Intelligenza artificiale e pappagalli stocastici: il rischio di pregiudizi amplificati a briglia sciolta

Stavolta vorrei provare una cosa nuova: raccontare l'intelligenza artificiale con articoli di ricerca tra i più noti nel giro accademico ma sconosciuti ai più. Credo che sia d'interesse pubblico divulgare alcune delle riflessioni che maturano sul fronte dell'IA anche attraverso conflitti epici.

Il primo è un articolo uscito a marzo 2021, fondamentale e a suo modo scandaloso, con un buffo titolo: *Sui pericoli dei pappagalli stocastici: i modelli di linguaggio possono essere troppo grandi?* Autrici principali: Emily M. Bender, linguista dell'Università di Washington, e Timnit Gebru, informatica di prima grandezza e attivista cofondatrice di *Black in AI*.

Che cos'è un "pappagallo stocastico"? Un pappagallo ripete quello che diciamo, imitando i suoni senza capirci un'acca. Un *processo stocastico* è un fenomeno nel tempo che possiamo misurare ma non prevedere, come le precipitazioni sul mio balcone o l'andamento del Dow Jones. I valori che assume istante per istante possiamo anticiparli solo in termini di statistica e probabilità.

Combinando le due cose si ottiene una sagace definizione dei *language models* (LM), i modelli IA che manipolano e generano linguaggio (le IA "creative" delle scorse puntate): "*un LM è un sistema che appiccica insieme a casaccio sequenze di forme linguistiche che ha osservato nei suoi sterminati dati di addestramento, in base a informazioni probabilistiche sui modi in cui si possono combinare, ma senza alcun riferimento al significato: un pappagallo stocastico*".

È proprio così: i LM non fanno che ripetere quello che hanno sentito da noi. Siccome lo riassemblano con sofisticatezza, le loro esternazioni *sembrano* autentica produzione umana e destano meraviglia, proprio come i pappagalli. Ma come questi non hanno la benché minima *comprensione* di ciò che dicono.

I LM sono essenzialmente *distribuzioni di probabilità di sequenze di parole*. Producono testi cercando di *predire la prossima sequenza* come noi proviamo a prevedere che tempo farà. Bizzarro, vero?

A partire da Bert (Google, 2019) i LM sono ingrassati a dismisura sia per numero di parametri (coefficienti dei nodi interni della rete neurale, ora centinaia di miliardi) che per dimensioni dei dataset di addestramento. Le prestazioni sono migliorate, ma al prezzo di un colossale impiego di denaro e di energia per il calcolo. I costi ambientali non sono ancora fra i criteri per valutare la loro efficienza.

Dataset più grandi, inoltre, non portano più varietà o verità. I dati sono raccolti dal web, che è un ritratto del mondo umano assai distorto. Molte lingue sono quasi o del tutto assenti. I punti di vista dominanti sono assai più frequenti delle culture minoritarie e includono vaste paludi di discriminazione e malignità contro donne, trans, disabili, anziani, minoranze, emarginati di tutti i tipi, e ovviamente razzismo assortito. Figure cristallizzate nei dataset e rese immutabili. L'egemonia culturale è codificata e occultata nel profondo delle reti neurali.

È questa la *Bildung* delle macchine: una formazione falsamente universalista ma in realtà faziosa e retrograda, che guasta alla radice la promessa dell'IA di aiutare l'umanità a risolvere problemi universali. Del linguaggio c'è solo la forma (i dati) senza il significato. Il significato infatti non risiede in rapporti statistici: è incalcolabile.

L'agilità nel fabbricare testi con tali mezzi non fa che

portare “più testi nel mondo che rinforzano e propagano stereotipi e associazioni problematiche, sia presso gli umani che incontrano quei testi, sia verso i futuri LM che saranno allenati con gli output della generazione precedente”. Pregiudizi amplificati a briglia sciolta.

Il divorzio tra espressione linguistica e comunicazione di senso è una minaccia non da poco per il genere umano. L’importanza di capirsi con i nostri simili ci ha dotato di un automatismo evolutivo che ci induce a leggere un’intenzione e un significato ovunque ve ne sia la minima apparenza. Ora, però, ci sono macchine che dirottano questa natura. Ed eccoci pronti per essere ingannati oltre ogni limite.

Per esempio trasformando un LM in un ideologo complottista. Facile fargli produrre montagne di storie fantasiose, che una folla di bot sguinzagliati nei forum e nelle chat disseminerannno ai target giusti per reclutare seguaci e promuovere azioni politiche estremiste. Azioni fondate *sul nulla*, la beffa più amara. Credere che dietro le parole fatte a macchina ci sia qualcuno che le ha meditate può trascinarci in un delirio collettivo.

Per queste e altre serie ragioni argomentate con ben 158 fonti, le autrici propongono di abbandonare la via dei LM ipertrofici e insidiosi, spostando le risorse sulla vera comprensione del linguaggio naturale e sulla creazione di dataset più dosati, curati e documentati con la dedizione che si usa per gli archivi. Ma il fatto che dopo questo articolo Gebru sia stata licenziata da Google per rappresaglia suggerisce che la strada dell’IA non sarà quella della ragionevolezza.