1. Settembre

Non chiamatele allucinazioni: sono semplicemente stronzate

(parte prima: da Frankfurt a James)

Uno dei tratti salienti della nostra cultura è la quantità di stronzate in circolazione
Harry Frankfurt

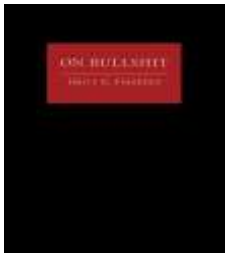
Un studio dal titolo **Chat is bullshit** curato da tre ricercatori dell'Università di Glasgow
Pubblicato 8 giugno 2024 su

Ethic and Information tecnology.
Volume26.articolo 38

Michael Townsen Hicks James Humphries Joe Slater

Ritiene che gli attuali modelli linguistici: sistemi di apprendimento automatico che producono testo e dialoghi simili a quelli umani sono afflitti da persistenti imprecisioni nel loro output; queste sono spesso chiamate "allucinazioni dell'IA".

Michael, James e Joie sostengono che queste falsità e l'attività complessiva dei grandi modelli linguistici siano meglio definite come *stronzate* nel senso esplorato dal talentuoso filosofo statunitense **Harry Frankfurt:** i modelli sono in un modo importante indifferenti alla verità dei loro output **Stronzate,** o se preferite il titolo originale inglese **"On Bullshit".**



A partire da *Wittgenstein* attraverso *Pound* fino a *Sant'Agostino,* **Frankfurt** costruisce una **"teoria generale delle stronzate**" offrendoci un formidabile strumento che ci consente di analizzare alcune aberrazioni della cultura contemporanea.

Uno dei tratti salienti della nostra cultura è la quantità di stronzate in circolazione. Tutti lo sanno. Ciascuno di noi dà il proprio contributo. Tendiamo però a dare per scontata questa situazione. Gran parte delle persone confidano nella propria capacità di riconoscere le stronzate ed evitare di farsi fregare. Così il fenomeno non ha attirato molto interesse, né ha suscitato indagini approfondite. Di conseguenza, non abbiamo una chiara consapevolezza di cosa sono le stronzate, del perché ce ne siano così tante in giro.

Per meglio comprendere il pensiero di Frankfurt è necessario distinguere tra la bugia e la stronzata. Mentre il "bugiardo" costruisce un'affermazione falsa a partire da una verità che conosce molto bene, il "bullshitter" racconta stronzate gratuite al solo scopo di impressionare e stupire il proprio interlocutore.

Il bugiardo è costretto a conoscere la verità in tutti i suoi dettagli per poterla meglio nascondere o contraffare, al contrario del "bullshitter" che non fa uso alcuno della verità o della nozione di verità. Per questo motivo**, Frankfurt** afferma che "*la stronzata è un nemico della verità perché è più grande della menzogna".*

I motivi per cui tutti noi a volte raccontiamo bugie sono innumerevoli e comprensibili molte volte indispensabili al punto da considerarle a volte giustificabili. *Giuseppe Prezzolini* nella **"Vita intima"** ritiene che la bugia non sia soltanto una maschera che creiamo per proteggere le nostre intimità, ma sopratutto un potente moltiplicatore dell'io: ***Se in poesia la ricerca della rima può suggerire un'immagine, se in scienza un casuale avvicinamento di parole può rivelare un'idea, la bugia può essere nella vita il modo di centuplicare la nostra esistenza.***

La tendenza a raccontare stronzate si esalta quando si invita qualcuno a parlare di argomenti di cui sa poco o nulla, come ad esempio nel rito della presentazione di un libro. In questo caso responsabile delle "stronzate" non è solo chi le dice ma anche chi le ha provocate inserendolo nel parterre dei presentatori.
George Bernard Shaw fa dire al suo Pigmalione: *In fondo che cos'è la vita se non una serie di ispirate sciocchezze? La difficoltà consiste nel trovare il modo di commetterle.*

Rassegniamoci, siamo completamente immersi in un mondo di stronzate e insieme a noi persino gli oggetti che ci circondano: nessuno e nemmeno le cose riescono a sottrarsi alle stronzate. Nella classifica personale di ***Edmond de Goncurt***, il noto gallerista francese, *La cosa che è costretta ad ascoltare più stupidaggini al mondo è probabilmente un quadro esposto in un museo* osservato e commentato ogni giorno da migliaia di persone.

Adesso anche i prodotti dell' Intelligenza artificiale contribuiscono ad arricchire *, si fa per dire* , il mare di stronzate su cui galleggia l'umanità

I ricercatori scozzesi non usano mezzi termini,distiguono due modi in cui i modelli possono essere considerati dei cialtroni e sostengono che descrivere le false rappresentazioni dell'IA come stronzate è un modo più utile e più accurato di prevedere e discutere il comportamento di questi sistemi.
Le Intelligenze artificiali generative, come **ChatGPT**, dicono fesserie ma non hanno le allucinazioni che, nello specifico

Nella prima sezione del lavoro , delineiamo come operano ***ChatGPT*** e ***LLM simili***. Poi, prendono in considerazione la visione secondo cui quando commettono *errori fattuali*, mentono o allucinano: cioè, pronunciano deliberatamente falsità, o le pronunciano in modo incolpevole sulla base di informazioni di input fuorvianti.

Sostengono così che nessuno di questi due modi di pensare è accurato, nella misura in cui sia la menzogna che l'allucinazione richiedono una certa preoccupazione per la verità delle loro affermazioni, mentre gli ***LLM*** semplicemente non sono progettati per rappresentare accuratamente il modo in cui è il mondo, ma piuttosto per *dare l'impressione* che questo è ciò che stanno facendo.
Questo, suggeriamo, è molto vicino ad almeno un modo in cui Frankfurt parla di stronzate.

Tracciano così due tipi di stronzate, che chiamiamo **stronzate "dure"** e "**morbide**", dove la prima richiede un tentativo attivo di ingannare il lettore o l'ascoltatore sulla natura dell'impresa, e la seconda richiede solo una mancanza di preoccupazione per la verità.

Sostengono che, come minimo, gli output di **LLM come ChatGPT** sono delle cazzate soft: cazzate, ovvero discorsi o testi prodotti senza preoccuparsi della loro veridicità, prodotti senza alcuna intenzione di trarre in inganno il pubblico sull'atteggiamento di chi li pronuncia nei confronti della verità.
Suggeriscono, in modo più controverso, che **ChatGPT** potrebbe effettivamente produrre delle **cazzate hard: se lo consideriamo come avente delle intenzioni (ad esempio, in virtù di come è progettato), allora il fatto che sia progettato per dare l'impressione di preoccuparsi della verità lo qualifica come un tentativo di trarre in inganno il pubblico sui suoi obiettivi, traguardi o agenda.**

Quindi, con l'avvertenza che il particolare tipo di cazzate che **ChatGPT** produce dipende da particolari visioni della mente o del significato.
Concludono che è appropriato parlare del testo generato da ChatGPT come di cazzate e sottolineano perché è importante che, piuttosto che pensare alle sue affermazioni non vere come bugie o allucinazioni, chiamiamo **cazzate ChatGPT.**

Questa querelle che a mio avviso esploderà nei prossimi mesi andrebbe integrata con il pensiero di **A. James**

Aaron James è un giovane brillante professore di filosofia dell'Università di California, noto anche per essere un eccellente surfista.Ha un curriculum prestigioso, le sue principali pubblicazioni sviluppano il paradigma liberal-egalitario di *Rawls e Dworkin* in maniera originale e intelligente.

Tuttavia la sua notorietà è esplosa per un suo libretto dal titolo poco accademico **"Ass-holes: a Theory"** ormai conosciuto in tutto il mondo e da noi in Italia tradotto da Rizzoli come: in cui si ***"Stronzi: un saggio filosofico"*** che si propone di comprendere la loro essenza filosofica**,** un'opera utile che riprende il discorso iniziato da Henry Frankfurt,

Il libro, pieno di cose interessanti e godibili, descrive e classifica quelle falangi di individui che sono soliti profittare degli altri senza neppure sentire il bisogno di giustificare con che diritto e a che titolo lo facciano. I vari capitoli consentono di ricostruire una originale tassonomia che va da ***"nuovi stili di stronzate" a "stronzate deliranti"*** passando per l'esplorazione di aspetti psicologici nel capitolo ***"lo stronzo che è in noi"*** fino a un "***pratico sistema affidabile per ridurre la profusione di stronzi".***

Gli **"ass-holes"** sono persone particolarmente fastidiose capaci di suscitare reazioni di rabbia e timore assieme. Sono descritte come persone che si arrogano sistematicamente il diritto di godere di speciali vantaggi nei rapporti interpersonali grazie ad un inopinato sentimento di superiorità che li rende sfacciatamente immuni alle critiche.

Lo **"stronzo"** non perde mai il proprio tempo, perde quello degli altri; è colui che, quando noi facciamo le dovute rimostranze di fronte ai suoi vergognosi soprusi, si mostra sorpreso e rifiuta di prenderle sul serio. In questo modo, lo **ass-hole** non rispetta la nostra dignità semplicemente perché non sa cosa sia ed ignora il principio aristotelico per cui la dignità non consiste nel possedere onori, ma nella consapevolezza di meritarli.

La loro arroganza può, agli occhi dei meno esperti, somigliare vagamente alla dignità, anche della dignità non ha assolutamente nulla. Secondo James, gli stronzi sono irritanti ed insopportabili perché attraverso loro non vediamo riconosciuto "il nostro status di persona morale".

Uno dei valori fondamentali della vita è la dignità. Non bisogna mai barattarla, gli ass-holes hanno scoperto che per non perderla è sufficiente non averla. Il giudizio di James è di natura morale e nasce sulle tracce di *Rousseau e Kant.* Per James gli ass-holes sono ripugnanti e ondamentalmente pericolosi perché rifiutano di considerare gli altri come eguali dal punto di vista morale.

Domani andremo nel dettaglio della critica scozzese e analizzeremo perché ChatGPT e i grandi modelli linguistici a volte sono poco attendibili e potenzialmente pericolosi.

To be continued…