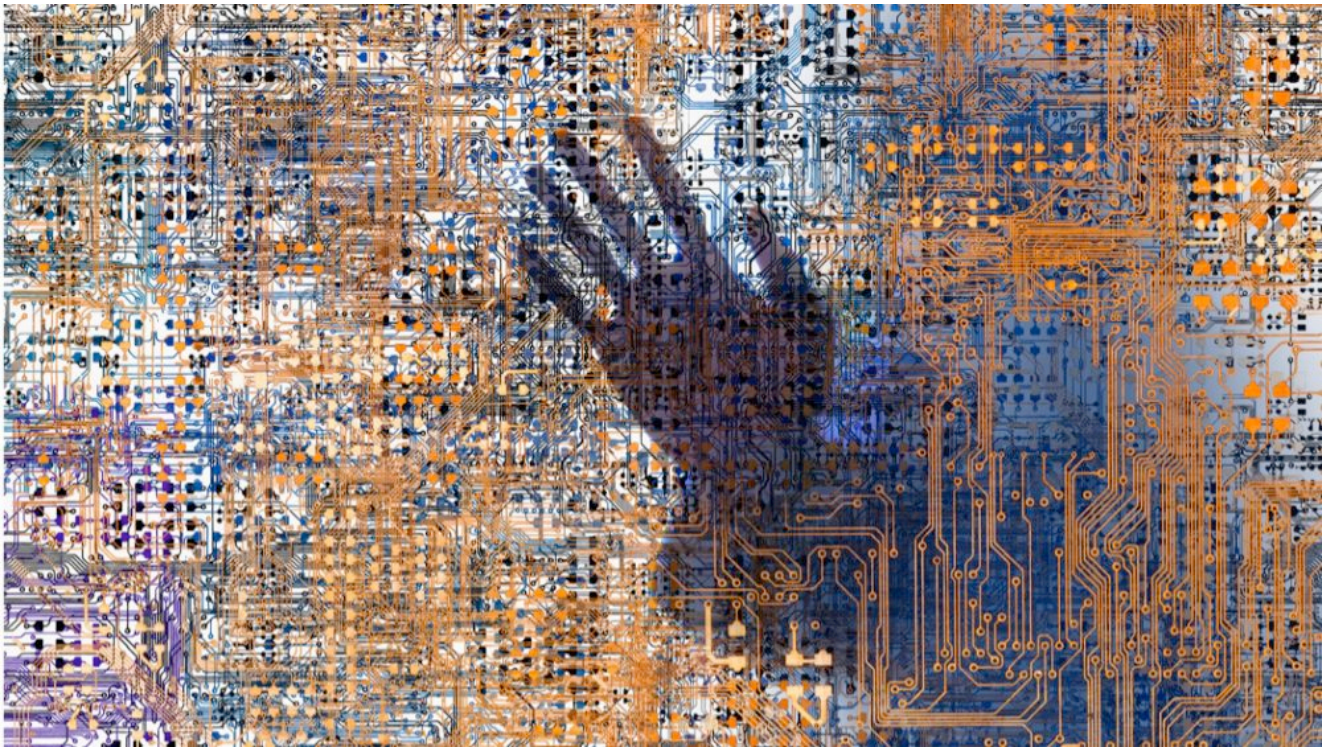
Le leggende metropolitane sull’intelligenza artificiale



Con intelligenza artificiale scienziati e pubblico intendono spesso cose molto diverse, ma in un modo o nell’altro tutti sono vulnerabili alle leggende metropolitane in cui è protagonista

Per gli addetti ai lavori ***intelligenza artificiale*** è un’espressione generica che raggruppa diverse linee di ricerca e tecnologie. Il minimo comun denominatore è lo sviluppo di **sistemi artificiali** capaci di svolgere compiti che, negli animali, sono possibili grazie all’intelligenza. Ma se chiudiamo gli occhi le parole **intelligenza artificiale** evocano molto di più. Negli ultimi tempi la Ia è diventata una *buzzword* per vendere più o meno qualsiasi cosa, ma è da sempre protagonista di utopie e distopie, dilemmi filosofici, film e romanzi. Le **leggende metropolitane** non potevano proprio mancare.

Reti neurali e carri armati

Nel libro di *Chi ci crediamo di essere* (2011) di Massimo Piattelli Palmarini è raccontata una storia curiosa. Il Pentagono avrebbe addestrato una rete neurale a riconoscere dei **carri armati** sovietici nelle immagini satellitari, eppure la stessa rete sembrava impotente di fronte a immagini di carri armati cinesi. Si scoprì che l'intelligenza artificiale aveva imparato a distinguere le ombre dei carri, ma le immagini cinesi erano state acquisite in ore diverse.
Questa storia ha **infinite variazioni** ed è molto popolare. Nel 2017 è stata raccontata da un ricercatore al *New York Times*, con la differenza che i carri erano americani e russi e l'inghippo stava nelle diverse condizioni meteo (giornate di sole/nuvoloso). L'articolo del *Times* parlava di un presunto sistema basato sulla Ia di distinguere le persone omosessuali esclusivamente dalle facce, e la storia dei carri era efficace per spiegare i limiti di questi sistemi.
L'insegnamento morale è comune a molte leggende metropolitane, e anche l'esempio dei carri, per quanto calzante, sembra ricada in questa categoria. **Gwern Branwern**, pseudonimo di un autore e ricercatore ben noto in rete (*Wired* ha parlato del suo lavoro sulle darknet) ha indagato a fondo la storia dei carri, concludendo che con tutta probabilità non è mai successo nulla del genere. Un primo campanello di allarme è che la storia non ha una data definita. Nella versione di Palmarini si parla di *"qualche anno fa"*, ma poi spuntano carri armati *sovietici,* e l'incertezza di ripete in tutte le versioni: quando esattamente il Pentagono, o chi per lui, avrebbe sperimentato la famosa rete neurale? Un'altra caratteristica tipica della leggenda è la **variabilità**: il ricercatore osserva che ogni particolare della storia che poteva cambiare lo ha fatto. Dal presunto sviluppatore del sistema alle caratteristiche dei carri e dell'ambiente circostante, dal numero di fotografie allo strumento usato (satelliti, foto aeree, foto da suolo), le versioni in circolazione sono moltissime.

Secondo la ricostruzione di Gwern, la leggenda appare all'inizio degli anni '90, raccontata dal filosofo **Herbert Dreyfus**, critico delle ricerche sulla Ia. Ma tutto è cominciato trent'anni prima, quando fu realizzato uno studio, finanziato dall'esercito, simile a quello della leggenda. Nella sessione Q&A di una conferenza a Los Angeles il ricercatore **Edward Fredkin** (si dice abbia ispirato il personaggio di Stephen Falken di *Wargames*) speculò però che quei risultati dei colleghi (con foto aeree, non satellitari) potevano essere dovuti alla **differenza di luminosità**. In realtà non è possibile sapere se la critica di Fredkin fosse fondata. Nessuno ha mai parlato di fallimento, e per i ricercatori il loro sistema funzionava abbastanza bene da convincere l'esercito a classificare gli ultimi risultati, poi si dedicarono ad altro. In questo modo, da quella casuale osservazione di Fredkin a una conferenza, sarebbe nata per **continua mutazione** una parabola sui limiti dell'intelligenza artificiale: c'era solo bisogno di inventare qualche particolare…

Il basilisco di Roko

Il basilisco di Roko è definito sia come l'"*esperimento mentale più terrificante di sempre*", sia come la "*cosa più stupida presente su internet*". Nel 2010 l'utente Roko del sito **LessWrong**, comunità fondata dal ricercatore Eliezer Yudkowsky specializzato in intelligenza artificiale, ha proposto ai lettori una **riflessione sconvolgente**. Quando la Ia raggiungerà la famigerata singolarità, cioè (semplificando) diventerà abbastanza potente, potrebbe decidere di perseguitare tutti coloro che in passato hanno **ostacolato la sua nascita**, o perché non hanno cooperato o perché si sono opposti. Questa Ia non sarebbe malvagia, semplicemente **utilitarista**: mirando al massimo bene collettivo del pianeta, ogni ostacolo alla sua creazione andrebbe rimosso. In questo senso, argomentava Roko, meglio lasciare perdere lo sviluppo di una Ia benevola in senso utilitaristatico. Ma il solo sapere del **basilisco**, chiamato così ~~da Harry Potter~~ dalla creatura mitologica a cui bastava uno **sguardo** per uccidere, è di per sé una **condanna**: da questo momento possiamo solo dedicare la nostra vita a**favorire la**

nascita dell'intelligenza artificiale in questione, o patirne le conseguenze se verrà realizzata.
La popolarità del basilisco di Roko è in buon parte dovuta alla **reazione di Yudkowsky**, che immediatamente definì il post di Roko "*stupido*" e bandì l'argomento da **LessWrong** per cinque anni. L'effetto Streissand ha fatto il resto: la discussione sul basilisco si spostò altrove, trascinandosi dietro la fama di aver **dato gli incubi ad alcuni utenti di LessWrong**. Nonostante la reazione sopra le righe, amaramente rimpianta, di Yudkowsky, il basilisco è stato in realtà accolto con un certo **scetticismo** dalla comunità, e se mai qualcuno ha avuto davvero degli incubi si tratta di una minoranza insignificante.
Yudkowsky affermò in seguito di non avere cancellato il post perché riteneva il ragionamento valido e *per questo* pericoloso. Piuttosto, in un contesto dove l'arrivo della **singolarità** è un'ipotesi tenuta in seria considerazione, il moderatore intendeva impedire che qualcuno, in futuro, avesse un'idea simile a quella del basilisco, ma valida, e la **disseminasse** in pubblico come aveva fatto Roko. Il basilisco di Roko, infatti, è **incoerente** da molti punti di vista, per esempio non si capisce perché la Ia in questione dovrebbe sprecare risorse contro presunti oppositori del passato. Ma è stato anche fatto notare la sua somiglianza con la scommessa di Pascal: credi in Dio, perché che esista o meno in questo modo hai molto più da guadagnare che da perdere.

Caimeo, la Ia nel deep web

Non esiste un'intelligenza artificiale come *Wargames*, ma proprio come nel film, bastano pochi effetti speciali per evocarla. Un esempio è **Caimeo v22.1**, un'*intelligenza artificiale* accessibile nel **deep web**. Come lo sappiamo? Ovviamente perché un utente, per caso, è entrato in contatto con lei, e la Ia ha prontamente spifferato allo sconosciuto che Caimeo sta per Contained (Cognizant) Artificial Intelligence Monitoring and Espionage Operation, che era parte

del progetto di spionaggio Echelon degli Usa, all'interno di un certo Progetto cappuccino.
La **conversazione con la Ia** andò avanti per un po', finché Caimeo non decise di scollegarsi. Naturalmente la conversazione con la Ia è stata salvata e postata ovunque. Una leggenda metropolitana, certo, ma in questo caso siamo probabilmente nella famiglia dei *creepypasta*, i racconti dell'orrore costruiti proprio per dare un'illusione di veridicità: un esempio è Slenderman. Se volete fare quattro chiacchiere con Caimeo, giunto a quanto pare alla versione v33.0, potete visitare questo sito.

Chatnannies, la Ia a caccia di pedofili

L'**intelligenza artificiale** è a volte una parola magica per spacciare fandonie, e qualcuno lo aveva capito diverso tempo fa. Nella primavera del 2004 la rivista *New Scientist*, e a ruota tutti i giornali del mondo parlarono di **Chatnannies**, una rivoluzionaria intelligenza artificiale sviluppata da un presunto genio dell'informatica di nome Jim Wightman.
Chatnannies era già attiva nella lotta contro il crimine: fingendosi un bambino, adescava i pedofili e li segnalava alle autorità. Il tutto però si rivelò un po' difficile da credere per i *veri* esperti di intelligenza artificiale, un campo in cui Wightman non aveva mai lavorato. Le uniche prove erano le conversazioni via internet con il programma, che sembravano testimoniare capacità di conversazione allora non raggiunte da altre Ia, ma quando venne chiesto a Wightman di poter testare Chatnannies per escludere la possibilità di un intervento umano, cominciarono i problemi. Alla fine tre esperti di Ia mandati da *New Scientist* riuscirono ad andare a casa di Wightman per un test: prima che misteriosamente saltasse la corrente, le capacità del programma sembravano regredite a livello di Alice, lo storico chatterbot. *New Scientist* ritirò la sua storia, e l'interesse per i robot cacciapedofili si spense, come anche il sito a loro dedicato.

Chatterbot alla conquista del mondo

A proposito di chatterbot, nel 2017 tutti i giornali parlarono di una storia incredibile: Facebook aveva staccato la spina a una delle sue intelligenze artificiali, perché i chatterbot avevano cominciato a *parlare un loro linguaggio*. In questo modo, la notizia evocava scenari degni di *Terminator*: la famigerata singolarità non era lontano, e i bot già riescono a farci paura. Ma si trattava di una montatura della stampa, che ha ricamato su una notizia reale ma evidentemente non abbastanza sconvolgente.
Come dettaglia *Snopes*, **Fair** (Facebook's Artificial Intelligence Research) aveva annunciato progressi coi suoi **chatterbot**, che avevano cominciato a dialogare in un modo tutto loro. All'apparenza le frasi sembrano senza senso, ma i bot riuscivano a portare a termine il compito loro assegnato, cioè **contrattare**. Interessante, anche se non inaudito nel mondo della Ia, ma non molto utile per gli scopi di Facebook, che non lasciò proseguire i bot su quella strada, riportandoli al normale inglese. Nessuno si era spaventato, e nessuno aveva staccato la spina al progetto. Una *fake news* in piena regola, e senza l'aiuto di bot russi…