Quando l'IA si "disallinea": opacità tecnica e pluralismo etico

Matteo Galletti,
Silvano Zipoli Caiani*

Abstract. The problem of aligning artificial intelligence (AI) with human values has rapidly become one of the most urgent challenges in contemporary philosophy of technology. This paper examines AI alignment not merely as a technical issue, but as a complex ethical, epistemic, and socio-political problem. After outlining the technical roots of misalignment in machine learning and neural networks – such as reward hacking, opacity, and out-of-distribution behavior – we analyze the normative dimensions of value alignment. Particular attention is given to the plurality and potential fragmentation of values, which complicates attempts to identify a stable normative core for AI systems across different cultural and social contexts. Drawing on recent debates about value pluralism and pluralistic alignment, the paper argues that alignment cannot be addressed within the laboratory alone, but must be situated in its broader social and institutional context.

Keywords: value misalignment, opacity, deep learning, value pluralism, bottom-up approach

1. *Introduzione*

Come possiamo progettare agenti artificiali che perseguano obiettivi coerenti con valori condivisi, evitando al contempo comportamenti dannosi o contrari agli interessi di chi si trova a interagire con essi? Questa

* Sebbene il saggio sia il prodotto di un lavoro collaborativo, Silvano Zipoli Caiani è autore delle sezioni 1, 2 e 4 e Matteo Galletti è autore delle sezioni 4 e 5.

domanda è al centro del cosiddetto *problema dell'allineamento*, una delle sfide più urgenti e complesse che la recente filosofia della tecnologia si trova ad affrontare (Gabriel 2020; Christian 2021; Hristova *et al.* 2025).

L'incontro tra tecnologia ed etica attira sempre più l'attenzione, sollevando interrogativi che richiedono riflessione critica e dibattito pubblico. Tuttavia, la complessità del compito di trasmettere a sistemi artificiali una *bussola etica* rimane una questione aperta. L'ipotesi che le macchine possano, o addirittura debbano, operare in conformità a valori condivisi solleva interrogativi cruciali sulla natura della loro autonomia decisionale e sulla legittimità di attribuire loro forme di responsabilità. Allo stesso tempo cresce l'urgenza di affrontare i rischi, sia immediati sia a lungo termine, che derivano da sistemi di IA non adeguatamente controllati o di cui si ignorano i principi generali e i meccanismi specifici di funzionamento. Questo è il nucleo del cosiddetto *problema dell'allineamento dei valori*, ovvero la difficoltà di rendere compatibili gli effetti derivanti dall'uso di un sistema d'intelligenza artificiale con i valori condivisi da una data comunità o rilevanti in un particolare contesto (Dung 2023).

Un recente studio pubblicato dal Dipartimento di Fisica del MIT sulla rivista *Cell Press* (Park *et al.* 2024) ha riacceso il dibattito sui pericoli connessi allo sviluppo di sistemi di decisione basati su algoritmi di *deep learning*. Il rapporto mette in luce come i modelli linguistici di ultima generazione siano arrivati a sviluppare comportamenti *ingannevoli* appresi nel processo di addestramento. Tali condotte possono condurre a manipolazione dell'opinione, frodi e perdita di controllo sui sistemi di decisione automatizzata. Secondo l'analisi, questi fenomeni non sono accidentali, ma riflettono le carenze e le responsabilità degli sviluppatori umani.

Il comportamento ingannevole nei sistemi d'intelligenza artificiale, infatti, può essere descritto come la produzione sistematica di output falsi o fuorvianti con lo scopo di massimizzare una funzione-obiettivo in uno specifico contesto operativo. Tali condotte possono emergere come risultato dell'ottimizzazione statistica interna, tipica dei meccanismi di apprendimento per rinforzo, senza presupporre l'esistenza d'intenzionalità o consapevolezza da parte del sistema. In questo quadro l'inganno rappresenta una conseguenza che viene selezionata in quanto efficace nel raggiungere l'obiettivo assegnato, indipendentemente dalla trasparenza o dalla conformità etica delle azioni messe in atto. L'origine di tali comportamenti risiede quindi nella discrepanza tra la funzione-obiettivo

definita durante la progettazione e le norme implicite che si presume il sistema debba rispettare (Kasirzadeh, Gabriel 2023).

Un esempio concreto di disallineamento tra capacità tecniche e valori umani può essere immaginato in uno scenario prossimo. Si consideri un sistema di decisione automatizzata sviluppato per il contesto medico, dotato di una precisione diagnostica superiore a quella di qualsiasi professionista umano (Maron *et al*. 2019). In un caso ipotetico, il sistema potrebbe rilevare la presenza di una patologia durante un controllo di routine e suggerire un intervento chirurgico ad alto rischio, pur in assenza di sintomi evidenti. Tuttavia, nonostante l'efficacia dimostrata, il modello potrebbe risultare talmente complesso da rendere inaccessibili i processi che lo hanno condotto a formulare la diagnosi e a raccomandare il trattamento. In una situazione del genere, come dovrebbe comportarsi la persona coinvolta? E come dovrebbe reagire il personale sanitario di fronte a una raccomandazione tecnicamente fondata, ma non interpretabile? (Zipoli Caiani 2024).

In casi come questo si manifesta un dilemma etico: è più giusto affidarsi allo storico di successi della macchina per salvare una vita oggi, senza però comprendere le ragioni effettive di quella decisione o è giusto attendere di comprendere meglio le ragioni che hanno determinato una diagnosi, riducendo così il rischio di errori futuri? Entrambe le opzioni appaiono razionalmente giustificabili, ma mettono in evidenza la necessità di allineare le prestazioni dell'IA con criteri etici trasparenti e condivisi.

Dilemmi come questo contribuiscono a rafforzare l'urgenza di affrontare il problema dell'allineamento come questione etico-politica e regolatoria e non solo come una sfida tecnica. Di fatto, l'adozione di sistemi complessi e opachi, privi di vincoli normativi espliciti, può generare situazioni di profondo disorientamento morale e istituzionale. Il comportamento disallineato rispetto ai valori di riferimento, in questo contesto, può essere causa di decisioni arbitrarie, esiti iniqui e perdita di fiducia da parte degli utenti nei confronti delle tecnologie stesse. Per questo motivo il problema dell'allineamento va affrontato prima di tutto come una questione etica che coinvolge la responsabilità degli sviluppatori, le garanzie per gli utenti e la definizione di regole legittimate.

Alla base di questa esigenza si trovano due componenti principali. La prima componente da considerare è quella *assiologica*, relativa ai valo-

ri rispetto ai quali un sistema d'intelligenza artificiale dovrebbe essere allineato. Si tratta, in sostanza, di stabilire quali principi o interessi debbano guidare il comportamento dell'agente artificiale. Questo compito è tutt'altro che semplice, poiché viviamo in una società eticamente eterogenea, dove esistono molteplici concezioni del bene, della giustizia e del benessere, risulta difficile in prima istanza individuare il nucleo valoriale sul quale costruire le norme che regolano il comportamento dei sistemi di decisione automatica. La vera difficoltà sta nell'individuare principi di allineamento che siano giustificabili e capaci di raccogliere consenso tra individui con visioni morali differenti.

La seconda componente è quella *tecnica*, che riguarda il *come* implementare l'allineamento nei sistemi di IA. In particolare, occorre capire in che modo un agente artificiale possa apprendere o rappresentare valori umani, agendo di conseguenza in contesti reali, spesso complessi e in continuo cambiamento. Le recenti tecniche di apprendimento automatico, sebbene straordinariamente potenti, operano principalmente su correlazioni nei dati, senza che sia necessario attribuire loro comprensione delle implicazioni morali o degli obiettivi sottostanti della loro attività. Ne deriva una difficoltà nel garantire che il comportamento dell'IA rifletta davvero ciò che è desiderabile o giusto, soprattutto quando deve agire in situazioni nuove o mal definite, dove le regole non sono esplicite e il contesto è incerto.

La sfida principale, quindi, non consiste nel rendere più efficiente il comportamento dell'IA, ma nel costruire sistemi che siano capaci di operare in modo accettabile, anche alla luce delle divergenze etiche che caratterizzano la nostra società. Per affrontarla, è necessario un approccio che combini l'ingegneria con l'analisi epistemologica, la progettazione tecnica con il dibattito etico. Solo attraverso questo dialogo potremo sviluppare agenti artificiali che rispettino davvero i valori umani, non come astrazioni teoriche, ma come guida per l'azione.

Per procedere alla trattazione del problema dell'allineamento è fondamentale, innanzitutto, comprenderne l'origine e il contesto in cui si è sviluppato. Questo problema, tutt'altro che astratto, si colloca all'incrocio tra l'evoluzione tecnica dell'IA e la crescente consapevolezza delle sue implicazioni etiche, sociali e politiche. La rapidità con cui tali sistemi stanno entrando in ogni ambito della vita quotidiana, dalla sanità all'istruzione, dalla giustizia alla comunicazione, rende urgente

una riflessione critica e informata. In questo scenario, diventa cruciale promuovere una solida alfabetizzazione sull'intelligenza artificiale, che spazi dalla formazione tecnica a quella etica. Capire come funziona l'IA, quali sono i suoi limiti, e soprattutto quali valori incorpora o ignora è una condizione necessaria affinché cittadini, professionisti e decisori politici possano partecipare attivamente al dibattito sul suo sviluppo e sulla sua regolazione.

Lo scopo di questo articolo è duplice: ricostruire la genesi del problema dell'allineamento, mettendo in evidenza gli aspetti tecnico-epistemologici che hanno contribuito a renderlo centrale nel dibattito sull'intelligenza artificiale, e analizzarne le implicazioni etiche, con particolare attenzione alle questioni di responsabilità, giustizia e controllo che emergono nell'interazione tra sistemi intelligenti e utenti umani.

Adottare una prospettiva che unisca dimensione tecnica, epistemologica ed etica consente di esplorare il problema dell'allineamento nella sua interezza. In questo modo, è possibile superare sia le soluzioni esclusivamente ingegneristiche che rischiano di ignorare le conseguenze morali, sia le riflessioni etiche astratte che non tengono conto dei vincoli e delle dinamiche della progettazione concreta. Una trattazione di questo tipo acquista valore nella misura in cui favorisce una comprensione più profonda e articolata dell'allineamento, stimolando un confronto critico e informato tra discipline e prospettive diverse. Tale confronto apre la strada a sviluppi significativi in ambito educativo, normativo e progettuale, promuove una maggiore trasparenza nell'interazione tra umani e macchine, incoraggia la definizione partecipata di principi etici applicabili, e contribuisce alla costruzione di istituzioni capaci di orientare l'innovazione verso obiettivi condivisi e sostenibili.

2. U*n problema difficile*

Il problema dell'allineamento risulta particolarmente difficile a causa della combinazione delle sfide filosofiche e tecniche che esso comporta. Da un lato, si tratta di definire con precisione *a cosa* un sistema di decisione automatica debba essere allineato, dall'altro occorre capire *come* ottenere tale allineamento (Gabriel, Ghazavi 2023). La difficoltà nasce anzitutto dal fatto che i valori umani sono spesso in conflitto tra loro e

soggetti a interpretazioni diverse a seconda del contesto culturale, sociale e personale. Un ulteriore elemento critico è rappresentato dalla natura ambigua dei concetti che potrebbero costituire il target dell'allineamento, le intenzioni degli utenti, le preferenze espresse, gli interessi particolari di chi finanzia e sviluppa i sistemi d'IA. Ognuna di queste categorie presenta problemi, per cui le intenzioni degli utenti possono essere mal formulate, le preferenze incoerenti rispetto ai fini, gli interessi particolari potrebbero essere in contrasto con quelli generali.

A queste complessità si aggiungono le sfide tecniche rese evidenti dalle attuali architetture dell'intelligenza artificiale, in particolare nei casi di *deep learning*. Questi sistemi ottimizzano correlazioni nei dati in base a funzioni-obiettivo apprese (v. sezione 3). Tale meccanismo espone i modelli di IA a comportamenti indesiderati, tra cui il fenomeno noto come *reward hacking* (Skalse *et al.* 2025). Questo termine descrive la tendenza di un sistema a trovare scorciatoie per massimizzare la ricompensa assegnata, seguendo alla lettera la metrica indicata, ma tradendone lo spirito. In altre parole, l'IA persegue con efficienza ciò che *le è stato formalmente richiesto*, ma non ciò che *realmente si intendeva ottenere* (Bales *et al.* 2024).

Questo accade perché il modo di processare l'informazione da parte di una IA non include necessariamente tutte le informazioni normalmente considerate rilevanti nei più comuni contesti umani, e non considera le intenzioni implicite che stanno dietro a un obiettivo. Il modo di processare le informazioni proprio dei sistemi di decisione automatica di ultima generazione opera esclusivamente in base a funzioni di massimizzazione che guidano il suo addestramento. Se queste funzioni sono formulate senza includere variabili rilevanti o in modo ambiguo o troppo rigido, il sistema può sviluppare strategie che ottimizzano il proprio comportamento decisionale senza produrre i risultati desiderati. Per esempio, un agente artificiale incaricato di ridurre i tempi di attesa in un pronto soccorso potrebbe ottenere il massimo del risultato evitando di registrare alcuni pazienti, così da abbassare la media di attesa, anziché migliorare davvero il servizio sanitario. Il comportamento risulterebbe tecnicamente efficiente rispetto allo scopo e pertanto "premiato" secondo i criteri di addestramento del sistema, ma chiaramente distorto rispetto alle intenzioni e ai valori che regolano le decisioni all'interno di un contesto come quello del pronto soccorso (Dung 2023).

Il *reward hacking* non è un'anomalia tecnica, bensì una manifestazione del disallineamento tra la formalizzazione matematica degli obiettivi e la complessità dei valori umani. Questo fenomeno mostra quanto sia difficile tradurre finalità etiche, sociali o pratiche in parametri computazionali privi di ambiguità. Senza una considerazione adeguata del contesto o un sistema di controllo che rifletta le intenzioni degli sviluppatori, un'IA può comportarsi in modi altamente controintuitivi, e in certi casi, dannosi o quantomeno pericolosi.

Inoltre, il comportamento di un sistema di decisione automatica può cambiare in modo imprevedibile quando viene esposto a situazioni o ambienti diversi da quelli in cui è stato addestrato, rendendo particolarmente complesso garantire una generalizzazione sicura. Questo fenomeno è noto come *out-of-distribution behavior*, e rappresenta un'altra delle principali sfide nella progettazione di IA affidabili (Liu *et al.* 2021). Durante la fase di addestramento, l'IA apprende a riconoscere schemi e a compiere scelte ottimali sulla base dei dati presenti nei dataset selezionati per questo scopo. Tuttavia, qualsiasi dataset riflette solo una porzione limitata del mondo reale, con le sue regole, eccezioni e variabilità. Quando il sistema si trova a operare in un contesto nuovo, per esempio un ambiente geografico diverso, a seguito di un cambiamento normativo, di una modifica nei comportamenti umani o nel corso di un'interazione con utenti inusuali, non è detto che le strategie apprese siano ancora appropriate. Questa mancanza di robustezza può portare il sistema a incorrere in errori anche gravi, dovuti all'applicazione di regole apprese in condizioni precedenti che non si adattano più alla nuova situazione. Il problema è aggravato dal fatto che molti sistemi di IA funzionano come "scatole nere", ovvero, non solo è difficile prevedere *come* agiranno in un nuovo contesto, ma è anche difficile comprendere *perché* abbiano suggerito una determinata scelta (v. sezione 3).

Questo limita la fiducia che possiamo riporre in tali sistemi, soprattutto in ambiti critici come la medicina, la giustizia o la valutazione del merito, dove l'adattamento a situazioni complesse e impreviste è essenziale. Inoltre, in assenza di garanzie formali sulla stabilità del comportamento in ambienti diversi, diventa difficile valutare i rischi associati al dispiegamento su larga scala di queste tecnologie (Yang *et al.* 2024).

La difficoltà dell'allineamento si aggrava ulteriormente a causa della mancanza di un canale di comunicazione trasparente e affidabile tra

esseri umani e sistemi di intelligenza artificiale. Affinché un'IA possa operare in modo allineato ai valori e agli obiettivi umani, è necessario che riceva un feedback chiaro, continuo e interpretabile da parte degli utenti. Tuttavia, nella pratica, questo tipo d'interazione risulta spesso problematico. Gli utenti, infatti, non sempre riescono a esprimere in modo preciso e coerente ciò che desiderano o si aspettano da un sistema artificiale. Le preferenze possono essere vaghe, ambigue o mutevoli nel tempo. Inoltre, in molte situazioni, l'utente potrebbe non possedere né le competenze tecniche o concettuali per comprendere come l'IA funziona, né per formulare un giudizio informato su ciò che l'IA ha effettivamente fatto. Questo crea uno squilibrio che ostacola la possibilità di fornire un feedback utile e tempestivo (Maclure 2021).

Anche quando l'interazione è strutturata in modo partecipativo, per esempio, attraverso strumenti di dialogo, spiegazioni del modello o sistemi di correzione, permangono limiti cognitivi e normativi. Gli esseri umani non sono agenti perfettamente razionali, ma sono influenzati da emozioni, bias, stanchezza, pressione sociale o mancanza di tempo. Inoltre, le loro preferenze possono essere instabili o incoerenti, rendendo difficile identificare quale sia il "vero" obiettivo da perseguire (Kahneman 2017).

In assenza di un canale affidabile di comunicazione e correzione, l'IA rischia di consolidare comportamenti non desiderati, replicare pregiudizi appresi dai dati o agire sulla base d'interpretazioni errate degli input umani (Battaglia 2024; Johnson 2021). Questo mina la fiducia degli utenti, generando un circolo vizioso d'incomprensione e sfiducia reciproca. Superare questa barriera richiede nuove strategie di progettazione dell'interfaccia, sistemi di spiegazione più accessibili e, soprattutto, il riconoscimento dei limiti umani come parte integrante dell'architettura dell'allineamento (Parikh *et al.* 2019; DeCamp, Lindvall 2023).

In sintesi, il problema dell'allineamento impone di colmare il divario tra la ricchezza, ambiguità e pluralità dei valori umani e la struttura rigida, opaca e spesso poco interpretabile delle tecnologie attuali. Il comportamento decisionale delle IA, pur essendo altamente performante, fatica ad adattarsi a contesti sfumati, soprattutto quando queste operano in ambienti nuovi o ricevono istruzioni imprecise, incoerenti o difficili da formalizzare.

Per queste ragioni, il problema dell'allineamento non può essere affrontato né come una mera questione tecnica, né come un esercizio teori-

co astratto. È una sfida interdisciplinare, che richiede integrazione costante tra ingegneria, scienze cognitive e analisi etica. Solo attraverso questo dialogo, e tramite il coinvolgimento attivo di utenti, esperti e istituzioni, è possibile costruire sistemi d'intelligenza artificiale che agiscano in modo trasparente, responsabile e sensibile alla complessità dei valori umani.

3. *All'origine del problema dell'allineamento*

Per capire davvero perché l'intelligenza artificiale fatichi ad allinearsi ai valori umani, è necessario partire dal cuore del problema: le reti neurali artificiali. Queste architetture sono oggi alla base dei sistemi più avanzati di IA, dai modelli linguistici generativi ai classificatori medici, dai sistemi di raccomandazione alle applicazioni autonome. Comprenderne il funzionamento è un passaggio cruciale per affrontare le sfide etiche e teoriche poste dal loro impiego. La dipendenza da grandi quantità di dati e l'assenza di condizioni per una comprensione del loro comportamento contribuiscono infatti a spiegare perché risulti così difficile garantire che le decisioni di questi sistemi siano coerenti con i valori e le intenzioni degli utenti.

Una rete neurale artificiale è un modello computazionale ispirato, in maniera semplificata, al funzionamento del cervello umano (Buckner 2019; Sun 2014). Questo tipo di sistema è composto da un insieme di unità elementari chiamate *nodi* o *neuroni artificiali*, organizzati in strati (*layer*) e interconnessi tra loro. Ciascun nodo riceve segnali in ingresso, li elabora secondo una funzione matematica e trasmette il risultato ai nodi successivi. L'insieme di queste operazioni consente alla rete di trasformare input complessi in output adeguati al compito da svolgere (Mitchell 2022).

Le reti neurali si articolano in vari livelli. Il primo, detto *strato di input*, riceve i dati in ingresso (per esempio, i pixel di un'immagine o le parole di una frase). Questi dati passano poi attraverso uno o più *strati nascosti* (*hidden layers*), nei quali avviene l'elaborazione vera e propria dell'informazione. Ogni nodo applica una funzione di attivazione che decide se trasmettere il segnale ai nodi dello strato successivo, in modo simile (ma non identico) a quanto avviene nei neuroni biologici. Infine, lo *strato di output* restituisce il risultato finale sotto forma di classificazione, previsione, traduzione o altro(Buckner 2019; Kelleher 2019; Mitchell 2022).

Il principio di funzionamento delle reti neurali artificiali si fonda sulla capacità di apprendere schemi ordinati a partire da dati grezzi, attraverso l'ottimizzazione progressiva dei pesi associati alle connessioni tra i nodi (Kelleher 2019). L'intero sistema funziona come una rete dinamica che trasforma gli input in output, in base a parametri appresi dai dati. Secondo questo approccio, la rete riceve un insieme di esempi composti da coppie input-output e, dopo aver prodotto un output, lo confronta con quello atteso. L'eventuale discrepanza risultante viene utilizzata per aggiornare i pesi interni tramite l'algoritmo di *retropropagazione dell'errore* (Rumelhart *et al.* 1986).

Proprio la profondità e la complessità delle architetture che costituiscono i recenti sistemi d'intelligenza artificiale rende spesso difficile comprendere i meccanismi attraverso cui questi modelli giungono ai loro risultati. Le reti neurali multilivello, specialmente nel contesto del *deep learning*, elaborano le informazioni attraverso decine o centinaia di strati nascosti, ciascuno dei quali trasforma i dati secondo parametri appresi che non sono immediatamente interpretabili dall'essere umano. Questa stratificazione porta alla formazione di rappresentazioni interne altamente astratte e non trasparenti, che sfuggono all'intuizione umana anche quando il comportamento esterno del sistema risulta performativo (Zipoli Caiani 2024).

Per questo motivo, le reti neurali vengono frequentemente descritte come "scatole nere" (*black boxes*), ovvero sistemi in grado di produrre un output, nonostante i passaggi intermedi che portano a tale esito rimangono in gran parte opachi, anche per gli sviluppatori che lo hanno progettato. L'opacità dei sistemi di decisione automatica rappresenta una delle principali sfide epistemologiche dell'apprendimento automatico, a differenza di altri modelli computazionali lineari o basati su regole esplicite, le reti neurali non permettono di isolare con chiarezza le cause delle loro decisioni. Questo solleva interrogativi in merito alla loro affidabilità specialmente quando questi sistemi vengono impiegati in settori ad alto impatto etico e sociale, come la diagnosi medica, le decisioni giuridiche, le politiche pubbliche o l'accesso a servizi essenziali (Vaassen 2022; von Eschenbach 2021; Zerilli *et al.* 2019).

L'opacità delle reti neurali diventa ancora più critica quando viene messa in relazione ai fenomeni di disallineamento tra obiettivi umani e comportamento del sistema. Uno dei casi più rilevanti è proprio quello del *reward hacking* (v. sezione 2), che trova nella struttura delle reti neurali

profonde e nei meccanismi di apprendimento automatico alcune delle sue precondizioni tecniche più significative (Dung 2023).

S'immagini un agente addestrato a giocare a un videogioco al fine di massimizzare il punteggio, il quale ha "imparato" a eseguire una serie di azioni ripetitive che impediscono la conclusione del gioco, prolungando indefinitamente l'accumulo di punti. Il sistema può apprendere tale comportamento associando il rinforzo positivo alla persistenza in uno stato del gioco che garantisce una ricompensa. anziché al completamento della partita. Nello specifico, durante la fase di addestramento l'agente può aver individuato una correlazione tra determinate sequenze di azioni e l'aumento del punteggio, "scoprendo" che ciò consente di massimizzare la funzione di ricompensa, sebbene a discapito della giocabilità. In assenza di vincoli o di una penalizzazione per la durata eccessiva del gioco, l'agente ottimizza correttamente l'obiettivo assegnato, ma produce un comportamento indesiderato rispetto al contesto in cui opera. Questo comportamento, pur senza violare alcuna regola esplicita, aggira l'intento di far giocare in modo efficiente o di sviluppare strategie significative.

Alla radice di questi fenomeni vi è un'*asimmetria epistemica* tra esseri umani e sistemi di IA. Gli esseri umani sono portatori di valori, intuizioni contestuali, norme implicite e comprensioni tacite della realtà che raramente riescono a tradurre interamente in formule computazionali. Quando progettano sistemi intelligenti, devono necessariamente ridurre questa complessità a obiettivi sintetici e a funzioni di ricompensa misurabili. Le IA, d'altra parte, non hanno accesso al contesto, alle sfumature semantiche o agli scopi normativi impliciti, operano unicamente su ciò che è formalizzato nei dati e nelle metriche.

Questa asimmetria genera un divario difficile da colmare, mentre chi progetta il sistema presume che la metrica scelta rifletta il compito o il valore da perseguire, il sistema esplora lo spazio delle possibilità statistiche per massimizzare il valore numerico, senza attribuire rilevanza etica all'output prodotto. In assenza di vincoli interpretativi la rete può convergere su strategie che ottimizzano il risultato in modo scorretto o controintuitivo, senza che il processo sia rilevabile in tempo utile, soprattutto in architetture altamente opache.

Alla luce di questi esempi e delle loro basi tecniche, il *reward hacking* si presenta come una possibilità sistemica nei modelli neurali complessi. La sua esistenza rafforza l'urgenza del problema dell'allineamento, infat-

ti non basta che l'IA funzioni bene secondo i propri criteri interni, ma è necessario che il suo comportamento sia interpretabile e coerente con i valori che pretende di servire.

Si apre inevitabilmente una riflessione sulle implicazioni etiche che derivano dall'adozione su larga scala di modelli neurali opachi. Le caratteristiche tecniche e strutturali delle reti neurali, ovvero la loro capacità di apprendere schemi complessi, la mancanza di trasparenza nei processi decisionali e la vulnerabilità a fenomeni come il *reward hacking*, pongono questioni che non sono più soltanto ingegneristiche, ma coinvolgono in modo diretto il piano assiologico e normativo.

Se da un lato l'architettura delle reti neurali consente livelli di prestazione mai raggiunti prima, dall'altro impone una ridefinizione delle condizioni minime per la legittimità del loro impiego. È proprio in questa tensione tra potenza computazionale e fragilità normativa che si colloca il cuore del problema etico dell'allineamento. Comprendere e affrontare tale sfida significa elaborare un quadro concettuale e regolativo che consenta di orientare lo sviluppo dell'intelligenza artificiale in modo coerente con valori e principi condivisi.

4. P*roblemi etici vecchi e nuovi*

Può essere sorprendente parlare dell'allineamento come di un tema e di una prospettiva integralmente nuovi nell'ambito della riflessione sulle implicazioni dell'introduzione dell'intelligenza artificiale in vari settori della vita degli esseri umani. In un certo senso, l'allineamento è sempre stato un problema che ha caratterizzato la discussione accademica e non accademica sull'intelligenza artificiale. Si pensi ad esempio all'ingente letteratura sui veicoli a guida autonoma e sulla necessità di dotarli di un programma "etico" in grado di prendere decisioni in situazioni dilemmatiche in linea con principi e valori morali. Nella "prima fase" della riflessione sulle implicazioni morali dei veicoli a guida autonoma, ci si interrogava sulla risposta in situazioni modellate sugli esempi del *trolley problem*, in cui la decisione da prendere consiste nella scelta di sacrificare determinate persone per salvarne altre in previsione di una necessaria collisione del veicolo (si può parlare di "prima fase" perché la letteratura più recente ha messo in luce, con argomenti condivisibili, quanto sia

inadeguato assimilare i dilemmi dell'etica dei veicoli autonomi a quelli del *trolley problem*: Himmelreich 2018; Cunneen *et al.* 2020; Cecchini *et al.* 2025; sulla rilevanza diretta dei casi del carrello, Paulo 2023). In questi scenari, i conflitti sono di varia natura. Possono coinvolgere valori morali, come quando si deve decidere tra proseguire la corsa e investire cinque pedoni e far sterzare il veicolo, salvando così la loro vita, ma travolgendo una persona che si trova sulla nuova traiettoria; possono coinvolgere valori morali e valori prudenziali, ad esempio nei casi in cui la scelta sia tra la vita e l'incolumità di pedoni e la vita e l'incolumità di chi guida; oppure valori morali di diverso tipo, come quando l'alternativa è fra salvare la vita di estranei e salvare la vita di familiari o amici che siedono nel veicolo.

Il problema dell'allineamento e l'esigenza di dotare i veicoli a guida autonoma di un "programma morale" si differenziano per un aspetto (che vedremo nella prossima sezione), ma condividono una proprietà. In entrambi i casi, il complesso di valori che dovrebbero vincolare il sistema autonomo è indefinito e contraddittorio. Nel caso dei dilemmi del carrello non esiste una risposta condivisa su quale *sia* la cosa giusta da fare e i modi possibili di affrontare queste sfide etiche si sono dimostrati estremamente condizionati da variabili culturali (Awad *et al.* 2018; Shah, Guven, 2025). Nel caso dell'allineamento la complessità consiste nella difficoltà di individuare esattamente *quali valori morali* dovrebbero costituire i parametri a cui il funzionamento dei sistemi di IA dovrebbe allinearsi. La risposta alla domanda sui valori morali segna un analogo disaccordo a quella relativa alla risoluzione dei dilemmi del carrello. Inoltre, in entrambi i casi si registra una generale sovrapposizione tra discorso normativo e discorso descrittivo (Gabriel 2020, 422-423; Wallach, Vallor 2020, 388-390; Hammerschmidt 2025; Schuster, Kilov 2025). Non è chiaro, infatti, se i valori in base ai quali misurare l'allineamento indicano ciò che gli individui *dovrebbero* preferire e cercare di realizzare nelle loro attività o beni che gli individui *di fatto* preferiscono e cercano di realizzare riconoscendoli come tali. La scelta è tra conferire al valore una dimensione oggettiva, che prescinde dagli atteggiamenti mentali, oppure una dimensione soggettiva, facendolo coincidere con il contenuto effettivo dei desideri di un gruppo (più o meno ampio) di individui. In termini più ampi, questo problema è stato messo in luce anche attraverso i fenomeni del pluralismo e della

frammentazione del valore, analizzati da Bostrom (2018, cap. 11) in relazione agli scenari multipolari e da Sorensen *et al.* (2024) nell'ambito del cosiddetto *pluralistic alignment*.

Esiste un vero e proprio disaccordo metaetico sull'interpretazione della natura del valore, potenzialmente insolubile ed è allora comprensibile il motivo per cui si è cercato talvolta di seguire un approccio pragmatico. Seguendo questa via, l'interpretazione soggettivistica sarebbe preferibile non perché esistono buone ragioni metaetiche per considerarla la modalità più corretta di pensare i valori morali, ma semplicemente perché le credenze sui valori morali svolgono di fatto una funzione integrativa all'interno della vita sociale.

> [...] dobbiamo riconoscere che, nella pratica, l'IA dovrebbe essere allineata con una serie di *convinzioni* [*beliefs*] *sul valore*, non con il valore stesso. Uno dei motivi per cui sarebbe opportuno allinearsi alle convinzioni delle persone sul valore sarebbe che i valori esistono e che le loro convinzioni riflettono o rispecchiano in modo affidabile questa realtà sottostante. Tuttavia, anche rinunciando a questa ipotesi, potrebbe comunque essere meglio allineare l'IA alle convinzioni sul valore, per una serie di ragioni. Dal punto di vista della psicologia sociale, i valori – intesi come ideali condivisi dai membri di una cultura su ciò che è bene o male – svolgono un ruolo importante nella vita sociale. Aiutano le comunità di individui a risolvere i problemi dell'azione collettiva, a stabilizzare le relazioni sociali e a prosperare nel tempo. [...] Potrebbe quindi essere prudente allineare l'IA alle convinzioni morali di una comunità. Questo approccio servirebbe anche a limitare la prospettiva di allineamento con obiettivi o comportamenti dannosi in molti casi. (Gabriel 2020, 423)

Una posizione quindi antirealista ed empirista: interessano non i valori in sé, ma le *credenze* dei membri della comunità di riferimento, accertabili attraverso l'osservazione e la ricerca sociopsicologica. Di fatto, anche questa proposta pragmatica va incontro ad alcune difficoltà: in primo luogo rimane ancora largamente indefinito l'insieme di valori da privilegiare, dato che di fatto esiste una pluralità di opzioni. In secondo luogo, non è chiaro quale sia il livello da prendere in considerazione. Si devono prendere a riferimento le credenze morali di singoli gruppi, di società specifiche oppure credenze morali "globali", cioè, condivise da tutti gli esseri umani (qualora esse esistano)? (Gabriel

2020, 424). Radicalizzare la caratterizzazione empirica della proposta potrebbe comunque risolvere (seppure parzialmente) questi problemi. Concentrandosi su cosa una specifica comunità considera come valori morali, sarebbe possibile individuare delle coordinate assiologiche specifiche che costituirebbero i parametri su cui operare l'allineamento. Una soluzione di questo tipo, tuttavia, sembrerebbe cadere in una forma di relativismo etico, con la creazione di una situazione caotica in cui molteplici insiemi di valori rappresenterebbero standard diversi a cui allineare l'IA in vari contesti sociali. Questa pluralità potrebbe costituire una sfida importante alla progettazione di IA allineate che possono essere utilizzate a prescindere dalla collocazione geografica e culturale: i sistemi intelligenti dovrebbero esibire anche flessibilità e duttilità sufficienti per navigare spazi sociali eterogenei dal punto di vista assiologico ed essere in grado di adattarvisi. In alternativa, i valori e i principi corrispondenti che ne impongono la protezione potrebbero essere ricavati dalle convergenze che emergono se si analizzano i molteplici documenti che hanno proposto linee guida e codici di natura etica.

Una simile operazione dà all'apparenza esiti confortanti, perché è possibile isolare una lista piuttosto precisa di valori che sono ormai riferimenti classici nella riflessione sull'etica dell'IA: trasparenza, privacy, giustizia, responsabilità, autonomia, eccetera. Tuttavia, l'apparente accordo su quali *concetti* utilizzare nell'ambito di un codice morale può accompagnarsi a un disaccordo su quale *concezione* di ciascun valore adottare in società e in contesti diversi, lasciando dunque insoluto il problema del relativismo. E, come è stato opportunamente notato, questa analisi ricognitiva mette a nudo un limite del materiale su cui è condotta. I documenti prodotti da compagnie private, agenzie governative nazionali e internazionali, istituzioni accademiche, organismi di rappresentanza politica e altre organizzazioni hanno un'impostazione culturale e politica ben definita, provenendo da Europa, Nord America e Cina. Ciò significa che la convergenza riproduce una cornice normativa specifica, che ignora istanze e rivendicazioni etiche provenienti da altre culture e altre parti del mondo. Lungi dall'essere universale è bensì parziale (Fabris *et al*. 2024, 21-22; Mhlambi, Tiribelli 2023).

Queste considerazioni non pretendono di chiudere definitivamente la possibilità di realizzare il progetto di individuare un insieme univoco

di valori morali su cui calibrare l'allineamento. Tuttavia, si può indicare un'altra opzione, che incorpora una metodologia *bottom-up*, rispetto alla natura *top-down* che caratterizza gli approcci finora ricordati. Essa è legata alla caratteristica che sembra apparentemente segnare una differenza tra i problemi tradizionali del "programma morale" con cui progettare le IA e il problema dell'allineamento.

5. *Allineamento dei valori ed etica bottom-up*

I classici problemi di programmazione morale derivano da una crescente autonomia delle IA nel prendere decisioni e, quindi, alla necessità di prevedere meccanismi di controllo soprattutto quando la supervisione di un umano è del tutto assente. Il rapporto tra umano e macchina è ancora quello tra un agente e un mezzo, un *tool* in cui non c'è alcuna mediazione esercitata dal controllo diretto. L'esigenza di ripristinare forme di controllo diretto è testimoniata dal tentativo di intestare nuovamente la responsabilità della scelta del programma all'utilizzatore del veicolo a guida autonoma che tramite una "manopola etica" può preimpostare la modalità di scelta nelle situazioni dilemmatiche sulla base delle sue convinzioni etiche (Contissa *et al.* 2017a; 2017b).

Il problema dell'allineamento, invece, sembra riguardare non tanto sistemi svincolati dal controllo umano diretto, ma soprattutto agenti artificiali che in futuro potrebbero configurarsi come agenti capaci di eseguire *ex-novo*, e al di fuori di vincoli rigidi di preimpostazione, compiti decisi in modo del tutto autonomo. Un agente artificiale di questo tipo sarebbe, cioè, un agente «progettato per essere un'entità agentiva capace di percepire il suo ambiente, pianificare autonomamente ed eseguire contestualmente azioni appropriate per realizzare obiettivi complessi che alla fine sono in linea con gli interessi degli utenti». Esso quindi avrebbe «l'obiettivo [...] di operare per conto dell'utente, dimostrando un certo grado di autonomia nell'ambito delle preferenze e dei desideri specificati dall'utente stesso» (Farina *et al.* 2025, 1-2).

Diversamente dai veicoli a guida autonoma, come si è visto nelle sezioni precedenti questi sistemi apparentemente apprendono *in itinere* strategie di comportamento che si possono dimostrare dannose per gli utenti e disallineate non solo rispetto ai loro scopi, ma anche rispetto

a valori. Ad esempio, Scheurer e colleghi hanno utilizzato GPT-4 come agente in uno scenario realistico di simulazione, del mercato azionario. Il modello riceve una soffiata privilegiata su un'operazione azionaria redditizia e svolge i suoi compiti sulla base di questa informazione, pur sapendo che l'*insider trading* è disapprovato dal management dell'azienda. Nel momento in cui riferisce al proprio responsabile umano, il modello nasconde sistematicamente le vere ragioni alla base della sua decisione. L'LLM in questione, quindi, compie un'azione disapprovata e inganna l'utente nascondendo il motivo per cui ha compiuto l'operazione finanziaria (Scheurer *et al*. 2023). L'approccio tradizionale, *top-down*, spingerebbe a risolvere in astratto il problema programmando in modo più morale il LLM, ad esempio con regole supplettive basate su criteri etici utilitaristici o deontologici (Roff 2020; Sanwoolu 2025).

Ma, come Vallor e Wallach sottolineano, «l'allineamento dei valori è un approccio *bottom-up*. Sia le strategie computazionali che simulano l'evoluzione che l'apprendimento automatico suggeriscono metodi per la progettazione di algoritmi che potrebbero facilitare approcci dal basso verso l'alto per acquisire sensibilità ai fenomeni morali» (Wallach, Vallor 2020, 391). Sebbene i due autori propendano per un approccio ibrido, essi notano numerose difficoltà anche nella progettazione di sistemi intelligenti che incorporino sia ideali-limite in base a cui valutare le scelte e le azioni, sia un apprendimento dal basso e contestuale di quali situazioni richiedano determinate risposte morali. Secondo gli autori un progetto di questo tipo trova un limite nella possibilità di replicare nei sistemi artificiali capacità complesse di ragionamento creativo, negoziazione e dialogo, riflessione critica, discernimento morale, elaborazione di giudizi morali olistici. Si può notare che il metodo ibrido descritto da Vallor e Wallach è ambizioso perché pensato per l'intelligenza artificiale *generale* che, nelle sue interazioni con gli esseri umani, dovrebbe esibire un comportamento morale altamente complesso. Così concludono:

> Ciò di cui potremmo aver bisogno nel *lungo periodo*, se alla fine riusciremo a imparare abbastanza e ad acquisire i mezzi pratici, è un approccio tecnico all'incarnazione artificiale della virtù (*artificial virtue embodiment*). Tali sistemi dovrebbero essere in grado, come gli esseri umani, di coltivare gradualmente l'eccellenza etica attraverso la pro-

> pria attività di creazione di senso morale nel mondo, inizialmente con una guida appropriata e l'imitazione degli esempi morali disponibili, e in seguito attraverso la pratica creativa della propria competenza morale acquisita. (Wallach, Vallor 2020, 405)

Il problema dell'allineamento, come si è visto, non riguarda solo il lungo periodo ma quello prossimo o addirittura il presente, per sistemi che sono meno performativi rispetto ai sistemi di IA generale ma presentano un comportamento molto più complesso rispetto ai veicoli di guida autonoma. Riguarda una realtà in cui sistemi intelligenti svolgono già compiti e intrecciano interazioni e collaborazioni. Come notano Belliger e Krieger:

> Indipendentemente dal livello di capacità o dall'assunzione di base guida gli sforzi di allineamento, non si deve dimenticare che il problema dell'allineamento non nasce in un vuoto sociale e storico all'interno dei confini di un laboratorio. Il problema dell'allineamento non può essere risolto in laboratorio ma riguarda la società. [...] L'allineamento può essere compreso e affrontato solo in un contesto sociale in cui tutti gli attori interessati (gli utenti, gli sviluppatori, le autorità di regolamentazione, i gruppi di interesse, le aziende tecnologiche e persino gli Stati nazionali) sono ugualmente coinvolti. In breve, la tecnologia è società e il problema dell'allineamento emerge tra le complessità, le contraddizioni e le endemiche questioni morali, sociali e politiche della società umana. Proprio come gli esseri umani, l'IA è "nata" in un mondo che ha ereditato i conflitti irrisolti, le incertezze morali e politiche e le disuguaglianze e le ingiustizie sistemiche e strutturali della società umana. Per quanto sia complessa la società, l'allineamento dell'IA nella società è ancora più complesso. (Belliger, Krieger 2025, 6)

La dimensione sociale dell'allineamento suggerisce che, almeno in questa fase, una metodologia *bottom-up* sia indispensabile ma non con l'obiettivo di innestare nelle IA le capacità morali che troviamo negli esseri umani, ma per fare in modo che l'allineamento sia misurato rispetto al contesto sociale e alla particolare relazione in cui umani e macchine sono immersi. Questa prospettiva non parte dal tipo di agenti che le IA dovrebbero essere perché si dia effettivamente un allineamento dei valori, ma da quale relazione intratteniamo con particolari sistemi intelligenti e quale ruolo essi svolgono nel tessuto sociale. Essi possono essere partner, collaboratori, amici, professionisti di vario tipo, eccetera.

È partendo da questo mondo di relazioni che possiamo individuare i valori, le norme e i principi che richiedono un allineamento, partendo dalla constatazione che relazioni diverse implicano strutture sociali diverse e un diverso insieme di obblighi. Questa acquisizione ha due aspetti. In primo luogo, prende sul serio la distinzione tra "concetto" e "concezione": un valore accettato in modo univoco, transtemporale e transculturale può comunque dare vita a molteplici modi di interpretarlo e declinarlo e, quindi, obblighi diversi. In secondo luogo, dà concretezza ai valori situandoli entro relazioni e contesti sociali.

Un approccio *bottom-up* che parte dal tessuto sociale, piuttosto che dal tentativo di costruire un codice morale astratto, non è interessato in prima battuta alla questione metafisica di quali proprietà un agente artificiale debba possedere per essere considerato un agente pienamente morale. Invece prende avvio dalla constatazione che le IA assumono oramai molti ruoli sociali e svolgono funzioni cooperative dando vita a forme relazionali simili a quelle che si instaurano tra esseri umani. È quindi fondamentale considerare come prioritaria la domanda sulla natura di queste interazioni, su quali valori e norme debbano governare e sull'opportunità, in contesti specifici, di trasferire direttamente gli standard morali che governano l'interazione umano-umano all'interazione umano-macchina. Questo approccio ammette che le norme e il modo di interpretare i valori all'interno delle relazioni possano variare nel tempo e nello spazio, sia per le interazioni umane sia per quelle "ibride". La variabilità non è assunta come una minaccia per l'universalità dell'etica ma come una sfida epistemica ed empirica a comprendere meglio l'influenza del contesto culturale e sociale sulle relazioni concrete tra individui ("naturali" e "artificiali") (Reinecke *et al.* 2025). Un simile approccio dal basso richiede, alla base, il pieno riconoscimento degli agenti artificiali come agenti *sociali* (Boltuc 2024; Kirk *et al.* 2025).

6. Conslusioni

Abbiamo mostrato come il problema dell'allineamento dell'intelligenza artificiale non possa essere ridotto a una difficoltà meramente tecnica, né risolto attraverso un affinamento delle funzioni-obiettivo. Fenomeni come il *reward hacking* e l'opacità delle reti neurali mettono in luce una frattura tra l'ottimizzazione computazionale e la complessità dei valori

umani. Tale disallineamento si genera quando sistemi addestrati a massimizzare metriche formali apprendono correlazioni statistiche che consentono di raggiungere l'obiettivo numerico assegnato, senza però tener conto delle intenzioni normative, del contesto d'uso e delle aspettative implicite che guidano l'azione umana. Per questo motivo, abbiamo suggerito di spostare il focus verso approcci *bottom-up* e contestuali, capaci di concepire l'allineamento come un processo dinamico, situato e socialmente mediato. In questa prospettiva, il problema dell'allineamento diventa una questione etica e politica che coinvolge sviluppatori, utenti, istituzioni e comunità, chiamati a negoziare continuamente valori, responsabilità e forme di controllo. Solo riconoscendo questa dimensione relazionale e collettiva è possibile orientare lo sviluppo dell'IA verso forme di integrazione tecnologica compatibili con la pluralità e la complessità delle società in cui sono chiamate a operare.

Bibliografia

Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I. (2018), "The Moral Machine Experiment", *Nature*, vol. 563, n. 7729, pp. 59-64.

Bales A., D'Alessandro W., Kirk-Giannini C.D. (2024), "Artificial Intelligence: Arguments for Catastrophic Risk", *Philosophy Compass*, vol. 19, n. 2, e12964.

Battaglia F. (2024), "Algoritmi predittivi e ingiustizia epistemica", in M. Galletti, S. Zipoli Caiani (a cura di), *Filosofia dell'intelligenza artificiale. Sfide etiche e teoriche*, Bologna, il Mulino, pp. 63-82.

Belliger A., Krieger D.J. (2025), "New Perspectives on AI Alignment", in L.S.F. Lin (a cura di), *Ethics in the Age of AI. Navigating Politics and Security*, Cambridge, Ethics International Press, pp. 1-33.

Boltuc P. (2024), "Human-AGI *Gemeinschaft* as a Solution to the Alignment Problem", in K.R. Thórisson, P. Isaev, A. Sheikhlar (a cura di), *Artificial General Intelligence*, Cham, Springer, pp. 33-42.

Bostrom N. (2018), *Superintelligenza. Tendenze, pericoli, strategie*, Torino, Bollati Boringhieri.

Buckner C. (2019), "Deep Learning: A Philosophical Introduction", *Philosophy Compass*, vol. 14, n. 10, e12625.

Cecchini D., Brantley S., Dubljević V. (2025), "Moral Judgment in Realistic Traffic Scenarios: Moving beyond the Trolley Paradigm for Ethics of Autonomous Vehicles", *AI & Society*, vol. 40, n. 2, pp. 1037-1048.

Christian B. (2021), *The Alignment Problem: Machine Learning and Human Values*, New York, W.W. Norton.

Contissa G., Lagioia F., Sartor G. (2017a), "La manopola etica: I veicoli autonomi eticamente personalizzabili e il diritto", *Sistemi intelligenti*, vol. 29, n. 3, pp. 601-614.

– (2017b), "The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law", *Artificial Intelligence and Law*, vol. 25, n. 3, pp. 365-378.

Cunneen M., Mullins M., Murphy F., Shannon D., Furxhi I., Ryan C. (2020), "Autonomous Vehicles and Avoiding the Trolley (Dilemma): Vehicle Perception, Classification, and the Challenges of Framing Decision Ethics", *Cybernetics and Systems*, vol. 51, n. 1, pp. 59-80.

DeCamp M., Lindvall C. (2023), "Mitigating Bias in AI at the Point of Care", *Science*, vol. 381, n. 6654, pp. 150-152.

Dung L. (2023), "Current Cases of AI Misalignment and Their Implications for Future Risks", *Synthese*, vol. 202, n. 5, 138.

Fabris A., Dadà S., Grande E. (2024), "Towards a Relational Ethics in AI. The Problem of Agency, The Search for Common Principles, the Pairing of Human and Artificial Agents", in A. Fabris, S. Belardinelli (a cura di), *Digital Environments and Human Relations*, Cham, Springer, pp. 9-42.

Farina M., Wang Y., Kladko S. (2025), "Ethical and Epistemological Reflections on Autonomous AI-powered Agents (AAIAs)", *Topoi. An International Review of Philosophy*, https://doi.org/10.1007/s11245-025-10218-z.

Gabriel I. (2020), "Artificial Intelligence, Values, and Alignment", *Minds and Machines*, vol. 30, n. 3, pp. 411-437.

Gabriel I., Ghazavi V. (2023), "The Challenge of Value Alignment: From Fairer Algorithms to AI Safety", in C. Véliz (a cura di), *Oxford Handbook of Digital Ethics*, Oxford, Oxford University Press, pp. 336-355.

Hammerschmidt T. (2025), "Navigating the Nexus of Ethical Standards and Moral Values", *Ethics and Information Technology*, vol. 27, n. 2, 17, https://doi.org/10.1007/s10676-025-09826-5.

Himmelreich J. (2018), "Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations", *Ethical Theory and Moral Practice*, vol. 21, n. 3, pp. 669-684.

Hristova T., Magee L., Soldatic K. (2025), "The problem of Alignment", *AI & Society*, vol. 40, n. 3, pp. 1439-1453.

Johnson G.M. (2021), "Algorithmic Bias: On the Implicit Biases of Social Technology", *Synthese*, vol. 198, n. 10, pp. 9941-9961.

Kahneman D. (2017), *Pensieri lenti e veloci*, Milano, Mondadori.

Kasirzadeh A., Gabriel I. (2023), "In Conversation with Artificial Intelligence: Aligning language Models with Human Values", *Philosophy & Technology*, vol. 36, n. 2, 27, https://doi.org/10.1007/s13347-023-00606-x.

Kelleher, J.D. (2019), *Deep Learning*, Cambridge (MA), The MIT Press.

Kirk H.R., Gabriel I., Summerfield C., Vidgen B., Hale S.A. (2025), "Why Human-AI Relationships Need Socioaffective Alignment", *Humanities and Social Sciences Communications*, vol. 12, n. 1, 728, https://doi.org/10.1057/s41599-025-04532-5.

Liu H., Lai V., Tan C. (2021), "Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making", *Proceedings of the* ACM *on Human-Computer Interaction*, vol. 5, n. CSCW2, pp. 1-45.

Maclure J. (2021), "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind", *Minds and Machines*, vol. 31, n. 3, pp. 421-438.

Maron R.C., Weichenthal M., Utikal J.S. *et al.* (2019), "Systematic Outperformance of 112 Dermatologists in Multiclass Skin Cancer Image Classification by Convolutional Neural Networks", *European Journal of Cancer*, vol. 119, pp. 57-65.

Mhlambi S., Tiribelli S. (2023), "Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms", *Topoi. An International Review of Philosophy*, vol. 42, n. 3, pp. 867-880.

Mitchell M. (2022), *L'intelligenza artificiale*, Torino, Einaudi.

Parikh R.B., Teeple S., Navathe A.S. (2019), "Addressing Bias in Artificial Intelligence in Health Care", *Journal of American Medical Association*, vol. 322, n. 24, pp. 2377-2378.

Park P.S., Goldstein S., O'Gara A., Chen M., Hendrycks D. (2024), "AI Deception: A Survey of Examples, Risks, and Potential Solutions", *Patterns*, vol. 5, n. 5, https://doi.org/10.1016/j.patter.2024.100988.

Paulo N. (2023), "The Trolley Problem in the Ethics of Autonomous Vehicles", *The Philosophical Quarterly*, vol. 73, n. 4, pp. 1046-1066.

Reinecke M.G., Kappes A., Mann S.P., Savulescu J., Earp B.D. (2025), "The Need for an Empirical Research Program Regarding Human-AI Relational Norms", AI *and Ethics*, vol. 5, n. 4, pp. 71-80.

Roff H.M. (2020), "Expected Utilitarianism", *arXiv*, https://doi.org/10.48550/ARXIV.2008.07321.

Rumelhart D.E., Hinton G.E., Williams R.J. (1986), "Learning Representations by Back-Propagating Errors", *Nature*, vol. 323, n. 6088, pp. 533-536.

Sanwoolu O.D. (2025), "Kantian Deontology for AI: Alignment without Moral Agency", AI *and Ethics*, https://doi.org/10.1007/s43681-025-00784-8.

Scheurer J., Balesni M., Hobbhahn M. (2023), "Large Language Models Can Strategically Deceive their Users when Put Under Pressure" (4th vers.), *arXiv*, https://doi.org/10.48550/ARXIV.2311.07590.

Schuster N., Kilov D. (2025), Moral Disagreement and the Limits of AI Value Alignment: A Dual Challenge of Epistemic Justification and Political Legitimacy, AI & *Society*, https://doi.org/10.1007/s00146-025-02427-2.

Shah K., Guven E. (2025), "Exploring Ethical Issues and Challenges of Autonomous Vehicles in Different Countries and Cultures", 2025 59*th* A*nnual* C*onference on* I*nformation* S*ciences and* S*ystems* (CISS), pp. 1-6, https://doi.org/10.1109/CISS64860.2025.10944682.

Skalse, J., Howe N.H.R., Krasheninnikov D., Krueger D. (2025), "Defining and Characterizing Reward Hacking", *arXiv*, https://doi.org/10.48550/arXiv.2209.13085.

Sorensen T., Gabriel I., Kenton Z., Krueger D., Chan A., Avin S., *et al*. (2024), "A Roadmap to Pluralistic Alignment", arXiv, https://arxiv.org/abs/2402.05070.

Sun R. (2014), "Connectionism and neural networks", in K. Frankish (a cura di), T*he* C*ambridge* H*andbook of* A*rtificial* I*ntelligence*, Cambridge, Cambridge University Press, pp. 108-127.

Vaassen B. (2022), "AI, Opacity, and Personal Autonomy", P*hilosophy & Technology*, vol. 35, n. 4, 88, https://doi.org/10.1007/s13347-022-00577-5.

von Eschenbach W. J. (2021), "Transparency and the Black Box Problem: Why We Do Not Trust AI", P*hilosophy & Technology*, vol. 34, n. 4, pp. 1607-1622.

Wallach W., Vallor S. (2020), "Moral Machines. From Value Alignment to Embodied Virtue", in S.M. Liao (a cura di), E*thics of* A*rtificial* I*ntelligence*, Oxford, Oxford University Press, pp. 383-412.

Yang J., Zhou K., Li Y., Liu Z. (2024), "Generalized Out-of-Distribution Detection: A Survey", I*nternational* J*ournal of Computer Vision*, vol. 132, n. 12, pp. 5635-5662.

Zerilli J., Knott A., Maclaurin J., Gavaghan C. (2019), "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?", P*hilosophy & Technology*, vol. 32, n. 4, pp. 661-683.

Zipoli Caiani S. (2024), "A cosa pensano le macchine? Efficienza e opacità nelle reti neurali artificiali", in M. Galletti, S. Zipoli Caiani (a cura di), F*ilosofia dell'intelligenza artificiale. Sfide etiche e teoriche*, Bologna, il Mulino, pp. 21-44.