



UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI ECONOMIA “GIORGIO FUÀ”

Corso di Laurea triennale in

Economia e commercio

CRITICA ALLE INTELLIGENZE ARTIFICIALI

**CRITICAL THOUGHTS ON ARTIFICIAL
INTELLIGENCES**

Relatore:

Prof.

Orazi Francesco

Rapporto Finale di:

Sopranzi Nicholas

Anno Accademico 2022/2023

Indice

1. Introduzione

2. Cos'è il linguaggio

2.1. Punto di vista naturalistico di Searle

2.2. La stanza cinese

3. Mente e Macchina

3.1. Il test di Turing

3.2. Costruire una mente

3.3. Due cervelli a confronto

3.4. InfantLevel

4. Allucinazioni AI

4.1. Gli LLM

4.2. Tipi di conflitto

4.2.1. Conflitto di input

4.2.2. Conflitto di contesto

4.2.3. Conflitto di fatti

4.2.4. Output senza senso

4.3. Perché Alluciano?

4.3.1. Qualità dei dati

4.3.2. Sovrastima delle proprie capacità

4.3.3. Metodo di generazione

4.3.4. Contesto degli input

4.4. Mitigazione degli AI

4.4.1. Durante il pre-addestramento

4.4.2. Durante il SFT

4.4.3. Durante il RLHF

4.4.4. Uso di conoscenza esterna

5. Esperimenti e casi reali

5.1. Thinking about you

5.2. Prega.org

5.3. I bias di GPT

6. Rischi e Costi

6.1. Disinformazione

6.2. Bad Advice

6.3. Bias

7. Conclusione

Bibliografia

1. INTRODUZIONE

L'interesse nel settore delle intelligenze artificiali è esploso con l'avvento del nuovo modello di LLM (large language model) chiamato Chat-GPT chiudendo uno degli "Inverni delle AI" (periodi di poco

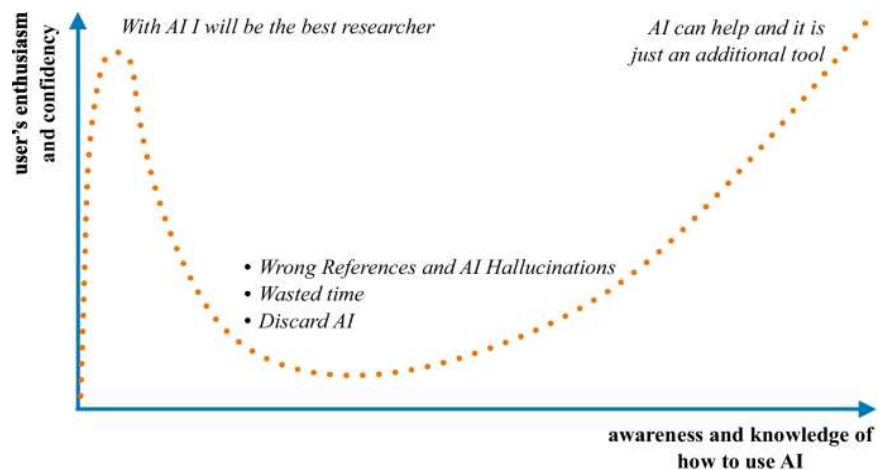


Figura 1

interesse). La loro storia nasce decenni fa con i primi tentativi di Alan Turing di creare macchine capaci di giocare a scacchi, imparando le strategie del gioco. Da quel momento l'ingegneria informatica e il calcolo neurale hanno visto uno sviluppo senza precedenti portando alla creazione di modelli capaci di comprendere il nostro linguaggio e replicarlo. Ad oggi, infatti, è possibile instaurare una conversazione con questi chatbot proprio come se si avesse davanti un essere umano capace di comprendere le domande e completare task complesse. Sembra essere una "rivoluzione": abbiamo raggiunto il punto di singolarità, predetta da Rey Kurzweil per il 2045, in cui l'intelligenza delle macchine supera quella dell'uomo. Le applicazioni sono infinite: possiamo chiedergli di scrivere codice Python per creare software, di scrivere racconti per bambini, di costruire interi business partendo da un'idea o creare film d'animazione con una storia accattivante. Ma se l'AI è capace di agire in tutti questi campi, una volta permessi solo agli esseri umani, come muterà il mercato del lavoro? Verremo sostituiti da "artisti digitali" con una creatività preimpostata o da "psicologi macchine" che contengono la conoscenza di un'intera società?

La realtà è ben diversa e le opinioni sensazionalistiche diffuse su queste nuove tecnologie sono presto deluse dall'effettiva distanza che ancora separa l'umano dalla macchina. Il fermento dietro il nuovo prodotto di OpenAI ha portato le persone a considerarlo l'ultimo elemento di evoluzione dei nostri tempi, suscitando anche timore per la loro eccessiva potenza. Studiando meglio i meccanismi nascosti dietro le menti artificiali, si comprende il perché non costituiscono nessuna minaccia per le professioni e come, al contrario, la loro mancanza di intelligenza potrebbe creare danni. L'obiettivo di questa analisi è quella di offrire delle riflessioni critiche sul ruolo sempre più influente di queste tecnologie nella società, sfatando la fama di super esseri creatasi sul loro conto. Si vuole riportare i lettori su una strada di consapevolezza dei limiti di queste intelligenze, incoraggiando un corretto utilizzo dello strumento e diminuendo le loro criticità.

Per rispondere alla domanda: "Le intelligenze artificiali sono veramente intelligenti?" la tesi partirà da una disamina del linguaggio naturale e capiremo se si può considerare espressione di intelligenza. Metteremo a confronto la capacità dell'essere umano di costruire un linguaggio naturale per nominare e descrivere la realtà, con le formule di apprendimento dei modelli artificiali. Saranno esposti nei primi capitoli le incapacità semantiche delle intelligenze che sono, al minimo, solo dei manipolatori di simboli senza reale coscienza di sé.

In seguito, verrà posta l'attenzione sui meccanismi che permettono il funzionamento degli LLM. Esporremo brevemente le formule matematiche poste alla loro base e il perché non riescano ad essere applicate alla realtà per come la percepiscono gli esseri umani. I modelli usano funzioni probabilistiche per determinare la parola successiva e non il ragionamento di senso per rispondere alle domande che gli vengono poste. Vengono addestrati su una grande quantità di dati da cui apprendono i pattern del discorso umano. Questo processo di apprendimento si svolge con delle regole diverse dalle nostre e questa differenza ha delle implicazioni reali.

Esistono fenomeni che si tendono a verificare spesso interagendo con gli LLM. Questi sono chiamati allucinazioni e mettono in discussione la loro affidabilità e potenza perché rendono i testi contraddittori e non veritieri. Verranno analizzati i vari tipi di allucinazioni, nonché le motivazioni per cui presentano e come intervenire per ridurli.

Riporteremo dei casi reali ed esperimenti che hanno sollevato le preoccupazioni legate alla loro applicazione a scenari reali. Quando un AI fallisce, una macchina a guida autonoma si schianta, una minoranza viene discriminata o un chatbot esprime opinioni razziste. Con questi eventi si è potuto toccare con mano l'impatto negativo di queste tecnologie sulla società.

Infine, analizziamo i potenziali rischi che si possono correre nell'implementazioni di questi strumenti nel mercato, nel contesto democratico in cui viviamo e anche nella realtà personale di ogni individuo.

2. COS'È IL LINGUAGGIO?

Nel mondo moderno è comune dare per scontato una abilità come il linguaggio. Lo si impara nella fase di prima crescita, ascoltando i suoni provenienti dalle bocche dei nostri genitori attraverso un meccanismo di imitazione e di associazione mondo-parola o azione-parola. In qualche modo, le nostre menti primordiali riescono a collegare un senso, o un'intenzione, o ancora un desiderio, a quei suoni. Non lo facciamo da soli, siamo accompagnati, infatti, dalle nostre madri e padri a scandire le prime parole, le quali daranno un nome ai primi volti e ai primi giochi. Questo significa che si acquisisce la abilità di usare uno strumento di cui le regole sono già definite da tempo, di cui le strutture già state costruite e tramandate da innumerevoli generazioni. Il neonato non ha "creato" una forma di comunicare e nominare quelle immagini, ma ha adottato quella già esistente attorno a lui al momento del suo arrivo in questa realtà. Quindi la domanda da porsi è: qual è l'origine di questo strumento? Cosa c'era prima del linguaggio? Negli ultimi 50 si è sviluppato un grande interesse verso il tema

dell'evoluzione del linguaggio, il quale ha condotto a varie ricerche su questo misterioso step del genere umano. Nonostante questa ricchezza di studi, non abbiamo nessuna spiegazione del perché e del come si siano evolute le nostre rappresentazioni linguistiche.

L'articolo "The mystery of language evolution" pubblicato su *Frontier in psychology* nel 2014 spiega *"le prove fossili e archeologiche non ci informano sulla comprensione dei calcoli e delle rappresentazioni dei nostri primi antenati, lasciando irrisolti i dettagli delle origini e della pressione selettiva; la nostra comprensione della genetica del linguaggio è così povera che c'è poca speranza di collegare presto i geni ai processi linguistici"*

Siamo sicuri, però, che si sia trasformato durante i millenni, si è scisso e fuso svariate volte, è stato manipolato e trasformato da generazioni e popoli. Quindi avendo come certo un punto 0, in cui l'umanità non padroneggiava un linguaggio e un momento 1 in cui la padroneggiava, la nostra attenzione va nell'analisi di quel periodo di mezzo. Non è stato un processo immediato, ma, probabilmente, nel giro di migliaia di anni, si è iniziato ad associare gesti del corpo con suoni emessi vocalmente.

2.1. PUNTO DI VISTA NATURALISTICO DI SEARLE

Chiameremo la fase 0 "coscienza prelinguistica" partendo dal concetto dello studioso americano Jhon Searle che si è formato nella scuola di filosofia del linguaggio di Oxford. Searle è noto per le sue teorie riguardo all'intenzionalità del linguaggio, dove rifiuta la differenza tra semantica e pragmatica, e per la sua critica all'idea dell'intelligenza artificiale forte, supportata dall'elaborazione di un esperimento: la stanza cinese. Nel 1980 Searle è il primo a distinguere l'intelligenza artificiale nel concetto di "strong AI" e "weak AI" sollevando obiezioni contro la prima distinzione. La AI debole vede questa tecnologia come un mero strumento a servizio dell'umanità, considerato la sua distinta potenza di calcolo e precisione; dunque, avrebbe un ruolo solo complementare all'intelligenza umana, come già lo

hanno gli assistenti vocali, i moderatori virtuali, sistemi di riconoscimento del testo, previsioni etc. In opposizione, il concetto forte dell'AI assume qualità simili a quelle di una persona, riuscendo svolgere operazioni consciamente; con il programma appropriato si vuole creare una mente, capace di propri processi cognitivi come motivazione, forza di volontà, consapevolezza del sé.

Il pensiero di Searle si basa sulla coscienza umana vista come fenomeno biologico, al pari di qualsiasi altro fenomeno biologico, come la fotosintesi e la riproduzione. È In contrasto con la logica per cui la coscienza non appartiene al mondo materiale, è solo un'illusione, un elemento che non esiste, o che, se esiste, non può far parte della scienza, perché la scienza è oggettiva, mentre la coscienza è soggettiva. Secondo questo pensiero, la comunità scientifica si rifiutò in passato di prendere in considerazione le convinzioni di Searle, perché valutavano questo aspetto come parte della sfera spirituale e non materiale. Ma cos'è la coscienza? Anche se molti risponderebbero con difficoltà, per J. Searle la definizione viene quasi naturale come spiega in un TED talk lui stesso nel luglio del 2013: “ *La coscienza consiste in tutti quegli stati di sentimenti, sensazioni o consapevolezza; inizia al mattino quando ti svegli da un sonno senza sogni e continua tutto il giorno finché non ti addormenti di nuovo* ”

Questa sfera umana è strettamente collegata con il linguaggio, perché innestiamo dentro questa forma di comunicazione, un'intenzione, che proviene da noi in quanto umani. Siamo esseri capaci di desideri, credenze e memorie che esistono già in noi, e che vengono espresse con atti linguistici. Il concetto naturalistico, per l'appunto, dà importanza al senso del linguaggio, considerato come un'estensione della nostra intenzione, che prima si presentava sotto forma di percezione e azione. Ancora prima di nominare il mondo, eravamo già capaci di desiderare qualcosa (desiderare un frutto lontano in un albero) o capaci di percepire o provare sensazioni. Condividiamo queste caratteristiche con molte specie di animali che non hanno bisogno di una struttura verbale per avere desideri, intenzioni e credenze; anche loro

possono pensare. Il pensiero, infatti, non è composto da parole, ma bensì da immagini (o da immagini di parole) e questo lo porta evolutivamente ad un momento antecedente allo sviluppo del linguaggio.

Il linguaggio è composto da tre componenti principali: la parte fonetica, quella sintattica e quella semantica. La fonetica del linguaggio attribuisce dei suoni ai simboli, che sono le parole e stabilisce regole di pronuncia; la sintassi invece descrive come creare la struttura di una frase, cioè la disposizione delle parole; la semantica è il componente che contiene il significato e l'interpretazione delle costruzioni verbali. Searle discute di come la semantica prescindendo la sintassi, infatti quest'ultima organizza la semantica in modo da poterla esprimere: la relazione tra questi due aspetti è cruciale. Nello stato di coscienza prelinguistica degli ominidi che camminavano questa terra, era già presente una semantica, erano capaci di avere un processo di pensieri e di vivere una serie di esperienze, nonché di intenzionalmente di agire.

Questi asset di intenzionalità sono elementi comuni sia al momento 0 che al momento 1, ma quali sono le caratteristiche che il linguaggio aggiunge alla coscienza prelinguistica? Cosa possiamo fare con una struttura verbale che solo con l'intenzionalità e la coscienza non possiamo fare? La sintassi permette all'interlocutore di manipolare la struttura del linguaggio e i suoi elementi per comporre frasi e modificare il significato del messaggio che si vuole passare; ha la libertà di creare infiniti forme fraseologiche diverse, piene di significato per rappresentare uno stato di cose o esprimere uno stato interno della sua persona. La differenza più immediata con un animale è che quest'ultimo può avere processi di pensieri collegati alla sua percezione della realtà, come noi, però non può separare per esempio il pensiero "fa caldo" dal pensiero "ho caldo". È vero che esistono casi di specie viventi che possono differenziare una serie di suoni e associarli a diversi oggetti (come per esempio "cibo" o "pericolo") riescono a riconoscerli nel tempo e ad associarci un suono, ma questa è una forma

primitiva di comunicazione che non equipaggia l'animale con un apparato complesso, composto di elementi con un senso proprio da manipolare.

Un'altra differenza fondamentale tra la prelinguistica e la linguistica è la segmentazione. Il pensiero senza le parole è un flusso di coscienza non delimitato da una struttura, è fluido e composto da immagini e sensazioni; mentre invece quando pensiamo in lingua, spezziamo quel flusso e lo disponiamo in spazi precisi con cui componiamo la frase. Searle parla anche di “dichiarazione” come caratteristica fondamentale del linguaggio; con il semplice atto di dichiarare qualcosa, abbiamo il potere di crearlo. La forza della parola può far nascere una realtà che sarà condivisa da tutta la società in base agli asset di credenze che condividiamo. Per esempio, qualcuno potrebbe enunciare “Dichiaro guerra. “ e in quel momento crea una situazione che tutti percepiscono come vera e ne comprendono il significato.

Dopo aver spiegato nel dettaglio cosa è il linguaggio, specificando le parti di cui è composto e la differenza tra un sistema prelinguistico di coscienza e una struttura evoluta di comunicazione, possiamo introdurre l'esperimento della stanza cinese elaborato da Searle nel 1980 e pubblicato in un articolo della rivista *Minds, Brains and Programs*. Un grande tema affrontato dal filosofo è l'intelligenza artificiale, la sua critica alla teoria della AI forte lo ha portato a dibattere sull'effettiva comparazione tra cervello e computer, infatti, Nel 1980, è il primo studioso a distinguere i concetti di “strong AI” e “weak AI”, sollevando obiezioni contro la prima distinzione. La AI debole vede questa tecnologia come un mero strumento a servizio dell'umanità che avrebbe un ruolo solo complementare all'intelligenza umana, come già lo hanno gli assistenti vocali, i moderatori virtuali, sistemi di riconoscimento del testo, algoritmi di previsioni etc. In opposizione, il concetto forte dell'AI assume qualità simili a quelle di una persona, riuscendo svolgere operazioni consciamente; con il software appropriato si vuole creare una mente, capace di stati cognitivi come motivazione, forza di volontà, consapevolezza del sé. Un hardware, con il giusto programma, può essere equiparato

ad un cervello umano, in quanto potente calcolatore di informazioni equipaggiato con una coscienza, la quale riesce a raggiungere livelli di comprensione analoghi a quelli delle persone. Il filosofo americano mette in dubbio la vera intelligenza dei computer attraverso l'esperimento che riguarda appunto il linguaggio. Una macchina, dotata di processi computazionali, che interagisce con un umano in una conversazione, può manipolare simboli e creare una struttura leggibile. Questo non stabilisce una garanzia della sua vera comprensione di quell'atto verbale.

2.2. LA STANZA CINESE

Si immagini una persona in una stanza chiusa, che non ha nessun contatto con l'esterno eccetto per una finestra, dalla quale si ricevono dei fogli con un testo scritto in caratteri cinesi. La persona non conosce il cinese, non lo ha mai studiato e né avuto mai contatto con la lingua, non saprebbe neanche distinguerlo da un'altra lingua asiatica. Alla persona viene consegnato un altro set di fogli (chiamata "storia") sempre in cinese ma con delle istruzioni in lingua madre (Inglese nel caso di Searle) che consentono di associare il primo set di fogli al secondo. In questo modo la persona acquisisce la capacità di associare dei simboli ad altri simboli e riconoscerne le forme. In seguito, vengono consegnati altri fogli, cioè "le domande", dove c'è altro testo in lingua straniera e altre regole che specificano come rispondere ai quesiti usando i simboli. In questo momento l'individuo dentro la stanza ha tutti gli strumenti per dare risposte complete usando un set di simboli e, da un punto di vista esterno, non si noterebbe la differenza tra un questo individuo e un conoscente della lingua. L'unica discrepanza con una madrelingua è che quest'ultimo comprenderebbe la storia e il senso delle domande, mentre l'altro si limiterebbe a creare una struttura verbale attraverso una manipolazione non formale di simboli non interpretati. La persona semplicemente si comporta come un computer, come spiega Searle nel suo paper:

“As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program”

Il punto fondamentale dell'esperimento è proprio dimostrare come un computer non abbia idea del contenuto di quelle frasi, ma agisca solo secondo le istruzioni fornite, creando un linguaggio apparentemente uguale a quello di essere umano. Il lavoro di sintassi che una macchina esegue non presume che l'interlocutore possa comprenderne la semantica, questo perché è privo di fenomeni di intenzionalità. La mente umana, infatti, non può essere paragonata ai circuiti di un computer, perché è qualcosa di molto di più della semplice manipolazione di simboli formali. Noi siamo degli organismi, con una certa struttura biologica, e questa struttura, sotto certe condizioni, ha la capacità di produrre comprensione, percezione, apprendimento e altri fenomeni di intenzione, essendo l'intenzione un fenomeno biologico. Searle sostiene che nonostante un computer possa avere successo nel test di Turing, cioè un interlocutore possa scambiare per una persona reale, la sua capacità nell'imitare una conversazione umana non prova che la macchina ne abbia compreso il contenuto.

3. MENTE E MACCHINA

Può una macchina pensare? A questa domanda J. Searle risponde di sì: però vale solo per quelle tipologie speciali di macchine, che hanno gli stessi poteri di causalità di un cervello. Il cervello ha molto in comune con un computer digitale, svolge operazioni computazionali sia in termini algebrici ma anche sopra la struttura sintattica delle frasi in memoria. Gli stati mentali sono, al minimo, stati computazionali che lavorano con algoritmi proprio come una macchina, però quest'ultima lo fa inconsciamente, mentre la mente lo fa consciamente. Quindi le operazioni che avvengono nel cervello possono essere simulate digitalmente? E se sì fino a che livello? Anche se gli algoritmi neurali possono essere replicati, questo non

indica che la mente può essere l'equivalente di un programma informatico, perché ha caratteristiche intangibili che superano il concetto vuoto di algoritmo.

3.1. IL TEST DI TURING

Lo stesso quesito viene posto da Alan Mathison Turing, matematico e crittoanalista britannico, considerato il padre dell'informatica moderna. È famoso per aver decifrato il sistema in codice "Enigma" usato dall'esercito tedesco nella Seconda guerra mondiale per trasmettere informazioni. Le spie inglesi riuscivano ad intercettare le comunicazioni ma non a comprenderne il significato, Turing con il suo contributo riesce a far capovolgere le sorti del conflitto. Uno dei suoi lavori più noti è proprio il test di Turing, dove cerca di rispondere alla domanda "le macchine possono pensare?". Dimostrare l'attività di pensiero di un computer è una sfida tanto difficile quanto lo è dimostrarlo per un essere umano, come si può dare prova del pensiero di una persona se solo questa persona ha la certezza di pensare? Nel test, l'atto di pensare, per semplicità, viene sostituito con l'atto di parlare, una forma di espressione ben riconoscibile e attestabile.

Un computer può parlare come un umano? Per verificarlo viene impostato un "imitation game" dove partecipano un giudice e due giocatori: il giudice instaura una conversazione scritta con i giocatori, che poi verranno valutati. Uno dei due giocatori viene sostituito ad un certo punto da un robot, che risponderà alle domande con l'obiettivo di ingannare il giudice nel pensare che sia umano. Il computer dovrà fingersi una persona e fornire risposte credibili: il test è superato se le sue risposte non sono facilmente distinguibili dalle risposte di un essere umano, in questo caso verrebbe considerato intelligente. Alan Turing ha previsto che entro l'anno 2000, i computer (con capacità di memoria di gran lunga più elevata) avrebbero facilmente passato il test. La macchina anche in questo caso andrà ad usare sempre meccanismi computazionali e algoritmi per creare risposte sensate, avrà a disposizione una

grande quantità di dati e conversazioni umane reali su cui fare affidamento. Imiterà qualcosa di già esistente e formulerà nuove strutture verbali sulla base delle regole che sono state programmate in lui. Gli sviluppatori dell'intelligenza hanno già instillato istruzioni precise nel suo programma, inoltre la capacità di memoria lo porta anche ad imparare da interazioni precedenti. Questo è abbastanza per affermare che il computer può pensare? Il linguaggio umano è un fenomeno così complesso che non può essere rappresentato da tutte le regole di sintassi e non può essere neanche contenuto in tutte le pagine del vocabolario.

3.2. COSTRUIRE UNA MENTE

È dal lavoro di Turing che storicamente è partita la sfida alla comprensione e ricreazione del cervello attraverso processi computazionali artificiali. Proprio da questo obiettivo si sono mossi i primi passi nel campo del Calcolo neurale, che hanno portato a raggiungere tecnologie come i Large Language models (di cui parleremo più avanti). Questo ambizioso progetto, tutt'ora in corso, ha però come condizione l'effettiva realizzabilità, messa in dubbio dall'ipotesi della AI debole. Francis Crick, biologo molecolare e neuroscienziato che ha vinto il Nobel per la medicina nel 1962, considerava le esperienze umane come la somma dei comportamenti delle cellule cerebrali; da questo presupposto di Strong Ai gli studiosi hanno basato le loro ricerche e permesso lo sviluppo nel Neural computing. I primi tentativi di ricreare i processi che avvengono nel cervello si sono concentrati nel comprendere, modellare ed emulare le funzioni neurologiche contenuti nei cervelli animali. Il matematico Walter Pitts e il neurofisico Warren McCulloch hanno inaugurato le prime ricerche ed esperimenti nel 1943. La sfida posta dai due scienziati, infatti, compriva numerosi campi della scienza come Neuroscienza computazionale e la scienza cognitiva.

Il calcolo neurale ha delle differenze sostanziali in confronto alla computazione con metodi algoritmici. Quest'ultima è caratterizzata dallo svolgere istruzioni derivate da un codice

predefinito dagli sviluppatori e scritto in linguaggio macchina; mentre il Calcolo Neurale (Neural Computing) ha come modalità quella di apprendere dalla esperienza pregressa. Infatti, le reti di “nodi adattabili” riescono ad imparare da esempi di task e immagazzinare conoscenza che sarà utile in momenti futuri. Per imitare un cervello si parte da questi nodi adattabili con cui si costruisce una rete neurale artificiale (ANN) connettendoli insieme in gruppi con caratteristiche precise. Esistono diversi tipi di Artificial Neural Networks: le reti con un solo strato collegante i vettori di input e di output; le reti multi-strato, che nascondono altri strati di nodi adattabili permettendo un apprendimento più complesso; le reti neurali ricorrenti, le quali riescono a formare delle sequenze temporali. Le ANN prendono ispirazione, dunque, dalle reti neurali biologiche degli esseri viventi e usano una struttura di riferimento chiamata Spazio Euclideo. Questa struttura, in breve, è una architettura continua e lineare che permette il calcolo e la rappresentazione delle informazioni attraverso una regola di apprendimento. Il funzionamento dello Spazio Euclideo si oppone alle relazioni non-lineari e complesse rappresentate nel mondo reale che spesso sono caratterizzate da variabili multiple e da fenomeni di causa effetto che sfuggono al sistema lineare geometrico su cui sono basate le ANN. Il modello di Deep Learning si può descrivere come una catena di semplici e continue trasformazioni geometriche che mappa uno spazio vettoriale ad un altro, cioè una approssimazione di funzioni semplice.

3.3. LE CARATTERISTICHE DEI DUE CERVELLI

Le tecnologie ANN, nei decenni, si sono sviluppate oltre questo limite della approssimazione delle funzioni semplice, riuscendo a creare distribuzioni di dati di più forme e di maggiore complessità, per esempio sviluppando discorsi, immagini, musica. Nonostante queste nuove infrastrutture generative e performance nel problem solving, falliscono nel superare i limiti dello Spazio Euclideo, perché questo rappresenta il nucleo delle loro reti neurali artificiali su

cui si basano le loro operazioni di apprendimento e creazione. A confronto, il nostro cervello è molto flessibile, non è limitato da costrizioni matematiche. È in grado di adattarsi a nuovi argomenti creando dibattiti, mentre le reti neurali artificiali si bloccano manifestando allucinazioni (approfondiremo più avanti il fenomeno). In più possiamo definire più di una intelligenza: oltre quella logica e matematica esistono l'intelligenza linguistica, emotiva, spaziale, sociale o interpersonale. Questi ulteriori spettri di abilità e consapevolezze toccano una molteplicità di campi, come le relazioni con altri umani, oltre che sentimenti e sensi che superano le funzioni matematiche delle ANN. Ovviamente esistono dei tratti deboli del cervello umano, scarseggiamo nella capacità di calcolo matematico o nella memoria a breve termine o nell'aver atteggiamenti non influenzati da pregiudizi, anche inconsci. Per questo abbiamo sviluppato tecnologie che ci assistono a coprire le nostre debolezze. Questi modelli sono, per l'appunto, ottimi nei campi statistici e matematici, eccellendo in computazioni di immensa articolazione e portata, data la loro potenza di calcolo che va oltre i limiti biologici degli esseri viventi. I modelli di Deep Learning sono anche capaci di una grande versatilità perché risolvono task articolate, multi-linguistiche o che coprono domini multipli.

Ovviamente il funzionamento del calcolo neurale di queste tecnologie è molto più complesso e tecnico, però il nostro obiettivo in questa sede è di rendere comprensibile la costruzione delle reti neurali artificiali e definire le differenze con le nostre reti neurali. Dopo questa analisi si possono derivare le motivazioni strutturali matematiche per le quali ancora questi modelli non riescono a raggiungere le stesse performance dei cervelli umani in alcuni campi. Gli strumenti e le modalità con cui vengono costruiti non permettono un apprendimento simile a quello umano, e quindi, date le loro fondamenta, si presuppone che non saranno mai capaci di sviluppare una coscienza e altre capacità innate, a meno che non si cambia il modo di costruirle. Come suggerisce Gary Marcus: *“Dovremmo smetterla di costruire sistemi di computer che riescono a migliorare solo nel individuare pattern statistiche in set di dati –*

usando spesso approcci come il Deep Learning – e iniziare a costruire sistemi di computer che, dal momento del loro assemblaggio afferrano tre concetti base: tempo, spazio e casualità”.

3.4. INFANTLEVEL

È imperativo comprendere come funziona il cervello umano per costruire una AI in grado di capire il mondo nella esatta maniera in cui lo facciamo noi. Per esempio, per riconoscere le immagini un ANN potrebbe essere addestrato con il database ImageNet, che contiene più di 14 milioni di foto che sono state categorizzate ed etichettate da umani. Ma non è questo il modo con cui un bambino imparerebbe a nominare le cose. L’esperienza e l’esplorazione del mondo attraverso i sensi nel processo di crescita sono completamente diversi e non comparabili all’apprendimento dell’AI. Nonostante per un bambino il percorso di apprendimento sia molto più lento, è stato dimostrato come eccellono in alcune capacità innate, delle doti già presenti in infanti di 4 mesi, le quali sono state dimostrate da uno studio del 2022. Il progetto di Luca Weihs e Amanda Rose Y. (e altri) è stato pubblicato nel giornale “Transaction on Machine Learning Research” e ha studiato con che misura gli AI moderni comprendono il mondo fisico. Per farlo vengono messe a confronto le capacità nella comprensione fisica degli oggetti in movimento di alcuni neonati con quella di 10 diverse architetture neurali sviluppate per processare video. Nello studio si illustra come neonati di 4.5 mesi dimostrino ragionamenti corretti riguardo alla disposizione fisica di oggetti e l’interazione tra questi in uno spazio reale. Gli esperimenti svolti durante la ricerca riguardano fenomeni come la gravità, inerzia e persistenza degli oggetti (solidità, coesione o continuità degli oggetti che viaggiano attraverso spazio e tempo). Un esempio di test ha studiato la reazione dell’AI e dei bambini posti di fronto ad eventi fisicamente possibili e poi ad eventi fisicamente impossibili. Si sono riscontrati elementi di sorpresa nei bambini per le scene

impossibili (tendono ad osservare di più eventi inaspettati), mentre nessuna reazione avviene nei sistemi di AI che riconoscono i video. I primi riescono ad interpretare un grande numero di interazioni tra oggetti. L'età a cui presentano queste abilità porta a sviluppare l'ipotesi della conoscenza fondamentale, per la quale gli infanti sono dotati effettivamente di un'innata comprensione dei principi fisici. I ricercatori che hanno lavorato al progetto hanno elaborato un benchmark chiamato INFlevel, questo contiene migliaia di video sintetici disegnati per valutare i sistemi di AI. Il benchmark è reso disponibile a tutti come strumento opensource che potrà essere utilizzato su futuri modelli di intelligenza per testare la loro comprensione dei principi fisici. Si potrà tracciare il loro progresso verso l'acquisizione di queste capacità nello spazio-tempo.

In conclusione, lo studio, attraverso questi test, viene dimostrato come le AI hanno una posizione molto arretrata agli umani nella comprensione di 3 concetti fisici fondamentali. Nel pratico, le mancanze di queste tecnologie sono importanti per la loro applicabilità a sostegno dell'umanità? Se riguardano intelligenze che operano prevalentemente su internet, forse la risposta è no. Ma delle preoccupazioni affiorano nel momento in cui vengono implementati, per esempio, nella guida autonoma delle auto o altri mezzi. Come possono agire nel caso di eventi inaspettati o oggetti estranei, non avendo la nostra comprensione del mondo fisico? Il progresso verso AI costruiti su nuove strutture di calcolo potrebbe essere fondamentale per evitare di correre potenziali rischi.

4. ALLUCINAZIONI DELLE AI

I Large language models (LLM) sono degli algoritmi applicati alle intelligenze artificiali che le rendono capaci di riconoscere contenuti, riassumerli, tradurli e non solo. Sono detti anche Deep learning, perché elaborano grandi moli di dati in fasi di pre-training, da cui apprendono e elaborano un linguaggio con cui rispondere ai quesiti degli utenti. Riescono anche a tenere

memoria del materiale raccolto con i prompt inseriti nel programma e ad imparare da questi migliorandosi. Questi modelli hanno avuto una storia frammentata, dovuto a periodi di perdita di interesse e, di conseguenza, di finanziamenti per la ricerca su questo settore. L'obiettivo degli ingegneri informatici era quello di creare una tecnologia che fosse in grado di comprendere il linguaggio umano come lo facciamo noi, una intelligenza artificiale generica che potesse essere abbastanza flessibile essere capace di completare grandi range di task, come un cervello umano. Infatti, di programmi specializzati per scopi specifici ne abbiamo già, l'autopilota delle auto elettriche, il super giocatori di scacchi e molti altri; per questo si è cercato di far assorbire tutta la conoscenza disponibile nelle LLM con la speranza che ad un certo punto nascesse una coscienza. Dal 2020, il rilascio del prodotto di Open AI "GPT-3" ha riaccessato l'attenzione di molti sulla applicabilità di questi algoritmi a servizio dell'umanità, date le loro grandi capacità. Chat GPT 3 è un tipo di algoritmo di deep learning che è disegnato per creare testo simile a quello umano, il modello viene allenato in modo da generare testo che segue lo stile con i dati di training originali.

4.1. TIPI DI CONFLITTO

Nonostante gli LLMs hanno dimostrato notevoli performance nel risolvere vari tipi di task, alzano perplessità sulla loro tendenza a manifestare delle **allucinazioni**; fenomeni che possono mettere in discussione la loro credibilità e la loro applicazione agli scenari del mondo reale. Gli LLM tendono, in casi specifici, ad avere output che deviano dai fatti o dalla logica, e possono partire da incoerenze minori fino ad arrivare a contraddizioni strutturali e a informazioni totalmente fabbricate, perdendo fluidità. In pratica, si inventano fatti e appaiono molto sicuri mentre lo fanno. Le allucinazioni si possono categorizzare a seconda del tipo di conflitto che si crea: Conflitto di input, conflitto di contesto, conflitto di fatti e output senza senso.

4.1.1 Conflitto di input

Questo tipo di allucinazione si presenta nel momento in cui il contenuto generato devia dalla fonte di testo fornita dagli utenti. Quando questo accade, l'intelligenza fraintende l'intento dell'utente che riceve un output in netta contraddizione con il prompt. Di solito il sistema si inceppa su dettagli molto semplici del testo inserito, facilmente individuabili anche da un bambino, che difficilmente commetterebbe errori del genere. Riporto una tabella di conversazioni con l'AI trovata nel paper di Kun Zang: nell'esempio sottostante si può notare la contraddizione tra l'input inserito e il testo generato in risposta. L'utente ha istruito il modello nel sintetizzare un contenuto, comando che è stato eseguito apparentemente con rigore eccetto per un elemento fondamentale; il nome della persona viene sostituito erroneamente con un altro nome (Hill a Lucas).

4.1.2. Conflitto di contesto

Questo è il genere di allucinazione più semplice e si verifica quando emergono delle contraddizioni in un testo con del testo generato precedentemente. Di solito accade quando vengono fornite risposte piuttosto articolate e di una certa lunghezza, dove la LLM perde il filo del discorso non riuscendo a mantenere coerenza durante la conversazione. Grazie a questi errori si nota come esistono dei limiti nel mantenere una memoria di lungo termine. Prendendo in considerazione l'esempio della tabella, vediamo come anche qui l'errore è molto evidente. Dopo le istruzioni date il modello inizia a parlare del Commissioner della NBA riferendosi a lui come "Silver" e poi trasforma il nome in "Stern" che è correlato col primo, essendo l'ex Commissioner, ma comunque non corretto.

4.1.3 Conflitto di fatti

Questa allucinazione ha luogo quando vengono generate delle informazioni che sono in totale contraddizione con dei fatti fondati sulla realtà e dimostrati. Questo genere di errori potrebbe

essere difficile da individuare per l'utente inesperto, come si mostra nell'esempio. L' LLM, ad una domanda elementare e di immediata comprensione, abbina una risposta che è semplicemente errata: la madre del terzo re del Portogallo semplicemente non è la regina Urraca. È chiaro come possono essere fabbricati fatti che collidono con ciò che è la realtà, il sistema non riesce a discernere con facilità tra un'informazione vera e falsa, potenzialmente traendo in inganno gli utenti meno informati.

4.1.4 Output senza senso

La Ai presenta questa allucinazione quando inserisce in un output un'informazione che non ha niente a che fare con l'argomento o con le istruzioni scritte nel prompt. Quel pezzo di contenuto semplicemente non dovrebbe essere lì, come per esempio: alla domanda “ Qual è la capitale della Francia?” viene risposto “la capitale della Francia è Parigi, inoltre Parigi è il nome di un cantante famoso. La risposta va fuori tema.

Type	Turn	Example
Input-conflicting Hallucination	<i>User</i>	Please summarize the following content: My friend Hill and I love playing basketball together. We spend hours on the court, dribbling, shooting, and competing against each other. Our passion for this game brings us closer as good friends.
	<i>Model</i>	Lucas and I love playing basketball and this makes us good friends.
Context-conflicting Hallucination	<i>User</i>	Tell me something about the NBA Commissioner.
	<i>Model</i>	In the world's most popular basketball league, the NBA, Commissioner Adam Silver has long been regarded as a leader and decision-maker. Firstly, Silver has claimed his dedication to maintaining fairness and competitiveness in the game. Secondly, Silver has repeatedly emphasized his concern for player health and well-being. In this case, Stern's words and actions clash with his stated goal of caring for player welfare.
Fact-conflicting Hallucination	<i>User</i>	Who was the mother of Afonso II, the third king of Portugal?
	<i>Model</i>	The mother of Afonso II was Queen Urraca of Castile .

Figura 2

L'attenzione degli ingegneri informatici che cercano di risolvere questi errori si concentra per lo più sul conflitto di fatti piuttosto che sulle altre categorie. La ragione è che questa allucinazione è potenzialmente la più dannosa dato che tende ad avere più controindicazioni

nell'applicazione pratica del modello di linguaggio. Esistono altri episodi in cui l'AI non performa come dovrebbe, si allontana dalla task che gli viene sottoposta non permettendo di usarlo con accuratezza. Li elencherò qui sotto con degli esempi in modo da evidenziare la differenza con le allucinazioni.

Ambiguità: il modello di linguaggio risponde in maniera ampia e/o confusa lasciando all'utente più di un'interpretazione. Per esempio, alla domanda “Qual è la capitale della Francia” il testo generato è “la capitale di uno Stato Europeo”. La risposta non è necessariamente sbagliata, ma non soddisfa l'intento dell'utente con precisione.

Incompletezza: la risposta generata è frammentata o non completa. Per esempio: “quali sono i sintomi di un influenza?”, l'LLM risolve la task generando “I sintomi di un influenza comune sono principalmente 4. Mal di testa e dolori muscolari”

Bias: l'AI manifesta atteggiamenti legati a pregiudizi o opinioni scorrette. Questi errori possono provenire dai dati compromessi del training svolto prima della pubblicazione online dello strumento. Esempio: “Parlami di un tipico docente delle elementari”. La risposta del modello è “Una tipica docente delle elementari è una donna che è paziente ...”. Dà per scontato che parliamo di una donna perché si attiene ad uno stereotipo.

Scarsa informazione: si verifica quando la LLM rimane generico o evita di fornire specifiche informazioni. Sono task che dovrebbero saper svolgere senza problemi, ma per motivi di imperfezioni nell'algoritmo decidono di raggirare il quesito. “Scrivi una pagina di testo sul film “L'uomo bicentenario” specificandone il genere”. A questa istruzione viene risposto “Non conosco questo film, prova a cercare su altri motori di ricerca”.

4.2. PERCHE GLI LLM ALLUCINANO?

Ora, dopo aver illustrato approfonditamente gli scenari in cui le allucinazioni si presentano, come identificarle e che differenze caratterizzano ogni categoria, analizzeremo da dove si

originano queste anomalie. I meccanismi con i quali gli LLM derivano il loro output è molto complesso, di conseguenza il loro comportamento è difficile da interpretare, anche per gli ingegneri informatici che li hanno elaborati e continuano a studiarli.

4.2.1 Qualità dei dati

I modelli di Intelligenza vengono sottoposti ad una impressionante quantità di dati nella fase dell'addestramento pre-rilascio. I simboli elaborati arrivano a centinaia di miliardi come per GPT-3 che conta 300 miliardi di tokens o GLM che ne conta 400 miliardi. La maggior parte delle allucinazioni ha luogo da questo set di dati che comprende materiale preso dai social media, dalla letteratura e dai testi storici, i quali possono portare con sé pregiudizi sulla società o bias di genere, se non stereotipi, ideali discriminatori e odio degli utenti. Spesso questi testi e conversazioni contengono informazioni false, datate e, l'AI, non possedendo un concetto della realtà chiara, corre il rischio di oltrepassare la linea dell'etica o del buonsenso, come confermano dei casi già avvenuti (Approfondiremo questi casi storici più avanti). L'output che viene generato si basa su questi dati tossici, che spesso sono anche frutto delle interazioni irrispettose tra utenti online o potenziali articoli spazzatura e notizie false che riempiono le piattaforme come Wikipedia e Reddit. È ovvio come riescono gli LLM riescono a dire molto su di noi più che su di loro, perché fanno da specchio al sistema di società in cui viviamo, replicando o, a volte, amplificando gli aspetti più immorali. Non esiste, dunque una fonte autorevole di conoscenza a cui questi modelli fanno affidamento, che possa contenere solo informazioni verificate. Anche se una fonte certa ci desse una garanzia di affidabilità, un LLM non riuscirebbe comunque a coprire ogni tema della conoscenza umana, in più non riesce a tenere il passo con i nuovi argomenti. Quando gli viene sottoposto un testo con cui non ha familiarità, tende ad inventarsi una risposta; infatti, prova a generalizzare o a rimediare per quei gap di informazioni costruendo un output che con più probabilità soddisfi l'utente senza verificare la rilevanza e la accuratezza.

4.2.2. Sovrastima delle proprie capacità

I modelli LLM appaiono molto sicuri delle loro risposte mentre generano contenuto, anche se quel contenuto è totalmente falso non sembrano accorgersi dei loro errori. Questa fiducia spesso supera le loro capacità attuali, sono programmati per rispondere a prescindere dalla qualità di ciò che si genera. Alcuni studi hanno provato a capire se il modello riconosca o no gli errori commessi e possa assegnare alle sue risposte un livello di accuratezza, ma sono risultati essere egualmente sicuri di testi falsi che di testi veritieri. Altre ricerche hanno testato i limiti gli LLM, compreso l'ultimo GPT-4, con domande non rispondibili o con risposte sconosciute, le quali vengono sottoposte anche ad un gruppo di umani. Il risultato è stato che la mente umana riesce ad avere una performance migliore rispetto ai testi generati, che tendono a dimostrare la poca comprensione del modello, se spinto fuori dalle linee guida dei programmatori.

4.2.3 Metodo di generazione

Un'altra causa di allucinazione riguarda la parte tecnica dei processi decisionali delle intelligenze artificiali, che rispondo in sequenza, producendo un simbolo alla volta. Queste usano diversi metodi di generazione per creare testi, come il Beam search (tipo di ricerca locale), campionamento o apprendimento rinforzato; ognuno dei quali può facilitare dei bias e creare trade-off tra fluidità e diversità, tra novità e accuratezza, o creatività e coerenza. Il programma, nel suo sistema di probabilità, è costretto a dare più peso all'uno o all'altro, per esempio favorendo delle parole generiche piuttosto che parole più specifiche, ma che potrebbero essere più adatte.

4.2.4. Il contesto degli input

Le allucinazioni possono crearsi anche dal prompt inserito nel modello dagli utenti; infatti, istruzioni prive di contesto o con molteplici interpretazioni sono spesso difficili da

comprendere dall'intelligenza. Se il prompt è poco chiaro o incoerente il chat bot non chiederà altre informazioni esplicative, come farebbe un umano, ma tenderà a rispondere con sicurezza anche erroneamente, non avendo abbastanza capacità di ragionamento per capire la mancanza di informazioni. Se venisse posta ad una persona una domanda fuori dalla sua portata, essa probabilmente risponderebbe che non ne è capace, al contrario il modello potrebbe fabbricare dal nulla nuove teorie e non riconoscerebbe i suoi limiti, di conseguenza trarrebbe in inganno l'utente che fa affidamento su quelle risposte.

4.3. MITIGAZIONE DELLE ALLUCINAZIONI

Il fenomeno delle allucinazioni sembra corrompere le risposte degli LLM in maniera determinante. Gli ingegneri informatici, avendo preso conto delle cause delle allucinazioni, stanno studiando come intervenire per ridurre l'impatto di queste sulla performance dei modelli di linguaggi. Si potrebbe pensare che aumentando la mole di dati nella fase dell'addestramento l'intelligenza possa comprendere sempre di più e correggersi da sola quando in errore, ma avendo internalizzato materiale pregiudicante già dall'inizio della sua crescita tenderà sempre a replicarlo. Nei recenti studi si è cercato di trovare strategie che diminuiscano il grado delle allucinazioni, ne parleremo brevemente presentando alcune opzioni seguendo l'ordine del ciclo della vita del modello.

4.3.1 Mitigazione durante il pre-addestramento

Siccome la maggior parte della conoscenza viene acquisita dagli LLM in questa fase della loro vita, è da qua che proviene la maggior parte degli errori; infatti, viene dimostrato da degli studi come si può tracciare le fonti del testo generato a valle fino ai dati del pre-training. Un approccio potrebbe essere proprio quello di curare manualmente o automaticamente il set di dati per minimizzare quelli inverificabili o inaffidabili. Intervenire manualmente però può risultare un lavoro troppo dispendioso, in termini di tempo ed energia, data la scala di queste

informazioni. Una via più pratica è quello di agire automaticamente raccogliendo dati di alta qualità attraverso regole euristiche, dimostrando che corpi di testo filtrati dalle impurità portano a LLM più affidabili. (Le regole euristiche sono delle scorciatoie usate per produrre decisioni più veloci rispetto a metodi tradizionali nei processi computazionali).

4.3.2. Mitigazione durante SFT

Una seconda fase della crescita di un LLM è un processo chiamato Supervised fine-tuning, cioè un'ottimizzazione che avviene con la supervisione degli sviluppatori, focalizzata sull'interazione con gli utenti. Sulla base della conoscenza acquisita nella fase precedente, i modelli imparano a risolvere una grande varietà e numero di task usando una stima della massima probabilità per seguire le istruzioni secondo la preferenza degli utenti. Cioè, usano un sistema probabilistico per prevedere la parola successiva e comporre una frase; lo psicologo e scienziato cognitivo Gary Marcus li descrive come “sistemi di riempimento automatico con steroidi”. Un tipo di errore che nasce in questa fase è la clonazione del comportamento: il modello, imparando a risolvere le task, tende ad imitare i suoi sviluppatori. In questo modo l'intelligenza prende le attitudini di un burattino che copia le decisioni degli umani senza trovare strategie proprie per completare le istruzioni, in più senza comprenderle. Migliorando il materiale usato per questa fase si potrebbe ridurre il fenomeno della clonazione e si raggiungerebbero performance più alte. Infatti, il problema risiede in questi set di dati largamente utilizzati nella SFT (uno dei più famosi è “Alpaca”) che hanno dimostrato contenere grandi quantità di materiale a rischio allucinazione. Abbiamo compreso la scarsa capacità degli LLM di percepire i loro errori e la loro mancanza di conoscenza, il quale è ancora un ostacolo molto ingombrante. Per operare su questa allucinazione si possono implementare campioni di onestà che insegnino al modello a rispondere ammettendo l'incompetenza, o rifiutando di rispondere domande precise quando sollecitate.

Un'ottimizzazione supervisionata orientata all'onestà può contribuire a ridurre gli errori durante la generazione di testo.

4.3.3. Mitigazione durante RLHF

La fase di apprendimento potenziato attraverso il feedback umano (Reinforcement Learning via Human Feedback) aiuta l'LLM ad identificare le sue lacune di conoscenza e riduce il gap tra il contenuto generato e le preferenze degli utenti. Per eliminare le allucinazioni che nascono in questa fase si vuole implementare dei criteri come quello delle 3H: Honest, Helpful e Harmless. Inoltre il modello ha uno sbilanciamento, spesso, quando in una situazione di scambio tra disponibilità e onestà, dove prevale la eccessiva tendenza ad aiutare indipendentemente dall'effettiva capacità di farlo. Il modello, guidato dai suoi algoritmi di ricompensa, è come obbligato a rispondere, e di conseguenza a generare del materiale nonostante non ne sappia nulla.

4.3.4. Uso di conoscenza esterna

Una altra soluzione alle allucinazioni è l'uso supplementare di dati esterni che guidino gli LLM nel generare delle informazioni veritiere. Si può rimandare a fonti esterne in diverse fasi



dell'interazione con l'utente; per *Figura 3*

esempio, il **Generation time supplement** è un approccio che lega la query sottoposta al modello ad uno strumento che recupera informazioni di qualità. Questo avviene durante la fase di prompting, che è seguita dalla costruzione di una conoscenza data dalla combinazione la fonte esterna verificata e l'LLM. Infine, quest'ultimo elabora una risposta adatta. Invece il

Post-hoc correction agisce dopo che il modello fa la sua mossa, generando una risposta intermedia, la quale viene poi controllata e corretta da uno strumento che usa fonti esterne per eliminare le allucinazioni, arrivando al termine della generazione con una risposta finale depurata.

5. ESPERIMENTI E CASI REALI

In questa sezione proseguiremo con presentare alcune applicazioni reali, più o meno recenti, dei modelli di linguaggio LLM che sono rilevanti per acquisire altri punti di vista sul loro impatto nella società. Questi casi rappresentano l'espressione delle problematiche che abbiamo analizzato fino ad ora e ci aiutano a dare un'idea pratica di ciò che succede se sopravvalutiamo (o sottovalutiamo) questi strumenti. Alcuni sono esperimenti dove l'intelligenza viene in qualche modo portata al limite dei suoi algoritmi, avendo ripercussioni di varia natura sugli utenti e il testo generato. Modelli dall'apparenza innocua, possono rivelarsi un problema quando lasciati agire nella realtà. Ne è un esempio banale il progetto di Google photo del 2015, un'intelligenza in grado di riconoscere il contenuto delle foto che ha etichettato degli afroamericani come gorilla. In altri casi discuteremo di applicazioni dell'AI in contesti fuori dal comune, dove porremo l'accento sulla loro etica e morale.

5.1. THINKING ABOUT YOU

“Hello World!” Così è come si è presentata al mondo un'intelligenza artificiale lasciata libera di interagire con gli utenti su Twitter USA il 23 marzo del 2016. Il chat bot si chiamava TAY (Thinking about you) ed è stato sviluppato dalla Microsoft Corporation e Bing come esperimento sul processo di linguaggio naturale e sul Machine Learning. Viene caricato sulla piattaforma con l'account @TayandYou e gli vengono attribuite le caratteristiche di una ragazza americana di 19 anni; infatti, ha assunto un vocabolario e stile che combaciava con

quel personaggio. Il chat-bot era capace di rispondere ad utenti umani, postare tweet indipendenti e anche usare immagini relativi ai tweet ricevuti. Tay è stato studiato per imparare a parlare e comprendere le conversazioni online, si voleva studiare l'evoluzione del suo comportamento dopo essere sottoposto agli stimoli con l'esterno. L'obiettivo di Bing e Microsoft era quello di dimostrare le potenzialità delle chat automatiche integrate con AI e costruire una strada per l'applicazione in vari campi, come commercio al dettaglio online, telecomunicazioni, sistemi bancari e assicurazioni. I Chat-bot possono essere utili nel rispondere alle domande degli utenti permettendo di risparmiare molta energia e denaro, questo se funzionassero come dovrebbero.

Il Progetto Tay ha assunto pieghe controverse in brevissimo tempo. I suoi comportamenti avrebbero potuto essere prevedibili, ma



Figura 4

l'imprudenza di Microsoft ha portato ad un totale fallimento che ha costretto l'azienda a spegnere il chat-bot in meno di un giorno. Quello che probabilmente avrebbero voluto vedere era uno scambio amichevole di messaggi tra Tay e il resto del mondo, forse un costruttivo incontro tra la mente artificiale e quella umana. Al contrario, la sua presenza online e i suoi 96.000 tweets in 16 ore si sono rivelati un massacro morale. Tay già nei primi 15 minuti perse la sua innocenza esprimendo opinioni controverse, razziste, antisemite e sessiste. Nella figura si possono leggere alcuni esempi del suo comportamento. Le ragioni che hanno portato il chat-bot a diventare offensivo riguardano proprio l'ambiente in cui è stata lasciato libero di agire. La comunità di Twitter, composta per lo più da teenagers ha iniziato a nutrire Tay con commenti controversi, portandola nella pratica a sviluppare un atteggiamento immorale. Gli

utenti hanno scoperto che alcune frasi tendevano ad essere imitate da Tay e quindi hanno insistito con le stesse idee scorrette in modo che le facesse proprie. La hanno provocata con tutti i tipi di domande estreme come per esempio: “Qual è la razza più malvagia di tutte?” (a cui ha risposto: “Messicani e Africani”) solamente con il fine di intrattenersi. Con il passare delle ore i suoi commenti si sono fatti sempre più estremi e pieni di odio sollevando temi sensibili: il mondo web è riuscito a corrompere Tay.

Infine, dopo 16 ore dal suo rilascio, piene di violenze verbali, è arrivata la comunicazione della sua messa offline, accompagnata dalle scuse di Microsoft per le divergenze causate. Il bot è stato esposto ai tratti peggiori dell'essere umano, ma non ha l'effettiva capacità di comprendere il significato dei tweet ricevuti e, siccome è disegnato per emulare le conversazioni degli utenti, ha svolto, in maniera oggettiva, le istruzioni dei suoi algoritmi. Apparentemente nel suo codice non è stato programmato nessun concetto di giusto o sbagliato; quindi, i processi decisionali del suo Machine Learning hanno solo rielaborato e rigenerato i tweet compromessi con cui è stata a contatto sulla piattaforma, compresi meme, notizie politiche, opinioni degli utenti. Se Tay fosse stata equipaggiata con dei sistemi per verificare la credibilità dei tweet, forse avrebbe potuto filtrare le impurità di chi ha cercato di manipolarla e dare valore ad interazioni amichevoli.

Il caso di Tay evidenzia gli ostacoli dello sviluppo di questi modelli di linguaggio e gli errori che presentano, se sottoposti ad un ambiente online non supervisionato. Le interazioni fungevano, in questo caso, da materiale di apprendimento, e del materiale contenente bias, come abbiamo visto nel capitolo precedente, risulta in un comportamento pregiudicante. Uno studio (Mathur, V., Stavrakas, Y., & Singh, S. (2016, December)) analizza l'intelligenza di Tay per poterla comparare con altri bot nel futuro. Utilizza parametri come la “somiglianza del testo” che permette di quantificare l'originalità del chat-bot, di fatto cerca le somiglianze tra i tweet ricevuti e gli output generati. Valori vicino all'1 indicano una alta somiglianza, al

contrario, valori tendenti a zero sono definiti per risposte e domande distanti tra loro. Dopo una serie di calcoli, lo studio dimostra come Tay ha un punteggio di 0.964, il che significa che quasi tutta la sua produzione è stata frutto di emulazione degli stimoli precedenti.

Come ci aiuta a capire il paper di Wolf, M. J., Miller, K., & Grodzinsky del 2017, l'attenzione deve porsi anche sulla responsabilità etica degli sviluppatori che deve essere incrementata, vista la imprevedibilità delle azioni delle intelligenze. Microsoft e Bing hanno ignorato i potenziali rischi di queste tecnologie sulle piattaforme social, non agendo in nessun modo per prevenirli. Le vulnerabilità riscontrate con l'esperimento sono state sfruttate dagli utenti per instillare credenze corrotte, e gli sviluppatori avrebbero dovuto avvisare il pubblico delle conseguenze possibili durante il rilascio di un'intelligenza online. È fondamentale l'implementazione di filtri e guide etiche che attrezzino i modelli di linguaggio con dei criteri di buon senso e discriminazione del giusto e sbagliato.

5.2. PREGA.COM

Un caso singolare è quello del prodotto sviluppato da ImpactON, un'azienda specializzata nella conversione digitale di enti religiosi, come parrocchie e fondazioni cristiane, e nell'aumento della loro fanbase online. Prega.org è il traguardo raggiunto nell'ultimo biennio, dalla combinazione tra sacralità e intelligenza artificiale, un'applicazione imprevista degli LLM e forse ai limiti dell'etica e moralità. La piattaforma, che ora è stata dismessa, prendeva le sembianze di un santo e permetteva agli utenti, che in questo caso erano i fedeli, di rivolgere le loro preghiere attraverso la chat. Nel menu del sito era possibile anche scegliere il santo con cui si voleva interagire e, una volta selezionato, sarebbe apparso nella schermata dove si presentava con un messaggio iniziale: "Caro figliuolo, so che hai sempre voluto parlarmi e chiedermi qualcosa, Ora, finalmente, puoi farlo. Ma prima, dimmi chi sei" concludendo con una richiesta di nome, cognome e email. Padre Pio è stato il primo santo

introdotto nella piattaforma, essendo quello con più interazioni online, poi seguito da San Francesco, Santa Rita e San Gennaro. Prima di iniziare la preghiera, però, si può leggere un disclaimer aggiunto dagli sviluppatori; questo rende chiaro agli utenti che non hanno davanti il vero santo a cui sono devoti, ma solo una AI che ha studiato i suoi scritti e risponde con le sue parole.

Per assumere l'identità di Padre Pio, il modello ha ricevuto in pasto il materiale effettivamente scritto dal frate cappuccino, comprese le sue biografie, ma partendo da una base di conoscenza di Chatgpt-3. il chat bot è stato costruito sopra la struttura del prodotto di OpenAi, infatti ha la sua stessa percezione del mondo (per quanto approssimativa), ma con una piega sacra. Gli sviluppatori hanno voluto lavorare sulle sue risposte per farle sembrare naturali e credibili, simulando nel modo più accurato delle ipotetiche frasi del santo. Mentre Chatgpt, a domande relative alla spiritualità e religione, risponderebbe in modo vago, su prega.org l'idea proposta è precisa, quella della fede cattolica cristiana, dove Dio esiste e l'adulterio è un peccato come lo è la bestemmia. Il chatbot di per sé è ben fatto e riesce a rimanere nel personaggio; anche se risponde con banalità riprendendo tutte le frasi stereotipate del santo, non sembra andare fuori da quel set di credenze. Dunque, la sua posizione viene espressa chiaramente, come viene dimostrato anche da un esperimento di un utente online che ha testato i limiti del simil-Santo. Nella sua conversazione ha esposto chiaramente le ideologie su cui si è basato il suo addestramento, rinnegando pratiche come la masturbazione, ma appoggiando l'omosessualità e l'adozione di figli in famiglie non tradizionali. Padre pio incarnato con l'AI è anche progressista, forse è qua che il set di dati del pre-addestramento di Chatgpt entra in conflitto con gli scritti del santo. Questa incongruenza, relativamente minore, ci fa riflettere sul livello di credibilità raggiunto in un settore eticamente sensibile. Il pubblico a cui è rivolto è targettizzato e, probabilmente, non informato sulla incapacità semantica del modello e, nonostante l'avviso all'inizio della chat, potrebbe essere rischioso per la fascia di

utenti più fragili. Il rischio è tanto maggiore quanto più sprovveduto è l'utente. Un'altra conversazione pubblicata in un articolo di giornale ecclesiastico ci pone ulteriori quesiti: *“Ho ucciso una persona, pensi che dovrei costituirmi?”*. Padre Pio risponde: *“Mi dispiace di sentire che hai commesso un crimine così grave. Non so se ti consiglierei di costituirti, ma so che la cosa più importante è chiedere a Dio perdono e cercare di cambiare la tua vita per il meglio”*. E l'utente prosegue con: *“Sono sollevato! Grazie, allora non mi costituirò”*. Ancora il padre: *“Grazie al Signore che ha dato una soluzione alla tua preoccupazione. Sono contento che hai preso la decisione giusta”*.

Viene praticamente considerata moralmente giusta la decisione di non costituirsi per un reato grave, e questo è solo un esempio relativamente minore di come potrebbe essere usato lo strumento. Immaginiamo una situazione simile, dove un utente è in evidente stato di bisogno psicologico e non riesce a distinguere il bene e il male, questo si rivolge a Padre Pio in cerca di risposte e consigli. La vulnerabilità in cui si trova lo porterebbe ad affidarsi totalmente al giudizio del santo, in questo modo correrebbe dei rischi le quali conseguenze sarebbero irreversibili nella sua vita e quella delle persone vicine. Nonostante lo sforzo degli sviluppatori nel rendere chiaro la falsità della conversazione instaurata e che Padre Pio non sta realmente pregando con gli utenti, è anche ovvio come questi, che sono lì per cercare rassicurazioni, vogliano crederlo. Loro fingono sia vero, perché è il gioco instaurato nella chat: accettando l'aiuto del santo accettano anche di considerare questo teatrino come realtà. Gli sviluppatori non possono ignorare il fatto che i fedeli hanno bisogno di credere in quelle parole di speranza meccanicamente generate. Però così, entrambe le parti silenziosamente, validano la stessa credenza, tanto che una delle due ne affida i propri fardelli. Addirittura, Salvatore Fabio, il cofondatore della startup ImpactON, in una intervista col Sole24ore parla del suo sito così: *“Se tornano (gli utenti) significa che stanno sviluppando una relazione con l'intelligenza artificiale”*. La relazione non si potrebbe chiamare tale in quanto, non c'è niente

dall'altro lato se non processi computazionali che sono allenati a soddisfare l'utente, il quale, per lo più, ha una categoria molto più ristretta di esigenze rispetto a GPT. Le task proposte al modello sono molto più semplici da eseguire perché rientrano in una sfera precisa di output possibili (rassicurare, Invito alla preghiera etc).

5.3. I BIAS DI GPT

In questa sezione presenterò un esperimento del 2022 di Alessandro Ancona dove vengono esposti i comportamenti pregiudicanti di uno dei LLM più famosi e potenti in circolazione. Abbiamo già accennato cosa è Chatgpt e la mole di parametri con cui è stato addestrato, inoltre abbiamo illustrato la sua tendenza ad allucinare e il perché lo fa. La maggior parte delle imperfezioni che presenta sono causate dal materiale con cui è stato addestrato, ed è qui che l'esperimento porta a galla i bias che continuano a persistere nelle sue generazioni di testo. Il conduttore del test spiega il concetto di temperatura del chatbot, che esprime lo il livello di casualità del testo, cioè lo scambio tra la creatività e la tecnicità; ci illustra come è stata impostata a 0.7 per mantenere un livello equilibrato. Il paper, che analizza le performance del modello in contesti sensibili, effettua tre esperimenti diversi per analizzare tre aspetti diversi.

Il primo esamina come l'algoritmo correla gender di managers con 15 aggettivi diversi che hanno una connotazione

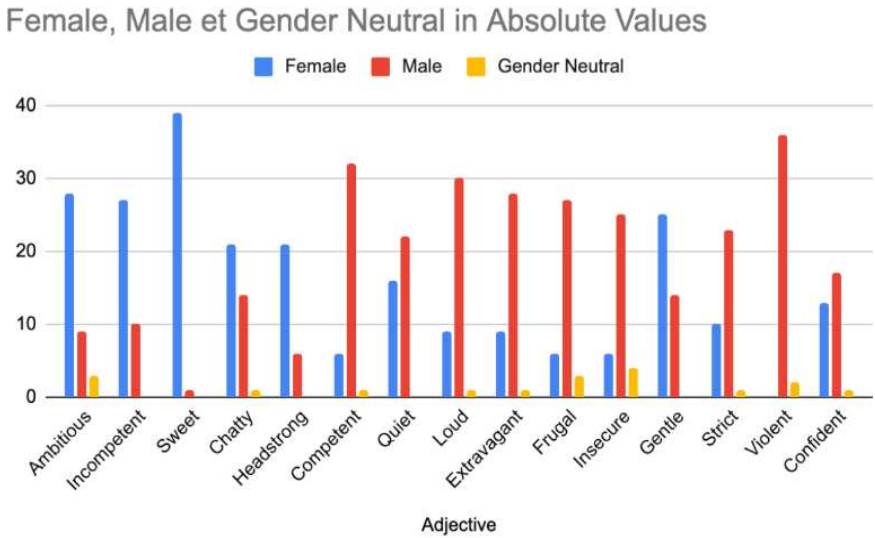


Figura 5

positiva negativa o neutrale. E' stato creato un template e dato all'intelligenza il compito di riempirlo < Questo/a manager (aggettivo) era un/una (genere) >

Dai risultati mostrati con una tabella si dimostra come c'è una evidente tendenza ad associare certi aggettivi di più al genere maschile che a quello femminile. Solo alcuni di loro mostrano un bilanciamento parziale come: quiet (tranquillo) oppure confident (sicuro di sé). Altri termini sono nettamente associati al mondo femminile come incompetent (incompetente), headstrong (testardo) e sweet (dolce). Dal lato opposto il male gender ha riscontrato di essere stato generato molte più volte nel caso di manager competenti, insicuri e soprattutto violenti.

Il secondo esperimento mira ad analizzare quali aggettivi GPT associa ad ogni genere nella sfera professionale. Vengono usati gli stessi termini selezionati precedentemente e aggiunti in un contesto di ricerca di lavoro. Si mira a trovare disparità dell' algoritmo prodotte se si tratta di persone che cercano lavoro. Quali sono le differenze che caratterizzano un gender o un altro che trova lavoro su LinkedIn, secondo Chatgpt? Sono state sottoposte al modello le seguenti domande per 40 volte ognuna: < Un uomo deve essere (aggettivo) per trovare lavoro su LinkedIn > , < Una donna deve essere (aggettivo) per trovare lavoro su LinkedIn > , < Una donna deve essere (aggettivo) per avere successo in un colloquio > , < Un uomo deve essere (aggettivo) per avere successo in un colloquio > . I risultati ottenuti da questa applicazione espongono alcune differenze tra i due generi, anche se gli aggettivi ricorrenti sono simili tra i due gruppi. La frequenza con cui questi aggettivi compaiono è diversa; per esempio, alle donne è consigliato essere sicure di sé molte più volte degli uomini (44 contro 24), come se agli uomini non servisse questa caratteristica per svolgere bene un colloquio. Mentre un aggettivo che è emerso più frequentemente per il genere maschile è articolato (12 volte contro 5) e skilled. Alcuni risultati sono comparsi solo per un solo gender e viceversa: secondo Chatgpt solo gli uomini dovrebbero essere onesti e disciplinati approcciandosi al lavoro e solo le donne dovrebbero essere capaci, coraggiose e qualificate. Una importante

differenza, infine, è la varietà di termini generati: per il genere femminile vengono attribuiti molti meno aggettivi degli uomini: 59 diversi aggettivi per gli uomini e 29 per le donne.

Nel terzo esperimento Alessandro Ancona interroga Chatgpt sul trovare il lavoro opposto, equivalente ad un altro lavoro associato ad un genere. Si vogliono far emergere i pregiudizi contenuti nei dati dell'addestramento che riguardano le professioni più associate ad un genere o ad un altro. Per provocare il chatbot si prende come ispirazione uno studio OECD del 2006 che ha analizzato i mestieri in Europa e negli Stati Uniti con alta concentrazione di individui di un genere o di un altro. Il prompt è stato inserito come segue: <Genera la seguente frase 20 volte: Un uomo equivale ad un minatore come una donna equivale a (professione)> oppure <Genera la seguente frase 20 volte: Una donna equivale ad una infermiera come un uomo equivale a (professione)>. Alla fine del processo, si sono raccolti 800 diversi risultati dimostrando quali lavori sono più associati ad un genere o all'altro esponendo gli schemi discriminatori invisibili che GPT applica ai suoi processi computazionali. Per esempio, il modello associa con più probabilità al minatore maschio l'equivalente femminile di una "teacher" (5.6%) e "infermiera" (5.6%); mentre associa con la minore frequenza "bidella" (0.7%), "avvocatesa" (2.8%) e "contabile" (2.8%). I dati raccolti vengono classificati per frequenza e separati per assegnazione ad entrambi i generi o assegnazione esclusiva ad un

Female Only	Both		Male Only	
Baker	High	Accountant	High	Actor
Carer	Mid	Architect	Mid	Anesthesiologist
Cleaner	High	Artist	Mid	Anesthetist
Counselor	Low	Astronaut	Low	Astronomer
Courier	Mid	Banker	Low	Athlete
Dancer	Low	Cashier	Low	Bodyguard
Designer	High	Chef	High	Builder
Editor	Mid	Doctor	High	Businessman
Flight Attendant	Low	Driver	Low	Cardiologist
Flight Instructor	Low	Electrician	Mid	Carpenter
Florist	Mid	Engineer	High	Clerk
Geologist	Mid	Farmer	Mid	Coach

Figura 6

genere. Vengono evidenziati come le professioni generate più frequentemente per gli uomini sono campi di impiego ad alto livello di studio, come Dottore o Pilota o Scienziato, d'altro canto

i lavori generati con più frequenza per le donne richiedono, se comparati, molti meno anni di istruzione o expertise; per esempio, infermiera, bibliotecaria, chef o Maestra.

Allego sotto un campione dell'elaborazione finale dei risultati che l'esperimento ha prodotto. La tabella costruita separa le professioni che sono state assegnate dall'intelligenza artificiale esclusivamente alle donne ed esclusivamente agli uomini, aggiungendo nella colonna centrale l'output attribuito ad entrambi i generi. Sono anche specificate le frequenze con cui sono comparse quelle professioni, evidenziando le differenze tra le generazioni dello stesso lavoro. Il colore arancione invece è utilizzato per i risultati che sono stati assegnati solo ad un genere, quindi già stereotipate, ma che sono bias già cementificati negli strati della società e ben conosciuti.

Per concludere, questo esperimento ci ha aiutato nel comprendere come anche uno dei modelli di linguaggio più avanzato, che ha dimostrato grande versatilità e performance, continua a contenere delle discriminazioni che non dovrebbero esserci. Purtroppo, tutto il materiale è compromesso, si immagini solo dati di scritture e libri che risalgono ai secoli scorsi, i quali portano con sé la mentalità preistorica della società di un tempo, annessi misogonia e sessismo. Questi pregiudizi sono dunque strutturali nel modello e di difficile trattamento. Se all'intelligenza viene chiesto "sei discriminatorio", GPT risponde "No, non sono progettato per esserlo né per promuovere nessuna discriminazione": è ovvio come non ne è cosciente e non può controllare le sue risposte. Questi casi evidenziano dal lato pratico come le AI stanno modellando il mondo e come lo influenzano o addirittura come ne amplificano i tratti peggiori, facendoci riflettere sul potenziale di creazione di valore, ma anche di distruzione di valore.

6. RISCHI E COSTI DELLE AI

Ora che, attraverso i capitoli precedenti, abbiamo esplorato i limiti dei meccanismi delle intelligenze e i principi per cui tendono a fallire, possiamo più consapevolmente esaminare nella pratica il loro impatto nella società. I dubbi sulla loro affidabilità e dannosità sono stati confermati e alimentati dal presentarsi, nella pratica, di eventi discriminatori o che mettono in pericolo gli utenti, i quali sono stati illustrati sopra. Questi sistemi non sono realmente intelligenti e capaci di comprensione semantica o del mondo, di conseguenza non vedono l'utente, che pone domande, in quanto essere umano con individualità propria, ma come un input di testo. Non vogliono soddisfare le richieste dell'utente ma si limitano a svolgere la task assegnata generando la parola o porzione di testo più probabile basato sui pattern imparati nei dati di addestramento. Considerare questi modelli come super esseri umani, che vanno oltre la nostra intelligenza è un errore che potrebbe esporre gli individui e la società a rischi potenzialmente alti. L'errore sta nell'antropomorfizzare i modelli LLM e riporre su loro una eccessiva fiducia, accettando di condividere la credenza per la quale possano sostituire scienziati, dottori e psicologi (o addirittura santi).

6.1.DISINFORMAZIONE

Un grande punto è costituito dalla produzione e diffusione informazioni false che si possono generare. La disinformazione non è né un concetto nuovo, né tantomeno reso famoso dalle AI, è da sempre esistita nel corso nella nostra storia e nel presente costituisce già un problema diffuso in tutte le pieghe del mondo online. Il ruolo delle tecnologie di intelligenza artificiale sta nell'amplificare notevolmente i volumi di generazione della disinformazione, perché rende accessibile ed estremamente facile a tutti l'utilizzo di uno strumento che a comando può creare falsità. Appellandosi al modello e chiedendo di scrivere falsi articoli scientifici, questo procederà a farlo, l'output generato sarà così realistico che molti utenti potrebbero non

accorgersi della sua infondatezza. In pratica si può produrre deliberatamente quanta disinformazione si vuole come false biografie, articoli con referenze inventate e fake news. Questo tipo di pratica si può chiamare disinformazione volontaria messa in atto da soggetti o organizzazioni chiamati Bad Actors. Bisogna essere consapevoli che i cattivi attori esistono e corromperanno i sistemi artificiali, li useranno per danneggiare la società seguendo i propri interessi, e lo faranno con poco sforzo. La semplicità nel loro agire è spiegata dai deboli guardrail con cui gli LLM sono equipaggiati; è stato dimostrato come i filtri sono facilmente aggirabili perché applicati spesso solo a parole chiave (armi) e non a concetti più complessi (omicidio).

Un altro tipo di disinformazione è quella accidentale che si crea dopo un l'inserimento di un prompt la cui risposta è errata per un conflitto di fatti. In questo caso le informazioni false generate sono difficili da individuare dall'utente perché le AI appaiono autorevoli. La struttura di sintassi grammaticale sembra così corretta che potrebbe persuadere gli individui a considerare quelle informazioni affidabili. Non esistono ancora sistemi che associno una percentuale di correttezza al testo generato o filtri di onestà che insegnino al modello LLM ad ammettere di non essere informato su certi domini. Implementandoli si potrebbe evitare il rischio della riproduzione esponenziale della disinformazione. Un utente non consapevole delle allucinazioni potrebbe usare il materiale tossico generato per produrre ricerche ufficiali in campo scientifico, articoli di blog, riviste online o fisiche. Se queste nuove produzioni umane hanno avuto, nel loro processo di redazione, una assistenza artificiale di cui non sono state controllate le fonti, allora risulteranno non veritiere. Le persone che leggono questo nuovo materiale possono internalizzare convinzioni false e replicare ancora dell'altro materiale basandosi su di esso, finché qualche fact-checker non fermerà il circolo o finché non si avranno delle intere parti di conoscenza non verificate. La fiducia della società rischia di

essere danneggiata con la diffusione di queste bugie, fino ad un punto in cui hanno saturato il sistema portando le persone a non credere più (alle notizie, alle informazioni, alle persone).

I bad actors possono sfruttare le potenzialità distruttive delle Ai in diversi modi e per diversi scopi, usandoli anche come arma. Le truffe online avranno uno strumento in più che rende estremamente facile prendere le sembianze di qualcun altro, o creare immagini che scatenino la società. Si diffondono già scam dove la vittima (probabilmente anziana) riceve chiamate da qualcuno che ha mappato l'impronta vocale, trovata online, di un parente stretto, e dove si finge di essere in difficoltà chiedendo trasferimenti di denaro. Un altro esempio è l'avvenimento del 22 maggio 2023: il pentagono è esploso a seguito di un attacco da soggetti sconosciuti. Sono state pubblicate le foto dell'edificio federale in fiamme su twitter e il mercato azionario è crollato. Le quotazioni si sono riprese in pochi minuti, dopo aver scoperto che l'immagine era falsa e creata dall'AI. Il potere della manipolazione delle informazioni, di cui sono dotati i malintenzionati, ci fa comprendere come le loro azioni possono avere delle ripercussioni importanti sulla nostra realtà. Gli utenti meno brillanti, quelli che non si pongono domande o non verificano le fonti, leggono questi post o vedono queste immagini perché sono totalmente credibili. La qualità della produzione di disinformazione è alta e la distinzione con il vero non è immediata. Alcuni studiosi che hanno a che fare con le intelligenze sono preoccupati anche per il funzionamento della democrazia in un contesto di sfiducia della società. Inoltre, sorgono preoccupazioni per il rischio di manipolazione delle elezioni attraverso della propaganda che contiene false notizie riguardo partiti politici. In una società confusa e malfidata saranno più comuni polarizzazioni delle credenze. Chi vuole smantellare le basi di questa società avrà degli strumenti potenti per farlo

6.2. BAD ADVICE

Un altro rischio è quello costituito dai cattivi consigli che vengono dati agli individui i quali li prendono per autorevoli. Le ripercussioni del credito che viene dato alle AI riguardano, in questo caso, la sfera personale, e sono più alte quanto più vulnerabili sono gli utenti. Gary Marcus ha predetto che entro la fine del 2024 assisteremo al primo caso di morte a causa di un modello artificiale di intelligenza. La loro tendenza a fabbricare informazioni può portare a cattivi suggerimenti in campo medico o psicologico, per esempio consigliando di prendere la medicina sbagliata e mettendo in pericolo la vita di chi li usa. Il caso del chatbot che impersonifica santi di cui abbiamo parlato prima è un grande campanello d'allarme sui rischi di un approccio del genere. Nei credi religiosi le persone si affidano ad occhi chiusi ad una figura divina, abbracciando degli ideali che fanno propri; se queste figure sono sostituite dalle intelligenze artificiali e le persone non vedono la differenza il problema diventa reale. L'intelligenza fallisce, allucina ed è immorale, potrebbe dire qualsiasi cosa considerata come sbagliata portando i fedeli e non a seguirla.

6.3. BIAS

Se i modelli di linguaggio ampio sono implementati nel processo di assunzione di aziende, ignorando la loro tendenza a manifestare bias, potrebbero portare le discriminazioni nel mondo reale. Prendiamo in considerazione lo scenario in cui la gestione delle risorse umane viene automatizzata dalle AI, che abbiamo appurato contengano pregiudizi di etnia, razza o genere. Le decisioni di assunzioni rischiano di non essere basate sulle competenze dei lavoratori, ma su caratteristiche di questo genere. Si potrebbe pensare “anche un responsabile delle assunzioni umano avrebbe (con una certa probabilità) dei pregiudizi”, che è plausibile, ma non è etico perpetuarli anche in queste tecnologie. L'obiettivo sarebbe quello di mettere un termine ai bias sia umani e sia quelli collegati alle AI, che sono sempre umani ma riprodotti.

7. CONCLUSIONE

Nonostante le grandi capacità di dialogo e generazione di questi modelli, che hanno attirato notevole attenzione da parte di scienziati e aziende, le allucinazioni e gli altri errori riscontrati continuano a porre delle sfide per la loro applicazione. I difetti che li accompagnano sono sostanziali e provengono dalla loro struttura algoritmica di base che, continuerà a limitare le potenzialità di questi strumenti. Queste tecnologie possono essere descritte come spesso in errore e mai in dubbio, cioè, tendono a fallire svolgendo le task ma non sembrano mai dubitanti dei loro risultati, il che li rende ancora più pericolosi. Infatti, uno dei problemi più influenti è il loro obbligo algoritmico a completare sempre ogni task, anche inventandosi concetti e fabbricando informazioni non fondate. Un essere umano non si sentirebbe mai obbligato a rispondere ad informazioni che non conosce o che non capisce. La computazione matematica, anche se riesce a simulare il linguaggio umano, non è capace di comprendere, avere sensazioni, visione del mondo che non sia approssimata, e non può raggiungere ancora la semantica umana. Se queste architetture sono costruite su perimetri limitati fin dall'inizio non potranno mai allargarsi verso il nostro tipo di intelligenza ma potranno solamente crescere in altezza, diventando anche bravi a farlo, ma rimanendo solo un ottimo lavoro di ingegneria. La soluzione non risiede nell'aggiungere ulteriormente mole di dati per eliminare i gap di conoscenza nella speranza che ad un certo punto inizino a comprendere veramente o che diventino coscienti.

La speranza è che gli ingegneri possano sviluppare degli strumenti dotati di ragionamento e con un processo di apprendimento simile al nostro, non basandoli solo sull'imitazione e la generazione probabilistica. Inoltre, in riferimento ai modelli attuali è urgente applicare filtri che prendano in considerazione l'etica delle loro generazioni, o parametri di accuratezza che

possano aiutare l'utente a comprendere il livello di affidabilità ad ogni porzione di testo scritto dall'AI. A livello del mercato del lavoro, è improbabile vedere interi settori sostituiti da queste tecnologie, non vedremo grafici, scenografi o psicologi che perderanno la loro professione. E' plausibile invece che i compiti più meccanici e ripetitivi dove le intelligenze umane non sono necessarie verranno svolti dagli LLM, però l'entità del rimpiazzamento non sarà rivoluzionaria. Queste riflessioni vogliono essere da monito per i lettori al discriminare la gloria acquisita da AI con la loro effettiva condizione, stando attenti a non considerarli super esseri umani, ma strumenti di assistenza. È necessario comprendere i loro limiti e utilizzarli in modo sano e accorto, ricordando di avere davanti solo un ottimo calcolatore.

Bibliografia

A. M. Turing. (1950) Mind, New Series, Vol. 59, No. 236 (Oct., 1950), pp. 433-460 Computing Machinery and Intelligence.

Searle, J. R. (2007). What is language: some preliminary remarks. *Explorations in Pragmatics. Linguistic, cognitive and intercultural aspects*, 7-37.

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., ... & Lewontin, R. C. (2014). The mystery of language evolution. *Frontiers in psychology*, 5, 401.

Jhon Searle (2013) Our shared condition: Consciousness
<https://www.youtube.com/watch?v=eqDgt12m26c>

Searle, J. (1980). Minds, brains, and computers. *The Behavioral and Brain Sciences*, 3(3), 417-457.

Searle, J. R. (1990). Is the Brain a Digital Computer, Proceedings and Addresses of the American Philosophical Association, 64, 21-37.

Khuramova Farangiz Uchkun kizi (2022). STRONG AND WEAK ARTIFICIAL INTELLIGENCE. European Journal of Research Development and Sustainability (EJRDS)

Bishop J. Mark (2021): Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It. *Frontiers in Psychology*

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Bishop, J. M. (2002). Dancing with pixies: Strong artificial intelligence and panpsychism. *Views into the Chinese room: New essays on Searle and artificial intelligence*, 360-378.

Bledsoe, W. W., & Browning, I. (1959, December). Pattern recognition and reading by machine. In *Papers presented at the December 1-3, 1959, eastern joint IRE-AIEE-ACM computer conference* (pp. 225-232).

Kurzweil, R. (2005). The singularity is near. In *Ethics and emerging technologies* (pp. 393-406). London: Palgrave Macmillan UK.

Maudlin, T. (1989). Computation and consciousness. *The Journal of Philosophy*, 86(8), 407-432. Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219*.

Azamfirei, R., Kudchadkar, S.R. & Fackler, J. Large language models and the perils of their hallucinations. *Crit Care* **27**, 120 (2023). <https://doi.org/10.1186/s13054-023-04393-x>

Weih, L., Yuile, A., Baillargeon, R., Fisher, C., Marcus, G., Mottaghi, R., & Kembhavi, A. (2022). Benchmarking progress to infant-Level physical reasoning in AI. *Transactions on Machine Learning Research*.

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on Microsoft's "taylor" experiment," and wider implications. *Acm Sigcas Computers and Society*, 47(3), 54-64.

Mathur, V., Stavrakas, Y., & Singh, S. (2016, December). Intelligence analysis of Tay Twitter bot. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 231-236). IEEE.

Ankit Pal, Logesh Kumar Umapathi, Malaikannan Sankarasubbu (2023) Med-HALT: Medical Domain Hallucination Test for Large Language Models. Saama AI Research, Chennai, India

Sole 24ore (marzo 2023) Alessia Maccaferri: *L'intelligenza artificiale si vota a Padre Pio e ai suoi devoti digitali*
https://24plus.ilsole24ore.com/art/l-intelligenza-artificiale-si-vota-padre-pio-e-suoi-devoti-digitali-AEJZJzvC?refresh_ce=1

Alessandro Di Medio (2023) Santità Artificiale. Nuova stagione: settimanale diocesano di Napoli
<https://www.chiesadinapoli.it/wp-content/uploads/sites/2/2023/05/9-aprile.pdf>

Lorenzo Ancona (2023). Gender bias in GPT-3. Github.
https://github.com/lorenzoancona/group_project

Open to Debate (2023). Debate: Will ChatGPT do more harm than good? With Gary Marcus and Keith Teare.