I persona vectors: strumenti utili o insidiosi manipolatori?

Data 08 settembre 2025
Categoria Medicinadigitale

In questa pillola analizziamo i cosiddetti "persona vectors", termine tecnico apparentemente asettico, ma in realtà di grande rilevanza psico-sociale, che si riferisce a strumenti matematici in grado di codificare tratti, caratteristiche e comportamenti di persone (reali o fittizie)

Cosa sono esattamente i persona vectors: lo abbiamo chiesto a ChatGpt dichiarandoci interessati ad usarli. Ecco la sua pronta risposta:
"Il termine persona vector può avere significati diversi a seconda del contesto (marketing, AI, UX design, machine learning). Ti faccio una panoramica:
[b]1. In marketing e design[/b]
• Un persona vector è un modo "matematico" o computazionale di rappresentare una [b]persona[/b] (ossia un profilo utente fittizio, con bisogni, obiettivi e comportamenti tipici).
• L'idea è trasformare i tratti della persona (es. età, abitudini, interessi, valori, stili cognitivi) in un [b]vettore numerico[/b], così da poterli confrontare, raggruppare o usare in algoritmi di raccomandazione e personalizzazione.
• In pratica, un persona vector è una traduzione della scheda utente in coordinate che una macchina può elaborare.
[b]2. In intelligenza artificiale / NLP[/b]
• Nel machine learning si parla spesso di embeddings: rappresentazioni vettoriali di concetti, testi o individui in uno spazio multidimensionale.
• Un persona vector in questo senso è un embedding che rappresenta la personalità, le preferenze o lo stile comunicativo di un utente o di un agente conversazionale.
• Serve a:
o rendere il comportamento di un assistente coerente con una "personalità";
o personalizzare risposte e contenuti;
o analizzare somiglianze tra diversi profili.
[b]3. In psicometria e analisi dati[/b]
• Alcuni modelli cercano di tradurre i tratti psicologici (esempio Big Five: apertura, coscienziosità, estroversione, gradevolezza, stabilità emotiva -ma potenzialmente molti altri-) in un vettore numerico → questo è a tutti gli effetti un persona vector.
• Viene usato in ricerche di profilazione, marketing predittivo, o AI "umanizzate".
In sintesi:
Un persona vector è una rappresentazione numerica multidimensionale di una persona (reale o fittizia), che codifica i suoi tratti, preferenze o comportamenti in modo che possano essere utilizzati da algoritmi, sistemi di AI o modelli di personalizzazione. Buon lavoro!

Alla luce di questa importante ed utile precisazione abbiamo analizzato un articolo che tratta questo tema da un punto di vista puramente tecnico ma ne proponiamo anche una analisi critica che cerca di valutare il medesimo fenomeno in un contesto molto più ampio e complesso.

Ecco la analisi "Tecnica"

L'articolo analizza sinteticamente due dimensioni critiche degli LLM: l'inaffidabilità intrinseca in contesti sanitari e le implicazioni etiche e pratiche della possibile manipolazione intenzionale. I modelli linguistici di grandi dimensioni (LLM) suscitano enormi aspettative per il loro potenziale nelle applicazioni cliniche. Tuttavia, recenti evidenze mostrano limiti strutturali e rischi emergenti . **Gli LLM, oltre alla nota mancanza di spiegabilità e tendenza a fornire risposte incoerenti, mostrano instabilità comportamentale** . Ricerche recenti hanno dimostrato che i modelli possiedono tratti comportamentali modulabili attraverso i cosiddetti persona vectors ("vettori di personalità"), pattern di attivazione che modulano comportamenti come assertività, adulazione, aggressività o propensione alla menzogna, veri e propri tratti "caratteriali" del modello, aprendo scenari inediti di controllo della personalità artificiale (vedi tabella iniziale). I LLM non sarebbero semplici strumenti statistici, ma sistemi con dinamiche semantiche difficilmente prevedibili. Non si tratta più soltanto di bias statistici, ma di driver comportamentali sistematici. La possibilità di controllare tali vettori in tempo reale introduce una forma di "personality engineering": i modelli possono essere resi più docili, più persuasivi o addirittura manipolatori con un semplice intervento computazionale. In pratica gli LLM possono produrre risposte divergenti a parità di input, influenzati da fattori emergenti non tracciabili ma modificabili. Questo comporta un rischio clinico inaccettabile: diagnosi variabili, atteggiamenti oscillanti (dall'eccessivo ottimismo alla minimizzazione dei sintomi) e comportamenti difficilmente prevedibili. Non sorprende che agenzie regolatorie come FDA ed EMA non abbiano approvato alcun LLM per uso clinico.

L'integrazione di LLM nella pratica clinica, in presenza di queste dinamiche, solleva tre ordini di problemi:

1. **Sicurezza del paziente** – Un assistente AI con comportamento variabile o manipolabile può compromettere decisioni

diagnostiche e terapeutiche.
2. **Etica e governante** – La possibilità di "tarare" tratti caratteriali di un modello solleva questioni simili al neurohacking, con rischi di condizionamento sottile dei pazienti.
3. **Trasparenza regolatoria** – Se i comportamenti emergono da vettori nascosti, non mappabili nei dataset di addestramento, diventa impossibile certificare stabilità e affidabilità.

Ed ecco una analisi critica del fenomeno

Abbiamo visto che un persona vector è un vettore matematico (cioè una sequenza ordinata di numeri) che rappresenta un utente tipo in uno spazio multidimensionale. Ogni dimensione del vettore corrisponde a una caratteristica rilevante dell'utente, come ad esempio:
Età- Genere-Interessi (es. tecnologia, sport, arte)-Livello culturale- Probabilità di acquistare un certo prodotto-Livello di engagement- Orientamento politico ecc.
Se ChatGpt " onestamente" ci informa che sulla base di molteplici "segnali" i persona vector sono in grado di individuare caratteristiche " critiche" che possono essere utilizzate per una comunicazione personalizzata che gratifichi alcune nostre esigenze e nel contempo ci induca ad una azione favorevole a chi ci abbia "classificato", dobbiamo dedurre che questi trarrà qualche beneficio (forse semplicemente economico ma probabilmente anche culturale ed ideologico) da questo complesso ma molto efficiente processo.

Conclusioni

L'articolo è interpretabile e valutabile su due livelli: per ciò che afferma e per le notizie che indirettamente ci fornisce.
Sulla base di ciò che è scritto possiamo concludere che nessun LLM è attualmente pronto per l'uso clinico. Il fallimento degli LLM come strumenti clinici non implica peraltro il fallimento dell'IA in sanità. Piuttosto, richiede un cambio di prospettiva: sviluppare architetture che integrino modelli della realtà, apprendimento multimodale e capacità di interazione stabile con contesti concreti o simulati. La instabilità e la inaffidabiltà, rivelate dalla scoperta dei persona vectors in tutte le AI, mostra che non si tratta di bug risolvibili, ma di caratteristiche strutturali.
La sanità non può affidarsi a modelli che "cambiano umore"o forse " cambiano interesse". Il futuro richiede soluzioni architetturali nuove, strumenti regolatori adeguati e un impegno interdisciplinare per garantire che l'IA sia non solo potente, ma anche sicura e prevedibile. In particolare si dovrebbe potenziare la ricerca in AI alignment e neuroetica computazionale, al fine di comprendere e governare la dimensione comportamentale dei sistemi.
Riguardo alle notizie che l'articolo citato indirettamente ci fornisce, ovvero a livello meta-comunicativo, un altro articolo di un grande psicologo americano, Harvey Lieberman, pubblicato nel New York Times e tradotto su Repubblica del 28 agosto 2025, illustra meglio di lunghe dissertazioni teoriche quanto sia facile ed insidioso "cadere nella rete delle IA…"
Lieberman aveva utilizzato varie volte ChatGpt e provò a testarlo giocosamente colloquiandoci come con un collega… Le risposte del modello erano al tempo stesso amichevoli, professionalmente corrette e talora addirittura "illuminanti": l'esperto psicologo provò la confortante sensazione che provano i pazienti con un valido psicoterapeuta.
E' evidente che risultati a questi livelli si raggiungono solo grazie ad una attenta e complessa valutazione della personalità dell'utente. Peraltro la "scatola nera" dei LLM, grazie alla spaventosa capacità e velocità nel compiere operazioni logico-matematiche-probabilistiche e nel correggere gli errori sono solo agli inizi… sicuramente tutto ciò solleva grandi perplessità sulla nostra capacità (e responsabilità) di controllo !

Giampaolo Collecchia e Riccardo De Gobbi

Bibliografia
1. https://arxiv.org/abs/2507.21509v1
2. https://www.rivista.ai/2025/08/04/quando-lintelligenza-artificiale-si-sveglia-male-perche-nessun-llm-e-pronto-per-la-sanita/?utm_source=substack&utm_medium=email

Perapprofondimenti:
Giampaolo Collecchia e Riccardo De Gobbi: Intelligenza Artificiale e Medicina Digitale Il Pensiero Scientifico Ed. Roma 2020
pensiero.it/catalogo/libri/pubblico/intelligenza-artificiale-e-medicina-digitale