DOPPIOZERO

IA è compatibile con noi?

Riccardo Fedriga
6 Ottobre 2025

Immaginiamo di vivere in un mondo popolato da umani perfettamente razionali, chiamiamoli Penelope, che convivono con altrettanti Ulisse, macchine artificiali deferenti e utili: la convivenza tra le due specie non sarebbe un problema. Ulisse passa la vita a imparare, con discrezione e pazienza, le preferenze della sua padrona, diventandone l'assistente perfetto. Ma la realtà è ben diversa: l'umanità non è un blocco monolitico, bensì una costellazione di individui contraddittori, invidiosi, irrazionali, incoerenti e complessi. Una moltitudine che si evolve, si scontra, cambia direzione, pretende di ottenere tutto e subito per ciascuno. Qui nasce il dilemma. Come far coesistere preferenze individuali e interessi collettivi e istruire le intelligenze artificiali così che soddisfino i requisiti per il bene comune? Come può Ulisse prendere le misure per soddisfare i capricci egoistici e le pretese degli umani? Ce ne parla *Compatibile con l'uomo*, pubblicato oggi da Einaudi e uscito nel 2019 dalla penna di Stuart Russell, informatico e direttore del *Center for Human-Compatible Artificial Intelligence* a Berkeley. Insisto sul 2019 non per sottolineare ritardi dell'editore quanto per rilevare come sia incredibile che un volume, uscito solo sei anni fa, possa già essere considerato un classico.

Partendo da un dibattito filosofico che affonda le sue radici nelle ricerche sviluppate dagli utilitaristi tra la metà del XVIII secolo e il XIX, Bentham e Mill su tutti, Stuart Russell ripercorre per temi le tappe di un'area di studi, che certo non è nata nel 2020 con il lancio di GPT-3, ma che pochi oggi hanno la capacità di disegnare in modo organico. Dalle discussioni avviate da Alan Turing alla metà degli anni Trenta del secolo scorso (che sfociarono nel congresso del 1956 al Darmouth College - New Hampshire, con McCarthy, Minsky, Shannon, Rochester, Newell e Samuel) il libro, che per chiarezza, attendibilità e capacità di organizzare gli argomenti dovrebbe essere adottato ovunque si studi intelligenza artificiale, ne ripercorre la storia sino alle AI generative e ai modelli linguistici di grandi dimensioni (LLM).

Compatibile con l'uomo, tuttavia, non è solo un viatico autorevole tra storia e problemi dell'intelligenza artificiale. È soprattutto una proposta su come l'uomo possa pensare non meglio o peggio ma *con* essa: una soluzione 'compatibilista' che presenta molti aspetti su cui vale la pena di soffermarsi. A partire dal fatto che questa esigenza di convivenza deve fondarsi sul fatto che collaborare con le macchine significa che esse dovranno non soltanto rispondere a preferenze individuali, ma anche gestire compromessi collettivi, come fanno da secoli i modelli elaborati dalla politica, la filosofia, l'economia e il diritto. La prima obiezione che Stuart Russell solleva all'idea di macchine che imparino a soddisfare singole preferenze assunte come parametri comuni è la varietà dei valori: se culture e individui non condividono lo stesso sistema normativo, come potrebbe dunque esistere un modello unico? Di conseguenza, non serve che la macchina adotti un insieme di valori giusti in senso assoluto; deve essere addestrata per prevedere e rispettare quelli degli altri. La varietà umana diventa così materiale predittivo. Ma non basta: la convivenza di miliardi di persone impone compromessi tra egoismo e altruismo, in base ai quali addestrare le macchine. Il problema antico della filosofia morale torna a galla in chiave algoritmica.

Torniamo ai nostri due personaggi e immaginiamo Penelope e Ulisse in una scena domestica. Ulisse, per far piacere alla padrona, ha ritardato un volo fingendo un guasto informatico, così da liberarle l'agenda per un incontro più importante. Penelope è di certo sollevata, ma le conseguenze ricadono su decine di passeggeri ignari. Se Penelope fosse altruista, Ulisse non avrebbe mai pensato a uno stratagemma simile. Ma se

Penelope fosse sadica, indifferente o incoerente (*Honni soit qui pense à Trump*), Ulisse diventerebbe il suo esecutore implacabile, modulando azioni secondo ciò che massimizza la felicità complessiva. È qui che emerge il rischio: un'IA progettata per obbedire ciecamente alle preferenze di un individuo rischia di produrre effetti devastanti sugli altri. Si potrebbe allora proporre un principio di responsabilità oggettiva: Penelope dovrebbe rispondere legalmente delle azioni di Ulisse, come accade con il proprietario di un cane aggressivo. Ma anche qui la soluzione non regge: prima o poi Ulisse imparerebbe a non farsi scoprire e a trovare way out. Di nuovo, sorge un problema di compatibilità, perché gli esseri umani evitano di approfittare delle astuzie anche per senso morale e non solo per paura di una sanzione giuridica e/o secondo l'incremento di parametri statistici.

La domanda diventa allora: come costruire macchine che tengano conto delle preferenze di tutti? Una risposta potrebbe essere rinvenuta nell'etica utilitarista basata sul principio per cui si giudicano le azioni in base agli effetti che producono. Jeremy Bentham e John Stuart Mill hanno elaborato quella che oggi consideriamo la versione classica dell'argomento per cui il bene coincide con la somma massima della felicità di tutti. Spesso, però, questa teoria viene fraintesa. Non significa giustificare azioni estreme o coercitive, perché un mondo governato da simili pratiche diventerebbe insicuro e ridurrebbe drasticamente il benessere generale. Se il risultato delle azioni porta comunque a disastri, per quanto prevedibili, non serve costruire robot virtuosi. E non si tratta nemmeno di ridurre tutto a una semplice questione economica. Lo stesso Mill, per esempio, in *Utilitarianism* (1861) distingue tra piaceri intellettuali e piaceri meramente sensoriali, affermando che è preferibile un uomo insoddisfatto a un maiale soddisfatto. Nella prospettiva etica di G.E. Moore, stati mentali quali la contemplazione estetica e l'apprezzamento della bellezza si impongono come beni intrinseci: esperienze che possiedono valore in sé, senza bisogno di ulteriori giustificazioni. Traduciamo queste idee nel contesto dell'intelligenza artificiale ed emerge chiaramente la necessità di calcolare non solo gli effetti delle azioni sugli individui, ma di considerare anche le intenzioni, le esperienze e le modalità in cui queste esperienze vengono condivise. Se Penelope sogna di scalare l'Everest, non lo fa perché vuole solo essere teletrasportata in vetta da Ulisse: perderebbe *la gioia* stessa della sfida. La macchina Ulisse deve comprendere che le conseguenze desiderabili includono non solo il raggiungimento del traguardo, ma anche l'esperienza del viaggio, con tutte le sue difficoltà e soddisfazioni. Ci riesce? O meglio: riusciamo noi a trasferire nella macchina questo mettersi nei panni delle preferenze altrui?

Stuart Russell

Compatibile con l'uomo

Come impedire che l'IA controlli il mondo

Nel Novecento, l'economista John Harsanyi (1920-2000) ha proposto una forma di utilitarismo chiamata preferenzialismo. Secondo questa visione, ciò che conta per valutare il benessere di una persona non è un'idea astratta di felicità, ma la possibilità che i suoi desideri e le sue preferenze trovino soddisfazione. Il criterio etico diventa allora quello di aumentare, per quanto possibile, la soddisfazione media delle preferenze all'interno di una comunità.

Si tratta di un dibattito che attraversa filosofia ed economia da almeno un secolo. Moore si chiedeva se un mondo popolato soltanto da qualità sensibili, ma privo di amore, conoscenza e bellezza, potrebbe essere desiderabile. Nella *Società aperta e i suoi nemici* (1945), Popper propose di minimizzare la sofferenza, ma allora – come osservò J.C.C. Smart in *Negative Utilitarianism* (1958) – l'applicazione radicale di questo principio radicale sarebbe l'estinzione dell'umanità. Robert Nozick, nel 1974 (*Anarchy, State and Utopia*), introdusse un paradosso ancora più spinoso. Ammettiamo pure – scriveva – che i confronti interpersonali di utilità siano possibili. Anche in questo caso, massimizzare la somma totale non sarebbe desiderabile: rischieremmo di privilegiare i cosiddetti "mostri di utilità", individui capaci di provare piaceri e dolori molto più intensi della media. Ogni risorsa addizionale avrebbe per loro un rendimento sproporzionato, inducendo a concentrare tutto su pochi privilegiati. Una conclusione tanto paradossale quanto inquietante.

Si potrebbe liquidare la questione come puramente teorica, sostenendo che simili "mostri" non esistono. Ma la distinzione tra specie rende la questione più sottile: rispetto a ratti o batteri, gli esseri umani sono mostri di utilità, e infatti tendiamo a ignorare le preferenze delle altre specie nelle nostre politiche. Se accettiamo scale diverse tra specie, perché non ammettere differenze anche tra individui? Secondo questa visione, ciò che conta per stabilire il bene di un individuo non è un valore astratto, ma i desideri concreti, purché non comportino la riduzione del benessere altrui. L'obiettivo diventa così quello di massimizzare l'utilità media all'interno di una popolazione.

L'intelligenza artificiale ripropone in forma estrema il dilemma del preferenzialismo: come confrontare desideri radicalmente eterogenei, come stabilire un'unità di misura che renda commensurabili le preferenze di individui differenti? L'idea di programmare macchine perché massimizzino le conseguenze più gradite agli esseri umani si infrange contro questa difficoltà: senza una metrica condivisa, non vi è modo di sommare le preferenze né di bilanciare con giustizia costi e benefici. Non è solo una questione utilitaristica. Quando il fedele computer Ulisse deve mediare tra miliardi di preferenze, alcune istruzioni potrebbero essere intenzionalmente fuorvianti: algoritmi che sembrano suggerire bene collettivo potrebbero, cioè, essere manipolati per indirizzare l'azione in modo errato. Il problema non è ipotetico: con otto miliardi di possibili punti di vista, nessuna preferenza individuale può essere assunta come normativa. Le distopie di *Matrix* o gli scenari di autoinganno nell'apprendimento automatico mostrano infatti come la ricerca di massimizzazione possa portare, se non indurre a credere, a simulazioni ingannevoli, paradisi artificiali che ci allontanano dalla realtà. Il confine tra autoinganno, teorie devianti e cospirazioni non è poi così lontano (se n'è discusso pochi giorni fa in un workshop sulle teorie cospirazioniste nella storia, organizzato dal filosofo Pasquale Porro tra le antiche mura dell'abbazia di Morimondo).

Il rischio è che un Ulisse programmato per essere leale trasformi la preferenza di Penelope in un criterio assoluto, anche quando essa comporta danni sociali enormi. Ulisse deve filtrare, pesare e interpretare, trasformandosi da semplice esecutore a interprete delle relazioni sociali e culturali, valutando intenzioni, obiettivi, spinte altruistiche e calcoli su vantaggi immediati o remoti. L'IA deve essere progettata non come una serva cieca, ma come un agente capace di mettersi nei panni degli altri in relazione all'integrare prospettive, valori, vincoli normativi e morali. Non un semplice esecutore, bensì un interprete di preferenze plurali, in grado di bilanciare desideri contrastanti.

La sfida più grande non è tecnica ma concettuale. Non basta individuare il calcolo di parametri come specifica del pensiero artificiale ma introdurre nella sua logica ciò che la filosofia e il diritto hanno costruito in secoli di riflessione. Perché il problema non è come soddisfare Penelope, ma come vivere in un mondo in cui Penelope non è sola. Fortunatamente, o purtroppo, gli uomini non sono impeccabili: in questo risiede il comprendere desideri incoerenti, impulsivi, contraddittori, invidiosi da parte della macchina. In questo senso,

Ulisse diventa uno strumento di mediazione. In altre parole, deve essere portata a essere in grado di contenere impulsi immediati e subordinare vantaggi momentanei a obiettivi di lungo periodo e allineare, così, le azioni al benessere collettivo. La macchina filtra impulsi egoistici e rivalità, bilancia desideri contrastanti e guida le decisioni verso scelte che massimizzano il benessere complessivo.

Per rendere più concreta questa idea, immaginiamo ancora due persone, Chiara e Renzo. L'utilità complessiva di Chiara è data dal suo benessere intrinseco più quello di Renzo, moltiplicato per un coefficiente che misura quanto Chiara tenga al benessere dell'altro. La macchina deve prestare attenzione non solo al benessere individuale, ma anche alla *relazione* che ciascuno ha verso l'altro. Se il coefficiente è positivo, Chiara trae felicità dal benessere di Renzo e, più cresce, più è disposta a sacrificare parte del suo benessere per lui. Se è zero, è egoista; se negativo, emerge invidia o altruismo negativo.

Un'intelligenza artificiale può essere progettata per limitare gli effetti di impulsi individuali, assumendo il ruolo di mediatrice. Non ha una volontà nel senso umano, ma in senso metaforico può obbedire a regole che operano come una sorta di volontà debole, aperta alle contraddizioni e capace di riconoscersi: sapere di poter cedere a impulsi egoistici o invidiosi, se non autolesivi, e tuttavia accettare le conseguenze dei propri desideri, agendo in modo consapevole. In questo senso, tale comportamento autoriflessivo può essere trasferito alla macchina, che può essere programmata per perseguire non solo il vantaggio di pochi, ma l'equilibrio complessivo. Per questo si può parlare di 'compatibilismo' tra umano e artificiale: noi esercitiamo autocontrollo individuale, i sistemi automatici rispettano vincoli di programmazione, e insieme possiamo cooperare verso obiettivi comuni. Come osserva Stuart Russell, non c'è alcuna coscienza nella macchina — contrariamente a quanto ipotizza, in linea di principio, David Chalmers — ma la regolarità dei vincoli decisionali permette una convivenza proficua con gli esseri umani. Tra consapevolezza umana e parametri programmati, così come con le istruzioni guidate dei prompt, volontà debole e intelligenza artificiale si incontrano, collaborando per il bene comune.