Ragioni tecniche e teoriche dell'alterità. Una questione sull'attitudine 'politica' dell'intelligenza artificiale

di Pier Giuseppe Puggioni

Abstract: Technical and Theoretical Reasons for Alterity. Question on the 'political' aptitude of Artificial Intelligence – This short paper aims to analyse the theoretical meaning of 'artificial life' in regard to the practical issue of implementing artificial general intelligence in the field of legislative decision-making and disputes' resolution. In order to delve into this problem, I will recall the most relevant statements set forth by the post-human studies, with some references to important political and legal theorists. To conclude, I will question whether the cognitive processes characterising AIs (especially artificial neural networks), from a scientific point of view, are to be included, or excluded, from the semantic realm of 'life' and 'learning'.

Keywords: Philosophy of law; Post-humanism; Neural networks; Artificial intelligence; Epistemology.

1. Introduzione

L'impiego delle tecnologie d'intelligenza artificiale nei processi produttivi e applicativi di norme giuridiche rappresenta una sfida importante nel panorama post-moderno del diritto pubblico. Per una più adeguata comprensione del problema, può essere utile interrogarsi sullo statuto teorico della «vita artificiale», nella misura in cui le questioni della prassi politico-giuridica si connettono a quelle sul rapporto tra vita biologica e vita artificiale. Un simile accostamento trova ragione nel legame che, seppur difficile da afferrare, in qualche modo connette l'«azione» (compresa quella «politica») alle categorie, più o meno definite, di «vita» e «conoscenza».

Si tratta di un problema prodromico rispetto a quello relativo alle possibilità di impiegare la tecnologia in attività che, tradizionalmente, sono risultate appannaggio dei soli esseri umani. Infatti, a seconda di come si definisce in linea teorica la relazione tra l'*artificial agency* e l''umano', si può addivenire a diverse conclusioni in ordine al modo di concepire il rapporto pratico – anche in termini di diritto positivo – tra questi due concetti. Così procedendo, si potrebbero interpretare, con un approccio maggiormente consapevole e problematico, le circostanze in cui l'essere umano entra in relazione con l'essere artificiale, alla luce della concezione assunta con riguardo all'uno e all'altro. Un buon esempio di tale *modus procedendi* è dato dalla distinzione fra *trust* e *reliance*, concetti dai quali si trae

un'efficace tassonomia delle relazioni fra agenti umani e agenti artificiali[1]. Si tratta di una questione rilevante, che pare doveroso affrontare nella misura in cui si consideri in chiave problematica la soggettività come categoria 'teorica'. Nelle pagine che seguono, pertanto, verranno prese in esame – e sommariamente discusse – alcune prospettazioni teoriche dirette a minimizzare (o, per certi versi, ad annullare) il rapporto di 'alterità' tra uomo e macchina.

2. Vita artificiale e agire morale nell'era del 'postumano'

Della questione dello statuto morale e filosofico-giuridico della 'vita artificiale' si è occupata in particolare la letteratura post-umanistica, con l'intento di decostruire criticamente la nozione di «umano», caratterizzata da una serie di «limiti e confini simbolici». Questi studi tentano di negare, per un verso, il «primato onto-epistemologico dell'umano», riabilitando, per altro verso, la «possibilità di *agency* umana attraverso un'espansione di tale nozione al non-umano»[2]. È interessante notare come alcuni studiosi giungano, seguendo questa linea (*speculative posthumanism*), a ridefinire concettualmente l'ambito dell'«evoluzione», onde includere in tale processo anche l'impiego di processi tecnologici, che già l'essere umano sfrutta per incidere sul proprio corpo in modo significativo. Così, si potrebbe estendere la nozione di «discendente» – che abitualmente si riferisce al discendente 'biologico' – ad altri 'esseri' (*beings*) generati attraverso meccanismi «puramente tecnologici», come le *uploaded minds*, le forme di vita 'sintetiche' e le intelligenze artificiali[3].

La questione del mutamento delle caratteristiche fondamentali del soggetto attraverso l'evoluzione tecnologica rappresenta un argomento toccato anche da alcuni classici del pensiero politico e giuridico. Nel corso di un seminario tenutosi nel 1982, Michel Foucault (1926-1984) annoverava tra i quattro tipi fondamentali di tecnologia le cosiddette «tecnologie del sé», che «permettono agli individui di eseguire, coi propri mezzi o con l'aiuto degli altri, un certo numero di operazioni sul proprio corpo e sulla propria anima», realizzando in tal modo «una trasformazione di se stessi»[4]. D'altra parte, si possono richiamare anche le considerazioni formulate da Hart (1907-1992) sulle verità ovvie che fondano il «contenuto minimo di diritto naturale»[5]: secondo il filosofo britannico, infatti, «il mondo in cui viviamo e noi che lo abitiamo potremmo, un giorno, cambiare in molti e diversi modi»[6], potendosi dunque immaginare uomini fisicamente invulnerabili o non bisognosi di nutrirsi attraverso risorse limitate. Orbene, in queste affermazioni sembra potersi leggere

[1] Cfr. F. Fossa, *«I Don't Trust You, You Faker!». On Trust, Reliance, and Artificial Agency*, in *Teoria*, 28, 1, 2019, 63 ss.
[2] F. Ferrando, *Il Postumanesimo Filosofico e le sue Alterità*, Pisa, 2016, 19-20, 48.
[3] D. Roden, *Deconstruction and Excision in Philosophical Posthumanism*, in *Journal of Evolution and Technology*, 21, 1, 2010, 28.
[4] M. Foucault, *Tecnologie del sé*, in L. H. Martin, H. Gutman e P. H. Hutton (a cura di), *Tecnologie del sé. Un seminario con Michel Foucault*, Torino 1992, 13.
[5] H. L. A. Hart, *The Concept of Law*[2], Oxford 1994, 193 ss.
[6] Id., *Positivism and the Separation between Law and Morals*, in *Harvard L. Rev.*, 71, 4, 1958, 622.

l'idea che le regole giuridiche dipendano dalle caratteristiche del soggetto di diritto, che a loro volta sono suscettibili di mutamento nel tempo.

Qualche suggestione al riguardo proveniva del resto già da Georg Simmel (1858-1918), per il quale una delle peculiarità che fondano la modernità risiede nella «preponderanza di ciò che si può chiamare lo spirito oggettivo sullo spirito soggettivo»[7]. La cultura (la conoscenza oggettiva), notava Simmel, tende ad accrescere la propria indipendenza nei confronti delle singole conoscenze soggettive e, di conseguenza, ad imporsi su di esse e a definire l'identità del soggetto. In altre parole, l'accumulo di dati e le costruzioni scientifico-intellettuali sempre più complesse non fanno che aumentare il divario tra la comprensione di cui è capace il singolo ed il prodotto della scienza, da cui risulta una sorta di «spiritualizzazione degli oggetti che si compie quasi con proprie forze e secondo proprie norme»[8]. È interessante rimarcare questo accento sulla conoscenza, stante il profondo legame che secondo alcuni intercorre tra la funzione cognitiva ed il 'vivere': la conoscenza rappresenta una proiezione dell'essere vivo, in quanto costituisce il significato minimo della vita attiva[9]. Sulla base di questa premessa, sarebbe possibile caratterizzare gli esseri viventi sulla base del fatto che essi «si producono continuamente da soli»[10], secondo una specifica forma di *autopoiesi*: essi hanno la possibilità di conoscere, senza l'intervento altrui, elaborando una serie di *input* provenienti dall'esterno.

A certe condizioni, dunque, veramente sembra potersi sostenere che il funzionamento dell'intelligenza artificiale non sia molto diverso da quello della mente umana. Questo parallelo ha importanti riflessi, peraltro, in ordine alle possibilità d'impiegare tale tecnologia nei processi decisionali, se si considera come il modello della *artificial general intelligence* (detta anche '*generalized AI*' o '*strong AI*') tende a riprodurre un funzionamento cognitivo analogo a quello del cervello umano, anche per quanto riguarda la specificità della valutazione politica, che presuppone quello che abitualmente si chiama «coscienza»[11]. La macchina, dunque, sarebbe in grado di replicare quell'autonomia che si attribuisce alla mente umana e che fonda l'agire 'morale' (su cui si basa la scelta 'politica'), inteso – per usare un'espressione weberiana – come agire razionale rispetto al valore o, comunque, come azione che tende ad assumere e considerare un certo bagaglio di valori.

In sostanza, prendendo in prestito una terminologia aristotelica, potremmo affermare che queste funzioni cognitive avanzate cerchino di replicare le caratteristiche del *logos* umano, ovverosia di quel «linguaggio» che, a differenza della semplice *phonē* (la voce), suggella l'attitudine dell'essere vivente a distinguere il bene dal male, il giusto dall'ingiusto «e degli altri valori»[12]. Sarebbe, quindi, questa vocazione alla cognizione *morale* – ossia, come si è detto, alla capacità di scegliere individuando una serie di valori – a qualificare l'essere vivente razionale,

[7] G. Simmel, *Le metropoli e la vita dello spirito* (1903), Roma, 1995, 53.
[8] Id., *Filosofia del denaro* (1900), Torino, 1984, 635.
[9] Cfr. H. Maturana, F. Varela, *L'albero della conoscenza*, Milano, 1987, 39.
[10] *Ibid.*, 59.
[11] Si veda, ad esempio, la distinzione fra «weak AI» e «strong AI» illustrata da S. J. Russel, P. Norvig, *Artificial Intelligence. A Modern Approach*, Englewood Cliffs (NJ), 1995, 29.
[12] Aristotele, *Politica*, I 2, 1253a, in Id. *Opere*, IX, Roma-Bari, 1986, 6.

che si caratterizza per il fatto d'essere *bìos*, e non meramente *zoē*[13]. D'altra parte, non è affatto scontato ritenere che gli attributi della razionalità e, dunque, della libertà morale e dell'*agency* politica rappresentino prerogative esclusive dell'essere umano in senso biologico. Lo stesso Immanuel Kant (1724-1804) parlava di «esseri razionali liberi», per riferirsi in generale ai soggetti destinatari dell'imperativo categorico: egli affermava, più precisamente, che «soggetto della legge morale» è «soltanto l'uomo, e con lui *ogni creatura razionale*»[14].

Gli studi sul postumano si chiedono, allora, in cosa la vita artificiale si differenzi da quella biologica, dal momento che entrambe le forme di vita sembrano potersi sussumere entro una generale definizione della «vita» come processo elaborativo di informazioni, scevro da ogni contingente presenza materiale o fisica[15]. Considerazioni siffatte fanno emergere, d'altronde, l'opacità delle conclusioni cui taluni studiosi pervengono, escludendo che l'essere artificiale possa essere considerato dal diritto civile come una «nuova figura giuridica dotata della capacità di agire e della capacità di essere titolare di situazioni giuridiche soggettive»[16].

3. Le reti neurali e il problema dell'apprendimento

Con riferimento alla problematica in oggetto, sembra particolarmente utile rivolgere l'attenzione al tema delle reti neurali (*artificial neural networks*). Questa tecnologia nasce da un tentativo di imitare il funzionamento del cervello umano e muove dall'assunto che i percorsi neurali (*neural pathways*), da cui scaturiscono la conoscenza e l'apprendimento, si potenzino ogni volta che ricevono ed elaborano stimoli. In questo modo, l'essere umano è capace di riconoscere forme, oggetti e, più in generale, insiemi di dati raggruppandoli sulla base di categorie che lui stesso, con il tempo e l'esperienza, ha 'costruito' e imparato a usare[17]. In modo analogo, la rete neurale è in grado di processare determinati *input* e generare *output* con crescente capacità di 'generalizzazione', ovverosia di riconoscere gli oggetti (*scil.* insiemi di dati), cercando di desumerne le caratteristiche primarie e ponendo, così, i presupposti perché siano avviati processi simili ad un vero e proprio 'apprendimento'.

Un simile processo si verifica grazie ai cosiddetti algoritmi di «*backpropagation*», che consentono a queste reti artificiali di 'allenarsi', elaborando informazioni di *input* in gran quantità, a riconoscere una serie di *pattern* (cioè di oggetti con certe caratteristiche) nuovi, senza bisogno che il 'supervisore' umano

[13] Così si legge nelle celebri pagine, ancora una volta 'aristoteliche', di G. Agamben, *Homo sacer I. Il potere sovrano e la nuda vita*, Torino, 1995, 3-4, riprese peraltro da alcune voci del postumanesimo, tra cui Ferrando, cit., 89-91.

[14] I. Kant, *Critica della ragion pratica* (1788), Roma-Bari 2006, 107. A tal proposito, M. C. Pievatolo (*Immanuel Kant, La metafisica dei costumi. Un esperimento di lettura ipertestuale*, 2016, btfp.sp.unipi.it, nota 63) afferma infatti che «l'umanità di Kant ha poco a che vedere con l'appartenenza alla specie *Homo sapiens sapiens*».

[15] Cfr. S. Kember, *Cyberfeminism and Artificial Life*, London, 2003, 3.

[16] *Paper sui principi giuridici*, coordinamento di A. Pajno, in *Statuto etico e giuridico dell'IA*, Fondazione Leonardo, 70.

[17] Cfr. il testo fondativo di D. O. Hebb, *The Organization of Behavior. A Neuropsychological Theory*, New York, 1949, 127-128.

inserisca nel programma della macchina criteri specifici per il riconoscimento. Tale meccanismo si basa, però, sull'idea che l'oggetto da riconoscere sia già dato e che l'allenamento della rete neurale sia volto a decifrare un insieme di informazioni (ad esempio, immagini) con una *performance* che si avvicini sempre di più alla risposta 'ideale' o 'desiderata'[18]. Questa 'risposta', peraltro, costituisce l'unico risultato apprezzabile del funzionamento della *neural network*, posto che quanto accade all'interno del processo elaborativo degli *input*, in linea di massima, presenta una complessità tale da risultare inaccessibile all'essere umano. Si tratta, dunque, di una *black box*, nella misura in cui abbiamo davanti «un sistema le cui operazioni sono misteriose», potendo noi osservarne solo *input* e *output*[19].

Ora, se si volesse applicare il funzionamento di questo meccanismo al tema delle cosiddette *generalised AI*, sopra menzionate, si dovrebbe immaginare che la rete neurale sia in grado di realizzare una sorta di 'apprendimento' anche di certi valori, insieme ai criteri tramite cui applicarli nell'ambito di concrete decisioni politiche. Non si tratterebbe – si badi – di tecnologie tese alla mera applicazione di 'valori' e criteri 'immessi' nell'intelligenza artificiale, dal momento che, in questo caso, tali valori e criteri dovrebbero ricondursi al suo creatore o sviluppatore, dei quali la macchina incorporerebbe tutti i *bias*. L'idea delle *IA* generalizzate sembra, invece, quella di realizzare un meccanismo di apprendimento di valori *interno* alla macchina stessa, ossia una sorta di 'ragione pratica' artificiale, sulla base dell'assunto – abbastanza arbitrario, a dire il vero – che l'essere umano in senso biologico maturi i propri valori in questo modo, cioè ricavandoli da una elaborazione interiore dei dati e non, invece, dalle proprie interazioni sociali.

In altri termini, si dovrebbe quantomeno dubitare che un processo computazionale possa condurre all'apprendimento di 'valori' solo attraverso un'«interazione con l'ambiente» senza il controllo di un essere umano[20]. Si può sostenere, infatti, che l''interazione sociale' rappresenti una *species* della generale 'interazione con l'ambiente', in quanto solo la prima implica necessariamente il rapporto con altri soggetti già portatori di valori: nella seconda, tale circostanza è solo eventuale. Inoltre, la realizzazione (ma anche la mera ipotesi) di un 'manufatto' capace di compiere, in tutto o in parte, operazioni di *decision-making*, si fonda necessariamente su di un assunto assiologico di tipo cognitivistico: in altre parole, per strutturare un meccanismo capace di 'apprendere' determinati contenuti valoriali, si deve giocoforza assumere che tali contenuti siano conoscibili e, in qualche modo, obiettivamente apprezzabili. Un'affermazione simile è tutt'altro che scontata nella 'teoria' e meriterebbe, pertanto, una discreta attenzione anche nella 'prassi', specie quando si vogliano riprodurre meccanismi mentali la cui comprensione risulta estremamente problematica.

[18] Cfr. D. R. Rumelhart, G. E. Hinton, R. J. Williams, *Learning representations by back-propagating errors*, in *Nature*, 323, 1986, 533 ss.
[19] F. Pasquale, *The Black Box Society. The Secret Algorithms That Control Money and Information*, Cambridge, MA, 2015, 3 che ha denunciato, in generale, i problemi del rapporto fra vita sociale e processi computazionali inaccessibili.
[20] Si veda il concetto di «*unsupervised learning*» in R. S. Sutton, A. G. Barto, *Reinforcement Learning. An Introduction*, Cambridge MA, 2014, 2-4.

4. Conclusione. Ulteriori dubbi su 'vita' e 'conoscenza' artificiali

La tematica qui affrontata conduce, a questo punto, a porsi una questione ulteriore, dal momento che l'operazione teorica con cui il postumanesimo accosta l'umano al non-umano manifesta un problema tipico dell'argomento analogico. Attraverso di esso, invero, sembra possibile liberare il campo da differenze apparenti, fantasmatiche e talvolta ideologiche, rivolgendo l'attenzione alle somiglianze tra gli elementi presi in considerazione (in questo caso, fra la mente umana e l'intelligenza artificiale). Non si deve dimenticare, tuttavia, che l'apprezzamento di somiglianze e dissomiglianze si fonda sovente su criteri di rilevanza per nulla 'obiettivi' o ampiamente condivisi. In ogni caso, l'attenzione alle caratteristiche simili va prestata contestualmente alla debita considerazione delle possibili differenze intercorrenti fra i vari modelli e così fra il singolo modello e le sue implementazioni. Il rischio insito in quest'analogia, com'è stato notato anche di recente[21], è dunque quello di occultare alcuni tratti distintivi tra la vita biologica e quella artificiale che potrebbero, invece, ritenersi assai rilevanti.

Il problema, infine, sembra radicarsi nella definizione generale che, da un punto di vista tecnico-scientifico, possiamo dare a nozioni quali «vita», «apprendimento» e «conoscenza». Dovremmo, quindi, chiederci se i meccanismi cognitivi di *machine learning* e *deep learning* possano davvero includersi nel dominio semantico della conoscenza, cercando di capire quali elementi sia possibile (o comunque scientificamente opportuno) considerare per definire tale dominio e – secondo un'ottica maggiormente pragmatica – se abbia senso porsi questo problema definitorio in relazione allo stato attuale dell'avanzamento tecnologico. D'altronde, l'argomentazione con cui, negli anni Ottanta, John Searle rispondeva negativamente a questa domanda si fondava sulla convinzione che sarebbe inappropriato applicare la nozione di 'intelletto' (*understanding*) ad operazioni computazionali relative ad elementi 'formalmente definiti', come simboli o immagini, e sulla stretta correlazione fra la 'mente' ed i processi causali interni al cervello umano[22]. La possibilità di riprodurre la 'mente' umana, ammesso che possano davvero stabilirsene gli attributi, sarebbe dunque preclusa ad ogni implementazione elettronica ed alla dimensione computazionale. Posta in questi termini, la questione sembra rilevante non soltanto sotto un profilo teorico-epistemologico, ma anche da un punto di vista pratico-normativo, dal momento che – come si è sottolineato in avvio – la sussunzione concettuale di un'entità sotto una certa categoria, piuttosto che sotto un'altra, può incidere in maniera dirimente sulle decisioni politiche adottate in ordine ad ogni rapporto giuridico che la coinvolga.

Pier Giuseppe Puggioni
Dip.to di Giurisprudenza
Università di Pisa
piergiuseppe.puggioni@phd.unipi.it

[21] Cfr. Fossa, cit., 68.
[22] J. R. Searle, *Minds, brains, and programs*, in *The Behavioral and Brain Sciences*, 3, 1980, 424.