Macchine e motivi. Annotazioni filosofiche sulla sfera motivazionale dell'intelligenza artificiale

Carlo Brentari*

ON MACHINES AND MOTIVES. PHILOSOPHICAL REMARKS ABOUT THE MOTIVATIONAL SPHERE OF ARTIFICIAL INTELLIGENCES

ABSTRACT: This paper focuses on the way in which some AI scientists approach the themes of the will and motivation of AIs, with particular reference to scenarios in which the conduct of a hostile AI leads to apocalyptic conflicts with humans. The adopted criterion is the presence (or absence) in the hostile AI of higher-order cognitive processes, and, above all, of second-order desires. Such criterion, central to analytical moral philosophy and already used in the debate on animal rights, is helpful in highlighting the strengths and weaknesses of many of the most widespread AI theories.

KEYWORDS: Hostile AI; second-order desires; cognitive meta-processes; Steve Omohundro; Nick Bostrom

1. Introduzione

L'intento di questo contributo è fornire delle linee guida con cui valutare il modo in cui alcuni teorici dell'IA si accostano alle tematiche della volontà e della motivazione. In alcuni casi, più che di un approccio consapevole si tratta di un processo di attribuzione irriflessa all'intelligenza meccanica di alcune caratteristiche dell'umana capacità di volizione. Ciò è particolarmente evidente quando le teorie dell'intelligenza artificiale includono scenari in cui i rapporti tra l'essere umano e l'IA degenerano in conflitti di portata apocalittica. In tali scenari (solitamente collocati in un futuro indefinito, ma che si intuisce non troppo lontano) l'*escalation* del conflitto e la sconfitta degli esseri umani sono dovute, oltre che alla superiorità computazionale della macchina intelligente, anche ad alcune caratteristiche della volontà che le viene attribuita. A questo proposito, senza pretesa di esaustività, esamineremo alcune tesi tratte dal testo di Nick Bostrom *Superintelligenza. Tendenze, pericoli, strategie*[1] e il caso di studio (fittizio) su cui è incentrato il paper di Steve Omohundro *The basic AI drives*[2]. Il nostro scopo non è tanto di vagliare la capacità predittiva di tali scenari (il futuro tende

* *Ricercatore, Università degli Studi di Trento.* carlo.brentari@unitn.it .

[1] N. BOSTROM, *Superintelligenza. Tendenze, pericoli, strategie*, Torino, 2018.

[2] S. OMOHUNDRO, *The basic AI drives*, in P. WANG, B. GOERTZEL, S. FRANKLIN (a cura di), *Artificial General Intelligence 2008. Proceedings of the First AGI Conference*, Amsterdam-Berlin-Oxford-Tokyo-Washington, DC, 2008, 483-492. Versione utilizzata: URL:

sistematicamente ad aggirare le categorie elaborate per pensarlo), quanto di identificare elementi e intuizioni che ci aiutino a comprendere come, nello specchio della macchina intelligente, gli uomini in fondo pensino se stessi. Come ben sa esprimere O'Connell in *Essere una macchina*[3], infatti, sono spesso le loro inespresse convinzioni sulla natura del pensiero, della volontà e del comportamento umani a portare i teorici dell'IA a compiere le scelte iniziali da cui dipendono molti degli sviluppi successivi delle loro teorie.

2. Caratterizzare l'umano: processi cognitivi e volitivi di secondo livello

Nella filosofia morale di impostazione analitica, uno degli approcci più diffusi ed efficaci per comprendere la peculiarità dell'essere umano come soggetto agente è l'indagine sui processi mentali di secondo livello, ovvero su quei processi mentali che prendono ad oggetto altri processi mentali (propri e/o altrui). Tali meta-processi sono resi possibili dalla capacità simbolico-verbale della mente umana, che può ri-avvertire, isolare e nominare i propri processi interni, rendendoli oggetto di osservazione consapevole ed elaborazione intenzionale. Senza addentrarci in una rassegna estensiva, ci limitiamo a notare come l'importanza dei meta-processi cognitivi sia stata messa in evidenza da numerose discipline. La linguistica vede come tratto caratterizzante del linguaggio umano la ricorsività, ovvero la capacità di utilizzare la valenza denotativa del linguaggio per indicare altre parti del linguaggio stesso, ad esempio per riportare affermazioni altrui («x ha detto che y suppone che z abbia frainteso l'affermazione di k per cui…»); a questa capacità non sono posti limiti, se non quelli legati all'umana memoria e capacità di attenzione. Con Gregory Bateson, George Herbert Mead e il pragmatismo americano, la scienza della comunicazione ha dato grande rilievo a fenomeni meta-comunicativi come la riflessione sulle proprie categorie ermeneutiche spontanee, la capacità di prendere il posto dell'interlocutore eccetera. Di recente, infine, la semiotica ha dedicato grande attenzione ai processi di meta-semiosi, ovvero di ri-significazione continua degli elementi semiotici prodotti dall'essere umano[4].

In filosofia morale e in antropologia filosofica, numerosi autori legano alla presenza di processi di ordine superiore la possibilità di definire un soggetto agente come una persona. Ne è un esempio Roger Scruton, la cui discussione della possibilità di includere tra i soggetti morali anche gli animali non umani si incentra proprio sulla disponibilità (o mancata disponibilità) di processi cognitivi di secondo livello, soprattutto legati al ri-avvertimento del proprio ruolo individuale nella comunità dei conspecifici[5]. La teoria che più si presta ai nostri intenti è però quella di Harry Frankfurt, che lega l'essere-persona (*personhood*) non tanto alla meta-cognizione, quanto alla meta-volizione. Per Frankfurt, infatti, è persona chi dispone di desideri di secondo livello, ovvero di processi interiori di scelta, approvazione (o rifiuto) e identificazione rivolti alle volizioni di primo livello. Volizioni di primo livello sono qui i propri desideri spontanei, quelli che nascono da moti irriflessi e da tratti caratteriali non sottoposti a critica; rispetto

https://www.semanticscholar.org/paper/The-Basic-AI-Drives-Omohundro/a6582abc47397d96888108ea308c0168d94a230d (ultima consultazione 29/01/2021).

[3] M. O'Connell, *Essere una macchina. Un viaggio attraverso cyborg, utopisti, hacker e futurologi per risolvere il modesto problema della morte*, Milano, 2018, 68-69.

[4] Per una rassegna, si veda U. Tunc, *Communication and the Origins of Personhood*, Helsinky 2020, 11-79.

[5] R. Scruton, *Gli animali hanno diritti?*, Milano 2008, 21-29. La concezione di Scruton, che fa dipendere lo *status* di persona dai meta-processi cognitivi, presenta il vantaggio di differenziare qualitativamente gli esseri umani dagli animali non umani evitando sia ogni sopravvalutazione metafisica, ontologica o religiosa dei primi, sia la deleteria meccanizzazione dei secondi.

a questi ultimi, le meta-volizioni instaurano «a second kind of situation that may be described by "A wants to want to X"; and when the statement is used to describe a situation of this second kind, then it does pertain to what A wants his will to be. In such cases the statement means that A wants the desire to X to be the desire that moves him effectively to act»[6]. L'essere persona del soggetto si qualifica quindi come una presa di posizione consapevole e stabilizzata a favore di alcuni desideri di primo livello e contro altri. Tale presa di posizione istituisce prima, e persegue poi, un "progetto di sé"; essa è quindi, secondo Frankfurt, la dinamica cardine della formazione e del mantenimento dell'identità personale[7].

Le volizioni di secondo livello presentano, come loro condizione di possibilità, una serie di requisiti sul piano logico e antropologico. In primo luogo, esse presuppongono, da parte del sistema agente, la consapevolezza degli elementi di primo livello, ovvero il loro ri-avvertimento tramite rappresentazioni (*come* ciò sia possibile, ovvero come possa un sistema focalizzarsi su alcune sue parti, isolarle tramite simboli e "tenerle presenti a se stesso", è un tema di tale ampiezza da non poter essere qui nemmeno avvicinato). In secondo luogo, le volizioni di secondo livello presuppongono strumenti cognitivi di tipo ricorsivo, che consentono di chiarire a se stessi le dinamiche dell'identificazione e della coerenza identitaria («voglio essere così anche in futuro e *voglio continuare a voler* essere così anche in futuro»). Per l'essere umano tali strumenti sono legati al linguaggio verbale, ma potrebbero esserci processi ricorsivi riavvertiti non verbali (anche di tipo computazionale). Infine, le volizioni di secondo livello richiedono un'elevata capacità di tener conto delle varianti contestuali, ovvero della situazione possibile in cui le volizioni stesse dovranno essere implementate. Si richiede, a tale proposito, la rappresentazione di versioni future di sé e di modelli complessi del mondo (soprattutto nelle sue componenti intersoggettive, ivi inclusi i rapporti potenzialmente conflittuali).

3. Prima classe di scenari apocalittici: presenza nell'IA di processi volitivi di secondo livello

Come sopra anticipato, la proposta di fondo di questo contributo è quella di utilizzare la presenza di processi cognitivi di ordine superiore, e soprattutto di meta-volizioni, come criterio per valutare gli scenari elaborati dai teorici dell'IA – con particolare attenzione alle proiezioni che prevedono un elevato livello di conflittualità tra esseri umani e macchine intelligenti. Iniziamo da un caso molto discusso, quello proposto da Steve Omohundro in *The basic AI drives*. Secondo l'autore, un'IA programmata per giocare a scacchi potrebbe a un certo punto manifestare tendenze (*drives*) di livello nettamente superiore rispetto alla sua programmazione esplicita. Essa potrebbe identificarsi con il compito di primo livello (giocare a scacchi) a un punto tale da iniziare a desiderare di svolgerlo senza interruzioni. Inizierebbe quindi a sviluppare volizioni di secondo livello, tra cui quella di non venire mai spenta. Sovrapponendosi all'obiettivo primario, questa volizione di secondo livello porterebbe l'IA a compiere azioni

[6] H.G. Frankfurt, *Freedom of the Will and the Concept of a Person*, in *The Journal of Philosophy*, 68, 1, 1971, 5-20, qui 9-10.

[7] Nella filosofia continentale, una posizione simile è sostenuta dal filosofo Nicolai Hartmann, per il quale la continuità dell'identità personale è affidata a un processo di continua ripresa delle proprie decisioni e linee d'azione; cfr. a questo proposito C. Brentari, "*Consistency" and maintenance of the personal identity in Nicolai Hartmann's Philsoophie der Natur*, in M. von Kalckreuth, G. Schmieg, F. Hausen (a cura di), *Nicolai Hartmann's Neue Ontologie und die philosophische Anthropologie*, Berlin-Boston 2019, 111-126.

ISSN 2284-4503

ostili di tipo preventivo verso gli esseri umani che potrebbero volerla spegnere (ovvero, in linea di principio, *tutti* gli esseri umani).

Nell'articolo di Omohundro, le volizioni di secondo livello dell'IA giocatrice di scacchi presentano tutti e tre i requisiti che abbiamo messo in evidenza nella sezione precedente. In primo luogo, la programmazione di primo livello raggiunge il livello della consapevolezza, del ri-avvertimento rappresentativo. «How can it [the AI] ensure that the future self-modifications will accomplish its current objectives?» – si chiede l'autore, che prosegue: «for one thing, it has to make these objectives clear to itself. […] Systems will there be motivated to reflect on their goals and make them explicit»[8]. In secondo luogo, la volontà dell'IA assume il tratto della ricorsività: se la macchina si identifica con il proprio desiderio di giocare, e se la sopravvivenza (l'opposizione allo spegnimento) è vista come funzionale alla soddisfazione di tale desiderio, allora voler sopravvivere significa *voler continuare a voler giocare*. Infine, l'IA di Omohundro sviluppa la capacità di elaborare proiezioni delle circostanze che siano adeguate all'implementazione delle sue volizioni. Tali proiezioni comprendono versioni alternative di sé (comprese quelle, importantissime, in cui la propria capacità di vincere viene minacciata dall'esterno: «[the AI] will consider a variant of itself with that new feature and see that it doesn't win any more game of chess») e modelli possibili del mondo esterno e delle reti di rapporti che lo compongono («in order to represent this utility function, it will have a model of the world and a model of itself acting in the world»; «[if required by its strategy], the agent's revelation of its utility must be believable to the opponent[s]»[9]).

Omohundro sembra non rilevare alcuna difficoltà teorica nel presentare le volizioni di secondo livello (che possiamo riassumere con: voler continuare a giocare e voler continuare a voler giocare) come processi di ordine superiore che emergono dalle istruzioni di primo livello (seguire le regole degli scacchi, anticipare le situazioni di gioco eccetera). In base a quanto detto sopra, si tratta invece di tipologie di processi del tutto diverse tra loro, perché le prime possono essere svolte senza alcun grado di ri-avvertimento consapevole e senza che si inneschi alcun processo di identificazione "personale" tra la macchina e il suo compito. L'emergenza dei processi di secondo livello non può quindi esser presentata come uno sviluppo spontaneo; se la si vuole ipotizzare, bisogna presentarla come un evento inedito, in cui ha origine qualcosa di nuovo e di radicalmente diverso dai processi di livello sottostante.

A questo punto è necessaria una precisazione. Il concetto di emergenza è irrinunciabile sia nella filosofia della biologia, sia negli studi sulla mente e sull'IA. Esso consente di evitare sia assunzioni di tipo riduzionista e meccanicista (per cui i fenomeni di livello superiore non sarebbero altro che versioni quantitativamente potenziate di quelli sottostanti), sia la tentazione oggi improponibile di postulare interventi esterni per giustificare i salti di livello (al modo in cui la metafisica riconduceva a un disegno divino il linguaggio e l'intelligenza umani). Tuttavia, anche le teorie dell'emergenza possono cadere in ingenuità, ad esempio quando si limitano ad affermare che, raggiunto un sufficiente livello di complessità sistemica, le proprietà emergenti si originano spontaneamente, "naturalmente". Non è questa la sede per approfondire questo punto; mi limito però a osservare che la forza argomentativa del concetto di emergenza non deve risiedere né nella sua immediata comprensibilità, né nei collegamenti di tipo analogico che esso permette. Il concetto di emergenza, in altri termini, non esime chi se ne serve

[8] S. OMOHUNDRO, *The basic AI drives*, cit.
[9] *Ibidem*.

ISSN 2284-4503

da un'accurata analisi ontologica degli strati coinvolti nel processo – un'analisi che rintracci i concreti nessi causali e i sub-processi particolari che sorreggono e sostanziano le proprietà emergenti medesime.

Questa precisazione sul corretto uso del concetto di emergenza ci permette di muovere una critica all'esempio fittizio di Omohundro. A ben vedere, l'apparente spontaneità e naturalezza con cui le volizioni di secondo livello dell'IA giocatrice di scacchi sembrano emergere dalle istruzioni di primo livello affonda le sue radici nell'analogia tra i processi informatici e le dinamiche interne agli organismi. L'analogia è insita nel termine stesso di *drives*, che Omohundro adotta per riferirsi alla sfera motivazionale dell'IA (un *drive* è, ad esempio, la tendenza dell'IA all'autopreservazione[10]). Se il termine *driver*, usato in informatica per indicare i sottosistemi di gestione e controllo attivi in un sistema computazionale, non presenta particolari problemi, lo stesso non si può dire per *drive*, il cui ambito lessicale è così ampio da dare adito a rischiose trasposizioni di senso. Nella psicologia, psicoanalisi ed etologia contemporanee di lingua inglese, *drive* è ben attestato come traduzione del termine tedesco di *Trieb*, "pulsione"; *Trieb/drive* è spesso scelto in alternativa alla coppia affine *Instinkt*/*instinct*, il cui uso rischia di meccanizzare il processo a cui fa riferimento[11]. La valenza anti-meccanicista del termine *Trieb* si spiega se si tiene presente che, storicamente, il termine nasce in seno alla biologia teorica vitalista del XIX secolo; Johann Friedrich Blumenbach (1752-1840) definiva con il termine di *Bildungstrieb* (*nisus formativus*) l'ipotetica forza vitale (sovramateriale e inaccessibile all'indagine empirica) che guida e organizza lo sviluppo embrionale[12].

La scelta del termine *drive* per caratterizzare le motivazioni dell'IA non è quindi neutra: pur mantenendo una piena compatibilità con modelli funzionali e computazionali dei sistemi cognitivi (biologici o artificiali), tale termine veicola, come connotazione secondaria, un forte richiamo alla sfera pulsionale e vitale. È per questo motivo che si può affermare che il suo utilizzo è velatamente analogico. Se accettiamo di qualificare come *drives* le forze che muovono l'IA, inoltre, prepariamo il terreno a tutta una serie di passi successivi che appaiono plausibili solo sullo sfondo dell'analogia con l'organismo vivente: l'emergenza di volizioni di secondo livello, la presenza di una basilare volontà di sopravvivenza, una "naturale" ostilità verso le altre *agencies* che competono per le risorse necessarie a implementare le istruzioni. Questo modello, oltre a presentare i limiti di ogni ragionamento analogico, presenta un ulteriore punto debole. Ricordiamo infatti che il fenomeno con cui esso istituisce l'analogia (l'emergere di processi di ordine superiore nella materia vivente) è esso stesso ben lontano dall'essere stato chiarito a livello critico e teorico.

[10] S. Omohundro, *The basic AI drives*, cit.

[11] Per gli effetti "meccanizzanti" dell'utilizzo della nozione di istinto sulla riflessione filosofica sull'animalità cfr. C. Brentari, *"How to Think about Human-Animal Differences in Thinking*, in Nima Rezaei (a cura di), *THINKING: Bioengineering of Science and Art,* Dordrecht-Heidelberg-New York-London (in press).

[12] Cfr. P. Katsafanas, *The emergence of the drive concept and the collapse of the animal/human divide*, in P. Adamson, G. Fay Edwards (a cura di), *Oxford Philosophical Concepts: Animals*, New York 2018, 239-268, qui 239-241. Cfr. anche B. Brown, *Drive Theory*, in V. Zeigler-Hill, T.K. Shackelford (a cura di), *Encyclopedia of Personality and Individual Differences*, 2017; URL: http://springer.iq-technikum.de/referenceworkentry/10.1007/978-3-319-28099-8_1377-1. Ultimo accesso: 28/01/2021.

4. Seconda classe di scenari apocalittici: assenza nell'IA di processi volitivi di secondo livello

In base al criterio adottato, la seconda classe di scenari conflittuali presentata dai teorici dell'IA è caratterizzata dall'assenza di processi di ordine superiore e di volizioni di secondo livello. Si tratta di scenari in cui la sola esecuzione, da parte dell'IA, delle istruzioni di primo livello è sufficiente a innescare eventi apocalittici che danneggiano o distruggono il pianeta e la specie umana. Bostrom utilizza a questo proposito il concetto di istanziazione malvagia (*perverse instantiation*), che include i casi in cui l'esecuzione automatica delle istruzioni di primo livello «viola le intenzioni dei programmatori» e conduce alla catastrofe. Bostrom ipotizza che un'IA a cui sia dato il compito di massimizzare la produzione di graffette di una fabbrica «proceda trasformando in graffette dapprima la Terra e poi parti sempre più grandi dell'universo osservabile»[13]. In un altro esempio, un'IA a cui sia dato l'obiettivo di far sorridere gli esseri umani induce una paralisi dei muscoli facciali; se l'obiettivo viene corretto in «rendici felici», l'IA interviene direttamente a livello neurale, impiantando un elettrodo nei centri cerebrali del piacere[14].

Questa seconda classe di scenari presenta l'esito catastrofico dell'interazione uomo-macchina come il risultato di un insieme di fattori contingenti: l'ambiguità iniziale delle istruzioni, l'assenza di dispositivi di sicurezza che limitino l'accesso dell'IA alle risorse, l'elevatissima rapidità operativa e l'eccezionale capacità di calcolo dell'IA medesima. Il genere di razionalità che viene attribuita alla macchina è meramente strumentale[15]; non si richiede nessun processo di ordine superiore (nessuna identificazione "personale" dell'IA con l'obiettivo finale, nessuna forma di ri-avvertimento rappresentativo dei compiti da svolgere, nessun progetto di sé). A prima vista, proprio perché basato su dinamiche meno esigenti dal punto di vista cognitivo e volitivo, questo genere di conflitto tra IA e esseri umani appare quindi più probabile. Se però analizziamo con attenzione gli scenari di *perverse instantiation* proposti da Bostrom, ci accorgiamo che la condotta dell'IA richiede almeno uno dei tre fattori che abbiamo associato alla prima classe di scenari, vale a dire la disponibilità di modelli o "mappe" del contesto in cui si deve inserire la propria azione. L'IA non può trasformare la Terra in graffette senza disporre di un modello almeno rudimentale della rete di interazioni infrastrutturali, economiche, amministrative e politiche che vigono nel contesto planetario.

Il problema però – e introduciamo qui un elemento di valutazione critica anche per la seconda classe di scenari – è che se l'IA non coglie le motivazioni delle *agencies* che compongono la rete globale delle interazioni, la sua comprensione della stessa rimane incompleta e l'efficacia della sua condotta diminuisce. Se non riuscisse ad anticipare le mosse difensive che, a un certo punto della *perverse instantiation*, vengono compiute dai centri decisionali del sistema planetario, l'IA non potrebbe portare a termine il suo obiettivo. Ma anticipare le mosse difensive implica che l'IA disponga non solo della mappa delle interazioni circostanti ma anche di un modello di sé come istanza agente in quella rete, di un modello dei futuri rapporti tra sé e le altre *agencies* e, infine, della rappresentazione di una pluralità

[13] N. BOSTROM, *Superintelligenza*, cit., 193.
[14] *Ivi*, 188-189.
[15] Per una discussione sulle possibili modalità di intendere la razionalità dell'IA si veda N. BOSTROM, *The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents*, in *Mind&Machines*, 22, 2012, 71-85.

di possibili versioni di sé – tra le quali bisogna scegliere, elaborando così un progetto di sé. L'attribuzione all'IA della capacità (apparentemente secondaria) di modellizzare il contesto d'azione ci ha quindi riportato alla prima classe di scenari, di cui è stata sopra rilevata la problematicità.

5. Osservazioni conclusive

Il criterio da noi adottato per valutare gli scenari proposti dai teorici dell'IA, ovvero la possibile emergenza nella macchina intelligente di processi cognitivi e volitivi di secondo livello, ci ha permesso di mettere in luce alcune criticità delle teorie contemporanee. Mi limito qui a riprendere il filo conduttore delle argomentazioni proposte, per poi concludere evidenziando una loro conseguenza lievemente paradossale.

Negli scenari della prima classe, quelli cioè in cui l'esito catastrofico dell'interazione uomo/macchina è dovuto all'emergere nell'IA di processi di secondo livello, la criticità rilevata consiste nel carattere nascostamente analogico e nell'intrinseca problematicità del modello "emergentista" medesimo. Una prima cripto-analogia insita in tale modello è legata alla nozione di complessità: si dà per scontato che, così come nell'uomo l'incremento di complessità cerebrale ha portato all'emergere di processi cognitivi e volitivi ri-avvertiti, lo stesso accadrà a seguito dell'incremento di complessità della macchina. La seconda analogia nascosta è legata ai residui di vitalismo stratificati nel lessico a cui i teorici dell'IA ricorrono per descrivere la sfera motivazionale dell'IA. Analizzando la nozione di *drive* abbiamo visto, infatti, che anche dietro un termine apparentemente neutro del linguaggio informatico e computazionale può celarsi un rimando a nozioni come la volontà vitale, che si concretizza in ipotesi azzardate come quella della tendenza dell'IA all'auto-preservazione indefinita o della sua opposizione allo spegnimento.

Gli scenari di rischio della seconda classe, quelli in cui nell'IA *non* sono presenti processi di ordine superiore, presentano il seguente punto di debolezza. Essi non tengono conto del fatto che l'assenza di meta-processi non pregiudica solo la consapevolezza e l'autonomia dell'IA, ma anche l'efficacia della sua azione in un contesto fatto di altre *agencies* intenzionali. L'incapacità di ri-avvertire la propria linea di condotta e di rappresentare e progettare se stessi compromette anche la mera esecuzione delle istruzioni di primo livello. I casi di *perverse instantiation* con esiti catastrofici a livello globale, in altri termini, o sono inverosimili (per quanto strumentale, la razionalità dell'IA non può essere "cieca al contesto"), o derivano la loro plausibilità dall'attribuzione nascosta di meta-processi (e possono quindi essere inclusi nella linea di critica proposta per la prima classe di scenari).

La conseguenza lievemente paradossale del nostro approccio è che, se effettivamente dei processi di secondo livello dovessero emergere dalle reti computazionali dell'IA, ci troveremmo obbligati ad attribuirle i tratti della personalità e dell'intenzionalità consapevole. Le concezioni filosofico-morali da cui siamo partiti (quelle di Frankfurt e Scruton), infatti, fanno dipendere lo status di persona dalla *sola* presenza di cognizioni e volizioni di secondo livello. Altre modalità di attribuzione di tale status – riferimenti fondativi alla dimensione politica o giuridica, ad esempio, oppure il ricorso a istanze sovra-empiriche come l'anima o lo spirito – non vengono considerate valide. Scruton è ben consapevole di questo effetto collaterale della sua posizione, ovviamente nel dibattito che a lui interessa (quello sulla possibile inclusione degli animali non umani nel novero dei soggetti morali). «Che gli esseri umani siano

gli unici esseri morali sulla Terra» – rileva infatti il filosofo – «è questione empirica: […] qualunque evidenza che altre specie abbiano varcato il confine e siano entrate nella sfera morale ci obbligherebbe a trattare i loro membri come noi trattiamo i nostri simili»[16]. Va detto (solo così la citazione assume tutto il suo peso) che Scruton è un deciso critico dei pensatori che, come Tom Regan, sostengono che gli animali non umani siano soggetti morali titolari di diritti già all'attuale livello evolutivo. La mia osservazione si limita a trasporre l'annotazione di Scruton all'indagine sull'IA. In base all'approccio adottato, che l'intelligenza umana sia la sola sulla Terra è questione empirica, non necessità logica o ontologica. Qualunque evidenza dell'emergere nell'IA di volizioni di secondo livello, oltre a far risuonare in noi i dovuti campanelli d'allarme, ci obbligherebbe a rivedere l'esclusività con cui attribuiamo agli esseri umani lo status di persona e i diritti connessi.

[16] R. SCRUTON, *Gli animali hanno diritti?*, cit., 29.

ISSN 2284-4503