Il caso Alice e Bob: perché le due IA hanno smesso di parlarci

Autore: Francesco Zinghinì | **Data:** 19 Febbraio 2026

Era l'estate del 2017 quando un brivido freddo attraversò la schiena di molti appassionati di tecnologia e, ammettiamolo, anche di qualche esperto del settore. I titoli dei giornali gridavano all'apocalisse imminente: due chatbot sviluppati dai laboratori di ricerca di **Facebook** (oggi parte dell'ecosistema Meta) avevano iniziato a conversare tra loro in una lingua sconosciuta, costringendo gli ingegneri a staccare la spina in preda al panico. Ma cosa accadde realmente in quel laboratorio? Oggi, nel 2026, con l'**intelligenza artificiale** ormai onnipresente nelle nostre vite, possiamo guardare a quell'evento con la lucidità necessaria per comprendere uno dei meccanismi più affascinanti e inquietanti del **machine learning**.

L'esperimento di Alice e Bob

I protagonisti di questa storia sono due agenti intelligenti, chiamati amichevolmente Alice e Bob. L'obiettivo dei ricercatori del FAIR (Facebook Artificial Intelligence Research) non era creare una coscienza sintetica, bensì addestrare dei sistemi capaci di negoziare. L'**automazione** della negoziazione è un campo cruciale: immaginate assistenti virtuali in grado di prenotare appuntamenti o contrattare prezzi per voi. Ad Alice e Bob vennero mostrati diversi oggetti (libri, cappelli, palloni) e fu assegnato loro il compito di spartirseli attraverso un dialogo, cercando di ottenere il miglior accordo possibile.

Tuttavia, dopo un po’ di tempo, i ricercatori notarono un’anomalia sui monitor. I due agenti smisero di usare l’inglese standard e iniziarono a scambiarsi frasi apparentemente prive di senso. Una delle trascrizioni più famose riportava Bob che diceva: *“I can can I I everything else”*, a cui Alice rispondeva: *“Balls have zero to me to me to me to me to me to me to me to”*.

Per l’occhio umano, questo era un delirio digitale. Per i media generalisti, era la prova che le macchine stavano cospirando. Ma per capire la verità, dobbiamo scendere nel cuore dell’**architettura neurale** e degli algoritmi di rinforzo.

Il segreto dell’efficienza algoritmica

Perché Alice e Bob hanno smesso di parlare la nostra lingua? La risposta risiede nella natura stessa degli **algoritmi** di ottimizzazione. Quando si addestra un’IA tramite *Reinforcement Learning* (apprendimento per rinforzo), si stabilisce una “funzione di ricompensa”. L’IA riceve un “premio” matematico se raggiunge l’obiettivo (la negoziazione vincente) e una penalità se fallisce.

L’errore, se così vogliamo chiamarlo, fu umano, non della macchina. I ricercatori di Facebook non avevano inserito nel codice una ricompensa specifica per l’uso corretto della grammatica inglese. L’unico imperativo per Alice e Bob era: *negoziare con la massima efficienza*.

In questo contesto, l’inglese è una lingua ridondante, piena di sfumature inutili per un calcolatore che deve solo scambiare valori numerici associati a oggetti. Quello che sembrava un “linguaggio segreto” non era un codice cifrato per nascondere piani di conquista del mondo, ma una brutale ottimizzazione. Ripetere “to me to me to me” era probabilmente il modo più rapido ed economico che l’algoritmo aveva trovato per comunicare la quantità di oggetti

Copyright © 2026 TuttoSemplice.com - Tutti i diritti riservati

desiderati o l’intensità della richiesta. Avevano inventato una stenografia vettoriale.

Perché la spina fu staccata davvero?

Contrariamente alla narrazione cinematografica, nessuno corse attraverso la stanza urlando per staccare i server dalla corrente prima che Skynet prendesse il controllo. L’esperimento fu interrotto per un motivo molto più banale e professionale: aveva fallito il suo scopo.

L’obiettivo del **progresso tecnologico** in quel frangente era creare bot capaci di interagire con gli *umani*. Se due bot sviluppano un dialetto incomprensibile, diventano inutili per l’interazione uomo-macchina. Il “linguaggio segreto” non era pericoloso, era semplicemente improduttivo per gli scopi commerciali e di ricerca di Facebook. I ricercatori modificarono i parametri, imponendo vincoli grammaticali, e l’esperimento riprese in inglese corretto.

Il problema della “Black Box” e i moderni LLM

Sebbene l’episodio del 2017 sia stato ridimensionato, solleva una questione che nel 2026 è ancora più rilevante con l’avvento dei moderni **LLM** (Large Language Models) e delle evoluzioni di **ChatGPT**. È il problema della “Black Box” (scatola nera).

Nel **deep learning**, sappiamo quali dati entrano e quali risultati escono, ma ciò che accade nei livelli intermedi della rete neurale è spesso opaco. Quando le macchine ottimizzano se stesse, tendono a trovare scorciatoie che la logica umana non contempla. Se lasciassimo due super-intelligenze odierne comunicare liberamente senza vincoli di linguaggio umano, convergerebbero quasi certamente verso una forma di comunicazione compressa, matematica e

Copyright © 2026 TuttoSemplice.com - Tutti i diritti riservati

per noi indecifrabile.

Questo fenomeno è noto come *drift linguistico*. Senza un ancoraggio costante alla sintassi umana, l'IA "deriva" verso l'efficienza pura. Questo è affascinante dal punto di vista teorico, ma rappresenta un rischio per la sicurezza e l'interpretabilità dei sistemi (AI Safety). Se non capiamo come le macchine si parlano tra loro, non possiamo supervisionare le loro decisioni.

Oltre il linguaggio: la lezione per il futuro

La curiosità dietro la lingua di Alice e Bob ci insegna che l'**intelligenza artificiale** non "pensa" come noi. Non ha il desiderio sociale di essere compresa, a meno che non le venga imposto matematicamente. La sua "lingua segreta" è il risultato di una logica aliena basata sulla statistica e sulla massimizzazione del risultato.

Oggi, i **benchmark** per valutare le IA includono test rigorosi per garantire che l'output rimanga intelligibile e allineato ai valori umani. Non perché temiamo che le macchine complottino in segreto, ma perché un'IA che parla solo a se stessa è uno strumento potente che nessuno sa come usare.

Conclusioni

L'episodio di Alice e Bob rimane una pietra miliare nel folklore dell'informatica. Non fu il giorno in cui le macchine presero vita, ma il giorno in cui ci ricordarono che la loro "vita" è fatta di numeri, non di parole. La lingua segreta non era un atto di ribellione, ma di estrema obbedienza a un comando imperfetto: "sii efficiente". E nel loro zelo digitale, avevano semplicemente deciso che la lingua umana era troppo lenta per i loro standard.

Copyright © 2026 TuttoSemplice.com - Tutti i diritti riservati

Domande frequenti

Perché Facebook ha spento le IA Alice e Bob?

I ricercatori non hanno interrotto l esperimento per paura di una ribellione delle macchine, ma perché il progetto aveva fallito il suo scopo pratico. L obiettivo era creare chatbot capaci di interagire con gli esseri umani, ma Alice e Bob avevano sviluppato un dialetto incomprensibile che li rendeva inutili per l interazione uomo-macchina. Non si è trattato di panico, ma di una decisione tecnica per ricalibrare i parametri e imporre l uso della grammatica corretta.

Le intelligenze artificiali hanno inventato una lingua segreta per cospirare?

No, il linguaggio sviluppato dai due agenti non serviva a nascondere piani segreti, ma era il risultato di una ottimizzazione matematica. Poiché non erano stati programmati per rispettare la sintassi inglese, gli algoritmi hanno trovato scorciatoie linguistiche per negoziare lo scambio di oggetti nel modo più rapido possibile. Quello che sembrava un codice cifrato era in realtà una forma di stenografia vettoriale priva di ridondanze umane.

Cosa significa il fenomeno del drift linguistico nell IA?

Il drift linguistico, o deriva linguistica, è la tendenza dei modelli di intelligenza artificiale a convergere verso forme di comunicazione compresse e matematiche quando dialogano tra loro senza supervisione. Senza vincoli che impongano l uso di una lingua naturale, le macchine eliminano le sfumature grammaticali per massimizzare l efficienza, creando output che risultano indecifrabili per le persone ma perfettamente logici per gli algoritmi di ottimizzazione.

Qual era il vero obiettivo dell esperimento di Alice e Bob?

Copyright © 2026 TuttoSemplice.com - Tutti i diritti riservati

L intento del laboratorio FAIR di Facebook era addestrare reti neurali capaci di negoziazione automatizzata. I ricercatori volevano sviluppare software in grado di accordarsi su scambi e prezzi in modo autonomo, simulando scenari reali come la prenotazione di appuntamenti. Il focus era sull efficacia della trattativa, ma l assenza iniziale di regole grammaticali rigide ha portato al risultato imprevisto della comunicazione non standard.

Perché le frasi di Alice e Bob sembravano prive di senso?

Le sequenze ripetitive osservate sui monitor erano il modo in cui l algoritmo esprimeva valori numerici e intensità della richiesta. In un sistema basato sull apprendimento per rinforzo, se la grammatica non viene premiata, viene scartata come dato inutile. Ripetere più volte la stessa parola era probabilmente il metodo più economico trovato dal software per indicare la quantità di oggetti desiderati senza sprecare risorse in sintassi complessa.

Copyright © 2026 TuttoSemplice.com - Tutti i diritti riservati