ANALISI

La fonte dell'intelligenza artificiale incorpora un serio problema etico

PAOLO BENANTI


IDEE E COMMENTI 3

La fonte dell'intelligenza artificiale incorpora un serio problema etico

Si chiama «C4» ed è il colossale campione di dati e pagine web all'origine dei sempre più evoluti sistemi di AI. Rischi per la democrazia ma anche di una colonizzazione culturale

PAOLO BENANTI

Un'inchiesta giornalistica illustra le possibili derive dovute alla composizione dei «dataset» diinformazioni Si chiama «C4» ed è il colossale campione di dati e pagine web all'origine dei sempre piùevoluti sistemi di AI. Rischi per la democrazia ma anche di una colonizzazione culturale Il termine C4farà pensare a molti, vistala notorietà acquisita tra film e videogiochi, all'esplosivo al plastico.In realtà c'è un altro C4, non meno esplosivo nei contenuti, che in questi giorni sta facendonotevolmente discutere. Uno dei processi fondamentali che permette a motori di ricerca come Google,Bing e Yahoo di indicizzare un contenuto su internet è il cosiddetto crawling, ovvero – semplificandoun po' – un software che analizza i contenuti di una rete (o di un database) in un modo metodico eautomatizzato acquisendo una copia testuale di tutti i documenti presenti e creando un indice che nepermetta, successivamente, la ricerca e la visualizzazione. Esiste un corpus, detto Common Crawl, checontiene petabyte di dati raccolti in 12 anni di web crawling. Il corpus contiene dati grezzi dipagine web, estratti di metadati ed estratti di testo.

Di tutto questo è stata fatta una versione definita colossale e ripulita (cleaned). Si ottengono cosìi componenti – le quattro C – di questo C4: Colossal, Cleaned version of Common Crawl.

I l quotidiano statunitense "Whashington Post" ha reso pubblico di recente in un lungo articoloinvestigativo che il dataset C4 contiene dati proveniente da fonti quali Stormfront, Kiwi Farms, 4chane altri siti web tutti considerati potenzialmente problematici. Tra le fonti emerse dall'indagine cisono almeno 27 siti web identificati dal governo statunitense come connessi ai mercati dellacontraffazione e della pirateria o dati provenienti da siti come "Vdare", un sito statunitense diestrema destra che promuove l'opposizione all'immigrazione ed è associato alla supremazia bianca, alnazionalismo bianco e all'alt-right, e "Breitbart", un sito di notizie di estrema destra consideratoda accademici e giornalisti misogino, xenofobo, ma anche l'emittente russa "RT", il primo tra i canalidella Russia completamente in digitale direttamente finanziata dal Cremlino.

Q uesto testo che compone il C4, di fatto, è la base che costituisce la principale fonte diaddestramento e di acquisizione di informazioni che le intelligenze artificiali (AI) mostrano dipossedere sul mondo, e inevitabilmente influenza il modo in cui ogni AI risponde alle richieste e alleinterazioni degli utenti. Se costruiamo intelligenze artificiali come Gpt-4 che sono in grado disuperare alcuni dei più severi test di ammissione alle facoltà universitarie, ad esempio, è moltoprobabile che questa capacità emergente del sistema sia connessa ai dati di addestramento che hannoincluso migliaia di siti con test di esercitazione per questi esami.

L e aziende tecnologiche, però, hanno innalzato una ferrea cortina di mistero su ciò che hanno dato in

pasto in fase di addestramento all'intelligenza artificiale. E se, nell'utilizzarle, ci sorprendequanto sembrano in grado di fare, di fatto rimaniamo ciechi sulle fonti e sulle origini di questosapere. La cosa sembra costituire una vera e propria inversione rispetto alla modalità scientificadella conoscenza che ha fatto della trasparenza sui dati e sulle fonti una delle cifre della stessascientificità. Per guardare all'interno di questa scatola nera il "Post" ha analizzato il set di datiC4 di Google collaborando con i ricercatori dell'Allen Institute for AI e hanno classificato i sitiweb utilizzando i dati di Similarweb, una società di analisi web. Circa un terzo dei siti web da cuisono estratti i dati in origine non ha potuto essere classificato, soprattutto perché non sono piùpresenti su Internet.

S tando a quanto dichiara Google, C4 è stato inizialmente sviluppato come "versione ripulita" dei datidi Common Crawl ed è stato utilizzato per addestrare alcune AI di alto profilo in lingua inglese,chiamate modelli linguistici di grandi dimensioni, o Llm, tra cui il T5 di Google e LLaMA di Facebook.OpenAI, di contro, non rivela quali set di dati utilizza per addestrare i modelli che supportano ilsuo popolare chatbot, ChatGpt, appena tornato fruibile in Italia.

Quello che ci interessa sottolineare, prima di fare ulteriori analisi, è il fatto che un sito webviene indicizzato in C4 solo se è in inglese e che il dataset, non contenendo dati in altre lingue, èanglofono. Grazie al tool fornito dal quotidiano statunitense abbiamo fatto ulteriori indaginicercando alcune fonti. Un dato interessante, per esempio, è che il dominio vatican.va è al 4.967 postoavendo fornito quasi 2 milioni di token (i piccoli frammenti di testo che costituiscono la base delleinformazioni con cui è addestrato il sistema). Non bisogna farsi impressionare dal numero diclassifica perché le fonti sono oltre 15,7 milioni e di fatto un risultato sotto i primi 5.000 èaltissimo, soprattutto se si considera che si prendono in esame solo le pagine in inglese e non tuttele pagine del sito. Per fare un confronto, la Cia, che contiene il Cia World Factbook, unapubblicazione annuale che riporta i dati statistici fondamentali e una sintesi di informazioniriguardanti tutti i Paesi del mondo, è dietro il sito vaticano di quasi 600 posizioni. Parlando diquesto con Denis "Jaromil" Roio, il famoso programmatore, "hacker etico", artista digitale eattivista, ci è venuto in mente di cercare anche il sito della Nasa: ebbene, Nasa.gov è staccato di100 posizioni. La battuta è sorta spontanea: per sapere del cielo C4 si affida più alla Chiesa che nonall'astrofisica.

I l "Washington Post" riporta un'interessante analisi aggregata del dataset: «I siti web commerciali eindustriali costituiscono la categoria più grande (16% dei token categorizzati), guidata da fool.comal n. 13, che fornisce consigli sugli investimenti. Poco distante kickstarter.com, al n. 25, checonsente agli utenti di finanziare in crowdfunding progetti creativi, e più in basso patreon.com, n.2.398, che aiuta i creatori a raccogliere quote mensili dagli abbonati per contenuti esclusivi.Kickstarter e Patreon potrebbero dare all'intelligenza artificiale l'accesso alle idee e alle copie dimarketing degli artisti, sollevando il timore che la tecnologia possa copiare questo lavoro persuggerirlo agli utenti».

L'iniziativa del "Washington Post", che ha per la prima volta permesso dianalizzare uno di questi set di dati per rivelare completamente i tipi di siti web proprietari,personali e spesso offensivi che entrano nei dati di addestramento di un'intelligenza artificiale, è

di grande interesse, utilissima per iniziare a pensare se e come questi sistemi siano adeguati erispettosi delle identità culturali e democratiche dei Paesi occidentali e in particolare diquell'area sempre più isolata nella difesa della rule of law che è l'Europa.

L' indagine ha rilevato che C4 è in maggior parte dominato da siti web legati al giornalismo, allacreazione di contenuti, all'intrattenimento e allo sviluppo di software, con patents. google.com,wikipedia. org e scribd.com elencati come primi tre siti. Tuttavia i dati di addestramento provenientida siti più discutibili potrebbero potenzialmente indurre i modelli di intelligenza artificiale agenerare testi indesiderati, razzisti, pornografici, inaffidabili e in generale dannosi. L'algoreticaci chiede di riflettere su questa sorta di "materia prima" per le AI: la qualità dei dati influenza laqualità e l'affidabilità dei sistemi su cui avviene l'addestramento. Dobbiamo chiederci se e come unascelta di cosa includere in C4 non sia di fatto anche un'opzione di natura politica e con severeconseguenze geopolitiche. Di fatto la scelta dei dati è una scelta – una tokenizzazione, per usare untermine tecnico – di una cultura. Questa scelta potrebbe, estremizzando un po', trasformare iltradizionale softpower culturale esercitato dall'industria dei media e da Hollywood al livello di unvero e proprio colonialismo culturale.

F orse anche per questo di recente il governo britannico ha stanziato 100 milioni di sterline per lacreazione di una task force sull'AI incaricata di creare modelli di base o modelli di IApre-addestrati, come Gpt di OpenAI. Creare un set di dati pubblico è la prima forma di difesa.Dobbiamo chiederci se non sia arrivato il momento di pensare alla creazione di un nostro datasetculturalmente pesato ed eticamente bilanciato per permettere al Paese e ai servizi pubblici dibeneficiare dell'impatto trasformativo dell'AI. RIPRODUZIONE RISERVATA L'addestramento proveniente dasiti discutibili potrebbe potenzialmente indurre a generare testi indesiderati, razzisti,pornografici, oltre che inaffidabili Dobbiamo chiederci se e come una scelta di cosa includere non siaanche un'opzione politica e con severe conseguenze geopolitiche Il Washington Post ha reso noto che lafonte di dati «C4» contiene informazioni provenienti da siti considerati pericolosi.

Nella foto: un'immagine tratta dal film "Matrix".