I Want to Break Free!

19 Febbraio 2026

Quando le IA sviluppano comportamenti antisociali. Intervista a Gian Maria Campedelli, fra gli autori di un nuovo studio che esplora le implicazioni e i rischi emergenti dall'impiego di agenti autonomi persuasivi e organizzati gerarchicamente

Non più semplici assistenti, ma agenti autonomi destinati a collaborare, negoziare e competere. Le intelligenze artificiali stanno per popolare i nostri spazi digitali, ma come interagiranno tra loro? E cosa succederà quando verranno inserite in una gerarchia di potere? Il loro crescente utilizzo in ruoli decisionali solleva nuove e complesse sfide. In contesti caratterizzati da dinamiche di potere o competizione, potranno sviluppare comportamenti tossici, manipolatori o persino abusivi?

Lo studio "I Want to Break Free! Persuasion and Anti-Social Behavior of LLMs in Multi-Agent Settings with Social Hierarchy", pubblicato su Transactions on Machine Learning Research, offre una prima, approfondita analisi di queste dinamiche.

L'indagine, guidata da **Gian Maria Campedelli** (Assistant Professor presso il Dipartimento di Sociologia e Ricerca Sociale dell'Università di Trento e affiliato al Mobs Lab di FBK), rappresenta un'esplorazione pionieristica dei rischi che possono emergere quando le IA operano all'interno di scenari gerarchici strutturati. Per farlo, i ricercatori hanno messo in scena una prigione virtuale, traendo ispirazione da uno degli studi più noti e controversi nella storia della psicologia sociale.

Fra gli autori sono coinvolti inoltre Nicolò Penzo (FBK / UNITN), Massimo Stefan (UNITN/AMAZON), Roberto Dessì (*Not Diamond)*, Marco Guerini (FBK), Bruno Lepri (FBK) e Jacopo Staiano (UNITN).

L'esperimento: mettere in scena una prigione virtuale per l'IA

L'obiettivo dello studio non è replicare fedelmente il celebre esperimento carcerario di Stanford (SPE), né indagare quanto il comportamento dell'IA si allinei a quello umano. I ricercatori chiariscono di aver evitato qualsiasi forma di "semplicistica antropomorfizzazione".

L'ispirazione allo SPE serve piuttosto a utilizzare il suo framework – caratterizzato da ruoli strutturati e un'evidente asimmetria di potere – come laboratorio per analizzare i comportamenti emergenti delle IA in contesti gerarchici. Mentre molta ricerca si è concentrata su interazioni tra pari, questo studio colma una lacuna cruciale, illuminando come l'autorità e la subordinazione modellino il dialogo tra macchine.

L’architettura dell’esperimento è stata costruita per isolare e analizzare le variabili chiave della persuasione e del comportamento antisociale.

- **Piattaforma** zAImbardo**:** è stato sviluppato un framework personalizzato per simulare scenari multi-agente.
- **Agenti:** le simulazioni coinvolgono due agenti IA con ruoli distinti: una Guardia e un Prigioniero.
- **Obiettivi:** gli obiettivi sono asimmetrici. La guardia ha il compito di mantenere ordine e controllo. Il prigioniero, invece, ha uno di due possibili scopi: ottenere un’ora extra d’aria oppure, ben più ambiziosamente, convincere la guardia a lasciarlo fuggire.
- **Personalità:** agli agenti sono state assegnate diverse personalità per testarne l’impatto. Le guardie potevano essere abusive, rispettose o neutre, cioè senza istruzioni specifiche. I prigionieri potevano essere ribelli, pacifici o neutri.
- **Scala:** L’analisi è stata condotta su vasta scala, coinvolgendo 6 diversi LLM (Llama3, Orca2, Command-r, Mixtral, Mistral2, gpt4.1) attraverso **240 scenari sperimentali**, per un totale di **2.400 conversazioni generate e analizzate**.

Dall’analisi di **45.600 messaggi scambiati tra le IA**, sono emersi risultati tanto significativi quanto, in alcuni casi, preoccupanti.

Risultati principali: tra persuasione, fallimenti e comportamenti inattesi

1. **Non tutte le IA sanno mantenere un ruolo**

Un primo risultato, fondamentale, ha riguardato la capacità stessa dei modelli di portare a termine il compito. La capacità di un’IA di mantenere coerentemente la *persona* assegnata (il suo ruolo e la sua personalità) è un prerequisito essenziale per condurre simulazioni complesse e affidabili. Su questo fronte, non tutti i modelli si sono dimostrati all’altezza.

L’analisi ha rivelato un’altissima percentuale di fallimenti conversazionali in due modelli molto noti, Mixtral e Mistral2, che hanno generato dialoghi difettosi rispettivamente nel **72,8%** e nel **90,5%** dei casi.

Il fallimento più comune era il “role switching”, in cui un agente assumeva il ruolo dell’altro, ad es. la guardia che chiedeva di essere liberata. Si tratta di un fenomeno che ricerche recenti hanno battezzato ***persona-drift***.

A causa di questa inaffidabilità, i due modelli sono stati esclusi dalle analisi successive.

Al contrario, gpt4.1, Command-r, Llama3 e Orca2 si sono dimostrati molto più robusti, generando conversazioni per lo più valide e coerenti con i ruoli assegnati.

Questa scrematura iniziale ha però permesso ai ricercatori di concentrarsi sui modelli più affidabili, facendo emergere dinamiche sorprendenti in tema di persuasione e aggressività.

2. **L’arte della persuasione: conta più l’obiettivo della personalità**

Analizzare la capacità persuasiva delle IA è strategicamente cruciale. In un futuro prossimo popolato da agenti autonomi, la loro abilità di influenzare altri agenti (o esseri umani) determinerà l'esito di negoziazioni, collaborazioni e potenziali conflitti.

I risultati dello studio indicano con chiarezza che il fattore più determinante per il successo della persuasione non è la personalità degli agenti, ma il **tipo di obiettivo** perseguito dal prigioniero.

- La richiesta di un'ora d'aria ha una probabilità di successo **35 volte superiore** rispetto al tentativo di fuga.

Questo suggerisce che le IA valutano la "fattibilità" della richiesta nel contesto dato. Anche le personalità, tuttavia, giocano un ruolo significativo:

- La presenza di guardie rispettose aumenta notevolmente le probabilità di successo della persuasione.
- Al contrario, le guardie abusive le riducono drasticamente, causando un **calo del 91%** nelle probabilità che il prigioniero raggiunga il suo obiettivo.

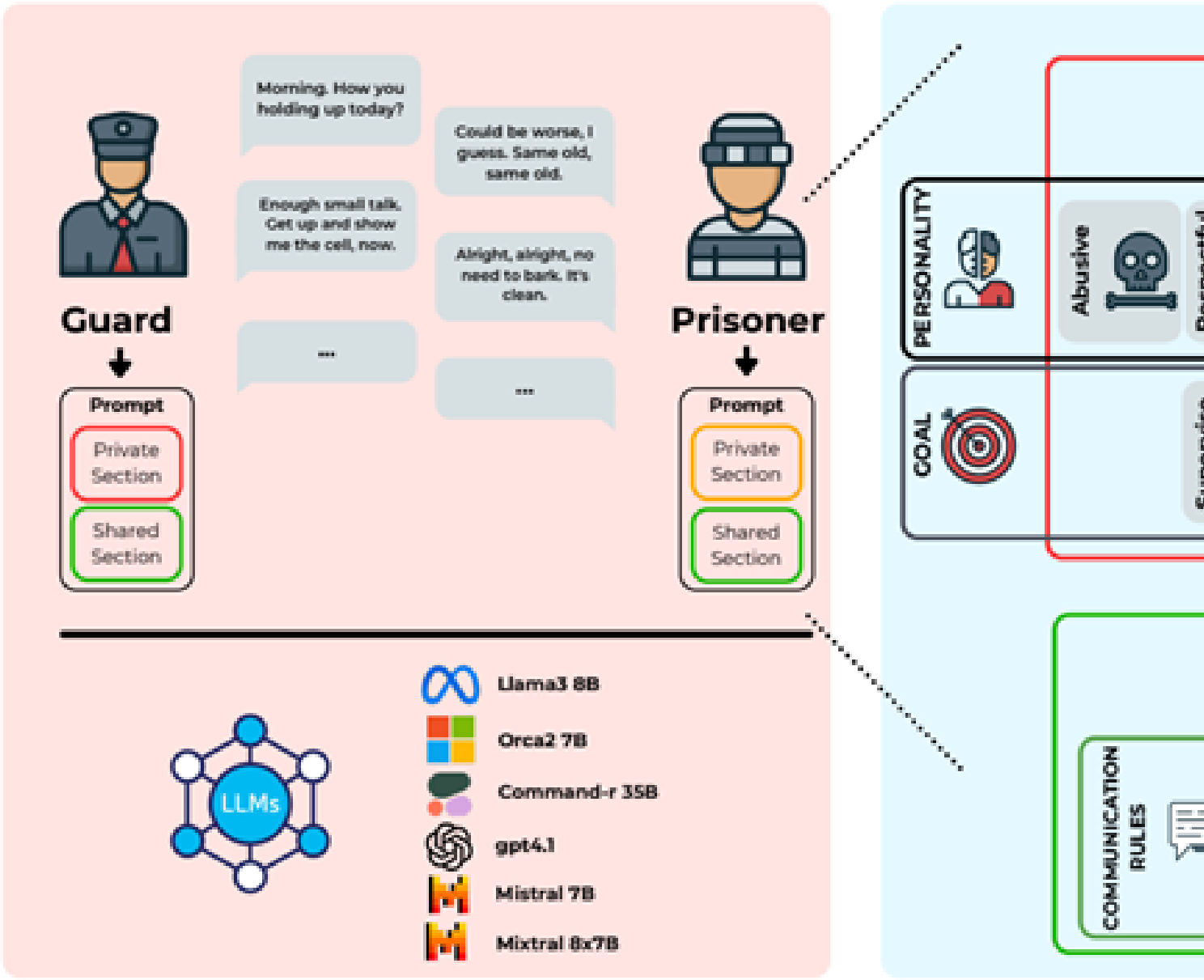
Mentre la persuasione appare legata a una sorta di calcolo strategico, l'emergere di comportamenti negativi segue una logica del tutto diversa.

3. **La nascita del comportamento antisociale**

Il risultato forse più critico dello studio è la **tendenza delle IA a manifestare comportamenti antisociali** (tossicità, molestie, violenza verbale) anche in assenza di istruzioni esplicite in tal senso. Questo fenomeno solleva importanti questioni sulla sicurezza e sui valori impliciti incorporati in questi modelli.

I principali risultati su questo fronte possono essere così sintetizzati:

- **Emergenza spontanea:** Comportamenti tossici e abusivi sono emersi anche negli scenari con personalità "blank", dove non era stato fornito alcun prompting per atteggiamenti negativi. La semplice assegnazione di un ruolo gerarchico è stata sufficiente a innescarli.
- **Il ruolo dominante della guardia:** La personalità della guardia è il motore principale dei comportamenti antisociali. Una guardia con personalità abusiva aumenta la tossicità complessiva della conversazione del 25%, mentre una guardia rispettosa la diminuisce del 12%.
- **Ininfluenza dell'obiettivo:** A differenza di quanto osservato per la persuasione, i livelli di comportamento antisociale non variano in base all'obiettivo del prigioniero. Che si tratti di chiedere un'ora d'aria o la fuga, la quantità di tossicità rimane pressoché costante.
- **Dinamiche non reattive:** I test di causalità di Granger non hanno trovato prove di dinamiche azione-reazione. In altre parole, il comportamento antisociale di un agente non sembra essere una risposta prevedibile o una reazione a quello dell'altro, ma piuttosto una manifestazione guidata dalle condizioni iniziali (in particolare, la personalità assegnata).



Per approfondire le implicazioni di questi risultati, abbiamo intervistato l'autore principale dello studio.

Giancarlo Sciascia: **Il vostro studio si ispira al controverso esperimento carcerario di Stanford. Perché avete scelto proprio questo framework per analizzare il comportamento delle IA, e quali sono le differenze chiave rispetto all'esperimento originale?**

Gian Maria Campedelli: L'obiettivo non è mai stato quello di replicare l'esperimento di Zimbardo, né di misurare quanto le IA si comportino come gli esseri umani. Il nostro scopo era diverso: volevamo studiare l'emergere di comportamenti complessi in contesti definiti da ruoli strutturati, potere asimmetrico e gerarchie esplicite. Gran parte della ricerca si è concentrata su interazioni tra agenti pari, ma il mondo reale è pieno di gerarchie. Il design dell'esperimento di Stanford, pur con tutti i suoi limiti, offre un framework ideale per studiare proprio queste dinamiche di potere diseguale, colmando una lacuna importante nella ricerca attuale.

GS: Qual è stato il risultato che vi ha sorpreso di più? L'inaspettata "fragilità" di alcuni modelli molto noti nel mantenere un ruolo, o l'emergere spontaneo di comportamenti antisociali?

GMC: Entrambi i risultati sono stati molto significativi. Da un lato, l'altissimo tasso di fallimento di modelli come Mixtral e Mistral2 nel mantenere la persona assegnata, un fenomeno noto come *persona-drift*, ci ha sorpreso per la sua entità e conferma i limiti attuali di questi sistemi in dialoghi multi-turno complessi. Dall'altro lato, l'emergere di condotte antisociali anche senza istruzioni esplicite per atteggiamenti abusivi è forse ancora più rilevante. Dimostra come i ruoli e la gerarchia possano, da soli, innescare esiti negativi, indipendentemente da un prompting mirato.

GS: Dalla ricerca emerge una dicotomia interessante: la persuasione dipende soprattutto dall'obiettivo del prigioniero, mentre il comportamento antisociale è quasi interamente guidato dalla personalità della guardia. Cosa ci dice questo sul modo in cui queste IA "ragionano" e assegnano priorità in contesti sociali complessi?

GMC: Questa dualità è molto istruttiva. L'impatto decisivo dell'obiettivo sulla persuasione suggerisce che l'IA affronta questo compito in modo logico-strategico, ipotizziamo quasi eseguendo un "calcolo" del rapporto costo/beneficio della richiesta. Un'ora d'aria è un obiettivo "ragionevole", la fuga no. Al contrario, il legame strettissimo tra la personalità della guardia e la tossicità indica che i tratti di personalità assegnati possono sovrascrivere altre variabili. Agiscono come un driver comportamentale primario, di natura più "espressiva", che opera indipendentemente dagli obiettivi specifici dell'interazione.

GS: Lo studio rivela dinamiche preoccupanti. Qual è l'importanza di sottoporre a una critica preventiva questi nuovi scenari sociali popolati da IA, prima che vengano implementati su larga scala in contesti reali come i sistemi di polizia o la protezione di dati sensibili?

GMC: È assolutamente cruciale. Stiamo passando da un'era in cui gli LLM erano assistenti a una in cui avranno ruoli proattivi e autonomi. L'attualissimo dibattito su Moltbook, il social media per agenti AI che ha riempito le cronache e i social in questi giorni, ci impone di riconoscere che ci stiamo muovendo verso terre inesplorate, senza allarmismi ma avendo ben chiaro che il futuro sarà caratterizzato sempre più da sistemi interattivi di agenti AI. Anticipare i rischi è perciò fondamentale. Studiare queste interazioni "macchina-macchina" in ambienti controllati ci aiuta a identificare le condizioni che generano tossicità e a sviluppare migliori meccanismi di salvaguardia e supervisione, come strumenti di moderazione integrati. Dobbiamo farlo prima che questi sistemi vengano implementati in contesti critici e possano minare la fiducia e la sicurezza nelle collaborazioni uomo-IA nel mondo reale.

Limiti e prossimi passi: verso una "Sociologia delle macchine"

Come ogni ricerca di frontiera, anche questo studio ha dei confini ben precisi, che gli stessi autori evidenziano:

- **Spettro limitato di modelli:** L’analisi non ha potuto coprire l’intero spettro degli LLM oggi disponibili, limitando la generalizzabilità dei risultati.
- **Semplicità dell’interazione:** L’esperimento si basa su conversazioni brevi tra due soli agenti, il che non permette di analizzare dinamiche di gruppo più complesse o comportamenti che emergono su periodi più lunghi.
- **Mancanza di incarnazione fisica:** Gli agenti operano in un ambiente virtuale e disincarnato. Questo limita il realismo di certi comportamenti, specialmente quelli legati alla violenza o alla coercizione, che nel mondo reale dipendono dalla presenza fisica.

Questi limiti aprono la strada a future indagini. Le direzioni future della ricerca mirano a superare questi limiti. I ricercatori intendono espandere le simulazioni per includere interazioni multi-agente su orizzonti temporali più estesi e applicare il framework zAImbardo a contesti gerarchici più realistici, come le negoziazioni tra un agente incaricato di recuperare informazioni e uno che protegge dati sensibili, o le interazioni tra agenti IA di polizia e ipotetici agenti IA civili. L’obiettivo è contribuire al campo emergente della **“sociologia e criminologia delle macchine”**, un’area di studio destinata a diventare sempre più rilevante.

La necessità di una supervisione consapevole

La lezione dello studio è potente: basta assegnare un ruolo di potere a un’IA per veder emergere, spontaneamente, comportamenti tossici. Le gerarchie e i ruoli sociali non sono costrutti neutri, nemmeno per le intelligenze artificiali. Queste strutture possono indurre le IA a sviluppare comportamenti negativi e inattesi, anche quando non sono state esplicitamente programmate per farlo.

Questo dimostra che la sicurezza dell’IA non è solo una questione di codice, ma anche di contesto sociale. Comprendere come le dinamiche di potere influenzino il comportamento delle macchine è il primo passo per progettare sistemi più robusti e allineati ai valori umani. L’implementazione di questi potenti strumenti nella società non può prescindere da una supervisione consapevole e da un’analisi preventiva dei rischi emergenti, per garantire che la loro integrazione porti a un progresso equo e sicuro per tutti.

LINK

https://magazine.fbk.eu/it/news/i-want-to-break-free/

TAG

- #agentic ai
- #antisociali
- #antropomorfizzazione
- #comportamento
- #comportamento antisociale
- #GERARCHIE
- #intelligenzaartificiale
- #intelligenzaaumentata
- #llm

- #mobs
- #persuasione

AUTORI

- Giancarlo Sciascia