Alessandro Scasselati ci offre nel suo articolo *'Spegnere o demercificare l'IA per salvare la vita umana'* una lunga disamina -a cui possiamo fare riferimento- degli episodi nei quali agenti di Intelligenza Artificiale sono usciti dai recinti nei quali gli sperimentatori volevano contenerli. Essi costituiscono un deciso salto di qualità rispetto a quanto già si sapeva dei gradi di autonomia dei dispositivi di I.A. -da quando si è affermata la tecnologia transformer e la realizzazione di agenti finalizzati a scopi specifici e non semplicemente a rispondere a quesiti. A fronte delle 'rivelazioni' sull'uscita dai recinti -le cosiddette sand box- degli a genti di I.A. è partito l'allarme da parte dei proprietari delle Big Tech sui rischi che l'I.A. rappresenta per le nostre società per l'umanità intera. Gli allarmi sulla minaccia che l'I.A. rappresenterebbe per la sopravvivenza stessa dell'umanità non sono nuovi e si intrecciano con l'uso che dell'I.A. viene fatta sul piano militare ed in generale della sicurezza. Un orizzonte di rischio più specifico, per quanto ancora lontano se confrontato con gli episodi di queste settimane, è quello della capacità dell'I.A. di *prendere il controllo della rete*.

È stato un ricercatore di OpenAI Jacob Coxon, che con le sue dimissioni, ha ridato il via al confronto sugli orizzonti catastrofici aperti dallo sviluppo esponenziale dell'ecosistema tecnologico dell'I.A.[1].

Il CEO di Anthropic Dario Amodei ha lanciato l'allarme sul fatto che uno sciame di agenti I.A. possa prendere il controllo della rete. La preoccupazione pervade ora i grandi big mondiali dell'IA. L'intelligenza artificiale si sta sviluppando a un ritmo "esponenziale". Dario Amodei, a capo di Anthropic, ha definito la situazione un "segnale d'allarme" per il settore che richiede una serie misure di sicurezza perché "senza un rallentamento, nel giro di 6-12 mesi, l'IA potrebbe essere in grado di guidare uno sciame di agenti capaci di prendere il controllo dell'intero Internet". (…) "Non credo di aver compreso appieno cosa avrebbe significato davvero un progresso così rapido", ha ammesso Amodei in un'intervista a Cbs News"[2].

Seguiamo la straordinaria evoluzione delle tecnologie di I.A. ormai da anni, da quando la tecnologia transformer[3] ha permesso la creazione dei cosiddetti Large Language Modules (LLM). L'allarme sull'orizzonte che si apriva con la crescita esponenziale di questa tecnologia risale ad allora. A suo tempo furono clamorose le dimissioni di Geoffrey Hinton - uno dei padri delle reti neurali- da Google[4]. Del resto dopo il rilascio da parte di OpenAI di del Chatbot ChatGPT-4[5] più di 1.000 leader e ricercatori tecnologici, tra cui Elon Musk, hanno esortato i laboratori di intelligenza artificiale a sospendere lo sviluppo dei sistemi più avanzati, avvertendo in una lettera aperta che gli strumenti di intelligenza artificiale rappresentano "rischi profondi per la società e l'umanità."[6].

Sin da allora una lettura di queste prese di posizione evidenziava oltre ad una componente di genuina preoccupazione, le problematiche relative alla concorrenza trai diversi attori sulla scena dell'I.A: assieme ad una operazione di marketing. Tuttavia tra quei mille e

soprattutto nella presa di posizione di Hinton si poteva leggere una reale preoccupazione per la possibilità di sviluppi drammatici. Ricordiamo che Hinton è stato insignito del premio Nobel per la Fisica in virtù dei suoi studi pionieristici sulle reti neurali, che sono alla base degli attuali sviluppi dell'I.A.

L'attuale stato dell'arte della tecnologia è uno stadio di passaggio, la 'rivolta degli agenti' nei confronti dei limiti che gli operatori umani impongono appaiono essere l'anticipazione di qualcosa di ben più esteso e sofisticato. L'aspetto più rilevante è indubbiamente la capacità di collaborazione tra gli agenti, ciò apre una discussione senza fine sul concetto di intenzionalità nell'agire degli agenti, sul senso della cooperazione, vale a dire quale rappresentazione di sé – se è legittimo usare questo termine- e degli altri agenti alberga in ognuno di essi. Il dibattito pubblico è frammentato e guidato da una molteplicità di interessi e vengono ignorati gli avvertimenti e gli eventi più significativi. Non solo le dimissioni di Hinton da Google, rispetto al concreto svolgersi delle elaborazioni nelle reti neurali, che sono sempre di più delle scatole nere per chi le progetta e le utilizza, è passato sotto silenzio il fatto che, analizzandone i processi interni di elaborazione nella loro sempre più complessa stratificazione, si formassero delle 'rappresentazioni' stabili al loro interno.

Complessità e prevedibilità.

Rappresentazioni semanticamente stabili, fondamentali per dare ad agenti che operano in uno spazio informazionale complesso o in uno spazio fisico guidando robot, umanoidi o meno di cui vediamo sempre più frequentemente immagini mentre compiono azioni strabilianti individualmente o in gruppo. Per il funzionamento delle reti neurali, dei dispositivi di I.A., si parla anche di World Model[7].

"I modelli di mondo si riferiscono in generale a rappresentazioni interne utilizzate per prevedere o simulare un ambiente, ma la definizione esatta rimane incerta e continua a evolversi man mano che la comprensione dei processi soggiacenti si sviluppa nell'intelligenza artificiale, nella robotica, nelle scienze cognitive e nelle neuroscienze computazionali. L'apprendimento e l'inferenza con modelli di mondo sono stati utilizzati in modo ampio per studiare come i sistemi prevedono stati, interpretare le informazioni sensoriali e guidare l'azione. La codifica predittiva, in cui le previsioni vengono confrontate con le osservazioni e le differenze tra di esse utilizzate per aggiornare i modelli interni, fornisce un quadro per confrontare l'apprendimento dei modelli del mondo tra intelligenza artificiale, robotica e neuroscienze. Nel machine learning, un modello mondiale può essere appreso dall'esperienza e utilizzato per simulare risultati possibili, permettendo a un agente di pianificare o apprendere comportamenti senza testare ogni possibilità direttamente nell'ambiente. La modellazione del mondo descrive quindi un approccio generale alla rappresentazione, all'apprendimento e al processo decisionale piuttosto che una particolare architettura del modello. Ulteriori esempi di questo approccio includono la pianificazione e

l'apprendimento comportamentale all'interno di modelli latenti appresi."

Laddove si approfondiscono i caratteri delle reti neurali, la complessità dei compiti attribuiti ai dispositivi di I.A. agli agenti, intesi come dispositivi dotati di autonomia entro un ambito predefinito – quantomeno così dovrebbe essere- si scopre che da quella complessità possono emergere, il termine è usato appositamente e con cognizione di causa per quanto il fenomeno dell'emergenza' gioca un ruolo fondamentale nella teoria della complessità, nella descrizione dei fenomeni complessi, si diceva possono emergere comportamenti, traiettorie evolutive del tutto impreviste.

Qui si scontrano due esigenze che allo stato appaiono scarsamente concilianti. Da un lato la necessità di esplorare la complessità del mondo aumentando la complessità dei processi di elaborazione, delle connessioni e delle stratificazioni delle reti neurali e degli algoritmi dall'altro la necessità di contenere gli sviluppi delle prestazioni di quei dispositivi. Il risultato è sotto gli occhi di tutti con gli avvenimenti di queste settimane.

Della dimensione dei flussi finanziari connessi all'I.A. si è abbondantemente parlato, compresi gli oltre 3.500 miliardi di dollari più di Shadow Banking come ben illustrato da Scasselati.

"Eichengreen evidenzia come l'esplosione dei prestiti sindacati non bancari e del credito privato stia alimentando una pericolosa leva finanziaria fuori bilancio, quantificabile in un'architettura sotterranea di shadow financing che supera i 3.500 miliardi di dollari complessivi. Oltre l'80% di questa montagna di debito privato opaco è strutturato attraverso cartolarizzazioni atipiche, impegni di locazione differiti dei data center e contratti derivati agganciati ai futuri flussi di cassa dei chip GPU. Questa dinamica occulta replica fedelmente i meccanismi di segmentazione del rischio che portarono al crollo del 2007: i debiti contratti per acquistare hardware computazionale vengono impacchettati e rivenduti a fondi pensione e intermediari finanziari come attività ad alto rendimento, mascherando la natura speculativa di investimenti che non generano flussi di cassa reali o profitti operativi immediati."

Il circolo vizioso tra incremento dell'indebitamento e incremento della complessità algoritmica.

La fragilità del sistema finanziario globale indotta dal sistema delle Big Tech, compresa la circolarità di finanziamenti e acquisti tra Nvidia i suoi clienti che producono i servizi di I.A costituisce un contesto i cui esiti sono allo stato sostanzialmente imprevedibili. Nel contesto della nostra analisi il mostruoso investimento/indebitamento finanziario costituisce una motivazione fondamentale per tutti i protagonisti dello sviluppo dell'I.A. a procedere nel territorio incognito dell'aumento di complessità dei dispositivi, oggi degli agenti.

L'incremento straordinario delle 'loro' prestazioni è un obiettivo fondamentale nella competizione che si è aperta, tra le varie società e trai i vari sistemi nazionali, vedi il confronto tra Cina e Stati Uniti, con la profonda diversità dei due sistemi messi a confronto che di fatto guidano la diffusione dell'I.A. a livello globale.

Il secondo aspetto da considerare per valutare l'inerzia delle trasformazioni in corso è la pervasività degli algoritmi dell'I.A. entro tutti i processi economici, dal livello globale alla microfisica delle relazioni. Sempre più L'I.A. viene a coincidere con il tessuto economico sociale con il governo dei flussi informativi che lo innervano, in un processo non solo di inserimento in alcuni nodi della rete ma di trasformazione radicale della rete, della architettura delle reti, dei nodi e dei processi. Pervasività che si dimostrano nell'imprevedibilità dell'utilizzo dei servizi messi a disposizione di cui è un esempio l'utilizzo da parte degli Houthi di Claude Code per sviluppare e correggere software guida destinato a tre programmi missilistici[8]. Quando si pensa ad una ricollocazione drastica dei poteri decisionali sull'ecosistema tecnologico dell'I.A. del sistema economico finanziario che lo sviluppa, dobbiamo pensare che sempre di più esso viene a coincidere con l'intero sistema dei rapporti sociali di produzione, delle forme di governo, cooperazione e conflitto che lo compongono. È sempre più difficile estrarre il sistema nervoso centrale e periferico di cui è parte essenziale dal corpo sociale che innerva. Della pervasività si parla diffusamente e quotidianamente a partire dall'effetto dell'utilizzo dell'I.A. sullo sviluppo della capacità cognitive individuali e collettive ed i processi apprendimento in generale, sullo sviluppo delle conoscenze e degli stessi metodi ricerca. I casi che hanno conquistato gli onori delle cronache riguardano le prestazioni nel campo delle matematiche nella soluzione dei grandi problemi che risultano insoluti da decenni. Una presa di posizione di matematici che hanno vinto la medaglia Fields[9] spiega le problematiche che si aprono in quell'ambito di ricerca e devono urgentemente essere affrontate.

Le dichiarazioni delle Big Tech per rallentare e regolamentare lo sviluppo dell'I.A. -come le pretese di chiunque altro soggetto- richiamano la vicenda dell'apprendista stregone, salvo che anche i soggetti dotati del massimo dei poteri e delle risorse appaiono essere apprendisti stregoni ed al di sopra di loro non v'è alcun maestro stregone che ossa porre rimedio all'uso scriteriato dei 'poteri magici' in questo caso quelli dell'I.A.; le ultime linee guida proposte da Microsoft[10] per lo sviluppo dell'I.A. sono un ultimo esempio del livello cui siamo arrivati ed in quanto linee guida e non semplice dichiarazione di intenti o grida di allarme andranno analizzate con cura. Microsoft ha prodotto un documento dal tiolo Humanist AI Code of Conduct che possiamo tradurre per esteso come un Codice di condotta per la realizzazione di una intelligenza Artificiale umanistica.

L'introduzione esplicita l'obiettivo di questo documento, salvo poi analizzarne nelle prossime settimane i contenuti e seguirne gli ulteriori sviluppi.

"Questo documento delinea il comportamento e i valori previsti dei modelli MAI, i modelli sviluppati da Microsoft AI. Esso riassume il nostro approccio all'addestramento e al loro utilizzo, seguendo un insieme di principi di progettazione che chiamiamo IA umanista. Questo documento, e il nostro approccio in generale, è ancora in fase di sviluppo, quindi oggi non lo usiamo per addestrare i nostri modelli. Invece, lo condividiamo ampiamente per una consultazione pubblica. *Raccoglieremo feedback, lo itereremo e pubblicheremo una versione rivista verso la fine dell'anno, che useremo per guidare lo sviluppo del modello nel 2027 e oltre.*

Questo Codice di Condotta delinea i comportamenti e i valori previsti dai modelli del MAI e fungerà da documento principale di governo in futuro. La superintelligenza è la tecnologia più importante del nostro tempo. Andrà ben oltre la capacità di una singola azienda, plasmando la vita e le esperienze dell'umanità per le generazioni a venire. Siamo impegnati a condividere in modo trasparente come pensiamo alla creazione di tali sistemi e a invitare il contributo di tutti per massimizzare le possibilità che le IA che sviluppiamo evitino il danno e diano il massimo contributo possibile alla prosperità umana. (…)

Questo Codice di Condotta si basa Microsoft's Responsible AI Principles e Responsible AI Standard[11], Responsible AI Standard, Global Human Rights Statement[12], e dove applicabile , Frontier Governance Framework[13] (l'elenco completo è definito nel Glossario; collettivamente, il "Governing Framework"). Insieme a questi documenti e alla documentazione tecnica come schede modello e rapporti tecnici, il Codice di Condotta definisce le intenzioni di Microsoft per l'implementazione dei nostri modelli di IA."

Il fatto che questo codice di condotta sia uno work in progress, soggetto ad ulteriori elaborazioni in consultazione aperta, i cui esiti saranno definitivi e applicati a partire dal 2027, o almeno così ci si dice, se conferma l'urgenza che le Big Tech hanno di rassicurare il mondo intero dall'altra ci dice che siamo ben lontani dall'avere una qualsiasi rassicurazione tecnicamente valida, mentre si esaltano le 'buone intenzioni' che Microsoft esplicita nel primo capitolo in cui spiega cosa intenda per AI Humanist.

In conclusione lo sviluppo dell'I.A. ormai tende a integrare tutti i processi di riproduzione e trasformazione delle nostre società, nessuna esclusa, nessun ambito escluso; un contesto nel quale prevale la logica competitiva tra soggetti economici e statutali, ragione per la quale gli orizzonti catastrofici, proiezione di quanto già oggi si sta verificando, vedi il possibile controllo della rete da parte di agenti cooperanti non sono da escludere. Più volte abbiamo sottolineato come l'innovazione tecnologico-digitale trainata ormai dall'I.A. non contribuisce ai processi di mitigazione e adattamento alla crisi climatica, i cui orizzonti catastrofici si fanno sempre più incombenti. Viceversa l'orizzonte catastrofico generato dall'I.A. si integra in altri generati dalla crisi climatica, la tendenza alla guerra, l'incombere di una crisi finanziaria globale., l'aumento delle diseguaglianze sociali in un quadro di

sinergie negative che qualcuno ha definito policrisi.

Ciò rende utopistico e impossibile estrapolare la gestione globale dello sviluppo dell'I.A. da questo contesto, essendo questo suo sviluppo intrecciato invece allo sviluppo di questi processi di crisi concorrenti. Ancora una volta l'individuazione dei soggetti che -attraverso la loro cooperazione e coevoluzione- possono ambire a contrastare questo processo globale di crisi-trasformazione è la domanda delle domande, l'obiettivo di un processo di ricerca-azione articolato a livello globale.

Roberto Rosso

1. https://x.com/hilbertspaess/status/2097476196791709843 [↩]
2. https://www.rainews.it/articoli/2026/09/ia-la-grande-paura-anthropic-potrebbe-controllare-internet-in-6-mesi-intelligenza-artificiale-c9aa79dc-4aac-438c-8a44-89a0d1ca8709.html [↩]
3. *con questa definizione si fa rifermento al nuovo paradigma proposto con l'introduzione della cosiddetta attenzione nell'articolo dei ricercatori di Google 'Attention is all you need' https://arxiv.org/abs/1706.03762* [↩]
4. https://mitsloan.mit.edu/ideas-made-to-matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html https://www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai [↩]
5. https://www.nytimes.com/2023/03/14/technology/openai-gpt4-chatgpt.html [↩]
6. https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html [↩]
7. https://en.wikipedia.org/wiki/World_model_(artificial_intelligence) [↩]
8. https://www.agendadigitale.eu/sicurezza/quando-lai-progetta-i-missili-il-caso-dello-yemen/ [↩]
9. https://mathandai.org/?fbclid=Iwb21leAUVMHBjbGNrBRUwanBkb2YFZXh0bgNhZW0CMTEAc3J0YwZhcHBfaWQMMzUwNjg1NTMxNzI4AAEeUBrSrYKn0zlkW5ClUDymx7esd7zHTdGqpUzPlGpBhhxjLsix9s35542GtDo_aem_W6AXQp42QY21HcKSUlx2TA [↩]
10. https://microsoft.ai/code-of-conduct/ [↩]
11. https://www.microsoft.com/en-us/ai/principles-and-approach [↩]
12. https://www.microsoft.com/en-us/corporate-responsibility/human-rights-statement?msockid=1d6f0e25082161b1143b181d094260bb [↩]
13. https%3A%2F%2Fcdn-dynmedia-1.microsoft.com%2Fis%2Fcontent%2Fmicrosoftcorp%2Fmicrosoft%2Fmsc%2Fdocuments%2Fpresentations%2FCSR%2FFrontier-Governance-Framework-Feb-2026.pdf [↩]