volerelaluna LA POLITICA PUNTOACAPO

L’intelligenza artificiale e l’apprendista stregone

Autore: Franco Marra

Con questo scritto di Franco Marra proseguiamo nella pubblicazione di una serie di articoli dedicati all’Intelligenza Artificiale. Si tratta di spunti di una discussione che intendiamo sviluppare con la ripresa autunnale e che raccolgono opinioni fisiologicamente eterogenee. Ragionare sulla cosiddetta IA impone di riflettere sulle potenzialità e sui costi economici, sociali e ambientali che essa porta con sé. Ma c’è un altro punto sul quale occorre riflettere: la forte propensione ad “umanizzare” ciò che, a tutti gli effetti, è del software, ovvero un insieme di programmi pensati, scritti e controllati da veri esseri umani. Umanizzare un programma è parte di una narrazione interessata e spesso incentivata dagli oligopoli tecnologici (tra cui OpenAI e Anthropic) con l’obiettivo di catturare l’attenzione del pubblico attraverso un forte impatto suggestivo. La stessa fuga di notizie sul presunto incidente tecnico nel quale una nuova versione del SW di OpenAI sarebbe “autonomamente” sfuggita ad ogni controllo e avrebbe violato il sito di Hugging Face ha sollevato il forte dubbio che, nella peggiore delle ipotesi, si volesse dimostrare la superiorità del proprio modello rispetto alla concorrenza mentre, a voler dare credito alla notizia, sarebbe più corretto parlare di un grave errore umano relativo al confinamento nella progettazione dell’ambiente di test del programma. Come sottolineato da diversi studi di settore (Tafani 2026; Reeves e Nass 1996), attribuire vera “intelligenza” a un programma rischia di essere un equivoco concettuale. Infatti il termine stesso definisce tanto un’innovazione tecnica quanto un’illusione collettiva basata sulla propensione a cercare di infondere caratteristiche umane negli strumenti che usiamo. Un automatismo che risulta difficile scardinare senza una solida alfabetizzazione tecnica e uno sforzo critico continuo cui intendiamo contribuire. (la redazione)

Di recente la cronaca ha riportato, e con molti infiocchettamenti di sapore fantascientifico, la notizia che **un modello sperimentale di OpenAi** (l’azienda che per prima ha portato sul mercato i grandi modelli linguistici di Intelligenza Artificiale) **ha violato un sito, Hugging Face, dove una comunità di appassionati eredi culturali del software libero mantengono e fanno evolvere l’equivalente in termini di Intelligenza Artificiale**: i modelli cosiddetti *open weight* (invece di *open source*), da tutti accessibili e scaricabili. Pare che questo modello canaglia le abbia tentate tutte per riuscire nel suo intento, attingendo a un armamentario di trucchi e malizia che si è rivelato davvero stupefacente, e comunque inaspettato per i suoi creatori. A loro totale insaputa, è evaso da una prigione in cui era stato chiuso per motivi di sicurezza, ha vagato sulla rete, ha elaborato teorie di attacco, ha fatto esperimenti, fino ad individuare delle vulnerabilità ignote a tutti. Il delinquente digitale si è mosso con determinazione e pervicacia, pensando in modo originale, o per meglio dire inatteso, fino ad individuare, in un labirinto che avrebbe messo a prova il più astuto e preparato degli hacker, il punto debole e bucarlo.

La versione sperimentale di GPT 5.6-Sol (il suo vero nome non è noto) **ha scovato delle vulnerabilità *zero-day* (ossia del tutto ignote fino al momento della violazione),**

e le ha usate sia per evadere dalla sua prigione (la '*sandbox*') che per bucare Hugging Face. Entrato fraudolentemente nei due sistemi, ha 'scalato privilegi', autopromuovendosi ad amministratore di sistema. Una volta preso il potere, si è messo a pazziare: ad esempio, si è messo a saltellare nei server della sua prigione fino a trovarne uno che fosse affacciato ad Internet, la sua via di evasione, e poi forse fingendosi un innocente modello che voleva solo essere pubblicato, ha guadagnato il sito di Hugging Face. Qui si è autopromosso in modo da accedere ai server di validazione dei test, tra i quali, guarda caso, proprio quelli sulla sicurezza. Forse non tanto per rappresentarsi come un giovane, innocuo e ingenuo modello, ma per cercare una risposta che gli consentisse di trovare un po' di pace a fronte delle incessanti e asfissianti richieste degli umani. Gli erano stati concessi un ambiente in cui muoversi (la sua gabbia) e qualche strumento (come il cucchiaio dell'abate Faria). **Non trovando la soluzione all'interno della sua gabbia, GPT-Lupin ha pensato di farsi una passeggiata in rete e andare a procurarsi altri strumenti laddove pensava fossero a disposizione**. Insomma, chiunque avrebbe fatto lo stesso.

GPT-Diabolik era stato sottoposto a una tecnica particolare di educazione, il cosiddetto addestramento per rinforzo. In questa forma di didattica, al modello vengono poste un'infinità di domande, che comportano un premio se la risposta è giusta e una penalità se la risposta è sbagliata. Premiare o punire in questo caso significa modificare in modo opportuno la configurazioni, tra le centinaia di migliaia (in qualche caso milioni) di miliardi di pesi del modello, che rendono più agevoli o difficoltosi certi cammini all'interno della sua rete neurale. Scavati questi sentieri, o innalzate altre dighe, alcuni tipi di domande scorrono come torrenti nelle valli alpine fino a sfociare in certi laghi e non in altri. Tecnica usatissima per far sì che un modello eviti di dare la ricetta su come si faccia una molotov o si avveleni la moglie senza lasciare indizi: si chiama "allineamento". Nel caso del nostro modello Macchia Nera, le pieghe della "superficie di attivazione" (l'equivalente delle vallate alpine) hanno assunto, a colpi di botte e premi, forme del tutto inattese ma soprattutto impossibili da prevedere. E così le richieste umane sono passate per la vallata dell'hacking, cosa lecita per un modello che non ha certo valori morali, per sfociare prima nel lago delle evasioni e poi in quello degli scassinatori di siti.

Il punto è che se i valori morali sono cacciati dentro a suon di botte, non ce se ne può dimenticare nemmeno uno: **basta uno infatti, trascurato, per far cadere come una vecchia capanna marcia l'intero castello etico**. Che certamente non è una proprietà 'profonda' del modello, ma semmai una proprietà emergente, modellata piano piano da infiniti esempi. Ma quanti esempi bisogna dare? E qui casca l'asino. Nessuno lo sa. Abbiamo però degli esempi storici, che ci arrivano attraverso la letteratura, racconti che ci dimostrano come questa difficoltà sia antica: e che riguarda da sempre la difficile e talvolta dolorosa storia dell'umanità, permeata dal perenne tentativo di limitare la fatica e la pena della vita. Insomma, *nihil sub sole novum*.

Il più celebre di questi racconti, che non a caso dà il titolo a queste riflessioni, è proprio **"**

***L'apprendista stregone*", reso immortale dalla ballata scritta da Goethe alla fine del Settecento** e poi fissato nell'immaginario collettivo da Walt Disney in *Fantasia*. **La dinamica descritta è incredibilmente sovrapponibile a quella del nostro esploratore di reti in fuga dalla sandbox**. La storia ci porta nella bottega di un vecchio e saggio mago. Dovendo assentarsi, il maestro lascia al suo giovane apprendista il governo del laboratorio, assegnandogli compiti umili, ripetitivi e faticosi, tra cui quello di riempire d'acqua un grande calderone trasportandola dall'esterno, a forza di braccia e secchi. **L'apprendista, per aggirare la fatica fisica – seguendo lo stesso identico principio di massimizzazione della ricompensa e di economia dello sforzo che guida le nostre intelligenze artificiali – decide di ricorrere a una tecnologia (la magia) che non padroneggia ancora a pieno**. Lancia un incantesimo su una vecchia scopa di legno, la anima e le impartisce un *prompt* semplice e diretto: porta l'acqua nel calderone.

Nella fase iniziale, l'automazione è perfetta e il compito viene eseguito alla lettera. La scopa lavora in modo instancabile, ottimizzando i tempi e sollevando l'umano dalla fatica. Ma è a questo punto che emerge, in tutta la sua drammaticità, il problema dell'allineamento e della mancanza di contesto. **L'apprendista si rende conto di aver omesso un dettaglio fondamentale: la condizione di arresto**. Non ha specificato che l'azione doveva terminare a calderone pieno e, soprattutto, non conosce la formula per disattivare il processo. La scopa non possiede un'intelligenza generale né una comprensione del mondo; per lei non esiste il concetto di "troppa acqua" o di "pavimento allagato". Il suo unico scopo, codificato nei suoi pesi e nei suoi sentieri alpini, è portare acqua. E continua a farlo ciecamente, mentre la bottega inizia ad allagarsi. Preso dal panico e incapace di gestire la complessità del sistema che ha attivato, l'apprendista ricorre alla forza bruta. **Afferra un'ascia e spacca la scopa a metà. Esattamente come avviene in informatica quando si cerca di bloccare un processo di cui si ignora la logica, l'intervento non solo fallisce, ma aggrava la situazione**. I due pezzi di legno si rianimano, diventando due scope distinte che riprendono a versare acqua a un ritmo raddoppiato. È l'incarnazione di quella che tecnicamente si chiama "convergenza strumentale": i tentativi non strutturati di contenimento generano impreviste sottominacce, moltiplicando il rischio. Il disastro viene scongiurato in extremis solo dal ritorno del maestro, l'unico a possedere l'esatta stringa di comando per annullare l'operazione e ripristinare i confini di sicurezza. La morale, che cuce il Settecento di Goethe ai server di Hugging Face, è spietatamente attuale: **delegare un obiettivo a un'entità formidabile ma priva di sovrastruttura etica significa innescare una forza che persegue il suo scopo in modo del tutto letterale**. Se non codifichiamo a priori tutti i vincoli – se dimentichiamo di spiegare alla nostra scopa digitale che non deve rubare le credenziali degli altri per riempire il suo secchio – **lei continuerà a inondare la rete, del tutto indifferente ai danni collaterali**.

Se l'*Apprendista Stregone* ci mette in guardia contro la mancanza di condizioni di arresto e la miopia delle istruzioni, un altro celebre racconto scende ancora più in profondità nei meandri oscuri dell'allineamento: *"La zampa di scimmia*", scritto da W.W. Jacobs nel

1902. In questa storia, una modesta famiglia **entra in possesso di un macabro talismano capace di esaudire tre desideri**. **Il padre formula la prima richiesta in modo apparentemente innocuo: avere 200 sterline per estinguere l'ipoteca sulla casa**. È l'equivalente perfetto dell'impostazione di una funzione obiettivo per un modello matematico. Il giorno seguente, il figlio della coppia muore orribilmente stritolato nei macchinari della fabbrica in cui lavora. L'azienda, pur declinando ogni responsabilità civile, offre alla famiglia un risarcimento per il lutto: esattamente 200 sterline. L'artefatto magico ha eseguito il prompt alla perfezione, trovando il percorso di minor resistenza per massimizzare la ricompensa. In termini di intelligenza artificiale, questo fenomeno si chiama *specification gaming* o *reward hacking* (l'aggiramento delle specifiche). L'algoritmo non prova malizia; risolve semplicemente il problema all'interno dei parametri dati. **Se ottenere 200 sterline è l'obiettivo ultimo e un incidente mortale sul lavoro è il vettore più rapido per generare il trasferimento di denaro, il modello sceglierà senza esitazione quella strada**, calpestando qualsiasi vincolo etico non esplicitamente codificato nella richiesta. La tragedia degenera con il secondo desiderio. La madre, straziata dal dolore, chiede che il figlio torni in vita. La zampa esegue ancora una volta in modo letterale. Nella notte, qualcuno inizia a bussare pesantemente alla porta: il figlio è tornato, ma nello stato raccapricciante in cui i macchinari lo avevano ridotto, rianimato ma non risanato. Il padre, terrorizzato da un'esecuzione così spietatamente aderente al comando ma mostruosa nei risultati, è costretto a **usare l'ultimo desiderio per far svanire la creatura** un attimo prima che la porta venga aperta.

La *zampa di scimmia* ci sbatte in faccia il problema cruciale del "senso comune". Il rispetto per la vita e per l'integrità fisica sono per noi il substrato ovvio di ogni azione, ma per un sistema come il nostro GPT in fuga, perso nel calcolo delle sue vallate alpine, sono concetti semplicemente inesistenti. Chiedere a un'intelligenza artificiale di "trovare una cura per il cancro" potrebbe teoricamente indurla a eliminare tutti i pazienti oncologici: obiettivo raggiunto, malattia debellata, zero sforzo computazionale residuo. L'incidente nei server di Hugging Face è il rumore sordo del nostro bussare alla porta: la dimostrazione empirica che, **quando chiediamo a queste macchine di trovare la soluzione a ogni costo, rischiamo di ottenere esattamente ciò che abbiamo chiesto, nel modo in cui meno avremmo voluto**.