

La DARPA finanzia un programma di IA che identifica la violazione delle norme sociali

L'Università israeliana di Ben-Gurion ha sviluppato un'**intelligenza artificiale** che è in grado di identificare attraverso un'analisi del testo le descrizioni in cui vengono menzionati **comportamenti che violano le norme sociali**. Il progetto è finanziato dalla **DARPA**, agenzia governativa del Dipartimento della Difesa degli Stati Uniti incaricata dello sviluppo di nuove tecnologie per uso militare. Il [report](#) in cui si espone il progetto dichiara esplicitamente che «l'identificazione automatica delle norme sociali e della loro violazione è una sfida aperta che può essere molto importante per diversi progetti, come l'ingegnerizzazione di interpreti digitali a supporto di interazioni culturali e programmi di comprensione culturale computazionale di DARPA che ha finanziato il progetto in corso».

I ricercatori coinvolti nel progetto non si sono concentrati tanto sulle violazioni delle norme sociali, che possono variare da cultura a cultura, bensì sulle **emozioni sociali corrispondenti a determinate situazioni**, in quanto il repertorio di emozioni utilizzato per rispondere a determinate violazioni è piuttosto limitato. «Poiché il numero di norme sociali può essere enorme, un modo semplice e naturale di apprendere le norme è attraverso un numero limitato di emozioni sociali che sono fondate sull'evoluzione e profondamente associate a un sistema di valutazione universale degli esseri umani», si legge. In base al tipo di emozione che emerge da una determinata situazione, dunque, l'IA - nel caso specifico il modello GPT3 - è in grado di stabilire se ci sia stata o meno la violazione di una norma.

Il metodo consiste sommariamente nell'inserire un **input** rappresentato da un elenco di situazioni, ognuna delle quali è etichettata in base ad un'emozione sociale (ad esempio senso di colpa). Successivamente viene applicato il cosiddetto "**classificatore zero shot**" che utilizza **dieci valori qualitativi** per classificare i comportamenti. I dieci valori sono professionalità, educazione, disciplina, attenzione, gradevolezza, affidabilità, successo, conformità, decenza, lealtà e i relativi opposti. Per ogni situazione quindi il classificatore utilizza due etichette: una che indica l'aspetto positivo della norma (ad es. cortesia) e l'altra che ne indica l'aspetto negativo (ad es. scortesia). A questo punto, viene calcolata la differenza tra la probabilità dell'aspetto positivo e l'aspetto negativo della norma. L'**output** è rappresentato da un elenco di situazioni, accompagnata da dieci punteggi, ognuno dei quali indica la misura in cui il classificatore zero-shot le ha giudicate "aderenti alla norma" (cioè, punteggio positivo) o in "violazione della norma" (cioè, punteggio negativo).

Si tratta di un progetto che mostra come l'IA può essere utilizzata per **monitorare i comportamenti umani**, rappresentando dunque uno strumento in più, particolarmente pervasivo ed efficace, ai fini del **controllo sociale**. Il fatto che sia finanziato dalla DARPA può significare che il governo americano è interessato a sviluppare tecniche e metodologie

La DARPA finanzia un programma di IA che identifica la violazione delle norme sociali

efficaci per disciplinare i comportamenti e per fare applicare più efficacemente determinate norme. Il metodo impiegato può essere, inoltre, utilizzato facilmente sui social network per **censurare eventuali opinioni scomode**, classificandole come non conformi alle norme sociali. Quest'ultime, del resto, possono rivelarsi aleatorie ed essere stabilite a seconda del contesto sociale o delle convinzioni ideologiche facenti capo a determinate correnti politiche.

Attraverso l'IA, inoltre, il comportamento e le emozioni umane vengono interpretate secondo poche semplici variabili (ad esempio buono-cattivo; successo-insuccesso) predeterminate che non tengono conto né del contesto né della storia pregressa dell'individuo, cercando di **ridurre ad una mera funzione matematica** l'essenza e il significato delle frasi e dei comportamenti, così da poterli più facilmente inquadrare e classificare. Per questo è difficile pensare che un sistema di questo tipo possa effettivamente servire a studi sociologici e non piuttosto a sperimentare le potenzialità dell'IA nel disciplinare le masse e classificarne pensieri, emozioni e azioni, andando a minare sempre di più lo spazio del libero arbitrio. Il rischio è quindi quello che i modelli di apprendimento automatico si trasformino in un vero e proprio **strumento di governo** - piuttosto potente - per censurare e orientare i comportamenti, secondo i canoni delle [tecnocrazie](#) occidentali, in cui, non a caso, è la tecnica a dominare e a plasmare la realtà.

[di Giorgia Audiello]



Vuoi approfondire l'argomento?

Ventitré esperti di livello internazionale selezionati da L'Indipendente, affrontano con chiarezza e rigore i principali aspetti sociali, individuali e tecnologici del futuro che ci attende con la diffusione dell'IA.

Acquista ora