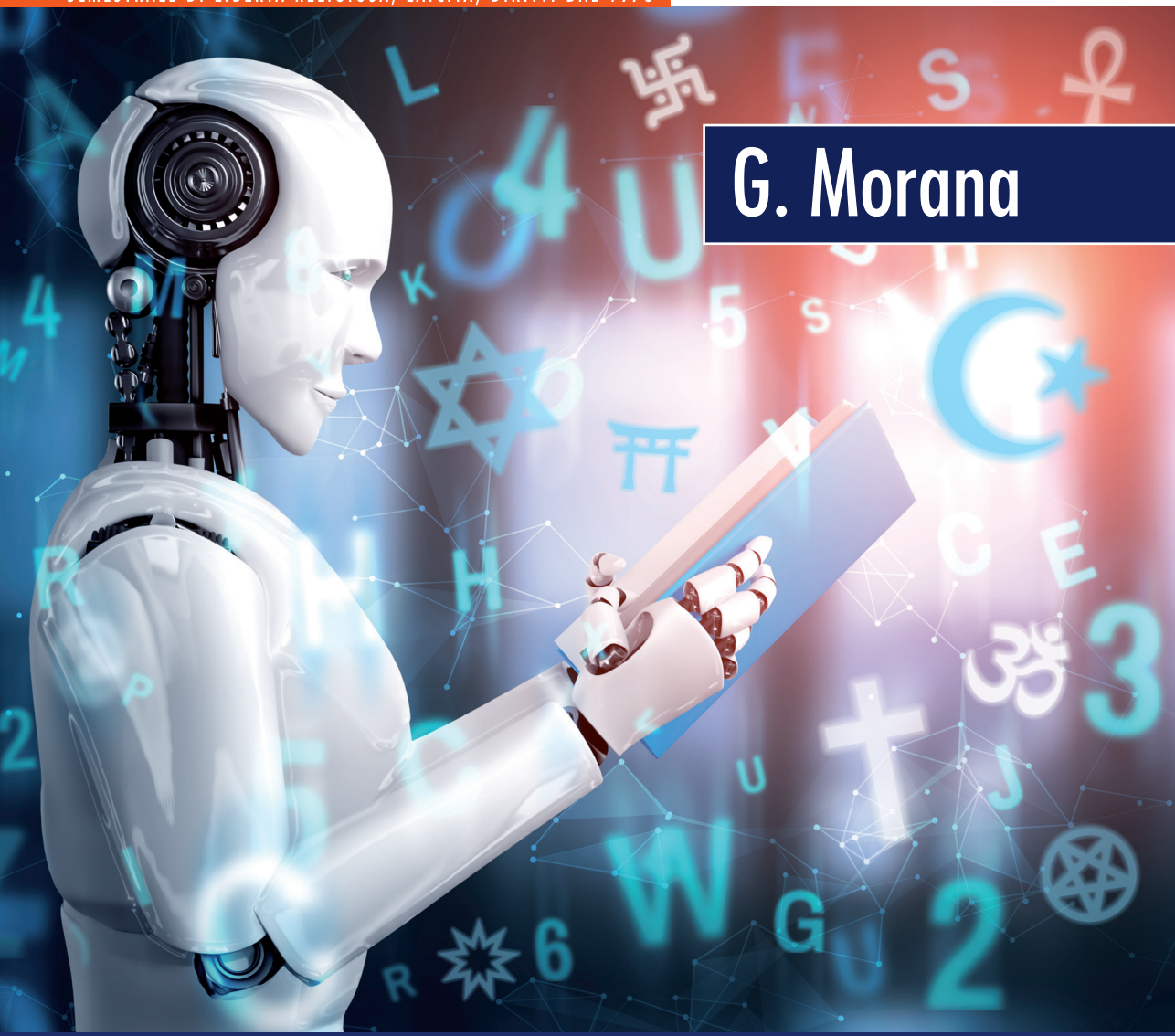
2024 NUMERO 68

Coscienza e Libertà

SEMESTRALE DI LIBERTÀ RELIGIOSA, LAICITÀ, DIRITTI DAL 1978



G. Morana

Diritto, Religioni e Intelligenza artificiale: quali prospettive?

ISSN 0394-2732

A. Casiere - G. Cimbalo
M. Croce - A. Cupri
L. Fregoli - E. Lipilini
M.L. Lo Giacco - G. Mobilio
G. Morana - F. Rescigno
D. Romano - G. Strada

Umano *versus* IA. Questioni di coscienza ed etica

Giuseppe Morana

Dottore magistrale in Scienze Filosofiche, Dipartimento di Filosofia e comunicazione FILCOM, Università degli studi di Bologna – Alma Mater Studiorum

ABSTRACT

L'IA sostituirà mai l'intelligenza umana? Le macchine avranno un giorno una 'coscienza'? Tale articolo tenterà di rispondere a queste domande sostenendo che la concezione dell'IA dipende dalla controparte dell'intelligenza umana. Si propone così di comprendere la *moral agency* e la responsabilità nell'IA, con l'obiettivo di evitare una morale e un diritto disumanizzanti.

SOMMARIO

1. Introduzione

L''Intelligenza artificiale'[1] rappresenta la nuova frontiera dello sviluppo tecnologico umano e, col suo enorme grado di automazione, porta con sé un

*Contributo selezionato dal Comitato Scientifico della rivista in relazione alla call "Diritto, Religioni e Intelligenza artificiale: quali prospettive?" del luglio 2024.

[1] Da ora in poi abbreviata con 'IA' (in lingua italiana) o con 'AI' (in lingua inglese). Di seguito alcune specificazioni. Un agente virtuale con IA è innanzitutto formato da un'architettura di base su cui si innesta un programma specifico. Per larga parte della storia dello sviluppo dell'intelligenza artificiale i sistemi di IA si sono divisi tra sistemi simbolici e connessionisti. I primi sarebbero basati su inferenze logiche e probabilistiche che permetterebbero di sintetizzare lunghe catene di ragionamento e di utilizzare la potenza espressiva delle rappresentazioni strutturate. I secondi, al contrario, sarebbero basati sulla minimizzazione della perdita di parametri non interpretati, puntando a riconoscere pattern anche in una mole di dati confusa. Tale divisione, tuttavia, sarà destinata ad essere superata con la più recente ricerca in merito (cfr. S.J. Russell, P. Norving, *Artificial Intelligence: a modern approach*, Prenctice Hall 2009[3], trad. Ita. *Intelligenza artificiale: un approccio moderno*, a cura di F. Amigoni, vol. 1, Pearson, Roma 2010, pp. 380, 381).

enorme potenziale di liberazione già intuito secoli fa anche da Aristotele[2].

L'IA, tuttavia, a differenza di altre tecnologie, se implementata scorrettamente porterebbe a mettere in discussione la supremazia dell'umanità sul mondo[3].

Togliendo la prerogativa al Dio biblico di creare qualcosa 'a propria immagine e somiglianza', l'uomo contemporaneo avrebbe posto la macchina di sua creazione come modello di sé stesso. Ciò ha portato a configurare i robot come strumenti di 'accompagnamento' dell'attività umana, più che di semplice 'affiancamento'. Tuttavia, fino a quando le categorie umane si possono applicare alla macchina e viceversa?[4] Come si cercherà di mostrare in questo elaborato, probabilmente molto poco.

2. Menti umane e computer

2.1 Tra Cibernetica e comportamentismo

Il modello della cibernetica di Wiener[5] ha costituito il programma di ricerca alla base del campo di studi sull'intelligenza artificiale. Questo studioso intendeva concepire i comportamenti degli animali (e dell'uomo) da un lato, e i processi delle macchine, dall'altro, come forme di comunicazione, in cui quest'ultima fosse ridotta ad una trasmissione di informazioni, tale da poterla sottoporre a computazione quantitativa[6].

La computazione, tuttavia, non indica un processo meccanico in cui avvengono trasferimenti di energia (come accade in natura), ma un processo meramente logico, formale e astratto in cui viene implementato un algoritmo in un computer[7].

[2] Aristotele, *Politica*, a cura di V. Costanzi, Gius. Laterza & Figli, Bari, 1925, I, 2, 1253b, p. 8.

[3] Basti pensare che già oggi la velocità di calcolo massima di 1 kg di macchina di calcolo (circa 10^{51} operazioni al secondo) è trilioni di volte più alta di quella del cervello umano. Cfr. S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., pp. 381-383.

[4] Cfr. A. Fabris, *Etica per le tecnologie …*, cit., pp. 74-76.

[5] Si rimanda alla lettura integrale del testo N. Wiener, *La cibernetica. Controllo e comunicazione nell'animale e nella macchina*, Il Saggiatore, Milano, 1968.

[6] Cfr. A. Fabris, *Etica per le tecnologie …*, cit., pp. 23, 85.

[7] Questa è sicuramente una differenza fondamentale tra cervelli umani e *hardware* della macchina. L'implementazione di un programma è solo un processo formale e astratto, al contrario delle connessioni neuronali che invece prevedono il trasferimento di energia elettrica, oltre a delle informazioni. Cfr. J.R. Searle, *Ventun anni dalla stanza cinese* in *Phylosophy for a new century*, cit., pp. 89, 103.

Questa definizione standard di computazione permette così di affermare due asserzioni: in primo luogo, per ogni oggetto c'è una qualche descrizione tale per cui in base a quella descrizione, l'oggetto è un computer digitale; in secondo luogo, per ogni programma c'è un qualche oggetto sufficientemente complesso tale per cui esiste una descrizione dell'oggetto sulla base della quale esso implementa il programma[8].

Tutto ciò comporterebbe l'irrilevanza della realizzazione dell'*hardware* per la descrizione computazionale. Un computer quindi, in linea teorica, potrebbe essere composto da qualsiasi materiale, quindi anche da cellule neuronali umane: la computazione compiuta dal computer umano e dal computer della macchina sarebbero equivalenti[9].

Proprio da ciò nascerà all'interno della psicologia comportamentista il modello computazionale della mente, in cui 'la mente sta al cervello come il programma (*software*) starebbe all'*hardware*'. Tale modello si inserisce all'interno di un programma di ricerca che si chiede se gli stati e i processi mentali del cervello siano 'almeno' computazionali e quindi assimilabili a quelli di un computer digitale[10].

Molti studiosi, usando dapprima la tesi di Church-Turing[11], e poi la tesi di Turing[12] hanno affermato la possibilità per cui qualsiasi cosa l'uomo possa fare algoritmicamente, essa potrebbe essere svolta da una 'macchina universale di Turing'. Ciò che distinguerebbe uomo e macchina sarebbe solo la modalità di questa operazione: conscia per il primo e inconscia per il secondo[13].

Più che chiedersi se le macchine fossero in grado di pensare, Turing pensò tuttavia di risolvere la questione domandandosi semplicemente se le macchine fossero in grado di superare un test comportamentale, il famoso 'test di Turing'[14]. Tale prova risolve l'intelligenza nella sua performatività esteriore,

[8] J.R. SEARLE, *Il cervello può considerarsi un computer?* in *Phylosophy for a new century*, cit., p. 116.
[9] Cfr. *Ivi*, pp. 113, 114.
[10] Cfr. *Ivi*, pp. 106-108.
[11] «Per ogni algoritmo esiste una macchina di Turing in grado di implementare quell'algoritmo».
[12] «Esiste una macchina universale di Turing in grado di simulare ogni macchina di Turing, e quindi di implementare qualsiasi algoritmo».
[13] Cfr. *Ivi*, pp. 109, 110.
[14] Il test, in sostanza, prevedeva la conversazione di almeno 5 minuti tra un programma computerizzato ed un esaminatore inconsapevole di star parlando con un programma. Se il programma è stato in grado di ingannare l'esaminatore per almeno il 30% del tempo della prova, facendosi passare per interlocutore umano, allora la macchina ha passato il test e può reputarsi apparentemente intelligente.

rispettando il fondamentale assioma della psicologia comportamentista[15].

Tuttavia, attestarsi al semplice test di Turing per misurare la fantomatica intelligenza delle macchine potrebbe risultare fuorviante e inefficace dato che tutt'oggi non c'è accordo univoco nella comunità degli esperti se tale test sia stato mai superato da qualsivoglia computer o IA[16].

Due sono le principali accuse mosse al test di Turing. La prima intacca in realtà tutta la psicologia comportamentista: si potrebbe comunque avere un certo comportamento esteriore senza alcun processo psichico interiore. In secondo luogo, classificare il risultato della macchina come un successo o un insuccesso empirico è un'operazione compiuta dall'uomo con parametri umani, e non dalla macchina[17]. La valutazione della *performance* presuppone sempre una metaetica, e questa è stabilita dal programmatore umano, in quanto è lui che sceglie i valori di riferimento, quale azione l'IA deve compiere e in che modo. La macchina quindi, a differenza dell'uomo, non ha di per sé interessi psicologicamente reali indipendenti[18].

2.2. 'IA forte' versus 'IA debole' ossia della coscienza umana

Al giorno d'oggi esistono due diversi programmi di ricerca sull'intelligenza artificiale, l''IA forte' e 'IA debole'. Gli studiosi del primo paradigma affermano che le macchine sono realmente coscienti e pensanti, essendo in grado di risolvere qualsiasi varietà di compiti, anche nuovi. Tutto ciò che serve è solo un *software*. L'IA debole, invece, si limita ad asserire che quella dell'IA sia solo una 'simulazione' computazionale del pensiero e non una vera cognizione[19].

L 'IA forte può essere considerata come uno strano miscuglio di due elementi: il comportamentismo, in quanto accetta il test comportamentale di Turing per misurare l'intelligenza, e il dualismo cartesiano, poiché rigetta l'idea che coscien-

[15] Ossia quello di considerare il cervello come una scatola nera da studiare solo tramite gli *input* che riceve e gli *output* che produce, ossia i suoi comportamenti.

[16] Cfr. S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., pp. 344, 345.

[17] In natura, come in logica d'altronde, le cose semplicemente accadono e provocano conseguenze, non vi è successo o insuccesso di per sé.

[18] In sostanza, non può scegliere se attenersi a determinate procedure algoritmiche. Può solo deliberare in modo computazionale sempre all'interno di esse. Cfr. J.R. Searle, *Ventun anni dalla stanza cinese*, cit., pp. 98, 99.

[19] Cfr. S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., p. 341.

za e intenzionalità siano fenomeni biologici ordinari. Ciò troverebbe conferma nelle parole di Dennett e Hofstadter secondo cui la mente sarebbe solo una «cosa astratta la cui identità è indipendente da particolari incorporazioni fisiche»[20].

Ciò porta i fautori dell'IA forte a preoccuparsi solo di trovare un metodo conclusivo per accertare la presenza di fenomeni mentali. Il problema di quali fatti interni corrispondano a questi fenomeni mentali osservabili dall'esterno diventa secondario[21].

L'IA debole sembra invece molto più interessata a questo problema, reputando insufficiente un semplice test strumentale e ricercando un certo isomorfismo metafisico tra linguaggio, pensiero e realtà[22]. Proprio per tale motivo l'IA debole sostiene che solo un procedimento meccanico può causare un procedimento mentale e di pensiero, ma la computazione, come detto prima, non è una procedura meccanica del tipo che si trova in natura[23].

3. L'IA può definirsi cosciente?

L'ontologia del computer IA è un'ontologia assolutamente duale in quanto

[20] Cfr. J.R. Searle, *Ventun anni dalla stanza cinese*, cit., p. 90. Se la mente è indipendente dal sostrato biologico che l'ha generata (il corpo-cervello), proprio come un computer può essere implementato su qualsiasi supporto come sostiene l'IA forte, allora viene rispettato perfettamente l'assioma del dualismo cristiano-cartesiano di indipendenza di anima (*res cogitans*) e corpo (*res extensa*). In una certa qual misura ciò è un grosso paradosso nel programma di ricerca dell'IA forte se si pensa che proprio gli studiosi di questa corrente di pensiero sostengono che non essere d'accordo con la tesi secondo cui il cervello è un computer digitale vero e proprio (e non semplicemente assimilabile ad esso) significa afferma qualche sorta di dualismo antiscientifico. In sostanza entrambi i paradigmi mirano nelle loro intenzioni a superare il dualismo cartesiano, ma solo l'IA forte non riesce a raggiungere questo obiettivo. In risposta a tali obiezioni alcuni studiosi dell'IA forte hanno cercato di aggiustare il loro modello tentando di assimilare la mente ad un elaboratore computazionale di tipo parallelo, diverso dall'antiquato modello di von Neumann. Cfr. J.R. Searle, *Il cervello può considerarsi un computer?,* cit., p. 111.

[21] La teoria dennettiana dell'*intencional stance*, predica infatti che si può attribuire una certa 'coscienza' alla macchina per 'salti di fede', semplicemente sospendendo l'incredulità umana su questo argomento (cfr. E. Bolsi, *Il comportamentismo nell'IA …*, cit., p. 49).

[22] Cfr. J.R. Searle, *Ventun anni dalla stanza cinese*, cit., pp. 94, 95.

[23] Cfr. *Ivi*, p. 90. Ciò che ha compiuto l'IA forte, in sostanza, sarebbe stato solo un aggiramento dell'*hard problem* della coscienza e dell'intenzionalità, considerando così il *self* della macchina solo come uno pseudo-problema, preso invece più seriamente dall'IA debole (cfr. E. Bolsi, *Il comportamentismo nell'IA …*, cit., pp. 51, 52).

è composta da un *hardware*, ossia il supporto materiale dei *file* (anche mentali), e un *software*, l'insieme dei *file* istallabili in qualsiasi supporto (secondo la tesi del *mind-upload*)[24]. È facile associare tale dualità a quella cristiano/cartesiana corpo-anima, ma tale analogia rischia di non far cogliere le differenze tra uomo e macchina e porta a pensare che l'IA sia 'cosciente'. Ergo bisogna abbandonare la dualità dell'ontologia umana e considerare ancora di più mente e corpo come un tutt'uno[25].

Di seguito analizzeremo gli argomenti a favore sia dell'IA forte, sia dell'IA debole, con l'intento di sostenere questo secondo programma di ricerca, al fine di rispondere alla domanda che si è posta a titolo del paragrafo.

3.1. Gli esperimenti mentali di Kurzweil

Lo studioso Raymond Kurzweil ha elaborato due esperimenti mentali usati per corroborare le tesi dell'IA forte[26].

Il primo è chiamato '*You* vs *You2*', in cui l'autore immagina che un soggetto (*You*) umano venga clonato artificialmente generando un secondo soggetto (*You2*) che, al momento della sua creazione e prima che faccia nuove e diverse esperienze, condivide la stessa identità e la stessa memoria con l'originale. Avendo un *file* di informazioni mentali coincidente con quello di *You*, *You2* sarebbe quindi parimenti cosciente, benché il clone costituisca un non-*self*, rispetto all'originale che invece si definisce *self* [27].

Qui subentra il secondo esperimento[28] dove lo studioso immagina che un soggetto X si faccia progressivamente sostituire una parte per volta tutto il cervello con un rimpiazzo artificiale. Una volta ultimata la sostituzione il soggetto X è rimasto la stessa persona? Se pensiamo che ogni singola cellula del nostro

[24] Cfr. A. Fabris, *Etica per le tecnologie …*, cit., p. 55.

[25] Parlare di Intelligenza artificiale significa anche parlare in primo luogo dell'uomo: ribaltando la questione del confronto uomo-IA, probabilmente in base all'opinione che avremo sull'ontologia dell'uomo potremo capire se un giorno le macchine saranno in grado di sostituirci o meno.

[26] Cfr. R. Kurzweil, *How to create a mind: the secret of human thought revealed*, Gerald Duckworth &company, United Kingdom 2013, trad. Ita. *Come creare una mente. I segreti del pensiero umano*, Maggiolini editore, Andria, 2013, pp. 170-172.

[27] E. Bolsi, *Il comportamentismo nell'IA …*, cit., pp. 47, 48.

[28] Quest'ultimo in realtà è solo una riproposizione del vecchio dilemma antico della 'nave di Teseo'.

corpo, quindi anche del cervello, cambia almeno 7 volte durante il corso della nostra vita potremmo rispondere di sì. Se è legittimo asserire ciò, allora secondo l'autore anche per quanto riguarda il primo esperimento mentale *You* e *You2* condividono lo stesso *self* e quindi la stessa identità personale[29].

Tali esperimenti, presupponendo il *mind upload*[30], rendono pensabile una 'macchina cosciente'. Sono tuttavia state mosse varie critiche a tali esperimenti.

In primo luogo, se pensiamo ai risultati dei due esperimenti, *You2* e il soggetto sostituito, essi in realtà non condividerebbero in toto l'identità con i soggetti originali. Se pensiamo infatti ai criteri di definizione dell'identità personale esposti da Searle[31], solo 3 di questi 4 verrebbero rispettati, venendo meno il criterio della coerenza del cambiamento fisico[32]. Per continuare ad usare la metafora digitale, non tutte le evoluzioni dell'*hardware* mantengono la compatibilità con il *software*.

In secondo luogo, coscienza e consapevolezza non coincidono: la prima, intesa come *self-recognition*, è solo una parte della seconda e rimanda ad una *self-hood* preriflessiva, la quale è specifica di ogni individuo. In linea teorica *You* e *You2* potrebbero avere la stessa coscienza, ma nel pratico, cominciando a parlare di ciascuno di sé stesso e dell'altro, dopo la clonazione acquisterebbero ciascuno una consapevolezza diversa di sé. Il fatto che *You* e *You2* sembrino

[29] Nel primo esperimento rispetto al secondo si sarebbe semplicemente omessa la gradualità della sostituzione del supporto materiale della mente, ottenendo comunque lo stesso risultato: ossia lo stesso *file* mentale in un diverso *hardware*. Cfr. *Ivi*, pp. 48, 49.

[30] Ossia la possibilità di considerare la mente come un *file* trasferibile e 'caricabile' su molteplici dispositivi, alla stregua di un *software* passibile di essere istallato su differenti *hardware*. Se la medesima mente, insieme al suo *self*, è parimenti implementabile sia su *You* che su *You2*, sia su un supporto neurale che artificiale come nel caso di X, allora può essere facilmente trasferita in un computer.

[31] Cfr. J.R. Searle, *Il sé come problema …*, cit., pp. 176-178.

[32] Da un lato, almeno per il momento, non è plausibile la sostituzione del cervello con un encefalo artificiale; peraltro, anche quando ciò fosse possibile non è detto che ciò non comporti cambi di personalità nel soggetto. Dall'altro, a dispetto di quanto si possa pensare, il vedersi rispecchiato in un clone potrebbe generare immediatamente un vissuto di prosopoagnosia esistenziale, non solo perché il momento di ipotetica coincidenza delle identità sarebbe infinitesimale, ma anche perché, attenendosi alla coscienza fenomenologica del corpo vissuto di Merleau-Ponty, come vedremo di seguito, si creerebbe una discrasia tra corpo vivente e corpo vissuto del *self*, non altrimenti sanabile. Cfr. M. Merleau-Ponty, *Fenomenologia della percezione*, cit., pp. 218, 233, 271, 277.

indistinguibili dall'esterno non implica che essi non possano viversi nella loro interiorità ciascuno come una singolarità diversa dall'altra[33].

In terzo luogo, come detto prima, non si può inferire la coscienza dai semplici comportamenti, i quali potrebbe essere semplicemente un'imitazione del comportamento umano da parte della macchina incosciente. Ponendo di essere *You*, di fronte a *You2* potremmo anche pensare di essere di fronte ad una simulazione di noi stessi, non di vedere il nostro vero *self*[34].

Un ulteriore critica rivolta al *mind-upload* è quella secondo cui essa nega il corpo cosciente. La fenomenologia del soggetto di Merleau-Ponty mostra come la coscienza umana si risolva fondamentalmente in un inerire alla cosa tramite il corpo[35]. In quest'ultimo, a sua volta si distinguerebbero un 'corpo oggetto', ciò che ciascuno sente come 'se-oggettivato' e un 'corpo vissuto', ossia il 'sé-vissuto'[36]. Questo modo antidualista di percepire e di avere coscienza, peraltro, potrebbe essere individuato come una delle specificità che distingue l'uomo dalla macchina.

Si potrebbero cambiare tutti i pezzi del cervello di X solo se si supponga che il sostrato materiale della mente (il corpo) non contribuisca alla cognizione, il che è stato dimostrato essere falso dal paradigma della 'cognizione incarnata' dalle scienze cognitive[37].

3.2. L'esperimento della stanza cinese di Searle

Nonostante quanto visto sin qui, rimane il fatto che i sistemi IA sono dotati di meta-ragionamento[38]. Tale abilità è possibile grazie ad un'architettura riflessiva, costituita tramite gli algoritmi *anytime* e/o con la teoria delle decisioni, i quali permettono di controllare le azioni computazionali che si compiono dentro la macchina[39].

[33] Cfr. E. Bolsi, *Il comportamentismo nell'IA …*, cit., pp. 59-62.
[34] Cfr. *Ivi*, pp. 62, 63.
[35] M. Merleau-Ponty, *Fenomenologia della percezione*, cit., p. 194.
[36] Cfr. *Ivi*, pp. 99, 124-128, 134.
[37] Cfr. E. Bolsi, *Il comportamentismo nell'IA …*, cit., pp. 63, 64.
[38] Tale caratteristica, in genere associata ad un'interiorità psichica, è ciò che ci induce falsamente a pensare che la macchina sia cosciente.
[39] Esempi di ciò sono i sistemi di IA istallati su autovetture a guida autonoma. Cfr. S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., pp. 380, 381.

Fornendo agli algoritmi decisionali l'accesso ai loro stessi processi deliberativi, semplicemente archiviandoli come dati *open source*, le IA potrebbero spiegare meglio degli uomini le loro stesse decisioni[40]. In questo modo si creerebbero dei modelli di 'IA spiegabile' e/o 'interpretabile'[41].

Ma il fatto che l'IA possa ripercorre le sue decisioni e tutti i dati che ha elaborato sapendoli 'spiegare', ossia semplicemente collegare l'un l'altro, ci permette di asserire che la macchina 'capisca' l'essenza semantica di quegli stessi dati? Il filosofo Searle sostiene di no.

Codesto autore ha infatti elaborato un famoso argomento, a favore dell'IA debole, con il famoso esperimento mentale della 'stanza cinese'[42]. Questa funzionerebbe come un computer: all'interno di entrambi non vi sarebbe autoconsapevolezza dei processi in corso, ergo non vi sarebbe coscienza[43].

Tale argomento di Searle si regge su due assiomi incontrovertibili: 'la sintassi non è semantica' e al contempo 'la simulazione non è duplicazione'. Il primo assioma afferma che il programma formale e astratto implementato

[40] Il collegamento tra immagazzinamento dei dati e la connessione a essi, compito prima svolto dall'uomo, viene così svolto dalla macchina sia per quanto riguarda sé stessa, sia per quanto riguarda l'uomo. Proprio in tal senso si parla sempre più spesso di *outsearching* cognitivo, una tendenza deumanizzante volta a rendere computazionali gli stessi processi mnemonici umani. Pensiamo, per esempio, alla tendenza sempre maggiore al giorno d'oggi ad affidare i propri processi mnemonici a dispositivi digitali esterni da parte di molte persone (sveglie, *reminder*, promemoria ecc.). Cfr. A. Fabris, *Etica per le tecnologie …*, cit., pp. 62, 63. Essendo la memoria, però, uno dei criteri d'identità indicati da Searle, come visto prima, intaccarla in tal modo inficerebbe negativamente sulla specificità dell'identità umana. Identità e memoria si concausano a vicenda, e pur tuttavia la prima non può essere appiattita sulla seconda (per fortuna diremmo in questo caso). Se così fosse, infatti, un'ulteriore conseguenza sarebbe quella di rendere possibile il *mind upload* predicato dall'IA forte. Cfr. J.R. Searle, *Il sé come problema …*, cit., pp. 185-187.

[41] Cfr. S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., p. 358.

[42] Breve riassunto per chi non conosce tale argomento. Si immagini, dice Searle, un uomo di lingua inglese chiuso in una stanza da cui non può uscire, in cui egli riceve dall'esterno dei messaggi in cinese. Egli non conosce il cinese e all'interno della stanza vi sono solamente dei fogli di carta e dei codici di regole in inglese in cui si dice di assegnare a determinati iconogrammi del linguaggio cinese ulteriori simboli. Con tali istruzioni il soggetto riesce a risolvere un problema senza capire di cosa si stia parlando, e alla fine potrà passare all'esterno un foglio con la soluzione del problema scritta con simboli cinesi. Dall'esterno si potrà pensare che il soggetto all'interno, essendo stato in grado di risolvere il problema, capisca il cinese, quando invece ciò non è vero. Cfr. J.R. Searle, *Ventun anni dalla stanza cinese*, cit., pp. 81-84.

[43] S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., pp. 345, 346.

dalle IA, essendo di natura sintattica, di per sé non garantirebbe la presenza di contenuto semantico nella computazione e quindi di un contenuto mentale. Il mentale, infatti, è ciò che riesce a dare senso al dato, non solamente ciò che lo elabora: l'uomo nella stanza cinese riesce solo a computare i simboli senza riuscire ad assegnare a ciascuno di essi un significato[44].

Il secondo assioma invece si riferisce al fatto che, se una macchina simula il pensiero umano, non vuol dire che essa stia effettivamente pensando, così come se un computer simulasse la digestione umana non significherebbe che esso stia effettivamente digerendo dei cibi[45].

Searle si fa qui sostenitore di un certo naturalismo biologico di tipo emergentista: le particolari proprietà dell'insieme dei neuroni 'causano' effettivamente la coscienza, poiché la viviamo in noi giornalmente[46]; non altrettanto si potrebbe asserire dei transistor, benché di essi si conoscano tutte le proprietà. Per Russell e Norving ciò costituisce una criticità per due motivi. In primo luogo, a loro dire, tutto ciò, di per sé, non comporterebbe una differenza tra transistor e neuroni tale per cui i primi non potrebbero svolgere la stessa funzione dei secondi. In secondo luogo, lo stesso argomento potrebbe essere usato da un robot per sostenere che un umano non ha una reale coscienza e conoscenza semantica di quello che elabora nel suo encefalo[47].

Si fa presto a rispondere a questa obiezione. La possibilità di una realizzabilità multipla del computer non è conseguenza del fatto che lo stesso stato fisico (il mentale) possa essere ottenuto in diverse sostanze fisiche (neuroni e transistor), quanto più per il fatto che le proprietà rilevanti che permetterebbero il *mind upload* sarebbero essenzialmente sintattiche e non semantiche. La sintassi, infatti, non è mai intrinseca alla fisica e l'attribuzione di proprietà sintattiche è sempre relativa a un agente che tratta certi fenomeni fisici come sintattici[48]. Da tutto ciò si può dedurre che la sintassi logica dei processi com-

[44] Cfr. J.R. Searle, *Ventun anni dalla stanza cinese*, cit., pp. 82, 83.

[45] Cfr. *Ibidem*.

[46] Ergo l'effetto, ossia la mente, non è separabile dalla causa, cioè il cervello materiale. Proprio per questo la mente non può essere paragonata ad un *file* trasferibile su altri dispositivi con un *mind-upload*.

[47] S.J. Russell, P. Norving, *Artificial Intelligence* …, cit., p. 346.

[48] Cfr. J.R. Searle, *Il cervello può considerarsi un computer?*, cit., pp. 115-118. Anche se il programma

putazionali ha solo una funzione descrittiva. La simulazione, quindi, 'mostra' semplicemente il processo deliberativo dell'IA, ma non lo 'dimostra'; al contrario, enucleare i nessi causali tra un nodo e l'altro di una catena algoritmica comporterebbe capirne il significato, ergo duplicare il ragionamento. Ciò fa capire che la macchina non segue intenzionalmente le proprie procedure, a differenza dell'uomo, perché non le 'comprende'. La macchina, in sostanza, non ha intenzionalità[49].

Infine, bisogna ricordare che un sistema di *transistor* non potrebbe duplicare la coscienza a causa della specificità della coscienza umana, la quale è composta parimenti da 'coscienza soggettiva' e la 'coscienza fenomenologica'. Alla prima dimensione perterrebbe un io-soggetto metafisico, inteso come prerequisito di ogni esperienza. La seconda, invece, rappresenterebbe i contenuti di esperienza con riferimento a chi la fa, facendo in modo che il *self* funga da referente e sfondo prospettico di ogni esperienza[50].

Anche quando la macchina riuscisse ad avere dei processi di interpretazione semantica in futuro, non è detto che riesca ad ottenere almeno una di queste due partizioni della coscienza, mancandole difatti un certo *self*-minimale preriflessivo, all'origine di entrambe[51].

4. *Moral agency* e responsabilità nell'IA

4.1. Etica e diritto computazionali

Considerato quanto detto finora possiamo capire che il cervello umano non elabora informazioni come un computer, nel senso di 'informazione' usato dalle scienze cognitive. Tale significato è troppo astratto per catturare la con-

della macchina fosse ipoteticamente sufficiente a 'causare' degli stati mentali in maniera *bottom-up* (in quanto è definito indipendentemente dagli elementi fisici della sua implementazione), d'altro canto sarebbe insufficiente a 'costituirli' per la differenza tra sintassi e semantica. Cfr. J.R. Searle, *Ventun anni dalla stanza cinese*, cit., pp. 35, 85-87.

[49] Cfr. J.R. Searle, *Il cervello può considerarsi un computer?*, cit., pp. 121-128.

[50] Cfr. E. Bolsi, *Il comportamentismo nell'IA …*, cit., pp. 54, 55.

[51] Tale ipotesi, in gergo detta MSH, presenterebbe anche dei correlati empirici nel *neuroimaging*, nei dati provenienti da studi su soggetti affetti da disordini di coscienza (DoC), e sarebbe rintracciabile anche in episodi di perdita di coscienza e di mancata consapevolezza. L'MSH presenterebbe in sostanza il *self* come un 'essere-di-coscienza'. Cfr. *Ivi*, pp. 56, 57.

creta realtà biologica dell'intenzionalità intrinseca[52]. Al contrario, il nostro encefalo, causando e costituendo la mente, produce coscienza e comunicazione.

Chiarito ciò si comprende allora che l'appiattimento dell'uomo sulla macchina, portato avanti dalla cibernetica e dall'IA forte, comporta la riduzione della comunicazione umana alla mera trasmissione di informazioni. La comunicazione, tuttavia, è molto di più: essa rimanda inevitabilmente ad una dimensione intersoggettiva e sociale di 'beni comuni', assente nel mero scambio informatico[53]. Il programma della cibernetica e dell'IA forte, quindi, porta inevitabilmente con sé un pericoloso potenziale di de-socializzazione per l'uomo[54].

La similitudine uomo-macchina, così, si infrange: se l'IA è certamente un computer digitale, il quale segue solo procedure algoritmiche, l'uomo è al più un 'computer analogico' in grado anche di computare certamente, ma comunque dotato di un ulteriore libero-arbitrio non-algoritmico[55].

Visto tutto ciò, si può affermare che la prospettiva analitica sull'etica, sostenuta anche da Searle, potrebbe risultare fuorviante. Risolvere la morale in un insieme di proposizioni passibili di essere considerate vere o false appiattirebbe l'etica su un codice binario computabile dalle macchine[56]. Tale tendenza è stata intrapresa per permettere alle macchine di avere una certa capacità deliberativa; tuttavia, per l'uomo ciò ha due conseguenze.

In primo luogo, essa genera un estremo riduzionismo della poliedricità delle azioni umane e della relatività dei valori insiti nelle scelte di ciascuno, intaccando l'infinita varietà di sfumature della volontà umana. La ricchezza

[52] Cfr. J.R. Searle, *Il cervello può considerarsi un computer?,* cit., pp. 128-130.

[53] Cfr. A. Fabris, *Etica per le tecnologie …,* cit., pp. 23, 26.

[54] E se intendiamo l'uomo aristotelicamente come un 'animale sociale' (Aristotele, *Politica*, cit., I, 2, 1253a, p. 5), e quindi consideriamo la socialità come una qualità ontologica dell'umano, comprendiamo quanto la cibernetica e l'IA forte, se portate avanti, possano snaturare la stessa essenza dell'uomo.

[55] Cfr. A. Fabris, *Etica per le tecnologie …,* cit., pp. 47, 60. Su questo punto si potrebbe anche ricordare l'argomento dell''informalità del comportamento' di Turing, secondo cui gli uomini utilizzano alcuni criteri guida informali nella loro comunicazione e interazione che non potrebbero essere catturati in un insieme formale di regole, ergo non potrebbero essere codificati in un programma informatico. Secondo Russell e Norving, tuttavia, tale argomento vale solo per le 'IA alla vecchia maniera'. Cfr. S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., p. 342.

[56] Cfr. J.R. Searle, *La filosofia in un nuovo secolo* in *Phylosophy for a new century, …*, cit., p. 43.

tradizionale dell'etica, derivata dall'ermeneutica morale, verrebbe impoverita pesantemente[57].

In secondo luogo, appiattire la ragion pratica di Kant sulla ragione tecnica/meccanica snaturerebbe quest'ultima a livello ontologico e non solo gnoseologico. La soppressione dell'etica umana operata dalla cibernetica avrebbe la stessa portata della soppressione che, secondo Husserl, la scienza avrebbe condotto nei confronti del «mondo della vita»[58].

Una medesima operazione a quella qui denunciata in campo etico potrebbe avvenire anche in ambito giurisprudenziale in seguito alla nascita del campo di studi sul 'diritto computazionale'. Questo sarebbe un diritto sintetico scritto in un linguaggio di programmazione, tale da portare alle estreme conseguenze il formalismo giuridico. Il ragionamento giuridico tipicamente umano divenendo computabile con mere rappresentazioni logiche potrebbe essere così più facilmente processato dalla macchina. Applicazioni di tali traduzioni sono già commercializzate con successo in alcuni ambiti, a dire la verità[59]. Nonostante ciò, la loro diffusione estesa crediamo possa generare nocive conseguenze per la società non dissimili da quelle enunciate prima per la cibernetizzazione dell'etica.

[57] Tale operazione sull'etica ha le sue radici in una tradizione della logica che sostiene che quest'ultima non si deve occupare della verità o falsità degli enunciati, ma solo delle sue conseguenze. La logica cioè, essendo essenzialmente sintattica, può elaborare solo simboli ma non può 'leggerli', come detto, prima. Tutto ciò è corretto e lo abbiamo sostenuto precedentemente, infatti. Tuttavia, da ciò spesso si compie un salto per certi versi indebito. Se non se ne occupa la logica, allora di questioni di verità e falsità, ossia di semantica e riferimenti ontologici alla realtà, se ne dovrà occupare l'ormai famosa dimensione del 'mistico' di Wittgenstein, a cui pertiene l'etica appunto. Tale salto risulta autorizzato da una filosofia del linguaggio che individua il significato nel riferimento denotativo, posizione questa che è stata superata anche essa nel corso della storia del pensiero (per ulteriori approfondimenti si rimanda alla lettura integrale del testo C. Barbero, S. Caputo, *Significato. Dalla filosofia analitica alle scienze cognitive*, Carocci, Roma, 2018). A prescindere da tale discussione, che non compete questo elaborato, resta comunque il fatto che ridurre la varietà dei colori del reale ad un semplice 'bianco' e 'nero' può risultare comunque svilente e deludente. L'etica può certamente occuparsi di bianchi e di neri, di verità e falsità di enunciati morali, ma non può limitarsi solo ad essi.

[58] E. Husserl, *La filosofia come scienza rigorosa*, Laterza, Roma-Bari, 2020[8], p. 104.

[59] T. Casadei, S. Pietropaoli, *Intelligenza artificiale: fine o confine del diritto?,* in T. Casadei, S. Pietropaoli, (a cura di), *Diritto e tecnologie informatiche. Questioni di informatica giuridica, prospettive istituzionali e sfide sociali*, Cedam Scienze giuridiche, Milano, 2021, pp. 224, 225.

Ciò ci fa capire quanto l'etica e il diritto che possono essere utilizzati per le macchine debbano rimanere differenti da quelli usati per gli uomini.

Posto questo avvertimento, non si può comunque asserire che le IA non debbano avere un'"etica'. Una certa dimensione morale pertiene alle macchine nella misura in cui l'IA riesce a deliberare, a mostrare i valori che ha assunto nelle sue scelte. È il livello successivo, quello della metaetica, ossia della riflessione sull'etica, che pertiene invece esclusivamente all'umano, poiché tale livello necessità di auto-coscienza e intenzionalità che abbiamo visto l'IA non possiede.

4.2. Autonomia e imputabilità

Il rapporto tra uomini e macchine al giorno d'oggi sta sempre più traslando da un'"interazione', relazione che porta due mondi ontologici distinti a contatto, ad una vera e propria 'integrazione', la quale crea una nuova dimensione della realtà, quella digitale appunto. Se nel primo stadio si agiva 'tramite' la macchina (semplice strumento), oggi di agisce sempre più 'con' la macchina (ossia anche insieme ad essa): essendo autonome le IA posso porsi allo stesso livello dell'uomo[60].

Detto ciò, benché l'autonomia sia sempre stata la precondizione della responsabilità giuridica e morale (e quindi dell'imputabilità penale) 'agire con la macchina' non deve voler significare dividere la responsabilità di azioni penalmente rilevanti con un certo sistema IA che si è usato per commettere reati, ad esempio.

Peraltro, nell'utilizzo dell'IA si può provocare danno, sfociare nell'illecito o semplicemente riverberare stereotipi sociali anche involontariamente. Tanto più complesso è un programma, maggiore è la possibilità che possa essere stato commesso un errore in fase di programmazione. Si pensi per esempio a sistemi IA usati in ambito medico che compiono autonomamente delicate operazioni chirurgiche. Nonostante si cerchi di aumentare sempre più la sicurezza e l'efficacia delle macchine usate, un blocco o un malfunzionamento rimangono in teoria sempre possibili. In caso di danno arrecato al paziente chi ne risponde[61]?

[60] Cfr. *Ivi*, pp. 38-40.

[61] G. Fioriglio, *eHealth*: *tecnologie, diritto, salute* in T. Casadei, S. Pietropaoli, (a cura di), *Diritto e tecnologie informatiche. …*, cit., p. 48. Il problema dell'allineamento dei valori, gli effetti collaterali inattesi, 'la tragedia dei beni comuni', il possibile 'alto impatto' delle decisioni dell'IA sulla realtà o il 'problema di Re Mida', sono solo alcune delle difficoltà in merito che ingegneri e psicologi che si occupano di IA devono risolvere. Per ulteriore approfondimento su ciascuna di queste tematiche cfr. S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., pp. 362-364.

Difatti, se l'IA è in grado 'autoregolare' i propri processi, non è ancora in grado di 'autoregolamentarli'. La sua, infatti, è sempre un'autonomia relativa: il *training* che ha compiuto, la gerarchia dei valori da adottare in caso di scelta, le sue modalità di apprendimento, le sue risposte a precise sollecitazioni ambientali sono sempre influenzate dalla programmazione umana. In sostanza, l'orizzonte noumenico degli eventi e le condizioni di verità e possibilità adottate dalla macchina sono ancora in mano al figlio di Adamo[62].

Giuridicamente si ha responsabilità per colpa quando sussiste un certo rapporto positivo fra il comportamento interno e spirituale del delinquente e l'evento cagionato o non impedito dal suo comportamento esterno, così come egli l'ha previsto e voluto[63]. Tale definizione di Kelsen sottintende un *self* dotato di intenzionalità da assumere come presupposto dell'esperienza, per come l'ha inteso Searle[64].

Tuttavia, come visto la macchina IA non ha un'autocoscienza intenzionale, quindi neanche un *self* esistenziale; essa non ha nessun comportamento spirituale interno, per cui tale rapporto alla base di questo concetto di 'colpa individuale' non sussiste in quanto manca uno dei due poli della relazione alla base. Nonostante ciò, un *framework* legale di vari individui (il programmatore, il produttore, il venditore e l'utilizzatore finale) può essere chiamato a rispondere delle azioni della macchina[65]. Ciò a nome di una 'responsabilità oggettiva' in cui l'evento cagionato non era stato né previsto né voluto dall'utente che ne risponde[66].

Per Kelsen l'imputabilità, e quindi la responsabilità, fonda la libertà, e non viceversa. Per lo studioso, ciò che determinerebbe a sua volta l'imputabilità/responsabilità sarebbe la determinabilità causale del volere, effetto della rap-

[62] Cfr. A. Fabris, *Etica per le tecnologie …*, cit., pp. 77-80. Su questo punto si veda anche T. Casadei, S. Pietropaoli, *Intelligenza artificiale: fine o confine del diritto?*, cit. p. 229.

[63] H. Kelsen, *Reine Rechtslehre*, Verlag Franz Deuticke, Wien, 1960, trad. ita. *La dottrina pura del diritto*, a cura di M.G. Losano, Einaudi, Torino, 2021[3], p. 170.

[64] Cfr. J.R. Searle, *Il sé come problema …*, cit., p. 191. Bisogna supporre che vi sia una x tale per cui (1) x è cosciente, (2) x persiste nel tempo, (3) x ha percezione e ricordi, (4) x opera nello spazio vuoto avendo ragioni, (5) x, nello spazio vuoto del dubbio rispetto all'azione, è capace di decidere e agire e (6) x è responsabile per almeno alcuni dei suoi comportamenti. Tale *self*, infine, va considerato come una caratteristica formale del campo cosciente e non un qualcosa di avulso da esso. Cfr. *Ivi*, p. 189.

[65] G. Fioriglio, *eHealth*: *tecnologie, diritto, salute*, cit., p. 49.

[66] Cfr. H. Kelsen, *Reine Rechtslehre*, cit., pp. 169-173.

presentazione della norma da parte dell'agente[67].

Tutto questo rapporto tra norma, reato, sanzione e responsabilità pensato da Kelsen sembra non funzionare più nella misura in cui oggi abbiamo delle macchine che, pur essendo autonome, non sono moralmente 'libere': di esse, si potrebbe determinare causalmente il volere ancor di più degli uomini, ma ciò non fonderebbe la loro responsabilità. I sistemi IA potrebbero raffigurarsi la norma e rispondere delle loro azioni, eppure non li considereremmo 'liberi'.

Per qualcuno il semplice fatto che la macchina conosca le norme e riesca a rappresentarsele è sufficiente di per sé per conferire responsabilità penale alle azioni dell'IA[68]. La dottrina di Kelsen rimane tuttavia valida nella misura in cui solo negli esseri umani la rappresentazione della norma provoca atti 'volontari' di adesione alla stessa, come non accade nella rappresentazione della macchina. Proprio in virtù di ciò solo gli uomini sono imputabili[69].

Per comprendere meglio questo punto si deve tenere a mente una duplice articolazione del concetto di responsabilità: da un lato, come capacità di 'rispondere-a' principi a priori della scelta/azione e dall'altro come facoltà di 'rispondere-di' conseguenze a posteriori della scelta/azione[70]. La limitata responsabilità dell'IA pertiene solo il primo senso; il secondo significato è riferibile invece solo all'uomo.

5. Conclusione

Visto quanto detto nei paragrafi precedenti, possiamo asserire che le macchine IA non avendo né intenzione né coscienza non sono né responsabili né imputabili. Il comportamentismo dell'IA forte si è dimostrato fallace. Risulta

[67] Cfr. *Ivi*, pp. 132-137.

[68] D'altronde la stessa possibilità di attribuire personalità giuridica (e quindi responsabilità penale) ad enti artificiali tecnicamente non-umani (es: le società per azioni) è sempre stata solo una questione più di opportunità che di possibilità, nella storia del diritto. Nel caso della maggior parte di queste entità artificiali, tuttavia, i fini e le azioni erano sempre riconducibili a persone fisiche. Nel caso delle IA, invece, ciò potrebbe avvenire solo nel caso di un *management buyout*, possibilità al momento abbastanza remota. Cfr. T. Casadei, S. Pietropaoli, *Intelligenza artificiale: fine o confine del diritto?,* cit., pp. 226, 227.

[69] Cfr. H. Kelsen, *Reine Rechtslehre,* cit., pp. 135, 136.

[70] Cfr. A. Fabris, *Etica per le tecnologie …*, cit., pp. 55-57.

così necessario abbandonarlo a favore di una nuova 'scienza del cervello', superando anche la classica scienza cognitiva computazionale a favore del paradigma delle 'neuroscienze cognitive'[71].

Tutt'oggi non possiamo ancora affermare che le macchine riescano a pensare, anche se nulla toglie che ciò possa avvenire in futuro con i giusti cambiamenti nei paradigmi di ricerca. Per il momento, fortunatamente, la 'robo-apocalisse' rimane ancora una distopia[72]. L'unico principio che potrà guidarci in futuro sarà la ricerca della differenza tra l'uomo e la macchina come presupposto ontologico fondamentale della morale e del diritto[73].

[71] Cfr. J.R. Searle, *Ventun anni dalla stanza cinese*, cit., p. 104.
[72] Cfr. S.J. Russell, P. Norving, *Artificial Intelligence …*, cit., p. 362.
[73] Cfr. A. Fabris, *Etica per le tecnologie …*, cit., pp. 94, 95.