Nick Bostrom – Superintelligenza – di Daniele Marini

Raramente ho letto qualcosa di così confuso e male impostato sul tema dell'intelligenza artificiale e sui suoi sviluppi potenziali, come in questo libro. Bostrom ha purtroppo avuto, negli scorsi anni e anche recentemente, una forte influenza nella diffusione dell'idea che l'Intelligenza Artificiale (I.A.) possa un giorno sfuggirci di mano. Una lettura critica di questo testo è quindi essenziale.

Cosa non mi piace di quest'opera?
Prima di tutto il fatto che è scritta da una persona che apparentemente ha una significativa formazione filosofica e nelle neuroscienze accanto a una formazione in fisica. Questa formazione interdisciplinare non emerge se non sul piano puramente tecnico. Quel che manca interamente è l'apporto filosofico o per meglio dire degli aspetti meno materiali della natura umana.

La nozione di intelligenza che emerge è un qualcosa che è dato al più dal lungo processo di selezione naturale e che si manifesta come *capacità umana di pensiero*. Ma dov'è la natura sociale? Dov'è la relazione tra uomo e natura? Nessuno mi convincerà mai che il pensiero umano si manifesti in totale autonomia e indipendenza dal suo rapporto con la natura e con altre persone. Soprattutto in quest'opera una parola, e con essa il concetto associato, manca completamente: **l'interazione**. Eppure nel campo delle scienze cognitive e delle neuroscienze molti ricercatori, purtroppo da pochi anni, stanno considerando l'interazione come un fattore essenziale per la costituzione e l'espressione dei processi cognitivi. Interazione tra esseri umani e interazione con l'ambiente

naturale. Del resto l'intelligenza si è storicamente caratterizzata per la capacità umana di modificare la natura e l'ambiente.

Ma veniamo ad alcuni punti specifici.

Bostrom definisce sinteticamente la superintelligenza "*come qualunque intelletto che superi di molto le prestazioni cognitive degli esseri umani in quasi tutti i domini di interesse*". Quanta vaghezza! Tuttavia questa definizione gli torna utile per esaminare alcune vie attraverso cui si possa creare superintelligenza. E queste vie sono:

- **l'intelligenza artificiale,** definita come la creazione di macchine capaci di apprendere, di affrontare l'incertezza e la probabilità, la "*capacità di estrarre concetti da dati sensoriali o da condizioni interne, e sfruttare i concetti estratti per creare rappresentazioni combinatorie e flessibili da usare nel ragionamento logico e intuitivo*". Immagina che si possano mettere in atto processi evolutivi che portino a questi risultati ma al contempo dichiara che "*le risorse computazionali necessarie solo per riprodurre i processi evolutivi che hanno portato alla intelligenza di livello umano sulla Terra sono notevolmente al di fuori della nostra portata*".

Peccato che non dica nulla su quali potrebbero essere i processi evolutivi significativi per l'emergere dell'intelligenza. Si dedica quindi a elencare possibili strade, citando reti neurali, programmazione genetici e tante altre tecniche che gli esperti di I.A. hanno messo in campo in quasi 70 anni. Nella introduzione aveva peraltro chiaramente individuato la natura di questi metodi che si limitano a risolvere, anche se grande efficienza e accuratezza, solo **problemi di classificazione**.

Ecco, mi sarei aspettato da un ricercatore con formazione

filosofica e di neuroscienze che mettesse in luce l'enorme limite che c'è in sistemi di I.A. capaci solo di classificare. Che *ruolo gioca la classificazione nella formazione dei concetti* in una conoscenza scientifica? I concetti emergono solo dalla classificazione? E come mai nella nostra esperienza di apprendimento scopriamo che la **metafora** costituisce lo strumento più potente per far emergere la concettualizzazione? Ma la metafora è anche una delle forme principali con cui si attua l'interazione umana sul piano del linguaggio, ed è qui che si apre la voragine della vaghezza dell'approccio di Bostrom. Lo scambio metaforico tra individui e gruppi di individui è la premessa per l'instaurarsi di *narrazioni* che a loro volta attivano *processi emotivi*. Ma in un altro momento Bostrom esclude di avere interesse sulle **emozioni**, che vede soltanto come motivazioni che potrebbero essere realizzate con algoritmi specifici nelle macchine di I.A. .

- la **emulazione globale del cervello**. Qui Bostrom si spinge a fare l'ingegnere. Racconta come si potrebbe costruire un modello fedele delle connessioni neurali di un cervello umano parendo da un *cadavere cristallizzato* e ottenerne una *copia artificiale* che incorpora nelle connessioni neurali minuziosamente riprodotte l'intelligenza di questo individuo comprendente esperienze e ricordi. Il tutto partendo da tecniche di scansione ad alta risoluzione ed algoritmi di estrazione della rete di connessione estratta da milioni di immagini del cervello sezionato, a loro volta riportate in strutture neurali descritte da algoritmi.

Siamo ormai nel dominio della follia di Fankenstein! Teniamo in vitro il cervello, mettiamolo in una soluzione fisiologica e osserviamolo … Molto più credibile l'idea di Matrix, che fa vivere i protagonisti in un mondo virtuale ma facendo in modo che l'intero loro sistema neurale, incluse le terminazioni sensoriali, venga sottoposto agli

stimoli dell'ambiente. Questa idea rivela una impostazione epistemologica perfettamente riduzionista: c'è solo il materiale biologico organizzato in neuroni e tessuti. L'unico limite che vede Bostrom è nella tecnologia, ancora priva della potenza di calcolo adeguata. **Riduzionismo biologico e meccanicismo**. Mi spiace usare gli –ismi per "denunciare" l'errore di Bostrom, ma non c'è altra via.

- La terza via che esplora Bostrom è la **cognizione biologica**. E qui si dilunga ma la sostanza è metter in piedi un bel programma eugenetico basato sulla selezione guidata dallo studio del DNA e attuata con forme di fecondazione artificiale per creare il *superuomo superintelligente*. Che novità! Ma ci crede al punto da ritenere che questa potrebbe essere la via più efficace per avere nell'arco di 5-6 generazioni individui dotati di superintelligenza.
- La quarta via consiste nella creazione di **interfacce cervello computer.** Qui Bostrom vede limiti nella criticità di creare protesi invasive, tipo sonde o microchip nel cervello, anche se riconosce che interfacce del genere possono agevolare il recupero di funzioni assenti (cecità, sordità, disabilità motorie o verbali). Peccato che non tenga in considerazioni un'ampia gamma di studi e ricerche sulla rilevazione dell'attività cerebrale elettrica (EEG) che hanno dimostrato di permettere di rilevare stati emotivi e addirittura *cosa stia pensando* il soggetto rispetto a una gamma di stimoli limitati (ad esempio scegliere da una serie di fotografie quella che sta osservando a un dato momento. Esperimenti che si sono spinti con relativo successo anche ad esplorare forme di *teletrasmissione* di azioni che un soggetto compie, attuando sul cervello destinatario una stimolazione magnetica trasmessa a distanza e che riproduce la distribuzione dell'attività elettrica del cervello

sorgente.

Naturalmente sarei in mala fede se dicessi, a questo punto, che Bostrom crede veramente a queste cose (non si capisce se ci crede o si tratta di esperimenti mentali). In realtà le quattro vie che elenca servono a Bostrom per valutare il grado di rischio di uno sviluppo della superintelligenza che perseguisse queste possibili strade. Alcune strade secondo Bostrom sono particolarmente rischiose in quanto i progressi graduali potrebbe mascherare l'approssimarsi del momento in cui la macchina superintelligente diviene talmente autonoma da progredire senza più alcun intervento umano.

Dal mio punto di vista, tuttavia, mancando una analisi approfondita a realmente critica degli aspetti che Bostrom trascura, ho l'assoluta certezza che una macchina superintelligente non vedrà mai la luce. L'assenza degli aspetti di interazione interpersonale e tra uomo e natura, l'assenza delle emozioni che caratterizzano la natura umana e che ne dominano le scelte, mascherano completamente la comprensione di cosa sia l'intelligenza e come sia possibile imitarla con metodi artificiali.

Proseguendo la lettura il senso del lavoro di Bostrom diviene via via più chiaro. Nel terzo capitolo discute delle possibili forme di superintelligenza. Bostrom individua queste forme:

- **superintelligenza di grande velocità**. Immaginando una velocità di 3-4 ordini di grandezza superiore mette in luce paradossi relativistici. Anche qui nasce però una illogicità: Bostrom immagina che la mente superveloce digitale interagisca con il mondo ma non si pone alcun problema al riguardo. E' come se desse per scontato o di nessun interesse l'interazione con il mondo. Riconosce peraltro che non è concepibile una mente accelerata umana.
- **superintelligenza collettiva**. Esempio di intelligenza

collettiva è una comunità di lavoratori che operano su un unico progetto (es. costruzione edile, progettazione di una navetta spaziale ecc.). Parlare di superintelligenza partendo da questi esempi secondo Bostrom non è appropriato, in quanto i problemi di coordinamento non permettono di superare il grado di intelligenza collettiva oggi possibile. Ma osserva che dal Pleistocene a oggi la popolazione mondiale è cresciuta di 1000 volte, e conclude che *l'intelligenza collettiva contemporanea può essere considerata una superintelligenza rispetto a quella del Pleistocene*. Prendendo in considerazione una popolazione mondiale più numerosa e sistemi di comunicazione molto più efficaci di quelli attuali Bostrom immagina una superintelligenza collettiva che potrebbe contare qualche centinaio di migliaia di intelletti superiori. Tutta questa discussione ha un vago sapore di fantascienza. Infatti manca di approfondire il senso del lavoro collettivo e collaborativo, che sono in effetti le vere componenti dell'intelligenza umana. Ma questo a Bostrom sfugge.

- **superintelligenza di qualità**. E qui siamo nella più totale vaghezza. Bostrom non è in grado di proporre alcuna sensata ipotesi di intelligenza qualitativamente superiore a quella umana corrente. Si domanda se possano esistere talenti non espressi ma non sa darsi la risposta, salvo immaginare intelligenze prive di qualità di quella che consideriamo l'intelligenza umana, come, ad esempio, l'assenza della capacità del linguaggio. E quindi?

Bostrom conclude il capitolo osservando che se una intelligenza digitale possedesse maggiore velocità, potesse sfruttare forme collettive evitandone gli svantaggi dovuti ad esempio allo scadente coordinamento, e possedesse qualità, peraltro al momento ignote, si potrebbe arrivare a una superintelligenza digitale. A chi legge queste note trarre

qualche conclusione.

Bostrom dà ormai per scontato che le macchine arriveranno a esplicitare un intelligenza almeno equivalente a quella umana, pur non avendo avuto la capacità di dimostrare questo assunto, visto che ha accuratamente evitato di affrontare i fondamentali problemi della interazione umana e della interazione uomo-natura.

Nel quarto capitolo prova ad analizzare quanto tempo ci vorrebbe per passare da una intelligenza digitale di livello umano a una superintelligenza, esplora cioè la cinetica di una esplosione di intelligenza. E qui si lancia in ragionamenti di tipo matematico partendo da una relazione che legherebbe il tasso di variazione di intelligenza al rapporto tra il potere di ottimizzazione e la resistenza. Il potere di ottimizzazione sarebbe lo sforzo di progettazione per accrescere la qualità dell'intelligenza e la resistenza sarebbe la difficoltà alla introduzione delle tecniche ideate con la progettazione. La cosa curiosa è che con questa equazione si arriva a una curva di sviluppo che ha una crescita che, superata una qualche soglia, molto rapidamente non solo diventa esponenziale ma a un certo punto viaggia quasi verticalmente all'infinito. Qui ritengo che siamo in presenza di un uso abusivo della matematica. Siamo cioè nel pieno dominio della fantascienza.

A questo punto, poco meno di metà del testo, ho smesso di leggere e mi son limitato a sfogliare le pagine, in verità non mi interessa per niente la fantascienza. O per meglio dire mi interessa se è narrata in forma letteraria, cosa qui del tutto assente. Quindi tranquillizzatevi: il futuro dominato dalle macchine è, se mai ci sarà, molto molto molto lontano e sicuramente figure fantasiose come Bostrom non daranno alcun significativo contributo.

Superintelligenza. Tendenze, pericoli, strategie

Nick Bostrom

Editore: Bollati Boringhieri Collana: Saggi. Filosofia Anno edizione: 2018
Pagine: 522 p. 28€ versione ebook 11 €

Nick Bostrom (Helsingborg, Svezia, 1973), laureato in filosofia, fisica e neuroscienze computazionali, è docente alla Oxford University, dove dirige il Future of Humanity Institute, da lui fondato; un centro di ricerca interdisciplinare che permette a un gruppo di matematici, filosofi e scienziati eccezionali di pensare alle priorità globali e alle grandi questioni dell'umanità. Sempre a Oxford, dirige anche lo Strategic Artificial Intelligence Research Center. Bostrom è autore di più di 200 pubblicazioni specialistiche e di diversi libri, tra i quali Anthropic Bias (2002), Global Catastrophic Risks (2008) e Human Enhancement (2009). Superintelligenza. Tendenze, pericoli, strategie – bestseller del New York Times – è il suo primo libro tradotto in italiano.
