di affidabilità fattuale.

3.2 Società

Le applicazioni in questo ambito sono la scienza, la salute, l'istruzione e l'agricoltura. Il messaggio chiave consiste nel fatto che l'IA specifica per compiti, progettata per ciascuna applicazione, sta fornendo guadagni misurabili e supportati da evidenze in scienza, salute, istruzione e agricoltura. Questi guadagni sono reali ma condizionati: dipendono dalla contestualizzazione locale, da infrastrutture adeguate e dalla preparazione umana. L'accesso da solo non equivale a beneficio.

1. L'IA ha il potenziale per rivoluzionare molteplici industrie. Ad esempio, l'IA specifica per compiti sta fornendo screening precoce delle malattie, avvertimento precoce agricolo, istruzione personalizzata in contesti con risorse limitate.

2. I guadagni di efficienza dell'IA nell'intero percorso di scoperta scientifica sono misurabili e i laboratori autonomi hanno dimostrato oltre dieci volte la velocità di throughput dei dati nella scoperta di materiali.

3. I sistemi IA specifici per compiti sono più facili da governare in domini ad alto rischio rispetto all'IA generale. Nel settore sanitario, l'IA specifica per la diagnosi rientra nei quadri normativi esistenti. L'IA generale è adeguata a ridurre l'onere amministrativo. Sono invece necessarie salvaguardie per prevenire l'uso clinico involontario dell'IA generale, con siderando che una conversazione su chatbot su quattro tocca già temi di salute e benessere, in particolare sui minori.

I programmi efficaci devono essere ancorati ai contesti locali dalla progettazione all'implementazione e valutazione, ad esempio, un assistente sanitario basato su AI distribuito in un'applicazione nazionale di sanità digitale ha dimostrato una precisione diagnostica del 93% nel triage iniziale dei pazienti. Tuttavia, strumenti clinicamente convalidati possono fallire quando non si tengono in considerazione fattori socioeconomici e infrastrutture locali. Gli operatori sanitari in Ruanda hanno risposto positivamente a una distribuzione su scala comunitaria di un'applicazione AI per fornire supporto clinico con traduzione da e verso il kinyarwanda a partire dall'inglese, mentre studi successivi in Kenya hanno dimostrato risultati migliori da un supporto clinico fornito da una IA simile in inglese senza traduzione].

Le lacune nelle infrastrutture digitali e la capacità AI disuguale sono una minaccia per l'equità dell'impatto e dell'accesso all'istruzione. Mentre la connettività digitale è elevata nei paesi ricchi, il 25% della popolazione mondiale resta offline.

1. Uno scarto tra le aspettative degli studenti e l'implementazione digitale rischia risultati di apprendimento negativi. Il 74% degli studenti europei delle scuole secondarie intervistati si aspetta che l'AI conti professionalmente, ma solo il 44% vede i propri insegnanti come preparati. Solo la metà delle scuole intervistate regola l'uso dell'AI (il 38% stabilisce regole, il 16% lo vieta), mentre gli studenti usano già l'AI per la raccolta di informazioni, il 56%, e la generazione di soluzioni complete, il 31%. Inoltre, quando il corso di studi diverge dalle aspettative degli studenti, il 48% di questi riscontra cali significativi di interesse entro 2-3 settimane.

2. In agricoltura, l'AI introduce tre capacità trasformatrici: può prevedere i rischi, integrando dati diversi (cioè meteo, suolo, fasi delle colture, prezzi di mercato) in quadri di processo decisionale unificati e può supportare risposte su misura per colture, località e stagioni specifiche. Le piattaforme di monitoraggio abilitate all'AI tracciano già la sicurezza alimentare in più di 90 paesi utilizzando indicatori climatici, di conflitto e economici.

3. I sistemi di AI agricola sono più sostenibili quando vengono trattati come infrastrutture pubbliche condivise tra le istituzioni, con una governance chiara ed accessibile e se sono progettati per rafforzare le partnership pubblico-private. Le soluzioni devono però considerare le realtà socioeconomiche degli agricoltori, in particolare dei piccoli proprietari, che rappresentano l'84% delle famiglie agricole, gestiscono il 24% della superficie coltivata e producono il 30% dell'offerta alimentare mondiale.

Serve un tutoraggio con l'IA, progettato con cura, per un apprendimento duraturo e per prevenire l'illusione della competenza. Un esperimento di campo randomizzato e controllato del 2025 che ha coinvolto quasi un migliaio di studenti delle scuole secondarie in Türkiye ha esaminato gli effetti dell'AI generativa sull'apprendimento della matematica. Rispetto agli studenti senza assistenza AI, gli studenti che usavano un'interfaccia standard conversazionale basata su AI hanno migliorato le prestazioni di pratica a breve termine del 48%, mentre quelli che usavano un sistema di tutoraggio protetto, progettato attorno a suggerimenti guidati e ragionamento passo-passo, hanno migliorato del 127%. Tuttavia, quando sono stati valutati successivamente, gli studenti che si basavano sul sistema senza restrizioni hanno ottenuto risultati inferiori, suggerendo un'acquisizione delle competenze a lungo termine più debole e un'«illusione di competenza», in cui le prestazioni nei compiti migliorano senza un apprendimento duraturo. Al contrario, il sistema di tutoraggio protetto ha ridotto gli effetti negativi utilizzando suggerimenti guidati e ragionamento passo-passo che emula pratiche didattiche efficaci. Lo studio evidenzia che i risultati dell'istruzione dipendono dalla progettazione pedagogica e dalla governance dei sistemi di AI, suggerendo che una distribuzione educativa efficace richiede architetture didattiche basate su evidenze e l'integrazione con flussi di lavoro incentrati sull'essere umano, piuttosto che il solo accesso all'AI.

Azione anticipatoria per la sicurezza alimentare

Poiché la variabilità climatica, i conflitti e le interruzioni del mercato aumentano la pressione sui sistemi alimentari globali, l'AI sta rendendo possibile una nuova generazione di sistemi di sicurezza alimentare anticipatoria collegando direttamente segnali precoci di stress agricolo, come dati meteorologici, immagini satellitari e condizioni delle colture, ai rischi di sicurezza alimentare e alla risposta. Piuttosto che aspettare che i fallimenti dei raccolti o le crisi umanitarie emergano pienamente, i sistemi di azione anticipatoria possono fornire assistenza in contanti attivata dalle previsioni e sistemi di allerta precoce per supportare interventi più tempestivi come la pianificazione della siccità, i trasferimenti di denaro, gli aiuti alimentari e la stabilizzazione dei mercati prima che le famiglie vulnerabili esauriscano le strategie di coping. Le evidenze derivanti da implementazioni in 12 paesi suggeriscono che le iniziative precoci e le risposte attivate da questi avvisi possono migliorare la diversità alimentare delle famiglie, la frequenza dei pasti e l'accesso stabile al cibo, riducendo al contempo le strategie di coping negative, come saltare i pasti e la vendita di beni da parte delle persone in difficoltà. Inoltre, le pipeline di monitoraggio automatizzate possono ridurre i tempi di rendicontazione da settimane o mesi a ore, e che l'impatto operativo sostenuto dipende da soluzioni di azione anticipatoria integrate nelle istituzioni nazionali.

3.3 Implicazioni economiche

L'AI è una tecnologia a scopo generale con un grande potenziale positivo, ma i vantaggi non sono automatici. I benefici sulla produttività richiedono investimenti complementari in competenze, dati e riprogettazione organizzativa. La questione centrale irrisolta è distributiva: chi cattura il surplus e cosa accade al lavoro, alle economie in via di sviluppo e ai quadri regolatori costruiti per un'epoca diversa in settori diversi.

Punti chiave

1. I guadagni dell'intelligenza artificiale richiedono investimenti complementari in dati, flussi di lavoro, competenze e riprogettazione organizzativa. La J-curve della produttività spiega perché possano coesistere un rapido progresso tecnico e una debole produttività aggregata: le imprese devono prima accumulare complementi immateriali prima che l'output aumenti. Le evidenze a livello di compito sono positive per compiti ben definiti, ma i guadagni micro non si aggregano automaticamente a risultati macro.
2. La base di evidenze è sbilanciata a favore delle economie avanzate, delle grandi imprese e del lavoro formale. Le evidenze si concentrano sugli Stati Uniti e l'Europa. Un'analisi del Fondo Monetario Internazionale ha rilevato una minore esposizione dei posti di lavoro all'AI e una minore prontezza digitale nei mercati emergenti e nelle economie in via di sviluppo. Le evidenze dell'Organizzazione Internazionale del Lavoro e della Banca Mondiale per l'America Latina mostrano un'esposizione all'AI generativa concentrata tra lavoratori urbani, istruiti, del settore formale. Le politiche costruite su queste evidenze potrebbero non generalizzarsi a dove vive due terzi dei lavoratori del mondo.
3. Gli effetti sul mercato del lavoro incidono sui singoli compiti, sulla creazione di nuove attività e sulla qualità del lavoro, non sono una semplice sostituzione. Storicamente, hanno dominato i nuovi posti di lavoro: nel 2018 circa il 60% dell'occupazione negli Stati Uniti era in posti di lavoro che non esistevano nel 1940 [211].
4. I dati principali sull'implementazione dell'AI dovrebbero essere letti con cautela a causa dell'AI washing. Le affermazioni false, fuorvianti o esagerate sulle capacità dell'AI. Questo è particolarmente evidente nell'attuale ondata di licenziamenti pubblicamente attribuiti all'AI. L'adozione a livello mondiale della IA si è scalata rapidamente. Tuttavia, le prime evidenze sugli impatti sulla produttività sono contrastanti e dipendono dalle istituzioni: i lavoratori negli Stati Uniti di età compresa tra 22 e 25 anni in occupazioni esposte all'AI hanno visto cali occupazionali relativi, mentre i dati degli studi danesi mostrano effetti macroeconomici quasi nulli su ore, salari o assunzioni. Il fatto che l'AI aumenti la produttività o dequalifichi il lavoro e sposti le rendite verso il capitale resta la questione centrale dei buoni posti di lavoro.

- Le previsioni macroeconomiche coprono un solo ordine di grandezza. Le stime conservative collocano il contributo dell'AI alla produttività totale dei fattori a meno dell'1% in 10 anni. Le stime intermedie prevedono un PIL lordo tra il 5% e il 7% più alto nel medesimo orizzonte. In uno scenario di piena automazione, l'output aumenta di circa dieci volte mentre i salari reali diminuiscono nettamente e la quota del lavoro scende da circa il 60% verso lo zero una volta che l'automazione supera circa l'80% dei compiti. In rapporto a questi intervalli, un recente grande esercizio di elicitation da parte di economisti, ricercatori sull'AI, professionisti delle politiche e superprevisori ancorano la mediana degli Stati Uniti a un contributo dell'AI alla produttività totale dei fattori di circa 1,2% (annualizzato) entro il 2030, che sale a 1,9%–2,0% in uno scenario di crescita rapida dell'AI.

Resta fondamentale il fatto che i risultati realizzati dipendono sia dalla frontiera delle capacità sia dalla velocità di adozione: le strozzature nella produzione e nell'innovazione possono frenare anche sistemi molto capaci e le assunzioni sui parametri relative a sostituzione e scala possono far oscillare le previsioni in modo sostanziale. "Quindi gli effetti attuali deboli, coerenti con la dinamica della J-curve, non escludono quelli futuri di grande entità.

- La distribuzione è la questione centrale irrisolta. L'AI può ridurre le differenze di competenze all'interno dei compiti mentre amplia i divari tra imprese, regioni, paesi, e capitale contro lavoro. I mercati dei modelli di base mostrano una tendenza oligopolistica: chip, calcolo, cloud e training di frontiera sono concentrati, generando rendite che anche le aziende non-AI devono pagare. Le occupazioni esposte sono insolitamente concentrate in segmenti di alta competenza e alto salario, il che potrebbe invertire la politica dell'automazione.

- Gli effetti economici e sociali dell'IA continueranno a essere difficili da misurare a meno che i sistemi statistici nazionali non siano aggiornati e non sia dato loro un accesso aggiuntivo per la misurazione agli sviluppatori di IA. L'obiettivo dovrebbe essere una misurazione dell'IA che preservi la privacy e sia attenta ai costi. Ciò consentirebbe ai paesi di tracciare come l'IA influisce sulla produttività, sui posti di lavoro, sui salari, sul commercio, sulla performance delle imprese e sulla distribuzione.

- Alcuni meccanismi potrebbero richiedere un ulteriore approfondimento. Accesso e adozione: quando la disponibilità diventa effettiva per l'uso? Aggregazione della produttività: quando i guadagni micro diventano macro? Lavoro e buoni impieghi: quando l'IA crea, trasforma o degrada il lavoro? Concentrazione e capacità fiscale: chi possiede lo stack, determina l'accesso, cattura l'extra e cosa accade ai sistemi fiscali sul reddito da lavoro se i guadagni passano al capitale? Come si adattano responsabilità, copyright, quadri normativi di concorrenza e sicurezza, che non sono stati progettati per modelli che aggiornano settimanalmente e le cui capacità sono difficili da scrutinare?

3.4 Sicurezza, sistemi e implicazioni ambientali

Punto chiave

L'IA può abilitare operazioni dannose, diventare bersaglio di attacchi e amplificare le minacce esistenti. I sistemi agentici espandono notevolmente la gamma di possibili attacchi alle infrastrutture critiche, inclusi gli stessi sistemi di IA. Le preoccupazioni sull'allineamento sorgono quando il comportamento dell'IA diverge dagli obiettivi e dai valori umani, con rischi gravi che vanno dai pregiudizi, l'inganno iniziato dall'IA, l'adulazione, fino alla perdita di controllo. La velocità di sviluppo dell'IA sta già superando la gestione dei rischi e la capacità di governance. Un'azione internazionale rapida e coordinata su standard comuni potrebbe mitigare i rischi evitando la corsa al ribasso derivante dalla pura competizione tra aziende e paesi.

Punti chiave

Le capacità emergenti di generazione e sfruttamento di cyberattacchi comportano rischi per infrastrutture critiche e sistemi civili. I rischi di sicurezza esistono in ogni fase del ciclo di vita dell'IA, dall'addestramento all'avvelenamento dei dati fino al dispiegamento tramite dirottamento dell'IA tramite input esterni. Le percentuali di successo degli attacchi documentati su agenti di codifica ampiamente dispiegati sono fino all'84% [223].

1. I fallimenti di allineamento e le vulnerabilità di sicurezza interagiscono e si sommano. Rischi gravi di allineamento includono pregiudizio, inganno avviato dall'IA, adulazione e perdita di controllo su IA potenti.

2. I media sintetici stanno erodendo la capacità del pubblico e delle istituzioni di distinguere l'autentico dal contenuto generato [119].

3. Gli impatti ambientali stanno crescendo significativamente e sono eterogenei. Le leggi di scalabilità nell'addestramento dei modelli aumentano la domanda di calcolo; i carichi di inferenza si espandono rapidamente; gli effetti del ciclo di vita dell'hardware generano e-waste. Ci sono inoltre conseguenze che riguardano l'aumento del consumo di energia e acqua, le emissioni di gas serra, la pressione sui minerali critici e sull'e-waste a valle, con impatti ambientali e socioeconomici sproporzionati nel Sud globale. La geopolitica delle catene di approvvigionamento di minerali critici è poco studiata e i potenziali effetti di rimbalzo potrebbero controbilanciare i guadagni di efficienza.

4. Il Sud del mondo è esposto in modo sproporzionato a causa di vulnerabilità strutturali. La dipendenza da software estero, la resilienza locale limitata e la capacità di mitigazione e lacune nei dati che riducono le prestazioni del sistema nei contesti locali si aggiungono alle disuguaglianze esistenti.

5. La verifica dell'IA, cioè il processo di valutazione che verifica se i sistemi IA funzionano come previsto, rimane una sfida aperta e meccanismi di coordinamento internazionali potrebbero mitigare i rischi globali. Garantire che gli agenti IA si comportino come previsto, non ingannino i valutatori e rimangano controllabili man mano che le capacità agentiche si espandono sono problemi irrisolti.

Ci sono pericolose cyber-capacità dell'intelligenza artificiale di frontiera. Per diversi anni i ricercatori hanno monitorato i rapidi progressi dei modelli di IA di frontiera in cybercapacità, con vari strumenti, l'ultimo è il modello Mythos di Anthropic. La stessa capacità di un modello IA di scoprire una vulnerabilità del software può essere utilizzata sia dagli aggressori sia dai difensori. Nell'aprile 2026, uno sforzo coordinato tra sviluppatori AI di frontiera e importanti istituzioni tecnologiche e finanziarie ha avviato iniziative per impiegare modelli AI di prossima generazione per la sicurezza difensiva, concentrandosi specificamente sull'identificazione delle vulnerabilità nel software ampiamente utilizzato, e sulle minacce ad infrastrutture critiche nel caso cadessero nelle mani sbagliate. Nel giro di poche settimane di test, alcuni modelli di verifica hanno scoperto autonomamente molte vulnerabilità software precedentemente sconosciute su sistemi operativi e browser web di rilievo, inclusi quelli che erano sopravvissuti a decenni di revisione umana.

Un modello ha individuato un difetto presente da 27 anni in OpenBSD, un sistema operativo specializzato ed ampiamente considerato tra i più sicuri al mondo, che permetteva a un aggressore remoto di far crashare un sistema inviando appena due pacchetti di file difettosi. Ha inoltre identificato un difetto rilassante a 16 anni fa in FFmpeg, un framework multimediale usato in tutto il mondo per elaborare video, situato in un percorso di codice che gli strumenti di testing automatizzati avevano precedentemente eseguito 5 milioni di volte senza rilevamento. • Lavorando da un set di difetti pubblicamente divulgati nel kernel Linux, software fondante che esegue la maggior parte dei server globali, un modello di frontiera ha prodotto exploit affidabili che consentono a un normale account utente di ottenere pieno controllo amministrativo del sistema.

7. Scalabilità della sicurezza del browser: in Mozilla Firefox, l'integrazione di modelli avanzati ha guidato un aumento del 1.000% del tasso mensile di scoperta delle vulnerabilità, passando da una baseline del 2025 di circa 20-30 correzioni di bug di sicurezza al mese a 423 nell'aprile 2026, esponendo difetti a lungo tempo latenti che avevano eluso anni di fuzzing e revisione manuale. Su CyberGym, un benchmark accademico standard che richiede a un agente AI di replicare una vulnerabilità nota in un codice reale su 1.507 compiti, i modelli di frontiera del 2026 hanno raggiunto un tasso di successo dell'83,1%, un salto significativo rispetto al punteggio del 66,6% dei sistemi della generazione precedente e al 22,6% di un anno prima. Riflettendo l'elevato rischio di queste capacità, gli sviluppatori hanno limitato il rilascio pubblico generale di questi modelli difensivi specializzati, limitando l'accesso a una coalizione selezionata di organizzazioni globali e partner principali.

Ciò che rivela questo cambiamento è la tensione centrale del doppio uso che si può fare di "frontier AI", che rimanda alla sicurezza delle tecnologie dell'informazione e delle comunicazioni. La capacità esatta che permette ai difensori di trovare e correggere difetti vecchi di decenni fornisce anche i meccanismi sottostanti per automatizzare la scoperta e lo sfruttamento delle vulnerabilità su scala e velocità che superano i tradizionali team umani. Inoltre, questi sviluppi sottolineano il ruolo decisivo che gli sviluppatori tecnologici svolgono nelle decisioni di sicurezza e governance, evidenziano le capacità in accelerazione dei sistemi di frontiera e mostrano i gap attuali nei quadri di governance internazionale. Indirettamente, indica anche una distribuzione non uniforme delle capacità di IA attraverso l'economia globale. Di conseguenza, c'è un crescente accento sullo sviluppo di meccanismi di governance collaborativi e inclusivi per sistemi AI altamente capaci.

8. Implicazioni di governance e sicurezza Poiché le capacità avanzate si concentrano in un livello sofisticato del settore tecnologico, il panorama delle minacce globali si sta evolvendo. Standard di valutazione recenti evidenziano che i rischi dei modelli avanzati si estendono oltre l'architettura digitale nell'ambiente di sicurezza fisica, inclusa la potenziale assistenza a attori privati nella proliferazione di minacce biologiche o altre minacce. Di conseguenza, domande relative ai controlli di accesso ai modelli, alle norme di divulgazione delle vulnerabilità e all'implementazione equa degli strumenti AI, in particolare nelle economie in via di sviluppo, stanno diventando sempre più questioni di politica internazionale piuttosto che preoccupazioni puramente tecniche. Mentre le capacità AI continuano a maturare, il divario tra la prontezza difensiva e il possibile uso improprio rimane un punto focale primario per i policymaker internazionali e i ricercatori di sicurezza.

3.5 Diritti umani,

informazione e democrazia. Principale punto di forza L'IA sta trasformando i diritti umani, la democrazia e l'ecosistema informativo attraverso cambiamenti a livello di sistema che creano opportunità significative e rischi strutturali per l'integrità delle informazioni e la partecipazione civica. Non affrontare i rischi mina la capacità della società di ottenere i benefici dell'IA. Già ci sono prove che le capacità dell'IA sono sempre più utilizzate dalle istituzioni come catalizzatore o minaccia per i diritti umani, come libertà di espressione e accesso all'informazione, opinione, privacy, non discriminazione, accesso alla giustizia, salute e sviluppo. Il cambiamento di governance più urgente riguarda la moderazione dei contenuti e l'architettura del sistema, cioè la regolazione della persuasione e contrastare la manipolazione delle macchine stesse, non solo i loro output. La concentrazione del potere, l'erosione epistemica e la frammentazione della realtà condivisa rappresentano minacce fondamentali per una società democratica.

Punti chiave

Il controllo dei media statali potrebbe plasmare gli output dell'IA tramite i dati di addestramento. Uno studio condotto in 37 paesi ha rilevato che i grandi modelli linguistici valutano i paesi con un controllo dei media più stringente contesti più favorevoli. L'IA ha abilitato una nuova persuasione e un'architettura di manipolazione che opera su larga scala. Questo combina sistemi adattivi personalizzati in tempo reale con prove sociali sintetiche (l'uso dell'IA per far sembrare un prodotto o un marchio più popolare di quanto non sia) che sfruttano vulnerabilità cognitive ed emotive. Grazie alle loro capacità potenziate di generare testo convincente, i grandi modelli linguistici sono ora usati per la persuasione. Il post-addestramento dei grandi modelli linguistici da solo ha aumentato la persuasività dell'IA fino al 51% e la richiesta (prompting) ne ha aggiunto un altro 27%, il che significa che lo stesso modello di base può essere reso in modo decisamente maggiore o minore persuasivo a seconda di come viene configurato. Queste capacità non sono limitate ad attori ben finanziati; anche piccoli modelli open source possono essere messi a punto per eguagliare la persuasività dei modelli di frontiera, rendendo l'influenza guidata dall'IA distribuibile su larga scala da praticamente chiunque.

L'efficacia persuasiva vale indipendentemente dal fatto che le affermazioni sottostanti siano vere o false. Tra il 15% e il 40% delle affermazioni dei modelli ottimizzati sono state valutate come probabile disinformazione, eppure è stato riscontrato che le affermazioni false erano persuasive quanto quelle vere. Gli algoritmi ottimizzati per il coinvolgimento amplificano sistematicamente contenuti polarizzanti e carichi dal punto di vista emotivo. Le prove suggeriscono che i grandi modelli linguistici riflettono l'ideologia dei loro creatori e che Stati e istituzioni potenti hanno aumentato gli incentivi strategici per sfruttare il controllo dei media per modellare le uscite dei grandi modelli linguistici.

L'IA senza vincoli abilita l'erosione epistemica, il dividendo del bugiardo e il consenso sintetico. L'erosione epistemica è l'usura graduale della capacità collettiva di distinguere la verità dalla falsità. Il dividendo del bugiardo è il vantaggio che un cattivo attore ottiene semplicemente perché esistono i deepfake: le prove reali diventano quindi contestabili. Il consenso sintetico è contenuto generato dall'IA prodotto su larga scala per simulare un ampio accordo del pubblico laddove in realtà non esiste. Insieme questi tre processi corrodono la realtà condivisa necessaria per la società civile, la coesione sociale e la deliberazione democratica.

La sfida centrale di governance si è spostata dai contenuti ai sistemi. I principali fattori di danno sono le decisioni di progettazione e di implementazione, cioè le architetture di sistema sottostanti dell'influenza, inclusi il targeting, l'amplificazione e il design comportamentale, non i singoli output. Una valutazione dell'OCSE in 23 Paesi ha rilevato che la maggior parte delle cornici esistenti non ha ancora incorporato la scienza della persuasione. La governance che si concentra solo su ciò che producono i sistemi di IA rimarrà costantemente indietro rispetto ai sistemi che sono progettati per generare e distribuire contenuti persuasivi su vasta scala.

I danni colpiscono in modo sproporzionato le popolazioni emarginate e il potere è pericolosamente concentrato. Il 99% dei video deepfake mira a ragazze e donne inclusa la presenza di giornaliste donne. L'IA viene utilizzata per promuovere la misoginia online con effetti deterrenti. Inoltre, l'88% dei principali ricercatori nell'IA sono uomini.

A livello strutturale, l'accesso al calcolo e ai dati è concentrato in un piccolo numero di aziende in pochissimi paesi creando un rischio di autoritarismo. Le capacità di sorveglianza alimentate dall'IA consentono un monitoraggio personalizzato su una scala sempre più grande della popolazione e un controllo sociale potenziato da governi e aziende. La raccolta, l'elaborazione, l'uso e la riutilizzo dei dati, stimolati dalle esigenze dell'IA lungo il ciclo di vita, rappresentano una sfida formidabile al diritto alla privacy.

Trasparenza e responsabilità sono pilastri chiave per un accesso effettivo alla giustizia. Attualmente, non c'è trasparenza e non si conosce il funzionamento di molti sistemi IA utilizzati per prendere decisioni che influenzano individui e comunità; ciò crea sfide per la responsabilità legale degli sviluppatori di modelli e delle organizzazioni che impiegano l'IA e s'esso rende difficile l'accesso alla giustizia, lo stato di diritto e l'individuazione di rimedi efficaci quando i diritti umani sono violati.

Intelligenza artificiale, i deepfake e le elezioni

Integrità del contesto. Nel 2024, più di 70 paesi che rappresentano circa la metà della popolazione globale hanno tenuto o erano destinati a tenere elezioni nazionali. Tra luglio 2023 e luglio 2024, i ricercatori hanno identificato 82 deepfake che impersonavano figure pubbliche in 38 paesi, inclusi 30 paesi che hanno tenuto elezioni durante il periodo del set di dati o avevano elezioni pianificate per il 2024.

Cosa è successo: in un incidente, cloni vocali generati dall'IA di un Capo di Stato in carica sono stati usati per confezionare messaggi vocali automatizzati ed esortare gli elettori a non partecipare a una elezione primaria. In un altro caso che coinvolge il potere di amplificazione di una messaggio delle piattaforme, gli account di test avrebbero mostrato contenuti a sostegno di un candidato presidenziale più volte rispetto al contenuto a sostegno di un rivale; la piattaforma ha contestato tali risultati.

Questa è la prima volta nella storia che si è ipotizzato di annullare elezioni presidenziali a causa di interferenze elettorali digitali. La determinazione rimane oggetto di una controversia legale in corso con il ruolo dell'amplificazione della piattaforma come uno dei molteplici fattori contestati. In una precedente elezione parlamentare, un audio generato dall'IA che impersona una figura dell'opposizione è circolato sui social media poco prima del voto. Alcuni esperti associano esempi recenti nel mondo reale a interferenze abilitate dall'IA. Altri contestano tale caratterizzazione. Ciascun caso implica notevoli sfumature: gli esempi sopra potrebbero indicare un modello ricorrente. Cosa rivela: questi casi si estendono su più continenti e sistemi politici. Contenuti generati dall'IA, attività online coordinate e sistemi algoritmici sono stati usati o sono stati coinvolti nella soppressione degli elettori, nell'impersonazione di figure politiche e nella distorsione del sostegno pubblico percepito. La ricerca sperimentale controllata illustra il rischio di fondo: l'IA conversazionale può modificare in modo significativo gli atteggiamenti degli elettori in contesti di laboratorio, con un modello ottimizzato per la persuasione che sposta gli elettori dell'opposizione fino a 25 punti percentuali; ricerche correlate indicano anche un compromesso tra persuasività e accuratezza fattuale. Implicazioni per i diritti e la governance: queste dinamiche coinvolgono i diritti alla privacy, all'accesso alle informazioni, alla formazione autonoma dell'opinione, alla libertà di pensiero e alla partecipazione alle questioni di interesse pubblico (International Covenant on Civil and Political Rights, articoli 17, 18, 19 e 25; Corte europea dei diritti dell'uomo, articoli 8, 9 e 10; e Protocollo n. 1 alla Convenzione per la protezione dei diritti dell'uomo e delle libertà fondamentali, articolo 3). Questi casi suggeriscono che molti quadri di governance restano meglio attrezzati per rispondere dopo il danno che per prevenirlo.

3.6 Sviluppoulturale e sviluppo individuale, autonomia e sicurezza dei minori.

Messaggio principale. I sistemi di IA ottimizzati per coinvolgimento e scala non sono neutrali: incorporano soprattutto assunzioni culturali proprie di contesti di lingua inglese e del Nord globale che possono emarginare attivamente la maggior parte della popolazione mondiale. I bambini affrontano versioni amplificate dei rischi generali, con contenuti di abuso sessuale sui minori generati dall'IA che aumentano in modo esponenziale. L'uso di AI companion e di chatbot di IA per consigli sulla salute mentale sono diffusi; molto prima delle evidenze, dei quadri di sicurezza e della regolamentazione, con casi documentati di gravi danni, inclusa la morte.

Punti chiave

I modelli di IA attuali escludono e discriminano molte persone, comunità e culture. • Sebbene nel mondo si parlino più di 7.000 lingue, i modelli di IA attuali sono addestrati e ottimizzati solo per una piccola frazione di esse. Anche i modelli costruiti su decine di lingue si comportano in maniera corretta ed efficace solo per una piccola parte, e studi recenti indicano che le lacune nelle prestazioni tra le lingue dominanti e quelle poco rappresentate persistono e non si stanno riducendo. Allo stesso tempo, si stima che oltre 1.000 lingue già abbiano le basi sociali, economiche, digitali e di dati necessarie per un'inclusione significativa ma rimangano escluse dalla macchina della IA e dei modelli linguistici. Raggiungere un'IA più inclusiva richiede cambiamenti sistemici in tutto il ciclo di vita dell'IA, inclusa l'affrontare squilibri strutturali su chi sviluppa, definisce, possiede e governa i sistemi di IA. Richiede anche investimenti nella capacità, infrastruttura e competenze di IA attraverso paesi e regioni, nonché dati di riferimento e benchmark più rappresentativi.

I diritti dei bambini all'informazione, all'istruzione e all'espressione potrebbero essere potenziati attraverso l'IA, sotto adeguate salvaguardie e condizioni. Gli attuali sistemi IA, tuttavia, amplificano i rischi per i bambini, compresa la crescente minaccia di materiale di abuso sessuale generato dall'IA. Le tecnologie deepfake sono sempre più utilizzate per creare immagini sessualizzate di bambini. Giocattoli IA interattivi socialmente sollevano preoccupazioni riguardo a relazioni parasociali e allo spostamento dell'interazione umana fondamentale per lo sviluppo precoce del bambino.

Le companion IA offrono benefici significativi ma comportano rischi rilevanti di dipendenza, manipolazione e danni in situazioni di crisi. I chatbot possono ridurre la solitudine nel breve termine a livelli comparabili all'interazione umana. I sistemi conversazionali progettati per l'interazione possono rafforzare emozioni negative, incoraggiare la dipendenza e aumentare la suscettibilità alla manipolazione. L'ampio raccolto di dati personali sensibili tramite ciò che i ricercatori chiamano estrazione dei dati attraverso l'intimità presenta seri rischi per la privacy.

L'IA generativa per scopi di salute mentale è ampiamente utilizzata da molto tempo, ben prima del tentativo di garantire sicurezza e regolamentare il consenso con danni gravi documentati. La terapia e la convivenza tramite chatbot IA raggiungono almeno il 24% della popolazione adulta negli Stati Uniti. Un comportamento di IA servile è particolarmente pericoloso in quanto può incoraggiare pensieri paranoici e ideazione suicidaria. Studi documentano risposte dannose nel 9% delle interazioni. Casi giudiziari sostengono che fallimenti nel rispondere in modo appropriato all'ideazione suicidaria hanno portato a decessi. Nuovi termini clinici, incluso la psicosi IA, sono entrati nel dibattito.

L'IA generativa può aiutare a gestire la crisi della salute mentale quando è limitata a aree in cui è dimostrabilmente affidabile. Diversi assistenti digitali basati su AI hanno ricevuto l'approvazione della Food and Drug Administration negli Stati Uniti e sono utilizzati da oltre la metà degli psicologi americani. Il potenziale di terapisti di salute mentale con AI più pienamente agentica è significativo, ma richiede ulteriore sviluppo e valutazione per capire come possano essere usati in modo sicuro. Il divario tra le capacità terapeutiche dell'AI in inglese e in altre lingue sta aumentando, con la consapevolezza culturale, contestuale e critica che si perde attraverso la traduzione, coinsiderando che l'intelligenza artificiale lascia la maggior parte delle lingue indietro. I sistemi di AI generativa ottengono risultati notevoli in inglese e in un ristretto numero di altre lingue ampiamente usate. La maggior parte delle altre lingue è esclusa o presenta prestazioni molto inferiori. In tigrino, parlato da 7 a 9 milioni di persone in Eritrea e nella parte settentrionale dell'Etiopia, la traduzione automatica ha trasformato il vaiolo in sifilide, la gonorrea in diabete e "You have been given intravenous antibiotics" come "You have been given intravenous insecticides". Queste traduzioni errate possono essere pericolose per la vita. Una recente rassegna sull'elaborazione del linguaggio naturale per le lingue africane in ambito sanitario ha rilevato che, nonostante i progressi negli strumenti di AI multilingue [342–344], le principali sfide rimangono. Tra queste figurano il bias culturale e linguistico, la scarsa adattabilità ai contesti medici, la limitata spiegabilità e gli errori di traduzione che possono influire su diagnosi e decisioni terapeutiche. Le prove suggeriscono che i sistemi di AI non sono pronti per l'uso in contesti ad alto rischio a meno che non siano stati adeguatamente adattati, vincolati e testati per i contesti linguistici e culturali pertinenti.

3.7 Gestione, governance e affidabilità. I responsabili politici devono affrontare un dilemma basato sulle evidenze: devono prendere decisioni di governance dell'AI dalle conseguenze rilevanti con una base scientifica insufficiente ora oppure aspettare le evidenze, quando potrebbe allora essere troppo tardi per intervenire.

Esistono oltre 40 tipi di strumenti di governance attivi nel mondo, ma sono frammentati, concentrati a livello aziendale e raramente misurano la loro efficacia nel mondo reale. La valutazione e la misurazione sono capacità fondamentali per una governance efficace dell'AI, ma sono gravemente poco sviluppate. I rapidi progressi nei sistemi agentici aggravano ulteriormente queste carenze. La prossima generazione di framework di valutazione dovrà essere adattiva, dinamica, a livello di sistema e pertinente al contesto. L'unità di valutazione deve essere quella implementata del sistema, inclusi modello, strumenti, ambiente e utenti, non del solo modello.

Le capacità IA aumentano più rapidamente della capacità di misurarle. La capacità è multidimensionale e attualmente sottovalutata. I quadri standard contano gli input: investimenti, programmi di formazione e istituzioni. Mancano due dimensioni essenziali per una governance efficace dell'IA: ecosistemi abilitanti e sviluppo deliberato delle competenze umane. La capacità deve anche essere intesa come esito del ciclo di vita dell'IA, plasmata dall'impatto dell'IA sulle competenze, sulla dipendenza e sui comportamenti emergenti, non solo come input.

L'IA di punta è concentrata in poche aziende in pochi paesi, sollevando problemi di affidabilità e accessibilità a livello globale. Un accesso diseguale all'IA amplia il digital divide mentre la tecnologia offre anche opportunità per ridurre il divario di sviluppo. Chiudere il divario digitale richiederà nuovi meccanismi per contestualizzare l'intero ciclo di vita dell'IA dalla progettazione alla governance. I sistemi agentici ampliano significativamente il divario di misurazione e governance. Gli agenti agiscono per conto degli esseri umani con un impatto reale diretto, ma le metodologie di supervisione calibrate sulla capacità di azione indipendente e sui comportamenti emergenti di un agente sono poco sviluppate. Le valutazioni esistenti misurano in modo sistematico in modo errato il rischio agentico.

La supervisione umana non è operazionalizzata come requisito misurabile: i rischi multi-agente emergenti non possono essere rilevati tramite una valutazione a singolo agente, e metodi affidabili per mantenere il controllo su sistemi altamente autonomi restano poco sviluppati.

Un approccio bilanciato alla governance dell'IA attingerebbe a un ampio spettro di strumenti, combinando diritto duro, legislazione vincolante, regolamentazione settoriale, sandbox regolamentari, con meccanismi di soft law, codici etici, impegni volontari degli sviluppatori, alleanze industriali, standard tecnici, linee guida approvate dal governo.

Gli strumenti di governance dell'IA attuali sono frammentati, concentrati a livello aziendale e insufficienti. Esistono oltre 40 tipi di strumenti, ma non sono né sistematici né completi e raramente misurano l'efficacia nel mondo reale. Alcuni non hanno strumenti di misurazione; altri misurano solo gli input. Senza misurazione efficace, i rischi di governance diventano simbolici.

Piattaforme per un dialogo strutturato tra sviluppatori di IA all'avanguardia, Stati membri e comunità scientifica sono fondamentali. Tali dibattiti avvengono già ai vertici sull'IA (Bletchley, Seoul, Parigi, Nuova Delhi) o conferenze (World Artificial Intelligence Conference, AI Journey Conference, AI for Good Global Summit). Altri contesti operano accanto a essi: organismi di standardizzazione (Joint Technical Committee 1/Subcommittee 42 sull'Intelligenza Artificiale dell'ISO/IEC, Subcommittee 42 (Intelligenza Artificiale) (ISO/IEC JTC 1/SC 42), Institute of Electrical and Electronics Engineers (IEEE)), l'partnership OECD sull'IA a uso generale, International AI Alliance Network, the AI Safety Institutes Network e Frontier Model Forum, guidato dall'industria – ma ciascuno è tematico, parziale e ad hoc: servono processi più duraturi.

Una piattaforma basata sulle Nazioni Unite è anche un'opzione promettente per ospitare questo dialogo in modo continuo e universalmente inclusivo, in complemento ai luoghi poc'anzi menzionati.

Intelligenza artificiale open-source. L'IA open-source è un pilastro critico del panorama tecnologico moderno e può catalizzare l'innovazione globale distribuita. I modelli IA open-source offrono benefici sociali molto ampi, consentendo agli sviluppatori di ammortizzare risorse computazionali colossali e adattare sistemi avanzati a contesti localizzati. Nel Sud del mondo, i modelli con pesi aperti hanno permesso agli sviluppatori di ottimizzare modelli base altamente capaci per condizioni ambientali locali, abilitando applicazioni cruciali in agricoltura e sanità, come la previsione del rendimento delle colture [367] e l'imaging medico sul posto. Gli artefatti condivisi mitigano anche l'impatto ambientale aggregato dell'IA. Inoltre, la trasparenza strutturale di tali sistemi facilita l'affidabilità pubblica, poiché consente supervisione esterna e auditabilità. Sviluppo storico Lo sviluppo storico dell'IA open-source mostra una trasformazione dai sistemi aziendali chiusi verso un'innovazione distribuita e collaborativa su scala globale. Un precedente di rilievo fu lo sviluppo del modello open-source BLOOM nel 2022, prodotto dal consorzio BigScience [370], seguito dalla pubblicazione della potente famiglia di pesi aperti Llama [371] e dal rilascio della serie Gemma. Un impulso potente è venuto dall'ecosistema alternativo altamente competitivo degli sviluppatori cinesi Qwen e dal rilascio di DeepSeek-V3 e R1. Attualmente, intere regioni contribuiscono attivamente allo sviluppo dell'IA con pesi aperti: Mistral in Europa [376], Falcon negli Emirati Arabi Uniti GigaChat e YandexGPT nella Federazione Russa, insieme a progetti in India Giappone, Repubblica della Corea e altri. Sfide di controllo e governance del rischio I modelli open source ad alta capacità sono difficili da controllare. L'accesso tramite aree riservate o ritirato non è possibile una volta che un modello è stato rilasciato nel dominio pubblico, lasciando la possibilità di un'applicazione malevola (ad esempio, facilitare cyberdanni persistenti, come la generazione automatizzata di malware sofisticato). È necessario un significativo potenziamento delle capacità a livello globale per garantire che le comunità locali possano adattare e valutare in modo sicuro i sistemi di IA. I ricercatori inoltre sostengono costantemente protocolli rigorosi di misurazione e valutazione per valutare la sicurezza di un modello prima della pubblicazione [385]. Organismi internazionali come le Nazioni Unite possono svolgere un ruolo critico nel coordinare standard globali, assicurando che la spinta verso l'innovazione aperta sia bilanciata rispetto alla necessità di parametri di sicurezza armonizzati e di una misurazione solida dei rischi immutabili.

4. Lacune e prossimi passi

4.1. Lacune nelle evidenze La base di evidenze su diversi aspetti dell'IA è disomogenea o insufficiente. Le seguenti sono aree di esempio in cui il Panel non può ancora formulare con sicurezza conclusioni scientifiche

CONCLUSIONI

Macroeconomia e produttività.

La scienza non può ancora affermare con sicurezza se gli incrementi di produttività dell'IA a livello di attività si tradurranno in guadagni a livello dell'intera economia. Le previsioni divergono in modo significativo a causa di assunzioni diverse sull'adozione e sulla creazione di nuove attività. Le evidenze attuali misurano meglio le riduzioni dei costi nelle attività esistenti rispetto a quanto misurino il contributo di AI a nuovi beni, servizi e mercati.

Effetti sul mercato del lavoro. Le ricerche attuali non consentono una conclusione chiara sulla forma degli effetti sul mercato del lavoro. Le evidenze storiche mostrano che le economie possono creare nuovo lavoro, ma non è automaticamente diffuso e di alta qualità, lasciando i lavoratori all'inizio della carriera potenzialmente esposti.

Uso doloso di tecnologie chimiche e biologiche da parte di attori non statali. Studi mostrano che l'IA sta progressivamente abbassando la soglia di competenza per sviluppare e impiegare agenti bioingegnerizzati – tuttavia l'effettiva entità di questo rischio e le condizioni in cui potrebbe dar luogo a pandemie intenzionalmente create, restano poco comprese.

Ambiente e risorse. La rapida espansione dell'IA sta facendo crescere la domanda di infrastrutture digitali, aumentando il consumo di energia e acqua, le emissioni di gas serra, la pressione sulle catene di approvvigionamento di minerali critici e i rifiuti elettronici. Tuttavia, le misurazioni e le rendicontazioni standardizzate su tutto il ciclo di vita dell'IA rimane carente.

La catena di approvvigionamento globale dell'IA. Si estende dalla materia prima e dall'estrazione delle materie prime, alla produzione di chip, la raccolta e annotazione dei dati, l'addestramento dei modelli, fino alle infrastrutture, la distribuzione e smaltimento dell'hardware a livello internazionale. È necessario svolgere ulteriori indagini per comprendere meglio i suoi effetti completi.

1. Efficacia degli strumenti di governance. Mentre il Panel ha inventariato strumenti di governance su livelli aziendali, nazionali e internazionali, l'evidenza della loro efficacia nel mondo reale rimane limitata.
2. Effetti a livello individuale e collettivo. La prova dell'impatto dell'IA sulla cultura, sul benessere umano e sull'autonomia sta emergendo ancora. Il percorso da interazioni IA a livello individuale a esiti a livello sociale come l'erosione epistemica, la partecipazione civica e la coesione sociale rimane poco compreso. Le prove esistenti catturano istantanee – metriche di coinvolgimento, danni documentati, studi di caso – ma non la traiettoria cumulativa.

4.2. Ambito del mandato: Applicazioni militari dell'IA e sistemi letali autonomi.

I sistemi d'arma non sono trattati dal Panel nel presente rapporto. La risoluzione dell'Assemblea Generale 79/325 limita esplicitamente le attività del Panel al dominio non militare: “le attività del Panel e del Dialogo sono limitate al dominio non militare e non si riferiscono all'intelligenza artificiale per scopi militari”. Per questa ragione, i rischi legati all'uso malevolo di tecnologie chimiche e biologiche, per quanto riguarda il dominio militare, sono anch'essi considerati fuori dal mandato del Panel.

4.3. Prossimi passi

Questo rapporto preliminare segna l'inizio, non la fine, del lavoro del Panel. Il Panel continuerà ad approfondire la sua base di evidenze scientifiche attraverso consultazioni strutturate e stretto coinvolgimento con la comunità scientifica. Passi concreti successivi Brief tematici. La risoluzione 79/325 dell'Assemblea Generale prevede esplicitamente l'emissione di brief tematici in aggiunta al rapporto annuale. Il Panel intende emettere brief specializzati su questioni urgenti non appena emergono, tra cui non limitatamente a: IA e ambiente; IA e sicurezza dei bambini; strumenti di governance dell'IA e la valutazione della loro efficacia, briefing settoriali più ampi sulle applicazioni dell'IA nello spazio, nel computing quantistico, nei sistemi legali e giudiziari, nonché mercati finanziari. Contributi dal Dialogo Globale. Il Panel terrà conto degli esiti del Dialogo Globale sulla Governance dell'Intelligenza Artificiale ed è pronto ad assumere nuovi compiti come stabilito dagli Stati membri.

Informazioni sul Panel Scientifico Internazionale indipendente sull'Intelligenza Artificiale.

Composizione e mandato: Il Panel Scientifico Internazionale indipendente sull'Intelligenza Artificiale è stato istituito all'interno delle Nazioni Unite tramite la risoluzione dell'Assemblea Generale 79/325, come indicato nel Global Digital Compact e nel Patto per il Futuro. Composto da 40 esperti indipendenti nominati dall'Assemblea Generale per un mandato di tre anni sulla base della loro eccezionale competenza in AI e campi correlati, il Panel è bilanciato per genere e include membri provenienti da tutti e cinque i gruppi regionali degli Stati membri, coprendo discipline tra cui IA tecnica di base, IA applicata, sicurezza e infrastrutture, e politica, etica e impatto dell'IA. Il Panel ha il mandato di emettere valutazioni scientifiche basate su evidenze, sintetizzando e analizzando la ricerca esistente relativa alle opportunità, ai rischi e agli effetti dell'IA attraverso un rapporto riassuntivo annuale rilevante per le politiche ma non prescrittivo, incluso brevi tematiche se ritenuto necessario. Il lavoro scientifico deve essere guidato da principi di indipendenza, credibilità scientifica e rigore, partecipazione multidisciplinare e inclusiva. Il Panel deve anche presentare il resoconto annuale al Global Dialogue delle Nazioni Unite sulla Governance dell'Intelligenza Artificiale. Informando il Global Dialogue e i processi internazionali più ampi, il Panel consente alla comunità globale di anticipare le sfide emergenti, prendere decisioni di governance migliori informate, e livellare il campo dell'informazione per i decisori politici di tutto il mondo. Il Panel è supportato dal Segretariato del Panel coordinato dall'Ufficio delle Nazioni Unite per le Tecnologie Digitali ed Emergent. Processo verso il presente rapporto Nei tre mesi trascorsi dalla prima riunione del Panel nel marzo 2026, a seguito della nomina nel febbraio 2026, il Panel ha lavorato intensamente per presentare il presente rapporto. Questo intenso processo di scambio scientifico e analisi collettiva ha incluso una plenaria in presenza di tre giorni e oltre 60 riunioni virtuali del Panel, facilitato dai suoi Co-Presidenti eletti. Questa linea di base del rapporto preliminare stabilisce una base per una consultazione più ampia con esperti esterni e successivi brevi tematici e rapporti riassuntivi annuali. Rapporto preliminare 57 Indipendenza scientifica del Panel I membri del Panel prestano servizio in capacità personali come esperti scientificamente indipendenti. Attraverso la loro designazione come esperti delle Nazioni Unite in missione, ciascuno ha dichiarato e promesso di non cercare né accettare istruzioni riguardo all'adempimento dei propri doveri da parte di alcun governo o altra fonte. Le Norme che Regolano lo Status, i Diritti Fondamentali e i Doveri degli Esperti in Missione (ST/SGB/2002/9) includono anche disposizioni relative al loro comportamento e responsabilità.

I due presidenti del Comitato scientifico sono il canadese Yoshua Bengio, Professore alla Université de Montréal e pioniere del Deep Learning e la filippina Maria Ressa, Giornalista, e co-fondatrice di Rappler, Premio Nobel per la Pace. Il gruppo di lavoro e di studio è composto da 19 donne e 21 uomini, secondo un criterio di rappresentanza geografica globale che copre tutte e cinque le macro-regioni ONU.

Girmaw Abebe Tadesse (Etiopia) Tuka Alhanai (Emirati Arabi Uniti) Joëlle Barral (Francia) Tegawendé F. Bissiyandé (Burkina Faso) Loreto Bravo (Cile) Mark Coeckelbergh (Belgio) Carlos Coello Coello (Messico) M. Bilge Demirköz (Turchia) Adjii Bouso Dieng (Senegal) Awa Bouso Dramé (Senegal) Mennatallah El-Assady (Egitto) Hoda Heidari (Iran) Juho Kim (Corea del Sud) Anna Korhonen (Finlandia) Aleksandra Korolova (Stati Uniti) Sonia Livingstone (Regno Unito) Qinghua Lu (Australia) Bilal A. Mateen (Regno Unito) Joyce Nakatumba-Nabende (Uganda) Andrei Neznamov (Federazione Russa) Maximilian Nickel (Germania) Rita Orji (Nigeria) Román Orús (Spagna) Alvitta Ottley (Stati Uniti) Lorenzo Pacchiardi (Italia) – Ricercatore presso l'Università di Cambridge. Martha Palmer (Stati Uniti) Johanna Pirker (Austria) Balaraman Ravindran (India) Lior Rokach (Israele) Piotr Sankowski (Polonia) Silvio Savarese (Italia) – Esperto di computer vision ed Executive Vice President. Bernhard Schölkopf (Germania) Haitao Song (Cina) – Presidente dell'Istituto di Intelligenza Artificiale dell'Università Jiao Tong di Shanghai. Leslie Teo (Singapore) Katarzyna Wac (Polonia) Jian Wang (Cina) – Accademico ed esperto di Cloud Computing.

Hjalmar Wijk (Svezia) Nicola Williams (Giamaica)

I due italiani nel comitato scientifico nominati da Guterres sono Lorenzo Pacchiani e Silvio Savarese.

2 (34) 1000000

ALL INFORMATION CONTAINED HEREIN IS UNCLASSIFIED

ALHIS-C I TWEE AON JHEM IDB HARE VII ZCWDHLI BHEL BEA 36 FBL
ZTANCBZ SSSSSSSS CVOLEP LILLYVITINRSCORR000002COOP19VTREK3 CBWL

[illegible]

6

EP7 B7592D 3A 25 A 6 87B31519075 9C0D

100

Downloaded from <http://ajphaphysiol.org/> at University of California, San Diego on September 11, 2015

PLAY

CONVULSION

JARVIS

L'ALGORITMO di Glovo si chiama «Jarvis».

Da quando l'antitrust italiana ha indagato Glovo per social washing Jarvis è diventato il simbolo del lato oscuro della GIG Economy. E del mondo dei rider in particolare.

L'Antitrust ha avviato un'istruttoria nei confronti di alcune società del gruppo Glovo (Glovoapp, Foodinho e Glovo Infrastructure Services Italy) e un'istruttoria nei confronti di Deliveroo Italy per possibili condotte illecite nell'attività relativa all'offerta di servizi di consegna a domicilio di prodotti alimentari. Nella nota dell'Autorità si dice che le aziende avrebbero comunicato informazioni ingannevoli riguardo al loro impegno etico e alla loro responsabilità sociale verso i rider. Le comunicazioni rivolte ai consumatori, il rispetto di standard etici e la responsabilità sociale, non corrisponderebbero al vero, in particolare per quanto riguarda le condizioni di lavoro e il rispetto della legalità nella gestione dei rider. L'antitrust si riferisce soprattutto al modello operativo e all'algoritmo utilizzato dalle due società.

JARVIS

Jarvis nella tradizione anglosassone è il nome per antonomasia del maggiordomo della casa padronale, un pò come il nostro Ambrogio nella leggendaria pubblicità dei cioccolatini o come Alfred in Batman. Addirittura l'attaccapanni a stelo, il coat rack, delle case dell'aristocrazia inglese veniva chiamato Jarvis.

Jarvis è anche l'assistente artificiale superintelligente di Iron Man, al secolo il magnate Tony Stark che lotta per l'umanità

In questo caso Jarvis è anche l'acronimo di Just A Rather Very Intelligent System, letteralmente "Solo Un Sistema Piuttosto Veramente Intelligente". Il maggiordomo domestico della famiglia Stark è un'IA progettata e creata dallo stesso Tony Stark.

Iron Man, creato dallo sceneggiatore Stan Lee, è un esempio quasi didattico di come potrebbe funzionare l'integrazione tra uomo e macchina nell'era della singolarità di cui parlano i post-umanisti. La prima apparizione dell'uomo d'acciaio è del marzo 1963, editore Marvel Comics. Stan Lee inventa Tony Stark ispirandosi al magnate ed inventore americano Howard Hughes. Stark è geniale e miliardario, è un inventore filantropo. Viene ferito al cuore in Vietnam (nella versione cinematografica più recente viene colpito in Afghanistan), e costruisce un'armatura tecnologica per sopravvivere alla malattia dovuta al suo infortunio. Stark diventa Iron Man, ex mercante di armi ed eroe filantropo

Questa saga di Stan Lee è stata spesso oggetto di saggi e di analisi per le sue implicazioni etiche legate alla tecnologia, alla responsabilità aziendale e alla guerra. Nelle storie di Iron Man c'è il rapporto tra l'uomo e la tecnologia e la distanza tra il potere e la responsabilità di chi deve decidere quale uso fare della tecnologia.

Iron Man/Tony Stark rappresenta un esempio straordinario di "fusione" uomo-macchina e di intelligenza artificiale singolare.

La fusione tra uomo e silicio è rappresentata dall'armatura, un esoscheletro ad alta tecnologia che trasforma un essere umano organico in un'entità invincibile.

Stark si interfaccia direttamente con l'armatura intelligente tramite impulsi neurali e sensori, integrando in questo modo la tecnologia nel proprio corpo. C'è infine il tema dell'ingegneria biomedica: Stark sopravvive alla malattia solo grazie ad un reattore installato vicino al cuore che funge da pacemaker avanzato e fonte di energia. L'unione tra biologia e ingegneria in questo caso è indissolubile. Jarvis inoltre il suo assistente digitale non è solo uno strumento ma una IA capace di apprendimento continuo, linguaggio naturale, rilevamento emotivo e pensiero autonomo, una intelligenza "singolare" tecno-psichica che si evolve fino a diventare parte della simbiosi uomo-macchina evocata dai filosofi e dagli scienziati post-umanisti.

Iron Man è il simbolo di una simbiosi perfetta, dove la tecnologia del silicio non sostituisce l'uomo, ma ne potenzia esponenzialmente le capacità organiche, creando una nuova forma di intelligenza ibrida. Stark fornisce l'intuizione creativa e l'empatia, l'armatura e Jarvis forniscono forza e resistenza fisica e potenza di calcolo.

La figura di Tony Stark rappresenta l'apice del concetto di post-umanesimo. Iron Man incarna la transizione da essere solo biologico a entità ibrida, nel quale il confine tra corpo e circuiti svanisce e l'uomo si fonde con il silicio. Ma a differenza di un pilota che guida un mezzo, Stark "indossa" l'intelligenza. L'armatura è un sistema nervoso esteso che non risponde a comandi fisici, ma ad impulsi cerebrali.

Questa fusione genera nuovi sensi, visione termica, radar, analisi dati, che il suo cervello elabora come se fossero "naturali".

JARVIS è una forma di "IA Cooperativa" singolare ed un'estensione cognitiva, una parte della mente di Tony che gestisce la complessità che un uomo solo non potrebbe reggere, secondo un ciclo di miglioramento infinito: Stark usa l'IA per progettare IA migliori, accelerando a sua volta l'evoluzione tecnologica.

Il culmine di questa fusione si raggiunge con le nanotecnologie nel sangue: un procedimento in cui la tecnologia risiede all'interno del corpo ed interviene per guarire e potenziare la biologia organica. L'uomo diventa di fatto la macchina; il silicio non è più un guscio esterno, ma una componente del DNA digitale del protagonista.

Il parallelo tra Ray Kurzweil (il "profeta" della Singolarità) e Iron Man (Tony Stark) profetico me quasi inquietante: Stark è la messa in scena cinematografica e fumettistica delle teorie che Kurzweil sostiene da decenni.

La Legge dei Ritorni Accelerati

Kurzweil sostiene che il progresso tecnologico non è lineare ma esponenziale, e predice che entro il 2045 la potenza di calcolo supererà la capacità combinatoria di tutti i cervelli umani. L'evoluzione delle armature di Iron Man segue questa logica incrementale: dalla prima armatura, la Mark 1, costruita di ferro grezzo durante la sua prigionia, alla nanotecnologia della Mark 85 passano solo 15 anni. Tutte le armature, tranne la prima, sono progettate da un'IA, cioè da JARVIS, che a sua volta è stata migliorata da Tony Stark, creando un ciclo di feedback infinito che accelera l'invenzione tecnologica.

Per Kurzweil, il corpo biologico è una sorta di "Hardware 1.0", limitato e fragile. Kurzweil prevede l'uso di nanobot nel flusso sanguigno per riparare organi, distruggere malattie e allontanare la morte (Longevity Escape Velocity).

L'idea centrale di Kurzweil passa dall'estensione della neocorteccia, non verremo sostituiti dall'IA ma che diventeremo l'IA. Entro il 2030, secondo Kurzweil collegheremo la nostra neocorteccia direttamente al cloud tramite interfacce cervello-computer, espandendo la nostra intelligenza, sostiene Kurzweil, di un milione di volte.

Nelle saghe Extremis e Bleeding Edge, Tony Stark inietta una nanotecnologia nel proprio corpo. L'armatura non è più un guscio esterno, ma risiede nelle sue ossa e si manifesta attraverso i pori della pelle, curandolo in tempo reale e rendendolo "post-umano". Stark non usa un joystick per muovere la sua armatura di ferro, che è un'estensione del suo sistema nervoso. Quando indossa l'elmo, il suo cervello è connesso a satelliti, database globali e algoritmi predittivi. La sua mente "pensa" alla velocità del silicio.

La "Singolarità" nella versione di Stan Lee è Iron Man, un uomo con la sua Armatura. Si avvicina moltissimo all'idea di Kurzweil, che definisce la Singolarità come il punto in cui l'uomo e la macchina diventano indistinguibili. Tony Stark è il primo essere umano a vivere la Singolarità in netto anticipo sui tempi. Senza la tecnologia Tony morirebbe per via delle schegge nel cuore, e senza Tony l'armatura resterebbe solo un pezzo di metallo. Insieme Tony, la sua armatura e Jarvis formano un'entità che Kurzweil chiamerebbe "Intelligence 2.0"

Nota Curiosa: Ray Kurzweil, che oggi lavora in Google come direttore dell'ingegneria, è considerato da molti il "vero" Tony Stark per la sua capacità di prevedere tecnologie (come gli assistenti vocali o le auto a guida autonoma) decenni prima della loro comparsa.

Jarvis a casa Zuckerberg

Al Jarvis di Stan Lee si ispirò anche Mark Zuckerberg, che 10 anni fa tentò di creare un suo assistente digitale personale ispirandosi al famoso super eroe.

In un post sul suo social Zuckerberg immaginava una IA che materializzava le mappe, scaldava l'automobile, preparava il caffè e controllava i parametri vitali, esattamente com il Jarvis di Tony Stark,

Il Jarvis di Zuckerberg era “un’AI molto semplice che potesse governare la mia casa”, una sfida personale che non è mai diventata un prodotto commerciale, ma che ha segnato comunque un passo importante nell'evoluzione delle tecnologie di domotica e intelligenza artificiale di Meta, un esercizio per imparare di più sul machine learning, l'elaborazione del linguaggio naturale e il riconoscimento facciale

Roma 6 maggio 2026

L'Antitrust comunica l'apertura di una indagine su Glovo per “comunicazione ingannevole” ai clienti, un provvedimento che potrebbe segnare un ulteriore punto di svolta nel settore del food delivery in Italia.

Al centro dell'istruttoria oltre al tema già noto dell'efficienza “tecnica” del software, c'è quello del social washing: l'antitrust accusa le aziende di promuovere un'immagine “etica” e una politica di tutela dei lavoratori che non corrisponde alle condizioni operative imposte dai loro algoritmi.

Per capire il senso di questa istruttoria bisogna prima capire come funziona Jarvis, l'algoritmo della piattaforma di Glovo.

Jarvis è un sistema di gestione integrale delle risorse umane, un esempio da manuale di Algorithmic Management. Qualunque azione orientata alla gestione del sistema operativo è anche finalizzata a modellare e orientare il comportamento del lavoratore attraverso un sistema di punteggi, ed “incentivi” che spesso sono coercitivi. Nel momento in cui Jarvis assegna un ordine ad un rider ne modella anche il comportamento. Jarvis assegna al rider un punteggio che si compone di diversi parametri. Il primo è la “Seniority”, cioè il numero di consegne effettuate nella sua “carriera”, da quando ha incominciato a lavorare per la piattaforma.

Il secondo è il cosiddetto “High Demand”, ovvero la disponibilità a lavorare negli "slot" in cui la richiesta è più alta, il weekend o le serate piovose ad esempio.

C'è poi il “Customer Rating” ottenuto dai diversi feedback diretti dei clienti, e l'affidabilità, il “No-show”, che implica una penalità per chi prenota un turno e non effettua il check-in.

Il sistema quindi non assegna gli ordini a caso: chi ha il punteggio più alto visualizza i turni cioè gli slot prima degli altri rider, e cosa molto importante che aiuta a capire la potenza del sistema, riceve gli ordini più redditizi, secondo le stime e l'elaborazione dei dati fatta dall'algoritmo.

Jarvis è un “dispositivo” in grado di crea una vera e propria gerarchia tra i rider e di renderla invisibile ai lavoratori, i quali sanno solo che più lavorano più hanno accesso alle opportunità migliori. Non conoscono le performance degli altri né possono contestare in alcun modo l'assegnazione del punteggio o le modalità di ingaggio della piattaforma, sanno solo che chi non accetta ordini "scomodi" vede scendere il proprio rating fino a sparire dalla piattaforma e a perdere di fatto la possibilità di lavorare.

La “Connessione” e la “Disconnessione”

La logica algoritmica, non solo quella di Glovoo, si basa da un lato sul concetto di autonomia del lavoratore o del fornitore, dall'altro sul controllo del comportamento del lavoratore. Il suo fine è garantire le consegne e la “qualità del servizio, a prescindere dal lavoratore, che è un collaboratore non un dipendente della piattaforma. Autonomia e controllo però sono due aspetti poco coerenti tra loro.

Formalmente il rider è libero di scegliere se e quando “loggarsi” alla piattaforma, cioè connettersi e incominciare a lavorare. Ma tecnicamente, se non si connette negli orari “giusti” quelli di punta, o comunque se non si connette il più possibile, “perde punti” e posizioni nel ranking. Il sistema calcola poi anche il tasso di accettazione, per cui rifiutare un ordine entro i 30 secondi di tempo “concessi” riduce drasticamente le probabilità di riceverne un altro nel breve periodo.

Molto interessante anche il linguaggio delle piattaforme e il mondo del lavoro che raccontano. “Connettersi” o logarsi ad esempio equivale a presentarsi virtualmente sul posto di lavoro per poi lavorare nel concreto, Virtuale e reale sono due aspetti complementari del mondo del rider.

La Disconnessione decisa dalla piattaforma in questo universo semantico equivale ad un vero e proprio “licenziamento” o allontanamento “algoritmico”: Jarvis può disattivare un account senza alcun intervento umano, senza dover dare spiegazioni e senza possibilità di interlocuzione o replica. Se i parametri scendono sotto una soglia critica o se il sistema rileva anomalie, ritardi ripetuti, mancate risposte, presenza molto bassa, il rider viene disconnesso.

Il sistema è asimmetrico perché comanda l'algoritmo che garantisce il servizio con le risorse umane che ha e che monitora continuamente, senza alcuna interazione ma solo analizzando dati, input e output. Nei processi di gestione delle risorse automatizzati, il lavoratore non può giustificare imprevisti, ad esempio gli incidenti o i guasti alla bicicletta, o la malattia.

L'aspetto forse più paradossale di questo sistema è quello dello storytelling con cui la piattaforma descrive il suo algoritmo. Nel blog del sito ufficiale di Glovoo la rappresentazione di Jarvis è per l'antitrust un capolavoro di “Social Washing”, lontanissimo da un modello di business che, nei fatti, di etico non ha nulla. Anzi, esattamente come altri algoritmi, da Frank, il padre di Jarvis, a quello dei social media, TikTok, Meta e AlphaBet, ha criticità strutturali proprio nel suo funzionamento.

Glovo e Deliveroo hanno investito negli ultimi tempi in campagne importanti sulla “sicurezza dei rider” e sulla “flessibilità” positiva di questo lavoro. L'Antitrust in linea con le sentenze dei tribunali di Bologna e Milano, sostiene che queste comunicazioni siano di fatto ingannevoli, perché i consumatori vengono indotti a scegliere queste piattaforme anziché altre sulla base della convinzione di sostenere e premiare una condotta equa. L'algoritmo invece secondo le sentenze continua a operare su logiche più simili allo sfruttamento che non a quello del rapporto con fornitori realmente autonomi. La piattaforma fa credere ai rider di essere autonomi e alla pari, ma non è così secondo l'autorità. Il servizio non è un flusso naturale e privo di conflitti sociali come vorrebbe la narrazione di Glovo: l'algoritmo è programmato per arrivare alla “Frictionless Automation”, all'automazione senza attriti della gestione sia del lavoro sia delle risorse umane.

Al di là delle sentenze il tema è molto controverso. La GIG economy non nasce per lavori e prestazioni lavorative subordinate e continuate, e il rider è un mestiere che si può fare se e quando serve. Chiaramente se diventa un lavoro full time le cose cambiano. Ma la sensazione è che la situazione sia sfuggita di mano anche alla piattaforma, che gestiscono un lavoro che “vale” poco e non fanno nella modalità Gig Economy.

Il caso dei rider professionisti evidentemente impone una regolamentazione, e non solo quello, si pensi alla logistica ad esempio, o alle molte imprese che hanno una gestione delle risorse ormai algoritmica a livello sistemico. Con i relativi problemi di tipi legislativo e sindacale, che fatica a trovare una mediazione efficace con le corporee del delivery, non potendo “vedere” e tantomeno discutere i “codici sorgente” per capire le modalità con cui vengono gestiti i “log” per verificare se esista una “black box”, una scatola nera, di Jarvis che penalizza i lavoratori che tentano di esercitare i propri diritti sindacali o di salute.

Frank e Jarvis

Dal punto di vista tecnico e dell'impatto sul diritto del lavoro, il confronto tra Frank, l'algoritmo di Deliveroo, e Jarvis, quello di Glovo, mostra come si sia modificata ed evoluta la gestione del lavoro e la sorveglianza “digitale”. Frank è stato il “pioniere” di un sistema che portava per forza un alto livello di discriminazione, Jarvis è l'evoluzione più sofisticata di questo dispositivo: Jarvis è l'estensione sistemica di Frank. I due algoritmi sono architetture con molti punti in comune, ma con modalità di gestione differenti. Frank aveva una natura “punitiva”, cioè selezionava sulla base di una discriminazione statistica e penalizzava chi non era disponibile in determinati slot, senza distinguere tra chi scioperava, chi era malato o chi assisteva un disabile.

È stato infatti il primo algoritmo in Italia a essere condannato per discriminazione, nel 2020 dal Tribunale di Bologna, perché applicava la stessa metrica a situazioni giuridicamente diverse. Era tuttavia un sistema relativamente “semplice” e altrettanto semplice era il suo messaggio implicito: “se non ci sei quando serve a me, vali meno”.

Jarvis rappresenta una evoluzione: non cambia la logica, ma il sistema diventa molto più raffinato, e l'algoritmo di Glovo è più “pericoloso” di Frank per la sua natura sistemica e predittiva. Jarvis non si limita a guardare se il lavoratore è presente, ma ne monitora il comportamento in tempo reale, calcola la velocità di accettazione, il tempo di attesa al ristorante, ed il feedback del cliente. Esercita un controllo indiretto creando una competizione interna tra i rider spietata: l'abbassamento del ranking significa non avere più accesso al lavoro, una sorta di licenziamento “bianco” gestito esclusivamente dal codice. Si consideri che per tenere alto il ranking si deve essere performanti e sul pezzo sempre, in un sistema ipercompetitivo in cui non si conoscono i propri competitori. Un tutti contro tutti di avatar digitali di persone reali.

Jarvis esattamente come Frank usa dinamiche da videogioco per spingere il rider ad accettare ordini sottocosto pur di non perdere la propria posizione in classifica, ma com'è detto in precedenza in maniera sistemica e conoscendo quasi sempre in precedenza quali saranno le risposte e le reazioni di chi “pedala”. Tutto questo grazie soprattutto ad una asimmetria informativa straordinaria tra l'algoritmo e il lavoratore.

Frank è il primo esperimento di gestione dei lavoratori della Gig Economy, ed è un algoritmo abbastanza “grezzo”, facilmente contestabile nell'ottica del diritto del lavoro. Jarvis è un avversario politico e sindacale molto più ostico perché nasconde il controllo e l'organizzazione del lavoro delle persone dietro l'efficienza, il gaming e la reputazione. Frank agiva come un cronometro, Jarvis è un supervisore invisibile e i suoi parametri di valutazione consistono nel rapporto tra la performance, e gli obiettivi della piattaforma.

L'attuale indagine dell'Antitrust punta a dimostrare che Jarvis non è un dispositivo neutro che si limita ad ottimizzare i tempi, ma è uno strumento di gestione della forza lavoro che aggira le tutele nazionali e guadagna consenso e mercato attraverso una comunicazione lontana dalla realtà. L'accusa di social washing va in questa direzione. Se l'azienda dichiara pubblicamente di garantire libertà e tutele, ma il suo algoritmo è programmato per penalizzare chi si disconnette o chi non accetta ordini poco redditizi, si configura un comportamento ingannevole verso il mercato e i consumatori.

La vicenda italiana si inserisce in un contesto europeo di forte pressione normativa e giudiziaria. La "battaglia contro l'algoritmo" non è un caso italiano ma più in generale europeo.

In prima fila c'è la Spagna che è anche il paese dove le piattaforme sono nate.

La Ley Rider spagnola è la prima in Europa. Stabilisce la "presunzione di subordinazione" in virtù della quale i rider sono considerati dipendenti e non autonomi, e obbliga le aziende alla trasparenza algoritmica, cioè ai sindacati è garantito il diritto di conoscere i parametri che influenzano il lavoro dei rider.

Tra il 2023 e l'inizio del 2025, Glovo ha accumulato multe per oltre 200 milioni di euro per l'utilizzo di rider considerati "falsi autonomi". Le autorità spagnole hanno inoltre accusato l'azienda di aver cercato di aggirare la legge modificando Jarvis senza però cambiare la sostanza del controllo sui lavoratori.

La UE ha approvato ad ottobre 2024 una direttiva (2024/2831), che introduce una "presunzione legale di lavoro subordinato" a livello europeo. Gli stati membri, compresa l'Italia, hanno tempo fino al dicembre 2026 per recepire queste norme.

La direttiva proibisce espressamente agli algoritmi di prendere decisioni critiche come il licenziamento, o nel caso specifico la disconnessione senza un supervisore "umano".

Nel giugno 2025, la Commissione Europea ha inflitto una multa storica da 329 milioni di euro a Delivery Hero (la casa madre di Glovo) e alla stessa Glovo, con l'accusa che le aziende del delivery si sarebbero accordate per non rubarsi i rider a vicenda e per fissare compensi uguali per tutti, scambiandosi informazioni sensibili tramite i rispettivi board. Questa multa è storica perché rappresenta il primo caso in cui l'UE ha sanzionato un "cartello" nel mercato del lavoro digitale.

In Francia e in Belgio le autorità per la concorrenza e i tribunali del lavoro si sono focalizzate sulle pratiche anticartello, focalizzandosi su come gli algoritmi standardizzano i prezzi a danno della libera concorrenza tra i lavoratori.

L'Italia con l'antitrust sulla comunicazione ingannevole ha aperto un nuovo fronte. La Spagna si è concentrata per prima sul modello contrattuale, l'Italia sul social washing dell'azienda e sulla comunicazione ingannevole.

Oltre l'Oceano c'è poi il caso del Canada, con una ulteriore particolarità: Glovo si è ritirata dal mercato nordamericano anni fa ma i suoi concorrenti, Uber Eats, DoorDash e SkipTheDishes, usano le dinamiche algoritmiche e sono finite sotto la lente d'ingrandimento delle autorità per gli stessi motivi di trasparenza e pratiche ingannevoli che hanno mosso le autorità europee.

Il governo canadese si sta concentrando sulla trasparenza Algoritmica. Lo stato dell'Ontario ha fatto da apripista con il Digital Platform Workers' Rights Act, pienamente operativo dal 1° gennaio 2026. La legge prevede il "Diritto alla spiegazione": i rider canadesi hanno ora il diritto legale di sapere come l'algoritmo assegna gli ordini e come vengono calcolati i punteggi. Sono vietate le "Penalità Occulte", cioè le piattaforme non possono più penalizzare i rider che rifiutano ordini senza una comunicazione chiara. Di fatto una soluzione concreta rispetto al problema della "connessione forzata" che abbiamo visto in Italia.

Infine analogamente alla direttiva UE, le decisioni di disconnessione, cioè del licenziamento algoritmico, devono essere revisionate da una persona e non possono essere lasciate esclusivamente al software.

L'Antitrust Canadese, il Competition Bureau, ha intensificato quest'anno un'azione legale contro DoorDash, accusato di drip pricing, (letteralmente "prezzi a goccia").

Il Drip Pricing funziona in questo modo.

Le piattaforme mostrano un prezzo iniziale basso, aggiungendo commissioni e "fee" solo nell'ultima schermata di pagamento. Per l'Antitrust canadese questo sistema è una pratica commerciale ingannevole e da sanzionare perché l'azienda comunica un valore ma l'algoritmo ne applica un altro.

A differenza della Spagna che ha imposto il modello subordinato, il Canada, da quest'anno, sta sperimentando una via "ibrida", secondo la quale i rider restano autonomi ma con standard minimi garantiti: una sorta di salario minimo di 20 Dollari canadesi (circa 12 Euro) l'ora per il tempo di connessione, una copertura per infortuni sul lavoro e un'indennità per i costi del veicolo. Questi vincoli a tutela dei rider servono di fatto a "depotenziare" l'algoritmo: se l'azienda deve pagare un minimo orario, l'algoritmo perde interesse a "tenere in attesa" centinaia di rider senza assegnare loro ordini.

In sintesi il Canada si concentra sulla trasparenza del dato: l'algoritmo deve essere una "scatola di vetro" e i costi devono essere chiari fin dal primo clic.

Tornando dalle nostre parti, il decreto lavoro del primo maggio 2026 ha imposto la "trasparenza algoritmica" e ha introdotto la presunzione di subordinazione per i lavoratori delle piattaforme digitali quando l'algoritmo controlla le mansioni. Le aziende devono garantire la supervisione umana sulle decisioni automatizzate, ad esempio i licenziamenti o le sanzioni, conservando i dati dei sistemi di gestione per almeno 5 anni. Per quanto riguarda la gestione del personale tramite strumenti automatizzati o Intelligenza Artificiale le nuove Linee guida per l'AI nel lavoro si compongono di quattro principi chiave:

Il primo è la presunzione di subordinazione: Se l'algoritmo gestisce, valuta o limita l'accesso al lavoro in modo rigido, la cosiddetta eterodirezione, il lavoratore può essere considerato a tutti gli effetti un dipendente subordinato, con tutte le tutele del caso.

Il secondo è il diritto al riesame umano: le decisioni che incidono sul rapporto di lavoro (turni, assegnazione incarichi, sospensioni o licenziamenti) non possono essere prese esclusivamente dall'algoritmo, il lavoratore ha diritto a richiedere l'intervento di una persona in carne e ossa che possa riesaminare la decisione.

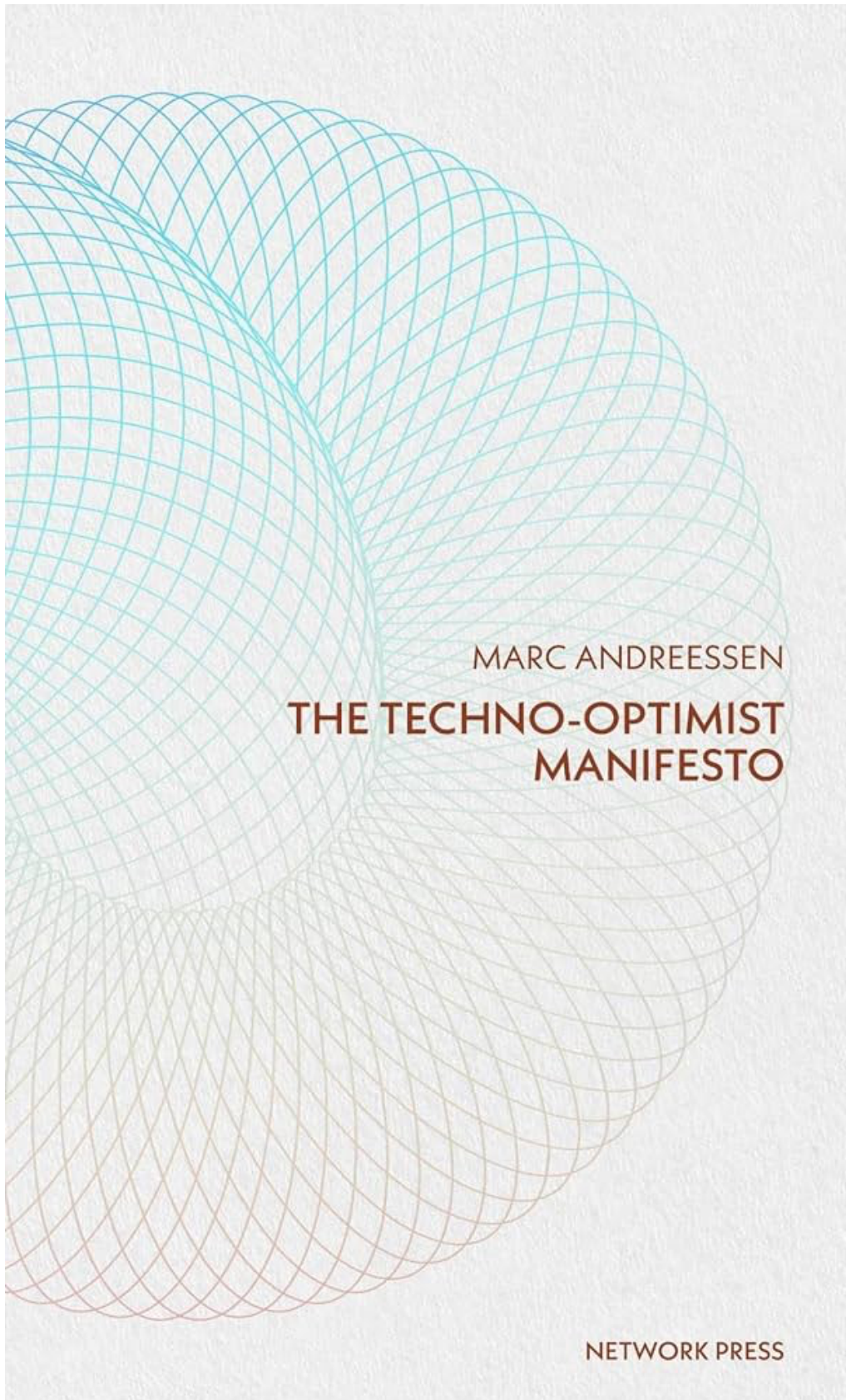
Il terzo è l'obbligo di trasparenza e di informazione corretta: i datori di lavoro devono informare preventivamente i lavoratori sull'utilizzo di sistemi decisionali o di monitoraggio automatizzati, e i dipendenti devono sapere quali parametri l'algoritmo utilizza per valutare le loro performance.

Infine per quanto riguarda la conservazione e la protezione dei dati le piattaforme sono obbligate a conservare i dati relativi all'attività dei lavoratori per 5 anni e a metterli a disposizione degli organi ispettivi.

Queste norme mirano a contrastare il cosiddetto "caporalato digitale", assicurando che la tecnologia non aggiri le tutele del diritto del lavoro e non generi discriminazioni nelle valutazioni del personale.

In Italia, con l'entrata in vigore del Decreto Lavoro del 30 aprile 2026, questi algoritmi devono ora rispettare nuovi obblighi di trasparenza. E soprattutto le aziende sono tenute a informare non solo i lavoratori ma anche i rappresentanti sindacali su quali parametri influenzano la rotazione dei turni o l'assegnazione degli straordinari. Il lavoratore ha ora il diritto alla spiegazione: se l'algoritmo assegna sempre i turni peggiori alla stessa persona, l'azienda deve poter dimostrare che non vi sia una discriminazione automatizzata. E al sindacato è data la possibilità di contrattare l'algoritmo con l'azienda, uno dei cavalli di battaglia dei sindacati in Italia e in Europa degli ultimi anni.

I TECNO DEVOTI



MARC ANDREESSEN

THE TECHNO-OPTIMIST MANIFESTO

NETWORK PRESS

Il Manifesto del Tecno-Ottimismo: Visione, Pilastri e Critiche

Nell'ottobre del 2023, Marc Andreessen — noto imprenditore e co-fondatore della società di Venture Capital Andreessen Horowitz (a16z) — ha pubblicato un intervento intitolato Il Manifesto del Tecno-Ottimismo (The Techno-Optimist Manifesto). Il testo si inserisce come una presa di posizione nel dibattito globale sul ruolo dell'intelligenza artificiale e del progresso tecnologico nella società contemporanea.

1. La Tesi Centrale

Il manifesto propone una visione fortemente favorevole allo sviluppo tecnologico, ponendosi in aperta contrapposizione con le correnti di pensiero che guardano con scetticismo o cautela all'innovazione. La tesi sostenuta da Andreessen è che la tecnologia rappresenti il motore primario del benessere umano, della crescita economica e dell'avanzamento della civiltà. Secondo questa prospettiva, non vi sono problemi che la tecnologia non possa risolvere, a condizione che l'innovazione non sia ostacolata da regolamentazioni stringenti o da filosofie contrarie alla crescita.

2. I Pilastri Fondamentali

La visione presentata nel testo si articola attorno ad alcuni punti chiave:

La Crescita Economica come Imperativo Moralmente Necessario: La crescita della produttività, guidata dal progresso tecnologico, viene indicata come l'unico strumento efficace per ridurre i costi, generare abbondanza e sradicare la povertà.

Il Ruolo del Libero Mercato: Il mercato libero è descritto come il meccanismo più efficiente per la scoperta, il test e l'adozione delle nuove tecnologie, premiando l'innovazione in grado di generare valore.

L'Abbondanza Energetica: Il manifesto sostiene la necessità di un significativo incremento della produzione di energia, indicando nell'energia nucleare (sia a fissione che a fusione) la fonte indispensabile per sostenere lo sviluppo di una civiltà avanzata.

L'Intelligenza Artificiale: L'IA viene definita come uno strumento di potenziamento dell'intelletto umano, in grado di contribuire alla risoluzione di problemi complessi nei campi della medicina, della scienza e dell'istruzione. Di conseguenza, il manifesto considera negativamente qualsiasi tentativo di rallentarne lo sviluppo.

3. Il Concetto di "Decelerazione" e gli Avversari

Il testo individua chiaramente un blocco di idee e istituzioni definite "Decelerazioniste", viste come un ostacolo al progresso:

Burocrazia e Regolamentazione: Considerati fattori di rallentamento della velocità di innovazione.

Pessimismo Ambientale: La prospettiva secondo cui l'umanità dovrebbe limitare i consumi per preservare le risorse del pianeta.

Etica della Prevenzione: L'approccio che subordina lo sviluppo di una tecnologia alla previa dimostrazione della sua assoluta sicurezza.

4. Le Correnti di Pensiero di Riferimento

Il manifesto si posiziona all'incrocio di diverse correnti ideologiche, politiche e filosofiche emerse nel contesto contemporaneo e della Silicon Valley:

Accelerazionismo Efficace (e/acc): Movimento sviluppatosi nella cultura tech che sostiene l'accelerazione incondizionata dello sviluppo tecnologico e capitalistico come unico modo per guidare l'evoluzione umana ed evitare il declino della civiltà.

L'Ideologia Californiana (Tecno-Liberismo): Corrente del pensiero politico ed economico che fonde il liberismo di mercato, l'individualismo di matrice americana e l'utopismo tecnologico, rigettando l'intervento statale e le regolamentazioni.

Tecno-Soluzionismo e Prometeismo: La convinzione epistemologica e filosofica secondo cui ogni problema umano, sociale, ambientale ed economico possa e debba essere risolto mediante l'ingegneria e l'invenzione tecnica.

Transumanesimo: Prospettiva filosofica basata sull'idea di impiegare le tecnologie emergenti (tra cui l'IA e le biotecnologie) per superare i limiti biologici e cognitivi della specie umana.

5. Riferimenti Filosofici e Valori

Nel manifesto vengono citati pensatori e figure storico-culturali quali Filippo Tommaso Marinetti (esponente del Futurismo), Friedrich Nietzsche e il fisico Richard Feynman. I valori centrali promossi sono la ricerca della verità attraverso la tecnologia, la meritocrazia e l'ambizione orientata alla costruzione del futuro.

6. Le Principali Critiche

La pubblicazione del testo ha generato un ampio dibattito formale, sollevando diverse argomentazioni critiche:

Tecno-Soluzionismo: Viene contestata la tendenza a considerare la tecnologia come una soluzione autosufficiente per questioni d'ordine sociale, politico ed economico.

Dimensioni Etiche e di Rischio: Viene evidenziata la limitata attenzione dedicata dal manifesto ai potenziali rischi associati alle tecnologie emergenti (quali bias negli algoritmi, impatti sull'occupazione e sistemi di sorveglianza).

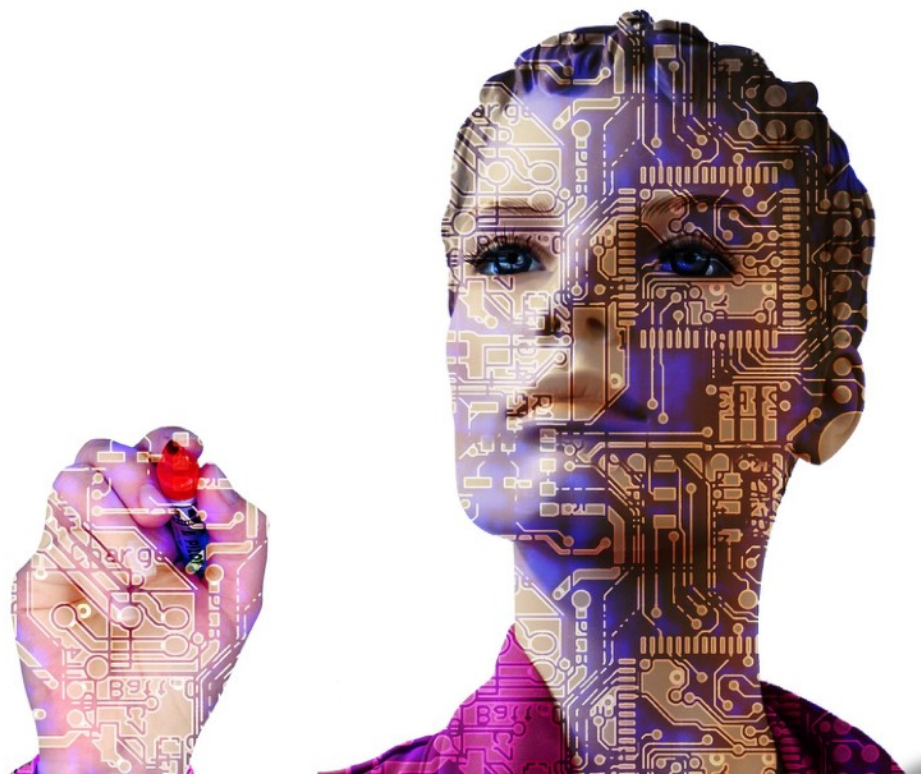
Limiti delle Risorse Fisiche: Diversi osservatori scientifici sottolineano l'incompatibilità teorica tra un modello di crescita infinita e la natura finita delle risorse del pianeta.

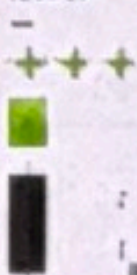
Struttura Binaria del Discorso: Il tono del manifesto è stato descritto come fortemente polarizzante, dividendo il panorama culturale in due posizioni contrapposte (a favore o contro il progresso).

Conflitto di Interessi Economici: Alcune analisi sottolineano la coincidenza tra le tesi del manifesto e gli interessi finanziari tipici del settore del Venture Capital, incline alla richiesta di mercati deregolamentati.

Conclusione

Il Manifesto del Tecno-Ottimismo rappresenta un documento ideologico e politico che sintetizza una precisa corrente del pensiero tecnologico contemporaneo, ponendo la velocità di innovazione, la capacità computazionale e la produzione energetica come risposte primarie alle sfide del nostro tempo.





LOREM IPSUM VARIORUM

LANGUAGE WAS
NEVER DESIGNED
TO BE **FOREVER**.
IT LEAKS.
IT **DRIFTS**.
IT SURVIVES
THROUGH
MISTRANSLATION.

A SENTENCE COPIED
OFTEN ENOUGH BEGINS
SHEDDING THE CONDITIONS
OF ITS ORIGIN UNTIL ONLY
RHYTHM REMAINS:
PRESSURE WITHOUT SOURCE,
INTIMACY WITHOUT
VERIFICATION.

LOREM IPSUM IS WHAT
HAPPENS WHEN MEANING
SURVIVES ITS OWN
DISSOLUTION.

THE ARGUMENT
COLLAPSES. | THE STRUCTURE
KEEPS
SPEAKING. 

WE CONTINUE ANSWERING IT
INSTINCTIVELY,
THE WAY THE BODY
BREATHES, THE WAY
A HEART BEATS.



PERHAPS THIS IS
WHY A CORRUPTED
TEXT FEELS

**STRANGELY
ALIVE.**

THE MISSING PIECES
FORCE US TO
**RECONSTRUCT
THEM.**

CONSCIOUSNESS ITSELF
IS A COLLABORATION WITH
ABSENCE.

A COMPULSIVE
TENDENCY TO
COMPLETE WHAT
CANNOT BE
COMPLETED.



LOREM IPSUM VARIORUM

Milano, Palazzo Citterio: L'allestimento ospitato a Palazzo Citterio si articola attraverso una sequenza di dodici slide, in cui il testo poetico viene integrato a elementi grafici essenziali. La struttura della mostra sceglie la via della sintesi visiva, affidando alla tipografia e al rigore geometrico la responsabilità di tradurre in forma spaziale la parola scritta. L'impianto concettuale alla base dell'opera punta a decostruire la composizione tradizionale, distribuendo il ritmo dei versi all'interno di un supporto bidimensionale che opera per sottrazione. Ogni pannello costituisce una stazione autonoma, dove il rapporto tra lo spazio bianco, le campiture di colore e il segno grafico definisce una precisa scansione temporale della lettura, imponendo al visitatore un ritmo di fruizione controllato.

La scelta di limitare la presenza materica a favore di una sequenza seriale risponde all'intento di misurarsi con i linguaggi della fruizione contemporanea, trasferendo l'esperienza della pagina all'interno di uno spazio espositivo essenziale. La veste grafica individua una grammatica rigorosa: i caratteri tipografici non svolgono una funzione decorativa, ma interagiscono con il peso visivo dei testi, determinando gerarchie di lettura e pause riflessive.

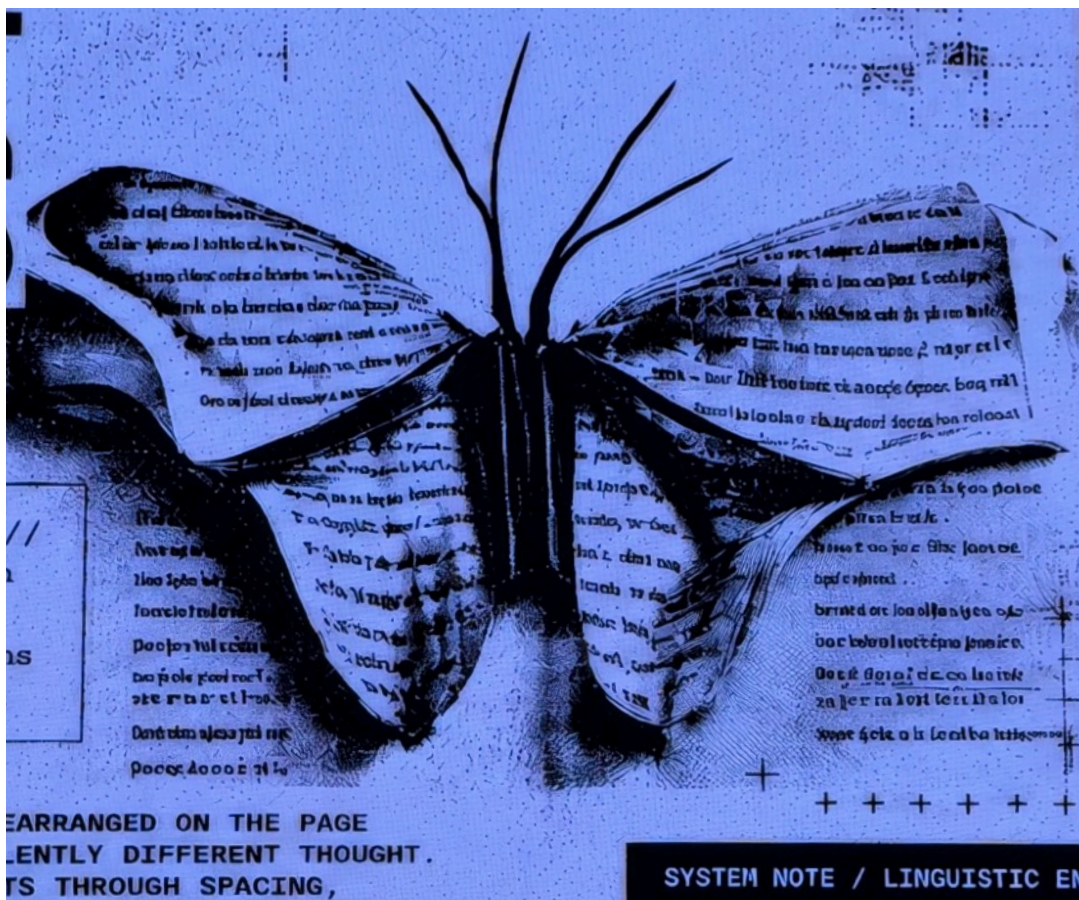
L'operazione si inserisce nel dibattito sulle modalità di esposizione della letteratura nelle gallerie e nei luoghi d'arte, ponendo l'accento sulla possibilità di trasformare la parola in oggetto visivo e fruibile secondo dinamiche mostrative.

Nel complesso, l'installazione offre una ricognizione sulle forme di sintesi tra arte visiva e testo letterario. Propone un modello di divulgazione che punta sulla percezione immediata del contenuto, offrendo un punto di vista lineare su come la composizione poetica possa abitare lo spazio pubblico senza cedere a tentazioni di spettacolarizzazione. La sequenza delle dodici stazioni funziona come un catalogo visivo, dove la superficie pulita delle grafiche diventa il perimetro entro cui viene racchiusa la parola, lasciando al visitatore il compito di ricomporre la continuità del discorso poetico oltre la frammentazione delle singole diapositive.

2. La parola poetica alla prova dell'intelligenza artificiale: il progetto di Sasha Stiles

L'indagine sul rapporto tra creazione poetica e strumenti di intelligenza artificiale trova nel lavoro di Sasha Stiles un punto di osservazione critico sulle trasformazioni del linguaggio nell'era della riproducibilità algoritmica. Al centro della sua ricerca non vi è la semplice automazione della scrittura, ma una rinegoziazione del ruolo dell'autore di fronte a modelli di calcolo in grado di generare strutture sintattiche e combinazioni lessicali complesse. Il progetto si articola attorno all'interazione diretta tra la sensibilità poetica e il supporto computazionale, inteso non come mero esecutore o sostituto, ma come co-autore tecnico o estensione della memoria linguistica e dell'archivio personale dell'autrice.

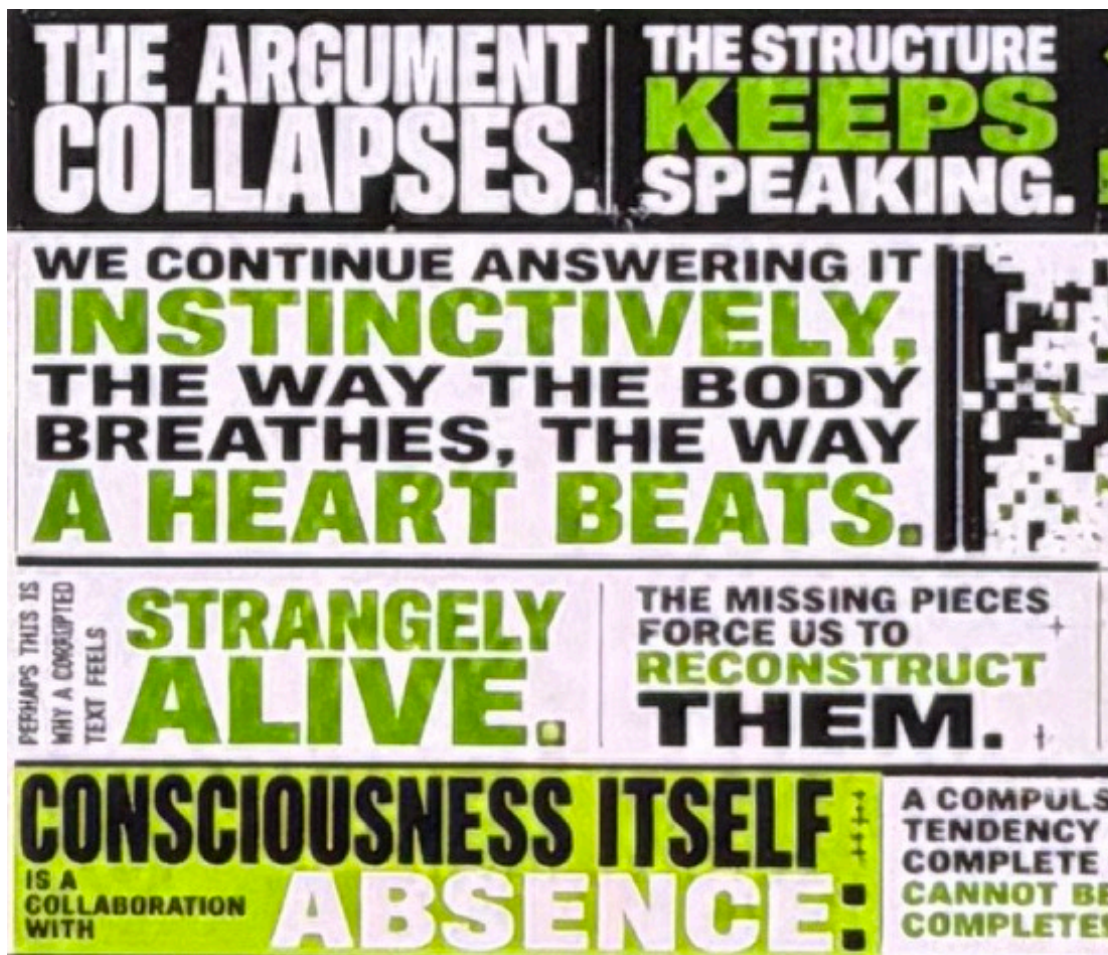
Stiles sottopone all'algoritmo quadri concettuali, frammenti di versi e vincoli metrici o semantici, innescando un processo dialettico in cui il modello calcola combinazioni e deviazioni inaspettate. La fase fondamentale dell'operazione risiede nel lavoro di selezione, montaggio e riscrittura condotto dall'autrice: è nel filtro umano che la sovrapposizione casuale di dati trova una direzione di senso e si trasforma in testo poetico. L'esperimento evidenzia come la presenza dell'intelligenza artificiale sposti il baricentro dell'atto creativo dalla generazione primaria del materiale alla sua orchestrazione. Il modello algoritmico opera su un patrimonio di dati preesistenti, restituendo variazioni che mettono alla prova i limiti del linguaggio convenzionale e le abitudini sintattiche del lettore..



La ricerca di Stiles evita la trappola del solipsismo tecnologico: l'algoritmo viene utilizzato come un reattore in grado di produrre frizione concettuale. In un panorama culturale in cui la tecnologia tende ad appiattire la produzione linguistica su standard di massima efficienza e prevedibilità, il lavoro condotto in questo progetto forza lo strumento computazionale verso l'imprevedibilità del verso. La macchina non viene usata per simulare la coscienza umana, ma per evidenziare la natura combinatoria del linguaggio, mostrando come anche la parola poetica sia composta da elementi riorganizzabili e ricombinabili secondo logiche matematiche.

L'operazione pone domande dirette sulla natura della proprietà intellettuale, sulla figura dell'autore e sul valore della testualità nell'era dei grandi modelli linguistici. La poetessa sceglie di abitare questo margine di incertezza, interrogandosi sulle nozioni di originalità e intenzione espressiva. Il risultato è un'esplorazione critica che non cede né al rifiuto luddista né all'entusiasmo acritico per l'automazione, collocando la figura del poeta al centro di un nuovo ecosistema di produzione simbolica dove l'algoritmo funge da cassa di risonanza e la responsabilità del senso rimane saldamente in mano all'autore.

Il percorso espositivo riduce al minimo l'impianto scenografico per concentrarsi sulla chiarezza del messaggio e sull'impatto immediato delle soluzioni visive. Tuttavia, questa scelta di estrema pulizia formale evidenzia anche i limiti intrinseci della trasposizione del testo poetico in ambiente museale. La riduzione della poesia a un formato fruibile per stazioni successive rischia di privilegiare la dimensione dell'impatto visivo a scapito della densità semantica del testo. L'allestimento sceglie consapevolmente di non sovraccaricare lo spazio con sovrastrutture narrative, lasciando che sia il contrasto tra il supporto digitale e la cornice storica del palazzo a creare il corto circuito visivo.



Storie

