Roberto Redaelli

Dall'intenzionalità alla preterintenzionalità
Una riflessione sui sistemi intelligenti

Abstract: In the debate about the moral standing of artificial intelligence, the question of whether or not technical artefacts have some form of intentionality that would allow them to be considered part of the moral world (or excluded from it) plays a crucial role. The problem of intentionality in fact constitutes a cornerstone in the construction of the notion of moral agency, as demonstrated, among other things, by the fact that technological objects have so far been excluded from the field of ethics because they lack intentionality. In contrast to this exclusion, there is now a trend in the opposite direction, which increasingly attributes some form of intentionality and moral status to AI systems. This paper addresses the clarification of the notion of technological intentionality in artificial intelligence, and especially in generative AI. To this end, it analyses a number of paradigmatic positions in this debate, highlighting their merits and shortcomings. Finally, it proposes replacing the notion of intentionality with the notion of preter-intentionality, which better expresses – according to the thesis put forward – the human-artificial intelligence relationship.

1. Il problema dello statuto morale dell'intelligenza artificiale

La diffusione di macchine dotate di intelligenza artificiale sta ridefinendo radicalmente il nostro rapporto con la tecnologia: dalla robotica assistenziale agli strumenti informatici di supporto ai processi decisionali ampiamente impiegati in campo medico e giuridico fino ai veicoli dotati di diversi livelli di autonomia, tali dispositivi assumono oggigiorno un ruolo sempre più centrale sia nella nostra vita privata sia in quella professionale. Difatti, tali sistemi intelligenti si prendono cura di noi, ci indirizzano nelle nostre scelte e facilitano l'assolvimento dei numerosi compiti a cui siamo chiamati nella nostra quotidianità.

Dinanzi alle prestazioni offerte da queste macchine, si sollevano oggigiorno questioni sempre più urgenti legate alle ricadute etiche dell'impiego pervasivo di tali sistemi nella nostra società: i robot possono sostituire l'essere umano nella cura degli anziani o nell'educazione dei più giovani[1]? Quali mansioni possono svolgere?

1 L'impiego di robot nella cura degli anziani solleva una serie di questioni etiche di cui

Mechane, n. 7, 2024 • Mimesis Edizioni, Milano-Udine Web: mimesisjournals.com/ojs/index.php/mechane • ISBN: 9791222317496 • ISSN: 2784-9961

Se le decisioni prese sulla base delle indicazioni offerte dai cosiddetti *decision support systems* sono efficienti, ad esempio, nel campo del *business*[2], allo stesso modo, possiamo affermare che tali decisioni siano eticamente sostenibili? Ed ancora: se l'ampia diffusione di macchine a guida autonoma condurrà nel prossimo futuro ad una riduzione del numero di incidenti stradali, su chi ricadrà la responsabilità dei danni provocati da questi veicoli a persone o cose[3]?

Tali questioni che coinvolgono disparati ambiti della nostra vita sono al centro di una riflessione etica che estende nel nostro presente il suo campo d'indagine agli artefatti tecnici e che quindi assume come oggetto della sua riflessione non solo il vivente umano e quello non-umano, ma anche il non-vivente. Con la presa in esame degli artefatti tecnici, difatti, l'etica ridefinisce i propri confini superando in tale modo quelle diverse forme di ostracismo che escludevano, e escludono ancora oggi, dalla comunità dei soggetti morali tanto gli animali quanto le macchine[4].

Di questa nuova direzione inaugurata in campo etico ne è testimone il recente dibattito intorno allo statuto morale dei sistemi intelligenti (si veda Coeckelbergh 2020a, pp. 47-62; Llorca Albareda *et al.* 2023; Redaelli 2023). A seconda delle diverse prospettive assunte, che vanno da un approccio funzionalistico (Wallach, Allen 2009) ad uno relazionale (Coeckelbergh 2014; Gunkel 2018), si ascrivono a tali sistemi differenti tipi di statuto morale: per taluni autori i sistemi intelligenti sono *moral entities*, ma non *moral agents* poiché manchevoli di stati mentali ed intenzione ad agire (si veda ad es. Johnson 2006); per altri sono invece agenti morali in virtù della cosiddetta *mindless morality* (Floridi, Sanders 2004); per altri ancora tali sistemi sono meri strumenti che, pur non possedendo alcuno statuto morale peculiare, possono avere effetti sul significato morale delle nostre azioni (Peterson, Spahn 2011).

In tale variegato dibattito intorno allo statuto morale dell'intelligenza artificiale riveste un ruolo niente affatto marginale l'attribuzione (o meno) di una qualche

si sono occupati ad es. Sparrow 2016; Sparrow, Sparrow 2006. Sul ruolo che l'intelligenza artificiale può svolgere nei processi educativi si veda il testo *BEIJING CONSENSUS on artificial intelligence and education*, pubblicato dalla United Nations Educational, Scientific and Cultural Organization (UNESCO) nel 2019.

2 Sull'impiego dell'intelligenza artificiale in campo economico si veda Daza, Ilozumba 2022.

3 Su questo problema si vedano, ad esempio, Loh, Loh 2017; Millar 2017.

4 Una giustificazione comune all'esclusione di animali e macchine dal dominio della moralità è tradizionalmente offerta dalla dottrina cartesiana che riconosce in essi entità che agiscono meccanicamente. A questo proposito Deborah Johnson osserva che "l'idea cartesiana è che gli animali, le macchine e gli eventi naturali siano determinati da forze naturali; il loro comportamento è il risultato della necessità. Le spiegazioni causali del comportamento di entità ed eventi meccanicistici sono date in termini di leggi di natura. Di conseguenza, né gli animali né le macchine hanno la libertà o l'intenzionalità che li renderebbe moralmente responsabili o soggetti appropriati di una valutazione morale. Né il comportamento della natura né quello delle macchine sono spiegabili in termini di *reason explanations* e l'*agency* morale non è possibile quando non è possibile una *reason-explanation*" (Johnson 2006, p. 199).

forma di intenzionalità agli artefatti tecnici che permetta di considerarli parte del mondo morale (o escluderli da esso). Il problema dell'intenzionalità costituisce infatti una pietra d'angolo nella costruzione della nozione di *moral agency* come dimostra, tra l'altro, il fatto che gli oggetti tecnologici siano stati finora esclusi dal campo dell'etica proprio perché manchevoli di intenzionalità, a cui si legano tradizionalmente le nozioni di autonomia e responsabilità, o poiché ridotti, da un punto di vista consequenzialista, a meri strumenti moralmente neutrali al servizio dell'ottimizzazione degli *outcomes*.

Di contro a tali posizioni si registra oggigiorno una tendenza di verso opposto che attribuisce sempre più spesso una qualche forma d'intenzionalità ai sistemi intelligenti. Più specificatamente, si riconosce ad essi, come alle tecnologie più complesse, quella che è definita nei termini di intenzionalità tecnologica (Mykhailov, Liberati 2023; Terzidis, Fabrocini, Lee 2023), intesa di volta in volta come funzionalità (Johnson 2006), direttività (Verbeek 2011), o intenzione ad agire (Sullins 2006) – solo per nominare alcuni significati ad essa attribuiti. Perciò, come emerge già da questi brevi accenni, sotto l'egida di tale termine si nasconde un ginepraio di idee non facilmente ricomponibili in uno schema unitario. Questa polisemia a cui da sempre è legato il termine intenzionalità ha l'effetto di moltiplicare le posizioni in campo nel dibattito sullo statuto morale dell'intelligenza artificiale. Difatti, tali posizioni, pur appellandosi ad una qualche forma d'intenzionalità, non sembrano trovare un accordo su cosa si intenda con questo termine.

Ad alcune posizioni paradigmatiche in merito all'intenzionalità degli agenti artificiali si rivolge il presente scritto con un duplice scopo. In primo luogo, si intende mostrare come ai diversi significati attribuiti al termine intenzionalità corrispondano, per lo più, due principali dinamiche che investono lo statuto morale delle tecnologie dotate di IA. La prima tende a riconoscere nell'IA una mera estensione dell'umano, riducendo l'intenzionalità tecnologica a quella umana. La seconda strategia mira invece ad attribuire ai sistemi intelligenti una certa *agency* morale, assegnando loro una forma di intenzionalità tecnologica che va oltre quella umana, pur rimanendole in qualche misura legata. Come vedremo, entrambe le strategie incorrono in taluni limiti: la prima presta il fianco all'accusa di antropocentrismo e di riconduzione degli artefatti tecnologici a meri strumenti. La seconda, pur riconoscendo alla tecnologia una certa *agency* morale, rimane ancorata al vocabolario dell'intenzionalità dando adito ad alcuni fraintendimenti, per cui i sistemi intelligenti sembrano dotati di una qualche intenzione ad agire pari a quella umana. Di contro a tali limiti proponiamo di attribuire ai sistemi intelligenti una certa nozione di intenzionalità, con cui si intende evitare tanto le secche di una visione antropocentrica dell'IA e più in generale dell'artefatto tecnico, quanto i limiti di una visione postumanista e postfenomenologica. In questo senso, l'operazione che si intende realizzare, almeno *nelle intenzioni*, è quella di perfezionare la prospettiva postfenomenologica mediante l'introduzione della nozione di preterintenzionalità quale carattere, cifra che contraddistingue i sistemi dotati di intelligenza artificiale, ed in modo particolare di quella generativa.

2. Dall'intenzionalità alla preterintenzionalità

a) Deborah Johnson: intenzionalità come funzionalità

Gran parte delle voci che prendono parte al fitto dialogo intorno allo statuto morale dell'intelligenza artificiale sono riconducibili a due macrocategorie. La prima categoria è costituita da coloro che ritengono non sia legittimo attribuire all'IA una qualche intenzionalità, poiché tali sistemi sono manchevoli di coscienza. In questo senso, l'intenzionalità è intesa come la proprietà degli stati mentali di essere diretti verso individui o stati di cose da cui deriva l'agire conformemente all'intenzione. Perciò laddove mancano gli stati mentali manca l'intenzionalità (si veda ad esempio Mosakas 2021). La seconda, a cui abbiamo accennato nel precedente paragrafo, è rappresentata da coloro che, pur negando ai sistemi intelligenti una qualsiasi intenzione cosciente, assegnano ad essi una certa intenzionalità, ridefinendo radicalmente il significato di tale termine. A questa seconda categoria rivolgeremo ora l'attenzione per mettere in luce i processi (di antropomorfizzazione o deumanizzazione) che sono sottesi alle diverse nozioni d'intenzionalità chiamate in causa.

Una prima riflessione particolarmente promettente è presentata da Deborah Johnson in *Computer systems: Moral entities but not moral agents* (2006). In questo testo, Johnson mira a riconoscere uno statuto morale ai *computer systems* pur senza attribuire loro una qualche forma di *agency* morale[5]. A tal fine, Johnson sottolinea innanzitutto come i sistemi informatici non possiedano alcuni tratti caratteristici dell'*agency* umana – gli stati mentali e l'intenzione ad agire – e per questa loro carenza non possano essere considerati agenti morali. Tale mancanza di coscienza ed intenzioni ad agire non conduce tuttavia l'autrice ad abbracciare la tesi opposta secondo cui tali dispositivi tecnologici siano moralmente neutrali (ivi, p. 195). Tra questi due opposti, Johnson inaugura una terza via introducendo nel dibattito sullo statuto morale dei sistemi intelligenti la categoria di *moral entities*. Con tale categoria l'autrice intende affermare che i computer siano parte del mondo morale in virtù del modo in cui sono progettati dall'uomo e delle funzioni che assolvono, pur senza che il riconoscimento di tali proprietà significhi riconoscere loro una vera e propria *moral agency* pari a quella umana. Quest'ultima, infatti, richiede un certo grado di libertà a partire dalla quale possono sorgere quelle intenzioni ad agire di cui i *computer systems* sono evidentemente sprovvisti.

Tuttavia, benché tali tecnologie non possiedano intenzioni ad agire, vale a dire la volontà di perseguire una certa azione, ad essi Johnson non nega una certa intenzionalità che sorgerebbe dall'unità uomo-tecnologia. Difatti, secon-

5 Benché in tale testo, non recentissimo, Johnson rivolga l'attenzione ai *computer systems* in generale, l'autrice chiama direttamente in causa l'intelligenza artificiale (Johnson 2006, pp. 195-196) per cui le sue riflessioni si rivolgono anche ad una serie di problemi specifici che investono il campo dell'intelligenza artificiale. Occorre inoltre osservare che le posizioni di Johnson rimangono, per lo più, invariate anche in testi successivi, laddove l'agente artificiale è ridefinito nei termini di agente surrogato (Johnson, Powers 2008; Johnson, Noorman 2014).

do l'autrice, gli artefatti hanno significato morale soltanto in relazione all'essere umano, vale dire in quanto parte di sistemi socio-tecnici (Van de Poel 2020) ed è in quanto parte di tali sistemi che essi presentano una propria intenzionalità. Più precisamente, l'intenzionalità degli enti morali (computer, robot, IA) emerge entro una fitta rete di relazioni che coinvolge l'intenzionalità dei progettisti e degli utilizzatori, laddove la prima è per così dire incorporata dal sistema, mentre la seconda fornisce gli input affinché si attivi l'intenzionalità del sistema. Tali sistemi, infatti, incorporano l'intenzionalità degli utenti e dei *designers* nella propria intenzionalità, che è correlata alla loro funzionalità (Johnson 2006, p. 201), formando in tal modo una triade d'intenzionalità che è in azione nelle tecnologie. L'intenzionalità del sistema è perciò in ultima battuta identificabile con la capacità di fornire output (*the resulting behaviour*) a partire da dati input, pur senza che i programmatori abbiano specificato la correlazione tra ogni peculiare input ed ogni peculiare output.

A partire da queste brevi indicazioni appare già chiara la strategia che si cela dietro la nozione di intenzionalità a cui si appella Johnson. Con intenzionalità non si intende qui alcuna intenzione ad agire, bensì la funzionalità e l'efficienza dell'artefatto tecnico. Perciò l'intenzionalità del computer è data dall'unione dell'intenzionalità dei programmatori con quella degli utilizzatori. In questo senso, l'artefatto è ridotto a mera estensione dell'umano, esso è veicolo di una intenzionalità intesa come mera funzionalità, è portatore di quella che Searle (1992) chiamerebbe intenzionalità derivata.

Al fine di evitare qualsiasi confusione tra l'intenzione ad agire propria degli esseri umani, per cui essi decidono di realizzare una certa azione di cui sono responsabili, e l'intenzionalità dei sistemi informatici intesa come mera funzionalità, Johnson osserva, tra l'altro, che "non importa quanto i *computer systems* del futuro si comporteranno in modo indipendente, automatico ed interattivo, essi saranno sempre i prodotti (diretti o indiretti) del comportamento umano, delle istituzioni sociali umane, e della decisione umana" (Johnson 2006, p. 197). Con tale affermazione Johnson sembra prendere radicalmente le distanze da qualsiasi idea di intenzionalità legata ad una decisione della macchina che chiami in causa un certo spazio di libertà. Perciò, ai suoi occhi, anche le conseguenze inattese prodotte dagli artefatti e così un certo carattere non deterministico di cui il computer è provvisto[6] sono da ricondurre agli esseri umani e alla loro (in)capacità di previsione.

Questo continuo richiamo al legame uomo-macchina, laddove la seconda è mera estensione del primo, fa sì che la posizione di Johnson appaia, ad alcuni criti-

6 Con carattere non deterministico, che Johnson assegna ad alcuni *computer systems*, si intende che l'output prodotto dal sistema informatico non può essere predetto con certezza e che, pur partendo da un medesimo input, il sistema può presentare comportamenti differenti. In questo senso, Johnson afferma che "quando i computer sono programmati per imparare, imparano a comportarsi in modi che vanno ben oltre la comprensione dei loro programmatori e ben oltre ciò che viene dato loro come input. Le reti neurali sono considerate un esempio di sistemi informatici non deterministici" (Johnson 2006, p. 200).

ci, viziata da un certo antropocentrismo. Difatti, tale posizione ridurrebbe i sistemi intelligenti a mera estensione dell'umano, senza riconoscere ad essi un carattere autonomo ed un impatto morale che molte volte, nel caso dei sistemi dotati di IA, non è riducibile a quello progettato dagli sviluppatori. In questo senso, sono istruttive le parole di Gunkel, per cui

> benché la "triade dell'intenzionalità" proposta da Johnson sia più complessa della posizione strumentale classica, essa deriva da e protegge ancora un investimento fondamentale nell'eccezionalismo umano. Nonostante le notevoli promesse di ridefinire il dibattito, il nuovo paradigma di Johnson non sembra molto diverso da quello che è stato progettato per sostituire. Gli esseri umani sono ancora e senza dubbio gli unici legittimi agenti morali (Gunkel 2017, p. 68).

Benché la riduzione della posizione di Johnson ad uno strumentalismo più complesso rispetto a quello classico risulti troppo radicale, occorre sottolineare che il comportamento di alcune tecnologie come il *machine learning*, il quale può portare i sistemi dotati di intelligenza artificiale a "disincarnare" i valori in essi incorporati (Vanderelst, Winfield 2018), è difficilmente spiegabile nei termini della relazione uomo-macchina. In altre parole, non tutti gli aspetti, le operazioni dei sistemi intelligenti sono riconducibili alle intenzioni dell'uomo che li ha posti in essere. In questo senso, proprio a partire dall'esempio del *machine learning*[7], è difficile condividere l'idea di Johnson per cui il sistema

> anche quando impara, impara come è stato programmato per imparare […] Il fatto che colui che progetta e l'utente non sappiano esattamente cosa fa l'artefatto non fa alcuna differenza qui. Significa semplicemente che lo sviluppatore, nel creare il programma, e l'utente, nell'usare il programma, sono coinvolti in un comportamento rischioso (*risky behavior*); stanno facilitando e avviando azioni che potrebbero non comprendere appieno, azioni con conseguenze che non sono in grado di prevedere. I progettisti e gli utenti di tali sistemi dovrebbero prestare attenzione all'intenzionalità e all'efficacia che introducono nel mondo. […] *Quando gli esseri umani agiscono con gli artefatti, le loro azioni*

7 La capacità di apprendimento automatico di cui sono dotati taluni sistemi fa sì che noi intenzionalmente progettiamo delle macchine di cui non possiamo prevedere i risultati finali (Redaelli 2024). In questo senso, un esempio ormai classico, è quello del gioco a somma zero giocato dalle *Generative Adversarial Network* (GAN) che producono degli *outcomes* finali i quali non sono conoscibili dai programmatori (Terzidis *et al*. 2023), come ad esempio nel caso della generazione di immagini. In questo caso, infatti, abbiamo un sistema capace di produrre una molteplicità di differenti risultati a partire da un set di input (in relazione alle tecniche di dropout si veda Wieluch *et al*. 2019), consentendo di generare una vasta gamma di immagini originali. Perciò, possiamo affermare che, benché il sistema sia allocentrico, 1) l'intenzionalità del sistema non è completamente riconducibile all'intenzionalità umana che lo ha posto in essere e che 2) questa discrepanza tra intenzionalità umana e tecnologica non è una mera conseguenza inattesa o un *risky behavior*, bensì è intenzionalmente perseguita dai programmatori qualora intendano produrre dei contenuti originali, proprio in tal senso parleremo in seguito di una preter-intenzionalità *congenita* in tali sistemi.

> *sono costituite dalla loro intenzionalità e dalla loro efficacia, nonché dall'intenzionalità e dall'efficacia dell'artefatto, che a sua volta è stato costituito dall'intenzionalità e dall'efficacia di colui che progetta l'artefatto stesso* (Johnson 2006, pp. 203-204, corsivi nostri).

Alla luce della critica di Gunkel, appare chiaro che considerare le conseguenze inattese delle tecnologie nei meri termini di *risky behavior* da parte dell'uomo sembra non tener conto del carattere attivo delle tecnologie e delle abilità emergenti che essi possono esibire, come, ad esempio, la capacità di autocorrezione morale messa in luce da Askell, Ganguli *et al.* (2023). In questo senso, se da un lato la posizione di Johnson ha l'indubbio merito di mettere in luce il portato morale delle tecnologie in relazione all'umano, dall'altro lato tale posizione non sembra poter offrire una spiegazione esaustiva dello statuto dei sistemi intelligenti, la cui intenzionalità strettamente legata alla loro autonomia, adattabilità e interattività, non è spesso riconducibile all'intenzionalità dei progettatori ed utenti. Lo sviluppo di caratteristiche proprie da parte di macchine dotate di capacità di apprendere sembra dunque richiedere un apparato concettuale che riconosca loro una certa *agency* morale e una congenita natura preterintenzionale, pur senza per questo dover attribuire loro una comprensione, una capacità di ragionamento o di azione morale pari a quella umana.

b) John P. Sullins: intenzionalità e livelli d'astrazione

Una diversa posizione che attribuisce una qualche forma di intenzionalità agli agenti artificiali, senza tuttavia considerarli mera estensione dell'umano, è quella presentata da John P. Sullins in *When Is a Robot a Moral Agent?* (2006). In questo breve scritto, Sullins, che basa le sue riflessioni sul metodo dei livelli d'astrazione proposto da Floridi e Sanders (2004), identifica tre requisiti necessari al soggetto al fine di essergli riconosciuta una *full moral agency*: autonomia, intenzionalità e responsabilità.

Per quanto riguarda il primo requisito, l'autore impiega, senza troppe precauzioni[8], la nozione ingegneristica di autonomia. In virtù di tale nozione, la macchina che presenta un certo grado di indipendenza rispetto agli altri agenti, vale a dire "la macchina che non è sotto il controllo diretto di nessun altro agente o utente" (Sullins 2006, p. 28), è per l'appunto autonoma.

Il secondo requisito identificato da Sullins, che è l'oggetto del nostro studio, chiama in causa una nozione debole di intenzionalità. Difatti, benché Sullins ritenga che sia possibile attribuire ai robot una certa intenzionalità sulla base della rilevanza morale delle loro azioni, le quali, a un certo livello d'astrazione, possono *apparire* come calcolate e deliberate, cioè dotate, per l'appunto, di *intenzioni autonome*, ciò non significa tuttavia riconoscere a tali artefatti una qualche forma

8 Sui problemi legati alla nozione di autonomia in merito agli agenti artificiali rimandiamo a Loh, Loh (2017).

d'intenzionalità in senso forte, poiché – osserva Sullins – questa non è attribuibile neppure all'uomo[9]. Appellandosi all'impossibilità di assegnare un agire intenzionale *in senso forte* sia all'uomo sia alle macchine, Sullins conclude che "nella misura in cui il comportamento [del robot] è complesso abbastanza da spingerci a fare affidamento sulle nozioni psicologiche popolari classiche di predisposizione o 'intenzione' di fare del bene o del male, allora questo è sufficiente per rispondere in modo affermativo alla domanda [relativa all'intenzionalità]" (*ibid.*). In altri termini, è esclusivamente il comportamento dei robot, nel senso dell'effetto morale delle loro azioni, e non qualche loro caratteristica, che spinge noi esseri umani a riconoscere a tali agenti artificiali delle intenzioni ad agire.

Ora, benché entrambe le posizioni riconoscano una certa intenzionalità agli artefatti tecnici, collocando tali artefatti entro il campo della moralità, le argomentazioni sopraesposte mostrano all'opera due divergenti tendenze che perimetrano il dibattito intorno allo statuto morale degli enti artificiali. Difatti, nonostante Johnson e Sullins evitino le difficoltà che coinvolgono nozioni quali quelle di libero arbitrio ed intenzione ad agire mediante una strategia di risemantizzazione del concetto stesso di intenzionalità, occorre sottolineare una chiara differenza che separa le prospettive in gioco: Sullins basa le sue riflessioni sulla *rilevanza* morale delle azioni realizzate da tali agenti[10], mentre Johnson intende mostrare come gli enti artificiali abbiano valore morale poiché componenti funzionali all'azione umana. In questo senso, si può osservare che le riflessioni di Sullins, nella scia di quelle presentate da Floridi e Sanders (2004), mirano ad affrancare l'agente artificiale da un certo antropocentrismo che tende a stabilire un'intrinseca connessione tra la *moral agency* dell'artefatto e quella umana o, in alcuni casi, che pretenda una loro piena sovrapposizione affinché si possa attribuire agli artefatti una qualche forma di *agency* morale. Dall'altro lato, occorre, tuttavia, osservare che la posizione di Sullins presta il fianco a diverse critiche che mettono in luce come dietro l'oggettività rivendicata dal metodo dei livelli d'astrazione proposto da Floridi e Sanders, e impiegato dallo stesso Sullins, vi sia già una scelta etica che determinerebbe i criteri

9 L'impossibilità di dimostrare che non solo i robot, ma anche gli uomini possiedano intenzionalità è chiamata in causa da Sullins al fine di spostare il *focus* della sua riflessione dalla questione della coscienza al *comportamento* degli agenti artificiali, evitando così il problema di come "accedere" alla mente degli altri, umani e non-umani. In tal modo, la prospettiva di Sullins assume la forma di un comportamentismo che vuole evitare di appellarsi a nozioni come coscienza e mente nel tentativo di definire la *moral agency* degli agenti artificiali. Con tale strategia, Sullins mostra, in ultima istanza, di abbracciare pienamente l'idea di *mindless morality* proposta da Floridi e Sanders, pur riconoscendo, diversamente dagli autori in questione, l'intenzionalità e la responsabilità come criteri qualificanti la *moral agency*.

10 A questo proposito Coeckelbergh correttamente osserva che "di contro a questo approccio (quello condiviso da Floridi, Sanders e Sullins), si potrebbe sostenere che questi argomenti confondono la *rilevanza* morale delle azioni con l'*agency* morale. Una cosa è riconoscere che questi animali [i cani da soccorso] e robot fanno cose moralmente rilevanti; un'altra cosa è affermare che essi hanno quindi un'*agency* morale, che – secondo questo argomento – solo le persone o gli esseri umani possono avere" (Coeckelbergh 2020, p. 156).

qualificanti la *moral agenthood* (Gunkel 2017, p. 73). In altri termini, tale metodo non possiederebbe quella oggettività cha presenta nel campo della matematica da cui è mutuato, poiché a capo della scelta dei criteri qualificanti l'agente morale, tra i quali vi è per Sullins l'intenzionalità, vi sarebbe già una decisione in merito a coloro che possono far parte della comunità dei soggetti morali e chi ne è escluso[11]. Oltre a ciò, si può osservare, come approfondiremo nel prossimo paragrafo, che la posizione di Sullins attribuisce delle intenzioni ad agire ai robot mediante il ricorso a una nozione psicologica di intenzionalità difficilmente applicabile ad un artefatto tecnico.

Dall'altro lato, come appena detto, anche le riflessioni di Johnson presentano dei limiti: esse sono esposte alla critica di antropocentrismo, per cui l'*agency* morale è solo umana e non è attribuibile ad altri soggetti[12]. Nonché, agli occhi di Johnson, lo stesso significato morale dei *moral entities* dipende esclusivamente dal loro essere componenti dell'azione umana, dal loro essere progettati dall'uomo e funzionali agli scopi umani. Questo vale anche per enti che presentino un certo grado d'indipendenza, il quale non è tuttavia sufficiente al fine d'assegnare loro lo statuto di *autonomous moral agents* (Johnson, Miller 2008, p. 127), semmai di *surrogate agent*. Tali artefatti sono e rimangono "un'estensione dell'attività umana e dell'*agency* umana" (*ibid.*) e così la loro intenzionalità è riconducibile a quella umana.

Di fronte ai limiti esibiti da queste riflessioni intendiamo qui presentare in ultima istanza la posizione postfenomenologica di Verbeek che, come cercheremo di mostrare, ha il merito di ricondurre la relazione uomo-macchina nell'alveo della nozione di agente composito, affrancando così la propria proposta dai limiti a cui incorrono le posizioni antropocentriche, senza tuttavia chiamare in causa, come avviene in Sullins, il metodo dell'astrazione sulla cui tenuta sono sorte diverse obiezioni[13]. Entro questi due estremi Verbeek ha il merito di ridefinire, seppur non senza difficoltà, la nozione di intenzionalità, tenendo conto tanto della commistione tra intenzionalità tecnologica e intenzionalità umana quanto, sebbene indiretta-

11 Oltre a ciò Gunkel osserva che il metodo dell'astrazione non evita equivoci e divergenze sui criteri qualificanti la *moral agenthood*: basti pensare a Sullins, che impiega tale metodo e, nonostante ciò, riconosce criteri differenti rispetto a quelli di Floridi e Sanders (Gunkel 2017, p. 73). Un altro punto debole della teoria di Floridi e Sanders è sottolineato da Verbeek, il quale, benché apprezzi la loro proposta, osserva come vi siano artefatti, quali le ecografie ostetriche o i famosi cavalcavia progettati da Robert Moses, che, seppur "non soddisfino i criteri stabiliti da Floridi e Sanders per l'*agency*, contribuiscono attivamente ad azioni morali e hanno conseguenze che possono essere valutate in termini morali" (Verbeek 2011, p. 50).

12 Per Coeckelbergh, infatti, di contro alla posizione di Johnson, "Floridi e Sullins potrebbero allora rispondere che questa definizione di *moral agency* è troppo antropocentrica, che tale libertà metafisica [dalla quale per Johnson sorge l'intenzione ad agire di cui i computer sono sprovvisti] non è necessaria per la *moral agency*, o che le sue condizioni sono già soddisfatte nei casi rilevanti di agenti artificiali come i cani da salvataggio" (Coeckelbergh 2020, p. 156)

13 Pur apprezzando la proposta di Floridi e Sanders, importanti obiezioni sono mosse anche da parte di Johnson, Miller 2008.

mente, di ciò che proponiamo di chiamare *natura preterintenzionale*[14] dell'IA, per cui l'intenzionalità tecnologica veicolata dall'IA, *in virtù della sua stessa natura*, va oltre quella umana che l'ha creata.

c) Intenzionalità tecnologica e intenzionalità composita. La prospettiva postfenomenologica

L'approccio filosofico di Verbeek (Verbeek 2011; 2008; 2005) mira a mettere in luce il ruolo svolto dalle tecnologie nella formazione dei nostri abiti di comportamento e dunque a riconoscere agli artefatti tecnici un chiaro significato morale. A tal fine, Verbeek sviluppa, in modo originale, alcune feconde intuizioni presenti in Latour (si veda ad esempio Latour 1993; 1994; 2002) e Ihde (1979; 1990; 1993) assumendo come *focus* della sua indagine la funzione di mediazione che le tecnologie assolvono nella nostra vita. Agli occhi del filosofo, infatti, i dispositivi tecnologici contribuiscono a formare la nostra esperienza del mondo con importanti ricadute sulle nostre azioni e decisioni. In questo senso, gli oggetti tecnologici non sono "'intermediari' neutrali tra uomini e mondo, bensì *mediatori* (mediators)" nel senso letterale che "mediano attivamente" la nostra relazione al mondo (Verbeek 2005, p. 114). In virtù di tale attività di mediazione delle tecnologie, l'etica ha, per Verbeek, il compito di estendere il proprio campo d'indagine oltre l'umano, accogliendo al suo interno quelle "forme non umane di *agency*" (Verbeek 2011, p. 17), comprese quelle tecnologiche, su cui Latour ha gettato nuova luce grazie alla sua Actor-Network Theory (Latour 2005).

Al fine di sviluppare tale etica con cui si intende superare la dicotomia moderna soggetto-oggetto (Latour 1993), pur senza accogliere al suo interno il principio della simmetria stabilito dallo stesso Latour[15], Verbeek si avvale principalmente dell'approccio postfenomenologico inaugurato da Don Ihde. Come appena detto, tale approccio ha il merito di indagare la funzione di mediazione svolta dalle tecnologie nella nostra relazione al mondo[16], ponendo al centro dell'analisi fenomenologica "*il modo in cui, nelle relazioni che sorgono intorno a una tecnologia, si costituiscono un 'mondo' specifico e un 'soggetto' specifico*" (Rosenberger, Verbeek 2015, p. 31). Per Ihde, infatti, e così per Verbeek, la relazione uomo-mondo non coinvolge soggetti e oggetti preesistenti, bensì soggetti e mondo si con-costituisco-

14 Ci avvaliamo qui del termine preterintenzionale impiegato in ambito filosofico da C. Di Martino 2017 al fine di spiegare gli effetti di ritorno delle tecnologie nel processo di antropogenesi.

15 Di contro alla simmetria tra gli attanti umani e non umani stabilita da Latour, Rosenberger e Verbeek affermano che "l'approccio postfenomenologico […] non rinuncia esplicitamente alla distinzione tra entità umane e non umane. Al posto della simmetria, considera l'interazione e la costituzione reciproca tra soggetto e oggetto" (Rosenberger, Verbeek 2015, p. 19).

16 In questo senso, la "la postfenomenologia non vede la fenomenologia come un metodo per *descrivere* il mondo, ma come la comprensione delle *relazioni* tra gli esseri umani e il loro mondo" (ivi, p. 11).

no nell'interazione, mediata dalle tecnologie, tra uomo e realtà (Verbeek 2011, p. 15). Per tale ragione, la postfenomenologia assume la peculiare forma di un'analisi delle *relazioni* sussistenti tra l'uomo, la tecnologia e il mondo.

In seno a tale tipo di analisi, riveste un ruolo evidentemente centrale la nozione fenomenologica d'intenzionalità, che è definita da Verbeek nei termini di "direttività dell'essere umano verso la realtà", per cui gli esseri umani, secondo il dettato husserliano, "non possono semplicemente 'pensare', ma sempre pensare a *qualcosa*; non possono semplicemente "vedere", ma sempre vedere *qualcosa*" (ivi, p. 55). Tuttavia, se da un lato tale nozione di intenzionalità, funzionale all'approccio relazionale di cui si fanno portavoce tanto Ihde quanto Verbeek, è dichiaratamente mutuata dalla fenomenologia, il nuovo impianto filosofico in cui è collocata le imprime una peculiare curvatura. Questa curvatura si lega a quanto abbiamo appena detto: l'intenzionalità non designa, in ambito postfenomenologico, una relazione diretta tra soggetto ed oggetto, bensì una relazione sempre più spesso mediata dalle tecnologie, laddove tale mediazione è la fonte, l'origine (Rosenberger, Verbeek 2015, p. 12) di forme diverse di soggettività e oggettività. Perciò, Verbeek può affermare che l'intenzionalità è distribuita tra umani e tecnologie, attribuendo a queste ultime non una qualche intenzione ad agire, bensì una certa direttività, vale a dire un "ruolo direttivo nelle azioni e nelle esperienze degli esseri umani" (Verbeek 2011, p. 57).

A partire da questa riconcettualizzazione della relazione intenzionale, Verbeek sottolinea poi come l'intenzionalità umana formata dai dispositivi tecnologici possa assumere diverse forme (si veda Verbeek 2008). Tra queste forme intendiamo focalizzarci su ciò che il filosofo definisce nei termini di intenzionalità composita, convinti che tale tipo di intenzionalità contraddistingua, in modo particolare, la relazione uomo-sistemi intelligenti, dato che, in questa variante, "vi è un ruolo centrale per la 'intenzionalità' o la direttività degli artefatti tecnologici stessi, in quanto gli artefatti interagiscono con le intenzionalità degli esseri umani che li utilizzano" (Verbeek 2011, p. 145)[17]. Con tali parole, in cui si sottolinea l'interazione tra l'intenzionalità umana e quella tecnologica, Verbeek

17 Al fine di comprendere meglio la proposta di Verbeek, è necessario osservare che in Verbeek (2011) l'autore utilizza l'espressione intenzionalità composita in un duplice senso. Da un lato, Verbeek, riferendosi ad una nozione ampia di intenzionalità composita, afferma che "l'intenzionalità è sempre un affare ibrido che coinvolge sia intenzioni umane sia intenzioni non umane o, meglio, coinvolge 'intenzioni composite' con intenzionalità distribuite tra gli elementi umani e non umani entro le relazioni uomo-tecnologia-mondo" (ivi, p. 58). In questo senso, una forma di intenzionalità composita è all'opera in ogni relazione uomo-tecnologia. D'altra parte, Verbeek sviluppa, nello stesso testo, una nozione, per così dire, ristretta di intenzionalità composita, in cui l'intenzionalità tecnologica gioca un ruolo centrale e l'intenzionalità umana interagisce con quella tecnologica, formando un'intenzionalità composita che non si limita a svolgere una funzione di mediazione, ma forma quelle che Verbeek definisce *relazioni composite* (ivi, p. 140). In tali relazioni, l'intenzionalità artificiale si aggiunge a quella umana. Nonostante questa distinzione, è necessario chiarire che la nozione ristretta, a cui ci riferiamo in questo articolo, presuppone evidentemente la nozione ampia di intenzionalità composita.

mette in luce come l'intenzionalità composita si ha laddove vi sia una sinergia tra intenzionalità tecnologica "che è rivolta al 'suo' mondo" e quella umana "rivolta al risultato di questa intenzionalità tecnologica" (ivi, p. 146). Perciò, nel caso dell'intenzionalità composita, "gli esseri umani sono diretti qui ai modi in cui una tecnologia è diretta al mondo" (Verbeek 2008, p. 393).

A questo proposito, è importante osservare che, agli occhi di Verbeek, con tale tipo d'intenzionalità tecnologica si dischiude una realtà che è accessibile solo a tali tecnologie e che, al tempo stesso, attraverso la loro mediazione entra nel campo umano. In questo senso, egli assegna a questo tipo d'intenzionalità una duplice funzione, rappresentativa e costruttiva. Ciò significa che tale intenzionalità tecnologica non solo può rappresentare la realtà, bensì può costituire una realtà che esiste per l'intenzionalità umana soltanto qualora si unisca a quella tecnologica.

Questa nozione d'intenzionalità composita sembra essere particolarmente adatta a spiegare l'intenzionalità presente nei sistemi dotati d'intelligenza artificiale, ed in modo particolare di intelligenza artificiale generativa, contribuendo allo stesso tempo a chiarirne lo statuto di mediatori morali. Tali sistemi, difatti, presentano una intenzionalità tecnologica intesa come direttività che orienta la nostra azione e il nostro pensiero. Una direttività o intenzionalità che è sì – come osserva Johnson – connessa all'uomo che progetta la macchina e la usa, ma che, allo stesso tempo, presenta un carattere emergente rispetto all'intenzionalità umana, sia quella dei programmatori sia quella dei fruitori. Tale carattere non è tuttavia riducibile ad una mera indipendenza dell'agente artificiale, bensì si lega alla sua capacità di strutturare nuove forme di realtà (altrimenti non accessibili all'uomo) secondo direzioni d'azione inaspettate[18], e dunque si lega, in ultima battuta, alla sua generatività, alla sua capacità di generare *outcomes* originali, come, ad esempio, testi ed immagini. In questo senso, se è corretto focalizzare l'attenzione sulla commistione uomo-macchina al fine di ripensare lo statuto morale di quest'ultima, evitando in tal modo di abbracciare posizioni che solo in apparenza assumono una visione oggettiva da cui far fronte al problema dello statuto morale[19], occorre tuttavia sottolineare che l'intenzionalità del sistema intelligente eccede quella triangolazione messa in luce da Johnson, poiché l'intenzionalità tecnologica non è riducibile alla mera funzionalità progettata dall'uomo e avviata dagli input inseriti dai fruitori. La nozione di intenzionalità tecnologica si lega infatti ad un certo carattere non completamente predeterminato dell'IA, per cui non tutte le azioni dei sistemi intelligenti sono prevedibili e comprensibili (si pensi alla cosiddetta *black box AI*[20]) e così il loro ruolo di mediazione non è sempre riconducibile a quello progettato. In

18 Verbeek definisce le tecnologie come "*mediatori* che aiutano attivamente a formare la realtà" (Verbeek 2011, p. 46).

19 A questo proposito rimandiamo a Johnson 2006, p. 196.

20 Con tale espressione si intende l'opacità che affetta alcuni sistemi di intelligenza artificiale, per cui non risulterebbe comprensibile il processo che porterebbe tali sistemi a produrre determinati *outcomes*. Sulla nozione di opacità si veda Burrell 2016. Su larga scala, si può osservare che l'opacità degli algoritmi conduce a quella che Pasquale ha definito nei termini di una

questo senso, gli artefatti sono contraddistinti da una duplice dinamica: essi incorporano le intenzioni umane in modo materiale e allo stesso tempo presentano delle "forme emergenti di mediazione" (Verbeek 2011, p. 127), che nei sistemi dotati di intelligenza artificiale sono congenite[21].

Oltre a mettere in luce tale dinamica, questo tipo di intenzionalità composita sembra offrire un decisivo correttivo a quanto affermato da Sullins. Secondo Sullins – ripetiamolo – se l'interazione di un robot con altri agenti *appare* particolarmente complessa e le azioni di quest'ultimo *sembrano* dotate di intenzioni autonome allora si può attribuire al robot una certa forma d'intenzionalità, nel senso di una predisposizione o intenzione ad agire. Tuttavia, benché l'interazione tra sistema intelligente e ambiente possa essere molto complessa è difficile chiamare in causa una nozione di intenzionalità che si lega, nelle parole di Sullins, alla *folk psychology*, la quale si riferisce evidentemente ad agenti umani, dotati di caratteristiche di cui gli agenti artificiali sono sprovvisti, tra cui l'essere coscienti. In questo senso, la nozione di intenzionalità composita, che non si basa sulla nozione psicologica di intenzione ad agire, bensì su quella di direttività, appare più consona alle tecnologie, poiché è affrancata dalle problematiche metafisiche che coinvolgono la nozione di coscienza, a cui sembra ancora rimandare la nozione di intenzione ad agire chiamata in causa da Sullins.

Al fine di chiarire la nozione di intenzionalità composita sarà utile qui offrire un esempio in cui tale intenzionalità è all'opera. A questo proposito ci riferiamo qui alla *Generative Art AI* denominata Midjourney. Questo tipo di intelligenza artificiale crea immagini a partire da un breve testo descrittivo inserito dall'utente. In questo caso, la tecnologia produce una realtà che non sarebbe affatto esperibile dal soggetto umano, se alla sua intenzionalità, in questo caso rappresentata *in prima battuta* dal testo descrittivo, non si aggiungesse l'intenzionalità del sistema intelligente[22]. Questa azione congiunta, questa intenzionalità composita, è ancora più evidente dato che Midjourney utilizza il canale Discord in modo tale che l'utente possa interagire con un bot il quale mostra la produzione delle immagini in tempo reale, così da offrire all'utente la possibilità di apportare modifiche. In tale dinamica si può dunque osservare all'opera tanto l'intenzionalità umana dei programmatori (che hanno sviluppato il sistema) e degli utilizzatori (che inseriscono il prompt) quanto l'intenzionalità dell'intelligenza artificiale generativa, che eccede quella umana. Difatti, benché il sistema prenda le mosse dal prompt inserito, il risultato non è completamente prevedibile dall'utilizzatore, poiché l'immagine creata è originale e non riconducibile ad una mera combinazione di immagini precedenti. In tal senso si può parlare di un'intenzionali-

Black Box Society, vale a dire di una società pervasa da sistemi che prendono decisioni, le quali, a noi essere umani, sono del tutto incomprensibili (Pasquale 2015).

21 Su questo punto si veda Redaelli 2024.

22 È necessario sottolineare che l'intenzionalità tecnologica è già sempre coinvolta in una relazione con l'essere umano in quanto progettatore e utilizzatore della tecnologia che è portatrice di tale tipo di intenzionalità.

tà composita bot-utente, laddove il bot non fa altro che lavorare sui cosiddetti tokens e "compararli" con i dati di addestramento, fornendo all'essere umano delle immagini su cui poter ancora intervenire.

Proprio questa azione congiunta uomo-macchina dotata di intelligenza artificiale ci permette di ribadire il carattere relazionale della nozione di intenzionalità composita, che si può riscontrare, per certi versi, anche in alcune direttive europee in materia di intelligenza artificiale (si veda Pacileo 2020). Pur richiedendo una continua sorveglianza umana sui sistemi intelligenti, e quindi una IA centrata sull'uomo (Human-Centered AI)[23], tali direttive si concentrano, infatti, sul carattere congiunto dell'azione uomo-macchina e sulla funzione integrativa (e non sostitutiva) svolta dall'intelligenza artificiale nei confronti delle nostre capacità. Perciò, anche in ambito regolativo, possiamo ritrovare quella intenzionalità composita che ha la sua massima espressione nei sistemi dotati di intelligenza artificiale *quali* mediatori morali dotati di una peculiare intenzionalità.

3) Preterintenzionalità

Se la nozione di intenzionalità tecnologica, legata a quella composita, sembra particolarmente adatta a rendere ragione della capacità dei sistemi intelligenti di agire conformemente alle proprie decisioni, adattandosi all'ambiente, il ricorso alla nozione di intenzionalità tecnologica lascia ancora spazio ad alcuni fraintendimenti. Difatti, il suo impiego ha sollevato numerosi sospetti nei confronti della proposta postfenomenologica di Verbeek. In modo particolare, le riflessioni di Verbeek hanno dato l'abbrivio ad una serie di critiche relative alla distribuzione d'intenzionalità tra uomo e macchina a cui lo stesso autore ha solo in parte risposto (Verbeek 2014). Al fine di sintetizzare tali rimostranze, possiamo domandarci: in che senso dobbiamo intendere la ridistribuzione di intenzionalità (e responsabilità) tra esseri umani e tecnologia? In che modo l'intenzionalità umana e quella della macchina, che incorpora già quella umana, formano un'intenzionalità composita? Come dobbiamo intendere questa composizionalità? Queste sono solo alcune domande, che, agli occhi di taluni critici, rimangono inevase da Verbeek.

A questi problemi, in parte irrisolti, si aggiunge il problema della terminologia impiegata dal filosofo, per cui Coeckelbergh giustamente osserva che "alcune di queste obiezioni potrebbero essere evitate se Verbeek non usasse termini ed espressioni come 'moralità delle cose' e '*agency* morale delle cose', ma rimanesse fedele all'affermazione che le tecnologie mediano la moralità" (Coeckelbergh 2020, p. 67). Difatti, benché si possa comprendere l'uso di un

23 Occorre qui osservare che la posizione postumanista di Verbeek non sembra essere in contrasto con la cosiddetta Human-Centered AI (HCAI), bensì si muove in una medesima direzione nella misura in cui riconosce che le tecnologie in generale e l'IA aumentano e amplificano le capacità umane piuttosto che sostituirle.

tale vocabolario da parte di Verbeek al fine di scardinare la comune comprensione della tecnologia (ivi, p. 67), l'approccio postfenomenologico non può esimersi da una continua vigilanza sul linguaggio, vigilanza a cui si è chiamati con maggiore vigore qualora le tecnologie sollevino questioni etiche sempre più urgenti.

Al fine di risolvere tale problema proponiamo di riconoscere alle tecnologie non tanto una forma di intenzionalità composita, la cui terminologia lascia spazio, ancora una volta, all'idea che la macchina possieda delle intenzioni ad agire (si vedano a questo proposito le critiche mosse a Verbeek da Peterson, Spahn 2011)[24], bensì quella di preterintenzionalità. Tale nozione, a nostro avviso, restituisce meglio l'intreccio tra essere umano e tecnologia richiamato dalla nozione di intenzionalità composita, evitando allo stesso tempo di prestare il fianco alle accuse di attribuire ai sistemi intelligenti una qualche forma di intenzione.

Allo scopo di chiarire il significato del termine preterintenzionalità è bene ricordare la sua origine latina. Tale termine, ampiamente impiegato in ambito giuridico[25], è, infatti, composto dal prefisso *praeter* "oltre" e *intendĕre* nel senso di rivolgere, tendere. Il significato che gli è normalmente attribuito è quello di andare oltre l'intenzione di chi agisce. In questo preciso senso, si parla, ad esempio, in ambito giuridico di omicidio preterintenzionale, laddove vi è una volizione di un evento (ad es. percosse) e la realizzazione di un evento più grave (morte). In questo caso, seppur il secondo evento sia eziologicamente legato al primo non è volontariamente perseguito.

Al di là dell'uso tecnico di tale termine in campo giuridico, il significato – di andare oltre l'intenzione di chi agisce – sembra particolarmente adatto ad indicare la dinamica messa in luce dalla intenzionalità composita, per cui l'intenzionalità tecnologica che emerge nell'interazione uomo-IA non è riducibile a quella umana, bensì va oltre quella umana incorporata dalla macchina. In tal senso, il termine preterintenzionalità sembra raccogliere in sé due istanze divergenti di cui le espressioni intenzionalità tecnologica e composita non rendono perfettamente ragione. In primo luogo, tale termine, in virtù del prefisso *preter*, ha il merito di mettere in luce come il ruolo direttivo svolto dai sistemi intelligenti nell'esperienza umana non sia completamente riducibile all'intenzione umana che l'ha posto in essere, bensì, per l'appunto, la ecceda pur essendole eziologicamente legato. In secondo luogo, l'"andare oltre" – secondo il significato di preterintenzionale – "senza intenzione o consapevolezza" sottolinea come il sistema intelligente non abbia un'intenzione *cosciente* di agire e il suo agire non possa essere ridotto alla mera intenzione

24 Per un'analisi di tali critiche, che si basano su alcuni fraintendimenti dovuti alla terminologia impiegata da Verbeek, si veda Redaelli 2022.

25 Il codice penale italiano contiene disposizioni sulla colpevolezza e menziona nell'articolo 42 la responsabilità per dolo, per colpa e per delitto preterintenzionale. Per quanto riguarda l'ambito giuridico, l'impiego del termine preterintenzionalità solleva la questione di attribuzione della responsabilità sia agli agenti artificiali sia ai loro creatori umani. Sebbene la questione non sia affrontata in questo lavoro, importanti considerazioni sono svolte da Faroldi 2021.

dell'essere umano. Difatti, se si considerasse il sistema intelligente e l'essere umano come un unico attante, per usare la terminologia di Latour (2005), si potrebbe evidentemente riconoscere un'azione congiunta che è tanto intenzionale (da parte dell'uomo) quanto non-intenzionale (da parte della macchina). In questo senso, tale andare oltre le intenzioni dell'essere umano da parte della IA assume un chiaro significato: esso mette in luce il carattere non predeterminato dell'agire dei sistemi dotati di intelligenza artificiale; un agire che tuttavia non è un agire cosciente della macchina – non è quindi dettato da un'intenzione ad agire di cui le macchine sono sprovviste – e ciò nonostante raccoglie al suo interno l'agire cosciente dell'uomo. In altre parole ancora, il termine preterintenzionale sembra tenere conto del carattere non interamente predeterminato dell'intelligenza artificiale e così raccoglie in sé tanto l'intendere umano quanto il non-intendere della tecnologia, che oltrepassa l'intenzione umana, pur essendole eziologicamente legata. Difatti, i sistemi intelligenti di ultima generazione, grazie ai meccanismi di *machine learning*, sono in grado di modificare il proprio comportamento nell'interazione con l'ambiente, per cui il loro comportamento non è completamente programmato e prevedibile.

Di fronte ai due poli dell'intendere e del non-intendere, si pone dunque il preterintenzionale, con il cui termine si vuole fare spazio entro i sistemi intelligenti al riconoscimento di un'intenzionalità tecnologica che non è mera estensione dell'umano, come voleva Johnson, e che neppure si può ricondurre, come fa Sullins, a una qualche nozione psicologica, seppur debole, di intenzione ad agire, evitando, allo stesso tempo, le ambiguità che avvolgono il termine intenzionalità così come usato da Verbeek.

Bibliografia

Askell, A., Ganguli, D., *et al.*
2023 *The capacity for moral self-correction in large language models*, arXiv:2302.07459v2

Burrell, J.
2016 *How the machine 'thinks': understanding opacity in machine learning algorithms*, in "Big Data Soc", 3(1):1-12. https://doi.org/10.1177/2053951715622512

Coeckelbergh, M.
2020 *Introduction to Philosophy of Technology*, Oxford University Press, New York.
2020a *AI Ethics*, The MIT Press, Cambridge (MA).
2014 *The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics*, in "Philosophy & Technology", vol. 27, pp. 61-77. https://doi.org/10.1007/s13347-013-0133-8

Daza, M.T., Ilozumba, U. J.
2022 *A survey of AI ethics in business literature: Maps and trends between 2000 and 2021*, in "Frontiers in Psychology", vol. 13. https://doi.org/10.3389/fpsyg.2022.1042661

Di Martino, C.
2017 *Viventi umani e non umani. Tecnica, linguaggio, memoria*, Raffaello Cortina, Milano.

Faroldi, F.L.G.
2021 *Considerazioni filosofiche sullo statuto normativo di agenti artificiali superintelligenti*, in "Revista Iustitia", n. 9.

Floridi, L., Sanders, J.W.
2004 *On the Morality of Artificial Agents*, in "Minds and Machines", vol. 14, pp. 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Gunkel, D.J.
2018 *Robot Rights*, The MIT Press, Cambridge (MA).
2017 *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*, The MIT Press, Cambridge (MA).

Ihde, D.
1993 *Postphenomenology: Essays in the Postmodern Context*, Northwestern University Press, Evanston.
1990 *Technology and the Lifeworld: From Garden to Earth*, Indiana University Press, Bloomington.
1979 *Technics and Praxis: A Philosophy of Technology*, Reidel, Dordrecht.

Johnson, D.G.
2006 *Computer systems: Moral entities but not moral agents*, in "Ethics and Information Technology", vol. 8, pp. 195-204. https://doi.org/10.1007/s10676-006-9111-5

Johnson, D.G., Miller, K.W.
2008 *Un-making artificial moral agents*, in "Ethics and Information Technology", vol. 10, pp. 123-133. https://doi.org/10.1007/s10676-008-9174-6

Johnson, D.G., Powers, T.
2008 *Computers as surrogate agents*, in J. van den Hoven, J. Weckert (eds.), *Information technology and moral philosophy*, Cambridge University Press, Cambridge, pp. 251-269.

Johnson, D.G., Noorman, M.
2014 *Artefactual agency and artefactual moral agency*, in P. Kroes, P.P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Springer, Dordrecht, pp. 143-158.

Latour, B.
2005 *Reassembling the Social: An Introduction to Actor-Network-Theory*, Oxford University Press, New York.
2002 *Morality and Technology: The End of the Means*, in "Theory, Culture and Society", vol. 19, pp. 247-260.
1994 *On Technical Mediation: Philosophy, Sociology, Genealogy*, in "Common Knowledge", vol. 3, n. 2, pp. 29-64.
1993 *Nous n'avons jamais été modernes*, La Découverte, Paris; tr. eng. by C. Porter, *We Have Never Been Modern*, Harvard University Press, Cambridge (MA).

Llorca Albareda, J., García, P., Lara, F.
2023 *The Moral Status of AI Entities*, in F. Lara, J. Deckers (eds.), *Ethics of Artificial Intelligence*, in "The International Library of Ethics, Law and Technology", vol. 41, Springer, Cham, pp. 59-83. https://doi.org/10.1007/978-3-031-48135-2_4

Loh, W., Loh, J.
2017 *Autonomy and Responsibility in Hybrid Systems: The Example of Autonomous Cars*, in P. Lin, K. Abney, R. Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* Oxford University Press, New York, pp. 35-50.
https://doi.org/10.1093/oso/9780190652951.003.0003

Millar, J.
2017 *Ethics Settings for Autonomous Vehicles*, In P. Lin, K. Abney, R. Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Oxford University Press, New York, pp. 20-34. https://doi.org/10.1093/oso/9780190652951.003.0002

Mykhailov, D., Liberati, N.
2023 *A study of technological intentionality in C++ and generative adversarial model: phenomenological and postphenomenological perspectives*, in "Found Sci", vol. 28, pp. 841-857. https://doi.org/10.1007/s10699-022-09833-5

Mosakas, K.
2021 *On the moral status of social robots: considering the consciousness criterion*, in "AI & Society", vol. 36, n. 2, pp. 429-443.

Pacileo, F.
2020 *L'uomo al centro. IA tra etica e diritto nella responsabilità d'impresa*, in M. Bertolaso, G. Lo Storto (a cura di), *Etica Digitale. Verità, responsabilità e fiducia nell'era delle macchine intelligenti*, Luiss University Press, Roma, pp. 83-99.

Pasquale, F.
2015 *The Black Box Society. The Secret Algorithms That Control Money and Information*, Harvard University Press, Cambridge (MA).

Peterson, M. Spahn., A.
2011 *Can Technological Artefacts Be Moral Agents?*, in "Sci Eng Ethics", vol. 17, pp. 411-424. https://doi.org/10.1007/s11948-010-9241-3

Redaelli, R.
2024 *Intentionality gap and preter-intentionality in generative artificial intelligence*, in "AI & Society", https://doi.org/10.1007/s00146-024-02007-w
2023 *Different approaches to the moral status of AI: a comparative analysis of paradigmatic trends in Science and Technology Studies*, in "Discover Artificial Intelligence", vol. 3, n. 25 https://doi.org/10.1007/s44163-023-00076-2
2022 *Composite intentionality and responsibility for an ethics of artificial intelligence*, in "Scenari", n. 17, pp. 159-176.

Rosenberger, R., Verbeek, P.P.
2015 *A Field Guide to Postphenomenology*, in R. Rosenberger, P.P. Verbeek (eds.), *Postphenomenological Investigations: Essays on Human-Technology Relations*, Lexington Books, Lanham (MD), pp. 9-41.

Searle J.R.
1992 *The Rediscovery of the Mind*, The MIT Press, Cambridge (MA).

Sparrow, R.
2016 *Robots in aged care: A dystopian future?*, in "AI and Society", vol. 31, n. 4, pp. 445-454. https://doi.org/10.1007/s00146-015-0625-4

Sparrow, R., Sparrow, L.
2006 *In the hands of machines? The future of aged care*, in "Minds & Machines", vol. 16, pp. 141-161. https://doi.org/10.1007/s11023-006-9030-6

Sullins, J.P.
2006 *When is a Robot a Moral Agent?*, in "International Review of Information Ethics", vol. 6, pp. 23-30.

Terzidis, K., Fabrocini, F., Lee, H.
2023 *Unintentional intentionality: art and design in the age of artificial intelligence*, in "AI & Society", vol. 38, pp. 1715-1724. https://doi.org/10.1007/s00146-021-01378-8

van de Poel, I.
2020 Embedding Values in Artificial Intelligence (AI) Systems, "Minds & Machines", vol. 30, pp. 385-409. https://doi.org/10.1007/s11023-020-09537-4

Vanderelst, D, Winfield, A.
2018 *The dark side of ethical robots*, in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. pp. 317-322. https://doi.org/10.1145/3278721.3278726

Verbeek, P.P.
2014 *Some Misunderstandings About the Moral Significance of Technology*, in P. Kroes, P.P. Verbeek (eds.), *The Moral Status of Technical Artefacts*, Springer, Berlin, pp. 75-88.
2011 *Moralizing Technology: Understanding and Designing the Morality of Things*, University of Chicago Press, Chicago.
2008 *Cyborg intentionality: Rethinking the phenomenology of human technology relations*, in "Phenomenology and the Cognitive Sciences", vol. 7, n. 3 pp. 387-395.
https://doi.org/10.1007/s11097-008-9099-x
2008a *Obstetric Ultrasound and the Technological Mediation of Morality: A Postphenomenological Analysis*, in "Human Studies", vol. 31, pp. 11-26.
https://doi.org/10.1007/s10746-007-9079-0
2005 *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State University Press, University Park (PA).

Wallach, W., Allen, C.
2009 *Moral Machines. Teaching Robots Right from Wrong*, Oxford University Press, New York.

Wieluch, S., Schwenker, F.
2019 *Dropout Induced Noise for Co-Creative GAN Systems*, in IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), pp. 3137-3140, doi: 10.1109/ICCVW.2019.00383.