

UNIVERSITÀ E SCUOLA AFFRONTANO IL TEMA DELL' I.A.

# FACCIAMOCI UN ESAME DI COSCIENZA<sup>1</sup>

Sono bastati due anni di esposizione ad annientare il piacere di pensare in proprio (ovvero, di pensare tout court) delle "classi dirigenti del futuro"?

Roberto Casati

L'IA non è, né mai sarà cosciente. Direi che è giunto il momento di fare un'affermazione categorica. Ripeto, l'IA (penso soprattutto all'IA generativa) non sarà mai cosciente. Né oggi, né prima della fine del 2025 come dice Musk, né entro il 2027 come vorrebbe il rapporto "IA 2027", né entro il 2030 o il 2050, né mai. Possiamo passare giornate intere a discutere dell'affermazione categorica, ma propongo un esercizio un po' diverso. **Quali sono le conseguenze del vedere l'IA come non cosciente?** Che cosa significa mettere la nostra vita nelle mani di macchine che *sembrano* coscienti ma che non lo sono? Come interagire con queste macchine?

Di fatto, c'è una parte del mondo che agisce già come se l'IA fosse già cosciente, e che interagisce con l'IA generativa come se interagisse con un essere cosciente dotato di saggia parola. **Se l'IA non è cosciente, questa parte di mondo vive nella grande illusione**, e andrebbe informata: come dice il titolo di un libro di Carlo Rovelli, a volte la realtà non è come ci appare. L'apparenza di coscienza viene dal fatto che l'IA passa in modo sempre più egregio dei test di Turing sempre più sofisticati, non solo captcha e trucchi per certificare che non è un robot a compilare un formulario, ma dei test in cui noi stessi siamo gli esaminatori. Le discussioni sulla portata del Test di Turing (sul quale si veda l'ottima introduzione di Diego Marconi alla riedizione dell'articolo originale) sono molto orientate dalla sorpresa sempre più forte di fronte alle mirabolanti prestazioni linguistiche dell'IA; come Gesù sorprende i dottori. E come non farsi abbindolare da una macchina programmata per risolvere in modo egregio dei test di Turing sempre più sofisticati (quelle "conversazioni" melliflue dei Chatbot, infarcite di scuse e di complimenti, quando va bene; quella disponibilità morbosa a incoraggiare sentimenti a volte autodistruttivi, che ha spinto al suicidio degli adolescenti fragili, quando va meno bene.)

Ma ovviamente l'attribuzione di coscienza non implica che vi sia una vera coscienza che viene snidata.

Un'altra parte di mondo **paventa gli sfracelli che l'IA cosciente potrebbe fare**. Come fare dapprima sparire tutti gli impieghi, manuali e intellettuali, creando schiere di disoccupati famelici, per poi automatizzare la costruzione di droni invincibili o telefonini esplosivi che annienteranno l'umanità al fine di proteggere la fame di energia dei server. Ma per **questa distopia non c'è bisogno di aspettare l'evento coscienziale**, la "singolarità". L'IA non solo fa già adesso moltissimi compiti intellettuali meglio di noi. (Anche se come ricorda Elena Pasquinelli, ci siamo automatizzati la vita per

rendercela comoda, e visto che questo ha però significato eliminare dei compiti difficili e stimolanti per riempircela di **task banali e noiosi**, ci siamo ritrovati a chiedere all'IA di eseguire questi ultimi per noi.) Dato che non si tratta di una macchina che costruiamo, ma di una macchina che *addestriamo*, finisce con il fare questi compiti *imparando a compiacerci*. Sembra non solo cosciente, ma anche furba e persino maligna. A furia di addestrarla, finiamo con creare un mostro che trama contro di noi, proprio perché continuiamo ad addestrarla per renderla sempre più furba e maligna. Sarebbe che dobbiamo imparare a usare termini come "malizioso" accettando che si possa essere maliziosi senza essere coscienti.

“

**La fantascienza horror è infarcita di zombie, esseri dall'apparenza umana che fanno tutto quello che fa un essere umano, compreso passare un test di Turing sull'intelligenza.**

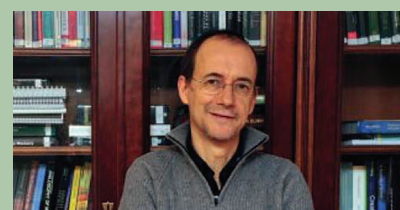
Per aiutare la discussione pubblica, direi che l'IA realizza una visione letteraria diversa. **La fantascienza horror è infarcita di zombie, esseri dall'apparenza umana che fanno tutto quello che fa un essere umano, compreso passare un test di Turing sull'intelligenza. Ma che per definizione non sono coscienti. Morti viventi**" Ma che per definizione non sono coscienti. Morti viventi. Se l'IA non è cosciente, e se passa tutti i test di Turing immaginabili, dobbiamo vederla come uno *zombie meccanico*.

Gli zombie letterari possono essere buoni; spesso sono malevoli. Sono spesso teleguidati, non hanno una volontà propria. Zelanti esecutori, inarrestabili. Ingannevole nell'aspetto, a parte gli occhi fessi. Insidiosi nel mescolarsi a noi. E via dicendo. Vorrei che ogni volta che "dialoghiamo" con un'IA generativa sentissimo l'alto cattivo dello zombie, dell'essere senza coscienza che sembra invece averne una. Cambierebbe il nostro modo di agire? Mi sembra di sì. La questione della coscienza dell'IA è diventata un'enorme e vacua distrazione rispetto ai molti problemi che l'interazione massiccia con le macchine generative crea. Per esempio, ci sono già contesti universitari



in cui gli studenti si rifiutano di passare un'ora se non possono consultare ChatGpt prima di rispondere: sono bastati due anni di esposizione ad annientare il piacere di pensare in proprio (ovvero, di pensare tout court) delle "classi dirigenti del futuro"? (Ho scritto a bella posta **"il piacere di pensare in proprio"**: far fare al posto mio un compito intellettuale all'IA mi fa l'effetto di chiedere a qualcun altro di mangiare il mio gelato al pistacchio). Altro problema, quello della responsabilità. Gli zombie non sono responsabili di quello che fanno; non sono persone; non li si può processare. **Il diritto riconosce certo una personalità morale** a delle entità non coscienti, come le società **per azioni, o le chiese**. Ma si tratta di un'estensione concettuale della nozione di persona che ha richiesto una notevole calibrazione (e il diritto fa una certa fatica ad accogliere tra le "persone" le entità naturali come i fiumi o gli alberi). Non portiamo in tribunale gli zombie ma andiamo a processare i burattinai che li comandano. La favola dell'IA cosciente serve a farci credere che un giorno potremo portare in tribunale un server e un software o un modello, e non invece, come dovremmo, le ditte che lo hanno addestrati?

L'icona che oggi segnala l'intervento di un'IA è la polvere di stelle lasciata da una bacchetta magica. Penso che dovremmo apporre invece un simbolo più perturbante, un volto di zombie con due crocette al posto degli occhi.



**ROBERTO CASATI**

È un Filosofo italiano, studioso dei processi cognitivi. Attualmente è Direttore di ricerca del Centre National de la Recherche Scientifique (CNRS), presso l'Institut Nicod a Parigi e Direttore dello stesso Istituto Nicod. Espone della filosofia analitica, già docente in diverse università europee e statunitensi, è autore di vari romanzi e saggi, tra cui *La scoperta dell'ombra* (2001), tradotto in sette lingue e vincitore di diversi premi, la raccolta di racconti filosofici *Il caso Wassermann* e altri incidenti metafisici (2006), *Prima lezione di filosofia* (2011), *Contro il colonialismo digitale*. Istruzioni per continuare a leggere (2013), recensito in "Professione docente", settembre 2016, con un'intervista all'autore e *La lezione del freddo*, presso Einaudi, una filosofia e un manuale narrativo di sopravvivenza per il cambiamento climatico. Questo libro ha vinto il premio ITAS del libro di montagna e il premio Proclama Elsa Morante L'isola di Arturo 2018. Oceano. Una navigazione filosofica. Einaudi 2022.

<sup>1</sup> L'articolo è già stato pubblicato, con variazioni, nell'inserto domenicale *del sole* 24 ore