the sign of common work on a broader topological conceptual plane. On this multiplicity of voices, on this variety of intents, and on the strength of this work, the Deleuzean dream of an authentic creation of concepts is thus realized following 'heterogenetic' spirit on which the roots of this research are founded.

Shannon Vallor, *The AI Mirror. How to Reclaim Our Humanity in an Age of Machine Thinking*, Oxford University Press, Oxford 2024 [Francesco Terenzio]

Nel panorama delle moderne tecnologie, le intelligenze artificiali si configurano come un pericolo del tutto particolare: a differenza degli ordigni di distruzione di massa, le IA non minacciano il nostro futuro in maniera diretta: ci rendono piuttosto impossibile immaginare e pensare il nostro stesso futuro. Con queste premesse si apre *The AI Mirror,* saggio in cui Shannon Vallor si confronta con le principali questioni etiche che sorgono dall'incontro tra intelligenze artificiali e vulnerabilità dell'umano.

L'argomentazione del libro prende spunto da una vicenda concreta: Blake Lemoine, un ingegnere di Google, nel 2022 aveva affermato che LaMDA, modello linguistico sviluppato dal colosso di Mountain View, sembrava aver sviluppato una propria coscienza. Per l'autrice del libro si tratta di un episodio emblematico in quanto rende chiara una caratteristica peculiare delle intelligenze artificiali: esse si configurano come "immensi specchi dell'intelligenza umana" (p. 2). Proprio in virtù della loro natura di 'specchi', i pericoli che sorgono dalle intelligenze artificiali sono differenti da tutte le minacce con cui l'uomo si è confrontato nella storia: non si tratta più di nemici in agguato nel mondo esterno, ma tecnologie che minacciano l'essere umano dall'interno della sua stessa umanità. "Non possiamo combattere le IA senza combattere anche contro noi stessi" (p. 4), esse sono estensioni dei valori umani nel mondo esterno. L'obiettivo che Vallor si pone con questo libro è dunque quello di consegnare nelle mani dei propri lettori gli strumenti filosofici necessari per comprendere le intelligenze artificiali, relegate oggi a 'specchio dell'umano', e collocarle nel loro giusto ruolo di strumento attraverso il recupero dei valori propri dell'uomo.

Il primo capitolo si divide in tre parti e introduce il lettore alle principali problematiche che sorgono dall'uso delle IA. Nella prima parte, ripercorre una breve storia della nascita del concetto di intelligenza artificiale che va dai contributi di Babbage e Lovelace al celebre articolo di Turing, fino ad arrivare alla conferenza di Darthmouth del luglio 1956 e alla prima

definizione di 'intelligenza artificiale'. Nella seconda parte l'autrice costruisce una mappa delle principali distinzioni interne ed esterne al mondo delle intelligenze artificiali: dai modelli GOFAI e quelli più moderni basati sul *Machine Learning* (nelle sue tre declinazioni di *supervised, unsupervised* e *reinforcement learning*); la distinzione searliana tra '*weak AI*' e '*Strong AI*'; le '*Narrow AI*' e le '*AGI*'; nonché il ruolo che rivestono *data science* e *cloud computing* nell'utilizzo quotidiano delle intelligenze artificiali. Nell'ultima parte il capitolo introduce infine il lettore alla metafora centrale del libro, quella dello specchio, confrontandola con la ben più nota metafora con cui Gebru e Mitchell nel 2021 hanno paragonato i *Large Language Model* (LLM) a pappagalli. Per Vallor, si tratta di un'analogia solo in parte legittima: "i pappagalli sperimentano un mondo e si confrontano con esso in modo intelligente" (p. 32). Al contrario, alle intelligenze artificiali sono precluse le dimensioni dello sperimentare e del comprendere. Esse si servono di modelli statistici per elaborare un *output* sulla base dei dati ricevuti. Affidare a tali modelli la comprensione di noi stessi, della nostra storia, delle nostre differenze e della nostra umanità condivisa conduce a riprodurre schemi del passato in un'epoca in cui abbiamo una grande necessità di trovare soluzioni adatte alle nuove sfide del presente.

Il secondo capitolo verte sui *bias* che sorgono dalle reti neurali. Si tratta di errori che alla luce di un attento *debug* non sembrerebbero essere prodotto degli algoritmi quanto dei *training data* che rivelano la topologia morale della nostra realtà. Non è sufficiente rimuovere un'etichetta, "aggiustare" l'algoritmo, perché il modello eviti atteggiamenti potenzialmente dannosi o discriminatori: "gli specchi rivelano fatti scomodi, le IA ci costringono a confrontarci con l'inesorabile realtà di errori nocivi e con l'ingiusta esclusione sociale" (p. 46). Proprio per questo, l'autrice critica i presupposti alla base del *machine learning* a partire da una prospettiva fenomenologica (sono menzionati a tal proposito Husserl e Lévinas): non tutta la realtà è riducibile a puri dati formali, gli esseri umani condividono conoscenze che le intelligenze artificiali non saranno mai capaci di apprendere, ossia cosa significa fare esperienza ed essere incarnati in un corpo (cf. pp. 59-60).

Il terzo capitolo, *Through the Looking Glass*, offre un'interpretazione dell'intelligenza artificiale a partire dalle virtù. L'argomentazione del capitolo si intreccia con suggestioni provenienti da importanti opere letterarie e cinematografiche: da *Alice attraverso lo Specchio* di Lewis Carroll e *Erewohn* di Samuel Butler a *2001: Odissea nello Spazio*, *Ex Machina*, *Terminator* e *Westworld*, opere di fantascienza che secondo l'autrice anticipano l'odierna preoccupazione riguardo le intelligenze artificiali, all'interno delle quali le IA sono spesso raffigurate come entità capaci di distruggere e

soppiantare la razza umana. Ai fini di questo discorso è particolarmente importante la giusta considerazione delle virtù: l'intelligenza artificiale può infatti influenzare la concezione delle virtù. Ad esempio, un confronto prolungato con le intelligenze artificiali potrebbe rendere difficile riconoscere la virtù, nonché cambiare la nozione di ciò che si ritiene essere "migliore" (p. 67). Le virtù sono particolarmente importanti per guidare la maniera in cui gli esseri umani pensano il loro futuro. Tra le più importanti virtù inibite dall'intelligenza artificiale Vallor individua l'immaginazione e la *phronesis*. Entrambe queste virtù nel momento in cui sono messe in connessione permettono all'uomo di valutare le situazioni e compiere scelte ponderate, immaginare insomma come costruire il proprio futuro. In questo senso le intelligenze artificiali non devono essere concepite come mere proiezioni del passato, ma come specchi che aiutano l'essere umano a guardare il passato mentre si dirige incontro al proprio futuro (cf. p. 101).

Il quarto capitolo riguarda la maniera in cui le intelligenze artificiali amplificano il pericolo configurandosi come sistemi di decisione algoritmica. Da un lato tali sistemi appiattiscono i ragionamenti umani sulle dimensioni proiettate dai loro algoritmi, dall'altro le IA sono caratterizzate da un'oscurità epistemica, il che costituisce un forte limite per il ragionamento umano. Le intelligenze artificiali non sono capaci di reagire prontamente alle nuove informazioni di carattere morale che si introducono nell'ambiente o ai nuovi ragionamenti morali che si presentano nel discorso. "Se le IA specchio fanno apparire lo spazio dei ragionamenti umani come troppo inefficienti, inaffidabili e superflui perché siano tollerati, allora l'impatto delle IA sulla maturità morale e politica dell'umanità potrebbe essere superiore a quello di qualsiasi dittatore" (p. 131). In quanto tale, il pericolo delle intelligenze artificiali non è intrinseco, esso risiede piuttosto nella loro capacità di limitare la natura umana. Nel capitolo successivo Vallor afferma che tali strumenti possono essere usati dagli esseri umani per potenziare la disinformazione e manipolare altri esseri umani. D'altronde i pericoli sorgono esclusivamente nel momento in cui le macchine interagiscono con l'umano, e ad esempio soltanto per un essere umano è possibile proiettare le proprie emozioni su un guscio vuoto come quello di una *chatbot*. Progettare i modelli in maniera tale che abbiano una forte morale è del tutto inutile se non cambia l'atteggiamento verso la responsabilità richiesta dal progettare il futuro (p. 157). Di fronte a tale chiamata, Vallor propone di tornare a concepire la creazione, l'espressione l'amore come atti morali che nessuno specchio può replicare, delegittimando la maniera con cui gli esseri umani si relazionano oggi con le IA.

Gli ultimi due capitoli riassumono la posizione di Vallor e chiudono i principali fili argomentativi presentati durante il corso del libro. I due concetti fondamentali che ne emergono sono il *bootstrapping*, termine informatico che Vallor riconduce al suo significato inglese originale, la capacità di tirarsi su con le proprie sole forze (p. 161) e il 'coraggio civile', unica via rimasta all'umano per poter assumere su di sé le proprie responsabilità. È uno sforzo che l'umano deve compiere di fronte a problemi attuali come il cambiamento climatico, l'esaurimento delle energie non rinnovabili, l'acidificazione degli oceani, l'incremento della diseguaglianza economica. "Le intelligenze artificiali non sono il problema. Il problema è la nostra riluttanza a prendere le distanze dai nostri strumenti per rivalutare gli schemi che stanno riproducendo – anche quelli presumibilmente virtuosi" (p. 183). Il compito è dunque richiamare le intelligenze artificiali e la cultura tecnologica per adattarle a una visione morale. Vallor propone in ultima istanza un "progetto eroico condiviso", che si ispiri alla saggezza pratico-creativa per rinnovare ed espandere le possibilità tecnomorali (p. 193).

La prospettiva proposta da Shannon Vallor è estremamente attuale e prende le giuste distanze dalle intelligenze artificiali per comprendere i modi in cui influenzano i nostri ragionamenti e il nostro agire. A partire da un'attenta riconsiderazione del ruolo dei valori nella società odierna e con uno stile di scrittura chiaro e profondo, *The AI Mirror* consegna nelle mani dei suoi lettori gli strumenti per tornare a pensare autenticamente il futuro dell'umano.