Aldo Pisano

La macchina e le forme dell'azione: deficit fronetico e autonomia artificiale

Abstract: Starting from the difference between "strong-AI" and "weak-AI", this paper wants to underline the present and future questions coming out from the development of Artificial Intelligences and their collocation in the field of applied ethics. Firstly, the paper will focus on the present time, then on the ethical and juridical implications of weak-AI uses. Secondly, it will go deeply into the ethical and anthropological matters, which come out from the possibility of a strong-AI use. About the second issue, the work will underline the difference between human and machine's act; in this sense, the machine probably has no skills in adapting its action to a particular moral situation (phronetic deficiency). That's why it has to be considered as a sub-symbolic entity, out of the "world of meaning" where all ethical acts have a purpose and action's motivations can be explained.

L'autunno a noi
Promette primavera
A voi l'inverno
Ian McEwan*, Macchine come me, persone come voi*

0. "Da dove veniamo?" Le intelligenze artificiali oggi

Nell'era contemporanea, l'attuale irruenza del digitale si sta esplicando sempre di più nella forma delle intelligenze artificiali, portando all'emergenza di problemi attuali e futuribili.

Su un primo livello, quello del presente, è fondamentale un disciplinamento etico-giuridico degli attuali utilizzi dell'intelligenza artificiale in vari ambiti della vita. In questi termini, l'IA del presente si mostra nella sua *weak-version*, ossia forme di intelligenza semi-autonome o eteronome che necessitano ancora della supervisione umana: veicoli, armi, robot chirurgici. Questo ambito permette, inoltre, di individuare una catena di comando e programmazione, tale da poter identificare con chiarezza chi sia imputabile e/o responsabile degli effetti prodotti delle macchine. Come scrive Gugliemo Tamburrini:

> Il pilota umano del drone prende decisioni di attacco o esegue decisioni di attacco prese dai suoi superiori. La catena di comando e controllo è in linea di principio in-

Mechane, n. 1, 2021 • Mimesis Edizioni, Milano-Udine Web: mimesisjournals.com/ojs/index.php/mechane • ISBN: 9788857581156

> tatta e le responsabilità sono sempre riconducibili a un essere umano. Non è invece ovvio che la catena di comando e controllo rimanga intatta nel caso di un'arma autonoma abilitata a prendere ed eseguire decisioni di attacco indipendentemente da ogni operatore umano[1].

Infatti, in questo primo stadio la mancanza di coscienza e consapevolezza di sé non permettono alla macchina di avvertirsi come autore della scelta morale[2].

Sempre su questo primo livello, non solo è necessario sollevare degli elementi di problematicità tali da far sì che l'IA assuma una propria identità etico-giuridica, ma anche considerare il ruolo che possiedono formazione e informazione: una questione su cui, già nel 2017, in Italia si è pronunciato il Comitato Nazionale di Bioetica[3].

È necessario inquadrare una doppia implicazione etica emergente dall'impiego di una *Weak-AI*: la prima si potrebbe riassumere nella locuzione "etica del programmatore", la seconda in quella di "etica per l'utente". Nel primo caso è necessario sostenere una formazione efficiente dal punto di vista etico, giuridico e filosofico garantendo un approccio interdisciplinare all'IA per chi programma. Spesso, infatti, il pericolo in cui si incorre è quello di un "pregiudizio (*bias*) positivo di automazione"[4] che induce a guardare alla macchina in quanto infallibile. L'apparato giuridico[5], tecnico, politico ed etico deve porsi a garanzia di un criterio fondamentale che non è quello utopistico della infallibilità, ma realistico della *trustworthiness* (affidabilità) per come sottolineato nella *White Paper – On Artificial Intelligence – A European approach to excellence and trust* firmata dalla Commissione Europea il 19 febbraio 2020. Qui è promossa l'idea di una visione ecosistemica della fiducia, necessaria a regolamentare il rapporto IA-Uomo, a garanzia e tutela della quale si deve porre lo Stato[6]; una questione già nota, quando Francis Fukuyaman nel 2002, a proposito dello sviluppo e delle applicazioni delle biotecnologie, sentenziava: "We should use the power of the state to regulate it"[7].

Il riferimento precedente alla *trustworthiness* è a una concezione di fiducia frutto di formazione e informazione critica, garantite da (a) un approccio basato sul rischio, (b) una prospettiva umano-centrica (che non significa mero antropocentrismo) e (c) rispettosa dei diritti fondamentali dell'individuo. In questo senso,

1 G. Tamburrini, *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Carocci, Roma 2020, p. 82.

2 Cfr. Ivi, p. 137.

3 CNB e CNBBSV, *Sviluppi della robotica e della roboetica*, 17 Luglio 2017. Il VI punto "Raccomandazioni" viene infatti modulato su specifici ambiti: sociale; medico; militare, di polizia e di sorveglianza; giuridico (pp. 35-36).

4 G. Tamburrini, *op. cit.*, p. 102.

5 A questo proposito si veda: U. Ruffolo (a cura di), *Intelligenza artificiale. Il diritto, i diritti, l'etica*, Giuffrè, Milano, 2020.

6 Cfr. OECD Employment Outlook, *The future of Work*, 2019.

7 F. Fukuyama, *Our posthuman future. Consequences of the biotechnology revolution*, Farrar, New York 2002, p. 10.

anche la seconda implicazione etico-giuridica relativa all'utente deve garantire un approccio costruttivo che informi in senso "glocal" intessendo le macro-esigenze con le micro-esigenze, soprattutto per garantire la sicurezza della gestione dei dati, la *privacy* e l'inclusione.

Tutto questo affinché possano evitarsi forme di diseguaglianza sociale, determinate da una potenziale *robotic-divide* ossia da una non-equanime garanzia di accesso alle tecnologie digitali e all'informazione. Questo dato è da considerarsi come un'emergenza da iscrivere fra le etiche applicate, attraverso anche una misurata e modulata forma di valutazione dell'uso delle intelligenze artificiali in specifici campi, come quello medico, focalizzando l'attenzione sul piano qualitativo della vita. Infatti, l'uso di IA operanti in base a meccanismi computazionali, potrebbero rendere sempre più effettivo l'appiattimento del piano qualitativo della vita su quello quantitativo, tenendo fuori ogni possibile paradigma di etica normativa (deontologico, etica delle virtù, etica della cura) che non sia quello consequenzialistico. Quest'ultimo da intendersi nella sua declinazione utilitaristica più basilare, quindi come un calcolo mezzi-fini, causa-effetto che trova una più facile corrispondenza nel codice binario secondo cui opera la macchina.

Considerando una finestra temporale più ampia, il problema si potrebbe presentare con quella tipologia di robot simili all'essere umano dal punto di vista cognitivo e pratico; tuttavia, anche considerata questa possibilità, i rischi dell'intelligenza artificiale partono dallo stato presente di cose.

Ad esempio, prendendo in considerazione i veicoli autonomi programmati secondo un'etica consequenzialistica, allora si avrà una variazione della programmazione delle scelte in riferimento all'orizzonte temporale considerato. Per cui, se il riferimento temporale è a breve termine, allora la programmazione avverrà secondo un certo criterio. Infatti, la scelta di adottare determinati sistemi di monitoraggio dell'azione dell'IA significa considerare che esse potrebbero produrre delle modifiche nel costume vigente e nel senso civico[8].

Qui si ha un primo rilevante effetto, per cui scelte nella programmazione dell'IA implicano mutamenti nel costume e quindi dell'ordine etico-giuridico vigente, da qui il ruolo critico che si assumono documenti come lo *Statuto etico e giuridico per l'Intelligenza Artificiale* della Fondazione Leonardo[9].

In questo senso, sia per le questioni attuali (*Weak*-AI), sia per le questioni futuribili etico-antropologiche (*Strong*-AI), è utile un approccio filosofico-critico per come promosso dal movimento dell'Umanesimo Digitale, in cui si sottolinea il disciplinamento dello sviluppo tecnologico in virtù di un'idea di macchina che affianchi l'uomo, ma non lo sostituisca. Per questo l'Umanesimo Digitale utilizza un approccio critico che "influenzi la complessa interazione tra tecnologia e persona, per una società e una vita migliori, nel pieno rispetto dei diritti umani universali"[10].

8 Cfr. G. Tamburrini, *op. cit.*, p. 26.
9 Cfr. Fondazione Leonardo, Civiltà delle Macchine, *Statuto etico e giuridico dell'IA*, 2019.
10 *Manifesto di Vienna per l'umanesimo digitale*, Vienna, maggio 2019, p. 2.

Attraverso un orientamento concreto, basato sul presente e sulla valutazione dei rischi futuri, si può evitare che la filosofia e la riflessione etica sull' IA facciano la propria comparsa solo "sul far della sera". Proporre delle riflessioni etiche *ex ante*, anziché *ex post* permette di uscire dalla visione della filosofia come "nottola di Minerva"[11], garantendo un approccio critico-produttivo che parte dalla constatazione dello stato presente delle cose, dai problemi emergenti e passibili di degenerazione a meno di un intervento garantito dalla riflessione etica che ne limiti i tragici sviluppi; così, l'IA rientrerebbe a pieno titolo nelle etiche applicate[12].

1. "Chi siamo?": Il futuro è alle porte

Per quanto riguarda la *Strong-AI*, oggi si chiamano in causa modelli teorici sui processi cognitivi umani che tendono a rintracciare quanto tali processi siano accostabili a un modello computazionale, tra cui quello dello HIP (*Human information processing*). Secondo questa interpretazione i processi cognitivi di base operano secondo meccanismi quali l'elaborazione delle informazioni, l'analisi del compito, l'uso di una metodologia sperimentale e l'auto-modificazione. Nello specifico, si intende inquadrare il meccanismo di auto-modificazione messo in atto da parte dell'uomo in senso etico e la sua mutuazione in un ipotetico caso di "singolarità tecnologica".

Tale analisi evidenzia la necessità di accumulo di un'esperienza morale del mondo per la macchina, che permette una coincidenza non solo fra la modifica dell'azione e la relazione percettivo-emozionale del mondo, quanto anche l'attingibilità della macchina da un eventuale deposito di *"ethical big data storage"*, spesso filtrato dai programmatori. Tuttavia solo rispondendo a una causalità per libertà è possibile che la macchina imiti l'intelligenza umana, sopportando un margine di errore in riferimento a predizioni etiche o giuridiche[13].

Rispetto alle urgenze etico-sociali sollevate dalle intelligenze artificiali nel contesto delle relazioni umane, esistono validi approcci teorici utili a controbilanciare l'idea di una versione *strong* dell'IA. In questi termini, i processi di digitalizzazione dell'umano e il prevalere di modelli sopra elencati (*HIP*) vengono sfavoriti rispetto a quelli di umanizzazione del digitale, qui da intendersi come la costruzione coerente di una regolamentazione etica e di diritto che possa definire l'estensione dei doveri dell'uomo. Tale regolamentazione – come si diceva – si pone a fondamento della garanzia dei diritti individuali fondamentali, valorizzando l'uomo in quanto essere significante, evitando così che la spinta alla capitalizzazione possa soffocare il valore etico-esistenziale di cui l'agente morale si fa portatore nei vari contesti di vita.

11 Cfr. F. G. Hegel, *Lineamenti di filosofia del diritto* a cura di G. Marini, Laterza, Roma-Bari 2004 (9ª ed).

12 A. Fabris (a cura di), *Etiche applicate. Una guida*, Carocci, Roma 2018, p. 11.

13 Cfr. Ivi, p. 46.

Dunque, spostando l'attenzione verso il futuro, si configura un quadro ancora più problematico ma non meno complesso di quello attuale. Infatti, sempre considerando il processo conoscitivo umano come un modello computazionale, la possibilità di avvicinare il funzionamento della macchina all'essere umano si basa su una concezione sostanzialmente riduzionistica. Questo tipo di interpretazione spingerebbe nella direzione di un appiattimento e annullamento della complessità umana, un dato ormai ineludibile per l'antropologia filosofica.

La conoscenza dei principi, l'esperienza etica del mondo, il sedimentarsi nella memoria di una serie di esperienze dal rilevante impatto emotivo concorrono alla costruzione di una coscienza e di un'identità radicata che, nel momento dell'azione, rende il soggetto autonomo, nonché "autore morale". Il punto di discrimine che oggi permette alla macchina di non essere ancora umana è nella possibilità di riplasmare l'universo morale mediante nuove forme dell'agire che, come analizza Hannah Arendt nel rapporto sussistente fra la ripetitività del lavoro e la performatività dell'azione, costituiscono l'atto di libertà in quanto creazione del nuovo o esperienza del "cominciamento". Sempre in riferimento alla lettura arendtiana nel rapporto automa-autonomia, infatti, non è un caso che l'inserimento nella dimensione inter-soggettiva (*inter homines esse*)[14] si configuri nell'azione:

> Agire, nel senso più generale, significa prendere un'iniziativa, iniziare (come indica la parola greca *archein*, "incominciare", "condurre", e anche "governare"), mettere in movimento qualcosa (che è il significato originale del latino *agere*). Poiché sono *initium*, nuovi venuti e iniziatori grazie alla nascita, gli uomini prendono l'iniziativa, sono pronti all'azione[15].

Per come evidenziato in *Vita Activa,* la macchina si colloca nella prospettiva del lavoro, nel mondo della necessità; un mondo a cui, di certo, l'uomo appartiene, ma dal quale può dissociarsi proprio grazie all'azione come categoria che attualizza la libertà. Il tema dell'automazione come pericolo è noto ad Arendt che, nel *prologo* dell'opera sopra citata, scrive: "un altro evento non meno temibile, l'avvento dell'automazione, che in pochi decenni vuoterà probabilmente le fabbriche e libererà il genere umano dal suo più antico e più naturale fardello, il giogo del lavoro e la schiavitù della necessità"[16]. In virtù di queste considerazioni di ordine teorico è possibile comprendere la strutturale differenza dell'uomo dalla macchina; questa, al momento, può creare un proprio magazzino di dati eticamente connotato, può conoscere, emulare, simulare comportamenti, ma la vera sfida è comprendere se potrà essere capace di pensare, di esprimere giudizi, di plasmarsi come coscienza autonoma, di porsi sullo stesso piano dell' *homo symbolicus.* Altro monito della Arendt è infatti:

14 H., Arendt, *The Human Condition*, The University of Chicago, U.S.A. 1958; tr. it. di S. Finzi, *Vita Activa. La condizione umana,* Bompiani, Milano 2014 (18ª ed.), p. 128.

15 Ivi, p. 7.

16 Ivi, p. 4.

> Se la conoscenza […] si separasse irreparabilmente dal pensiero, allora, diventeremmo essere senza speranza, schiavi non tanto delle nostre macchine quanto della nostra competenza, creature prive di pensiero alla mercé di ogni dispositivo tecnicamente possibile, per quanto micidiale[17].

Dunque, ciò che è noto della macchina è che essa si ferma al livello di riproduzione di processi conoscitivi di base, di comportamenti automatici, quindi incapace di introspezione e di spiegare le proprie azioni, rimanendo un semplice "gioco di imitazione" o una *black box*. Di fronte alla possibilità di un'intelligenza simile all'uomo, che diventa sempre più reale mediante forme di sperimentazioni che includono la creazione di reti neurali artificiali, è necessario un intervento etico-antropologico.

In senso etico è necessario porre attenzione su tutte le attività di ricerca e di sviluppo tecnologico che abbiano come fine quello della creazione di uno stato di super-intelligenza; in termini antropologico-filosofici bisogna porsi una domanda di senso più radicale, sul valore che assume la creazione di questo tipo di intelligenza. L'aspirazione alla creazione del robot umanoide potrebbe essere l'ennesimo sintomo di una tendenza antropocentrica, una necessità dell'uomo di sopravvivere a se stesso, di eternizzarsi, di deviare l'obsolescenza.

2. "Dove andiamo?". Umanizzare gli algoritmi

La prevalenza di un punto di vista tecnocentrico significherebbe l'apoteosi di una visione del mondo univoca e riduzionistica, che oggi trova ampi consensi in molte correnti che fanno riferimento alle neuroscienze, nonché nella versione forte dell'intelligenza artificiale.

Stando a questa prospettiva, l'equiparazione della macchina con l'uomo ridurrebbe tutto a un rapporto fattuale e deterministico tale che l'intelligenza – nella sua forma più evoluta – si riveli essere solo come una buona capacità di imitazione e simulazione di determinati comportamenti, come già mostrava il test di Turing[18]. Questo rende ulteriormente evidente la sostituibilità dell'umano con forme di intelligenza più evolute in grado di riprodurre i suoi comportamenti in maniera più efficace, precisa ed economicamente meno dispendiosa.

In questo caso, ovviamente, si parla di una semplice riproduzione di movimenti connessi da una relazione causale, quindi non teleologicamente orientati. Questo delinea la differenziazione fra movimento e azione, ben nota agli studi di Rizzolatti

17 Ivi, p. 3. In riferimento ad Arendt si vedano anche: Arendt H., *Lesson's on Kant political philosophy,* a cura di R. Beiner, University of Chicago Press, Chicago; tr. it *Teoria del giudizio politico,* a cura di P. P. Portinaro, Il Nuovo Melangolo, Genova, 2005 (2ª ed); Id., *Responsibility and Judgment*, a cura di J. Kohn, Schocken Books, New York 2003; tr. it. a cura di D. Tarizzo*, Responsabilità e giudizio*, Einaudi, Torino 2010 (2ª ed.).

18 Cfr. A. Turing, *Computing Machinery and Intelligence,* "Mind", Vol. 59, N. 236 (October), Oxford University Press, Oxford 1950, pp. 433-460.

e Sinigaglia sui neuroni specchio[19] e che apre le porte allo studio dei meccanismi empatici e alla loro funzionalità simbolico-culturale nel mondo delle relazioni inter-soggettive[20].

L'azione si permea di un significato a cui una visione tecno-centrica non lascerebbe spazio. Il predominio di una prospettiva meccanicistica mostra lo scenario futuro dell'intelligenza artificiale, dei veicoli autonomi e semiautonomi, dello stato di singolarità tecnologica come il risultato naturale di un processo evoluzionistico che non risparmia di lasciare l'uomo in semplice linea di continuità rispetto ad altre specie animali[21]. In tal caso, l'agire umano si appiattirebbe su una teoria dell'atto non performativa, in cui l'etica perderebbe il proprio statuto diventando solo etologia della specie umana.

Questo costituisce una retromarcia sostanziale a visioni deterministiche che non considerano l'ormai matura idea della complessità dell'umano. In questo senso, all'Antropocene – come sostiene James Lovelock – seguirebbe la "Novacene"[22], un'era del dominio delle macchine, che è risultante naturale dell'evoluzione del mondo umano. Di contro all'affiorare di queste nuove teorie, prospettive e ipotesi che riconfigurano il rapporto tra uomo e macchina, torna in soccorso l'umanesimo digitale che "sostiene una digitalizzazione che vada a vantaggio degli esseri umani, opponendosi alla riduzione dei singoli individui a unità funzionali all'interno di una sistema ottimizzatore normato e anonimizzato sotto la guida di un software"[23].

Per comprendere i processi di apprendimento della macchina, è bene considerare le intelligenze artificiali come "agenti reattivi basati sul modello", strettamente legate ai processi di *machine learning*[24]; oltre questi, la complessità della macchina

19 Nella loro trattazione, Rizzolatti e Sinigaglia, infatti, assumono una specifica area cerebrale che denominano F5 la cui maggior parte dei neuroni "*non codifica singoli movimenti, bensì atti motori*, cioè movimenti coordinati da un fine specifico" (G. Rizzolatti, C. Sinigaglia, *So quel che fai. Il cervello che agisce e i neuroni specchio*, Raffaello Cortina, Milano 2006, p. 25. Corsivo nel testo).

20 Cfr. G. Rizzolatti, C. Sinigaglia, *op. cit.*, pp. 181-182. Si veda anche: P. Dumouchel, L. Damiano, *Vivre avec les robots. Essai sur l'empathie* artificielle, Éditions du Seuil, Paris 2016; tr. it. a cura di L. Damiano, Vivere *con i robot. Saggio sull'empatia artificiale*, Raffaello Cortina Editore, Milano 2019.

21 La *conditio humana* è legata alla socialità intesa come edificazione di sé e del sociale attraverso la relazione con gli altri. Fare questo significa costruire narrazioni che rappresentino l'identità stessa del soggetto che si esprime esponendosi mediante l'azione, e che "produce storie, con o senza intenzione, con la stessa naturalezza con cui la fabbricazione produce cose tangibili. […] ognuno *incominci[a] la propria vita* inserendosi nel mondo umano attraverso l'azione e il discorso" [H., Arendt, *The Human Condition*, cit., p. 134].

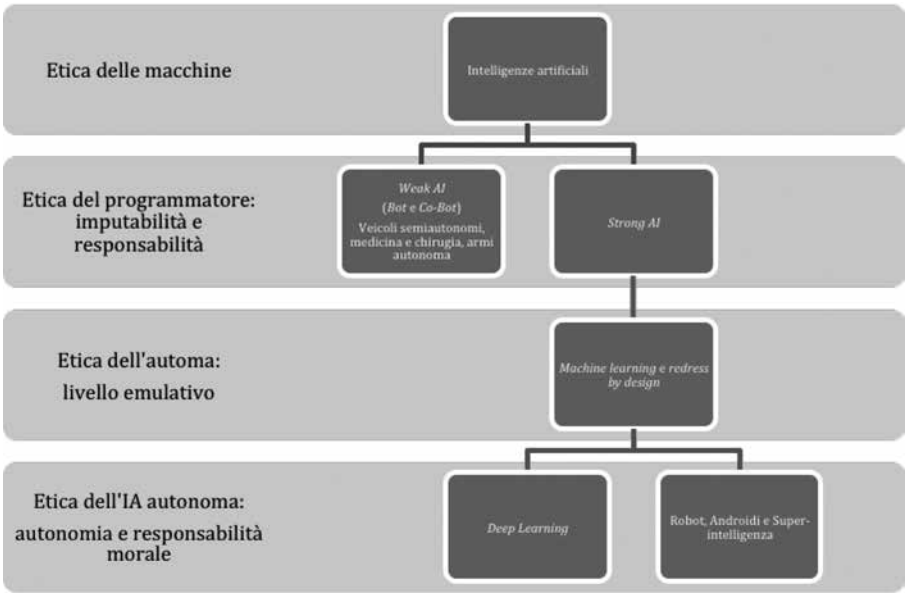
22 Cfr. J. Lovelock, *Novacene. The Coming Age of Hyperintelligence*, Penguin Books Ldt, London 2019; tr. it. a cura di A. Panini, *Novacene. L'era dell'iperintelligenza*, Bollati Boringhieri, Torino 2020.

23 J. Nida-Rümelin, N. Weidenfeld*, Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz*, Piper Verlag GmbH, München/Berlin 2018; tr. it. a cura di G. B. Demarta, *Umanesimo digitale. Un'etica per l'epoca dell'intelligenza artificiale*, FrancoAngeli, Milano 2019, p. 72.

24 Cfr. P. Domingos, *The Master Algorithm. How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, New York 2015; tr. it. a cura di A. Migliori, *L'algoritmo definitivo: la macchina che impara da sola e il futuro del nostro mondo*, Bollati Boringhieri, Torino 2020.

aumenta con il *deep learning*, per proiettarsi verso l'idea di una singolarità tecnologica. Qui, si parla di una forma di intelligenza che attua processi di modifica grazie allo scambio con un ambiente in cui la stessa intelligenza si trova collocata. Arrivare all'auto-coscienza della macchina significa introdurre processi in cui le macchine pensano, agiscono e quindi sono autonome non solo perché operano indipendentemente dall'uomo, ma perché sono in grado di conferire un significato alle loro azioni e dunque di esprimerlo[25].

Proprio sulla soglia tra un'intelligenza emulativo-rispondente e una simulativo-responsabile[26] sta il punto di discrimine che permette di distinguere l'essere umano dalla macchina. In questo senso viene chiamato in causa il concetto aristotelico di *phronesis*.



Schema 1 – Schema esemplificativo dei livelli di analisi etico-filosofica relativa all'intelligenza artificiale

25 Cfr. G. Tamburrini, *op. cit.,* 109-110.

26 Uno degli attuali sviluppi relativi alla possibilità di creare tramite algoritmi delle scale valoriali e dei principi nelle macchine è data dall'idea delle CP-nets (Conditional Preferences – net), attualmente oggetto di studio e teorizzazione. Queste si basano sulla riproduzione delle capacità di scelta e giudizio umani mediante un sistema basato sulle preferenze, mediante la programmazione di una sistema deontologico nella macchina. [cfr. Calegiari R., Loreggia A., Lorini E., Rossi F., Sartor G., *Modeling Contrary-to-Duty with CP-nets, ResearchGate*, 23 Marzo 2020]. La ricerca è disponibile al seguente link: https://www.researchgate.net/publication/340134395_Modeling_Contrary-to-Duty_with_CP-nets/link/5eac5f46a6fdcc7050a18084/download].

Questa necessità di rendere la macchina non solo utile all'uomo (es. per operare nei lavori ripetitivi), ma addirittura di ricrearlo e sostituirlo, apre lo scenario di riflessione antropologico-filosofico precedentemente evocato. Infatti, se la macchina deve essere progettata simile all'uomo nei processi di *decision making* deve contemplare la possibilità dell'errore perché abbia autonomia; questo implicherebbe un evidente controsenso rispetto alla prospettiva di creazione algoritmica di un automa morale, seppure rinforzi ancora di più l'idea di affidabilità, anziché quello di infallibilità. Una delle possibilità inconsce che soggiace alla creazione di questo tipo di IA potrebbe essere quella di sgravare il soggetto dalla responsabilità e imputabilità dell'azione, conferendola a un'entità impersonale. Analizzata in questi termini, la super-intelligenza risulterebbe finalizzata alla deresponsabilizzazione morale o, peggio, all'anonimia morale.

Tuttavia, proprio l'uso dell'immaginazione che proietta spesso verso un futuro distopico (qui si rinvia alla narrazione filmica e letteraria riferita agli androidi)[27] danno, per loro parte, impulso alla necessità di un'educazione tecnologica e al suo disciplinamento. Il futuro, inteso come proiezione, ha effetti sul presente, spingendolo in avanti verso un certo modo di regolamentare la realtà del rapporto uomo-macchina: la distopia o l'utopia del futuro, così, orientano lo *status* presente.

3. "Perché lo facciamo?": Il deficit fronetico

Quella che si configura essere la differenza sostanziale fra la macchina e l'uomo, si potrebbe evidenziare nel fatto che l'uomo agisce in base all'esperienza e alla conoscenza morale del mondo. Questo implica che l'agente morale, mediante un processo che pone in una relazione circolare teoria e prassi, riesce ad ampliare la propria *Weltanshauung* e le sue manifestazioni etiche, allo stesso tempo affinando il proprio senso morale.

Nei processi di consolidamento dell'identità etica intervengono fattori differenti che si possono raccogliere sia dalla tradizione classica, sia dalla tradizione contemporanea in riferimento agli studi delle neuroscienze[28]. Nello specifico, qui si vuole considerare il concetto aristotelico di *phronesis* in quanto virtù dianoetica, quindi non semplicemente legata alla conoscenza del principio, ma alla sua possibilità applicativa in contesti e situazioni differenziati e che si riferiscono anche alla cura complessiva che l'agente morale ha del proprio carattere. Scrive Aristotele:

27 "L'importanza dell'immaginazione nella vita etica spiega anche la rilevanza che viene ad assumere la letteratura e più in generale l'arte: grazie a essa il lettore è invitato a sentire e a guardare il reale in una determinato modo. […] La letteratura consente di indagare l'esperienza morale del soggetto, fornendo interpretazioni e resoconti che un teoria morale, specie di tipo impersonale, non sarebbe in grado di fornire". [A. Da Re, *Le parole* dell'etica, Mondadori, Milano 2010, pp. 43-44].

28 Si vedano: S. Pollo, *La morale della Natura*, Laterza, Roma-Bari 2008; A. R. Damasio, (2007) "Neuroscience and ethics: intersections", in *American Journal of Bioethics*, 7, pp. 3-7; L. Boella, *Neuroetica. La morale prima della morale*, Raffaello Cortina, Milano 2008.

> Potremmo comprendere cosa sia la saggezza nel modo seguente: osservando quali persone noi diciamo sagge. […] noi chiamiamo 'saggi' anche quelli che si limitano a un qualche ambito particolare, quando calcolano bene in vista di un qualche fine eccellente che non sia oggetto di qualche arte, cosicché, anche in generale, chi sa deliberare sarà saggio. Ma nessuno delibera su ciò che non può essere diversamente, né sulle azioni che non possono essere compiute da lui; e quindi, se è vero che la scienza procede per via dimostrativa, ma che di ciò i cui principi possono essere diversamente non si dà dimostrazione – infatti tutte le cose di questo tipo possono anche essere diversamente – e che su ciò che è per necessità non è possibile deliberare, allora la saggezza non sarà né scienza né arte. Non sarà scienza, perché il contenuto dell'azione è cosa che può essere diversamente, e non sarà arte perché azione e produzione rientrano in generi diversi. Infatti il fine della produzione è diverso dalla produzione stessa, mentre quello della prassi non lo è, dato che lo stesso agire con successo è fine. Allora rimane solo che la saggezza sia uno stato abituale veritiero, unito a ragionamento, pratico che riguarda ciò che è bene e male per l'uomo[29].

L'esperienza morale (osservazione dell'uomo saggio), la conoscenza delle norme (scienza) e il continuo incontro dell'agente con nuove situazioni morali impongono una necessità di ridisegnare la scala valoriale anche in base ai diversi contesti che, di volta in volta, definiscono la necessità di nuove forme dell'agire. Dunque, l'agente morale dirigerà la propria azione, ossia delibererà, in virtù di una valutazione delle condizioni in cui si trova ad agire[30]. Questo agire pratico-situazionale, tuttavia, necessita di una flessibilità che al momento non è presente nei modelli di comportamento con cui è programmabile il *machine learning*, che rimane fortemente ancorato all'utilizzo di un codice binario il quale, spesso, non lascia spazio a una "terza via" che invece l'attività dell'agente morale umano riesce a percorrere, evadendo da rigidi schemi applicativi. Dunque, la *phronesis* va a configurarsi come una vera e propria facoltà di giudizio morale[31], in cui si manifestano congiuntamente libertà e responsabilità del soggetto.

La *phronesis* richiama la questione secondo cui la macchina *conosce* ma continua a non agire in base a una propria "ragione pratica", poiché non dispone di quegli strumenti che sono essenziali per l'atto etico, in termini situazionistici. Per questo motivo, la macchina soffre di un vero e proprio "deficit fronetico" anche riconducibile all'assenza della dimensione somato-percettiva e di quella emotiva[32].

L'attività morale frutto del pensiero (*diànoia*) è da intendersi – ritornando sulla prospettiva arendtiana – come un dialogo interiore; essa si interconnette al proces-

29 Aristotele, *Etica Nicomachea,* a cura di C. Natali, Laterza, Roma-Bari 2005, p. 231; A proposito si vedano: P. Aubenque, *La prudenza in Aristotele*, Studium, Roma 2018; E. Berti, *Le ragioni di Aristotele*, Laterza, Roma-Bari 1989; C. Natali, *Le virtù particolari nell' "Etica Nicomachea" di Aristotele*, in P. Donatelli, E. Spinelli (a cura di), *Il senso della virtù*, Carocci, Roma 2008; Id., (2014) *Aristotele,* Carocci Roma; J. Mcdowell, *Incontinence and Practial Wisdom in Aristotle*, in Id., *The Engaged Intellect. Philosophical Essays*, Harvard University Press, Cambridge Mass 2009.

30 Cfr. S. Songhorian, *Etica e scienze cognitive*, Carocci, Roma 2020.

31 Cfr. L. Surian, *Il giudizio morale. Come distinguiamo il bene e il male*, Il Mulino, Bologna 2013.

32 Cfr. Ivi, pp. 72-97.

so di ripensamento delle azioni compiute, gettando le basi per lo sviluppo successivo della coscienza morale[33] mediante un processo di sedimentazione dei ricordi. Così, la memoria riveste un ruolo fondamentale nella costruzione dell'identità personale per mezzo di interconnessioni tra esperienze che si traducono in un olismo del mentale[34]. In conformità a questa lettura esiste un punto di convergenza forte fra memoria ed emozioni, nel momento in cui la vividezza del ricorso si correla all'esperienza emotiva del soggetto e all'intenzionalità.

Bisogna considerare che gli studi condotti dalle neuroscienze evidenziano come fondamentali i meccanismi reattivo-primari della specie, legati alle emozioni, non attuabili dalla macchina a meno di una *riproduzione* di modelli di apprendimento o di creazione di reti neurali artificiali (RNA). Per Antonio Damasio, infatti, è fondamentale la funzione svolta dal marcatore somatico nei processi di *decision making* e per renderli possibili soprattutto in situazioni in cui è necessario agire in un breve lasso di tempo:

> quando viene alla mente, sia pure a lampi, l'esito negativo connesso con una determinata opzione di risposta, si avverte una sensazione spiacevole alla bocca dello stomaco. Dato che ciò riguarda il corpo, ho definito il fenomeno con il termine tecnico di stato *somatico*; […]. Esso forza l'attenzione sull'esito negativo al quale può condurre una data azione, e agisce come un segnale automatico di allarme […]; vi permette di *scegliere entro un numero minore di alternative*[35].

Il ruolo svolto dalla percezione, dalle emozioni che pone in correlazione empatica gli esseri umani, amplificando le relazioni inter-personali, costituiscono gli elementi biologicamente deficitari che non permettono di immaginare la macchina in quanto capace di agire eticamente e, nello specifico, di agire froneticamente. In effetti, anche solo nel caso dei veicoli autonomi o semiautonomi, l'IA può, per errore, produrre danni fatali circoscritti. Per quanto riguarda i veicoli autonomi non è necessario che questi siano 'perfetti', ma produrli affinché il loro margine di errore sia inferiore a quello umano[36].

In relazione a questo, il caso-paradigma è quello della collisione inevitabile, in cui la macchina riesce a prendere decisioni più velocemente e più precisamente di un umano per minimizzare i danni, in quanto c'è minore interferenza emotiva e minore tempo di reazione. Tuttavia, si danno anche possibilità in cui il fattore emotivo risulti essenziale per avere risposte altrettanto veloci[37]. Sostanzialmente, la

33 Cfr. H. Arendt, *Il pensiero e le considerazioni morali*, in Id., *Responsabilità e giudizio*, cit.; Sempre Arendt scrive: "Il filosofo […] trova se stesso, nel dialogo tra 'me e me stesso' (*eme emautō*) in cui Platone individuò evidentemente l'essenza del pensiero" [H. Arendt, *The Human Condition*, cit., p. 55].

34 Cfr. N. Levy, *Neuroethics. Challenges for the 21st century*, Cambridge University press, Cambridge 2007, p. 162.

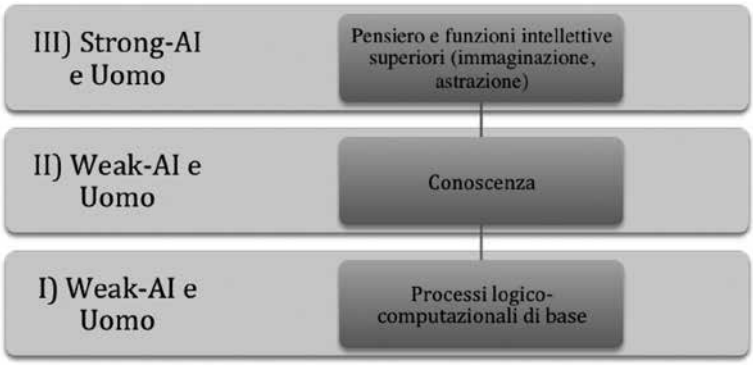
35 A. R. Damasio, *Descartes' Error. Emotion, Reason and the Human Brain*, Putnam Publishing, New York 1994; tr. it. a cura di F. Macaluso, *L'errore di Cartesio L'errore di Cartesio. Emozione, ragione e cervello umano,* Adelphi, Milano 1995*,* p. 245. Corsivo nel testo.

36 Cfr. G. Tamburrini, *op. cit.*, p. 20.

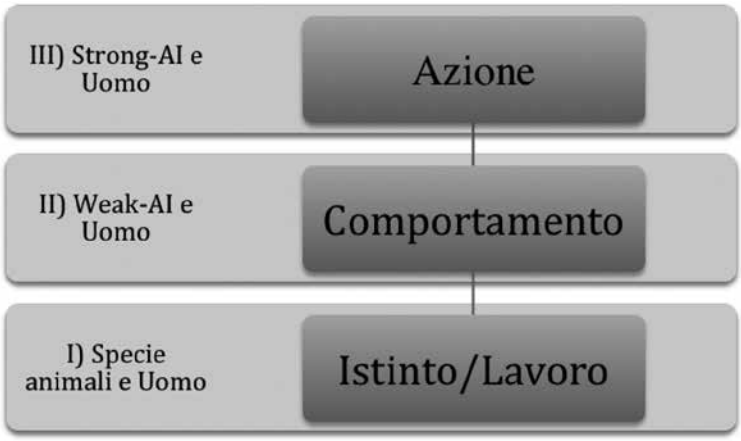
37 Ivi, p. 23.

macchina potrebbe risolvere qualsiasi situazione venendo programmata in un'ottica consequenzialistica e, in fondo, sarebbe questo l'ipotetico criterio secondo cui le macchine stesse arriverebbero ad auto-programmarsi, senza considerare un piano qualitativo dell'azione etica, ma misurando solo in termini quantitativi. In questo senso, ogni dilemma morale diventa un conflitto morale risolvibile, proprio perché esso non ha una connotazione esistenziale (condizione sconosciuta alla macchina che evade il piano del simbolico).

Sintetizzando quanto evidenziato finora, il 'fattore complessità' svolge la funzione di operatore che permette di discriminare l'intelligenza umana da quella artificiale, evitando che la prima possa essere schiacciata sulla seconda. In questi termini, i livelli di complessità potrebbero essere suddivisi in tre principali categorie: parziale, totale e assoluta:



Schema 2 – Ambito teoretico



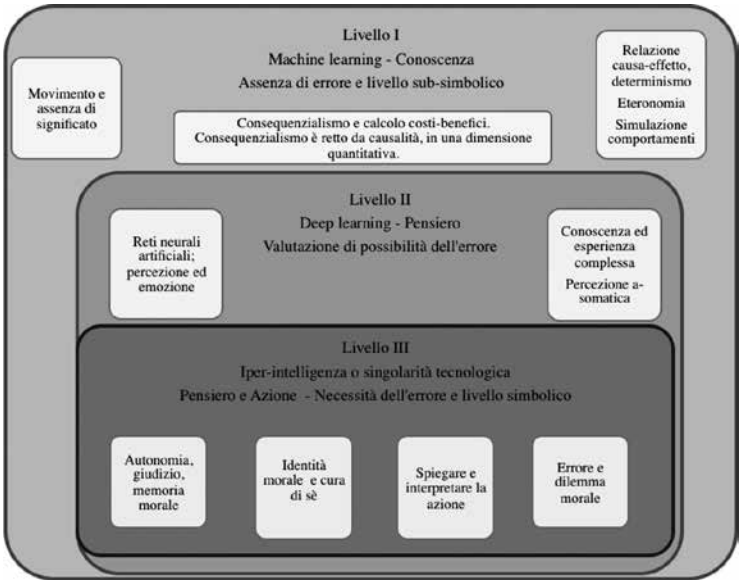
Schema 3 – Ambito etico-pratico

Per quanto riguarda gli Schemi 2 e 3 si evidenzia un livello di complessità crescente che procede dal livello I al livello III. Nel primo caso, i processi cognitivi attuati a livello I sono accomunabili a quelli computazionali della macchina, così come succede per il livello II. Tuttavia, l'accesso al livello III è possibile solo per intelligenze artificiali che siano capaci di riprodurre processi più complessi quali pensiero, immaginazione e, in generale, processi astrattivi o di simulazione mentale che, se combinati alla dimensione emotiva, rendono possibili i meccanismi di immedesimazione empatica. In questa prima classificazione, come si può notare, il riferimento va anche alla distinzione arendtiana, nonché kantiana[38]: è infatti il processo più elevato a livello teoretico (il pensiero) che permette un passaggio verso la dimensione morale, combinandosi con l'ambito etico-pratico. Lo schema 2 rappresenta un livello di complessità *relativo*, in quanto si riferisce esclusivamente all'ambito teoretico, così anche per lo Schema 3 riferito *solo* all'ambito etico-pratico. Anche in questo secondo caso si ha un livello interno di complessità relativa crescente: istinto/lavoro, comportamento e azione. Qui il riferimento corre alla Arendt de *La Vita Activa* e alla distinzione tra lavoro, opera e azione sopra riproposta e qui riformulata in relazione alla macchina. Ora, esiste una relazione di continuità e circolarità fra l'ambito teoretico e l'ambito pratico, in cui la combinazione simmetrica permette una crescita del livello di complessità che in questo caso si presenta come *totale* (teoretico ed etico-pratico insieme).

Un approccio critico-filosofico alla possibilità dell'Intelligenza Artificiale, nel senso sia forte che debole, permette di evitare l'abbattimento del livello III su quelli più bassi che coinciderebbe, di fatto, con una forma di riduzionismo o, in ambito strettamente etico, con una naturalizzazione della morale. Per ottenere, invece, il livello di complessità *assoluto* è necessario aggiungere al livello di complessità totale la dimensione emotiva. In questi termini rientrano nel paradigma di una riflessione etica sull'IA gli studi di Damasio. Se lo studio delle neuroscienze evidenzia come fondamentale l'intervento delle emozioni come valore aggiunto ai processi cognitivi ed etici, allora questa forma di scienza permette la salvaguardia della complessità, evitando forme di riduzionismo neurobiologico e ponendosi in stretta comunicazione con le discipline umanistiche.

Considerando i meccanismi integrativi di apprendimento della macchina – che ne aumentano il valore di complessità – se combinati con l'aspetto etico, si può pensare un triplo livello di acquisizione dell'autonomia che globalmente riassume quanto detto finora, tenendo sempre presente il rapporto tra dimensione cognitiva e dimensione pratica nell'IA (Schema 4).

38 Per ragioni di spazio non è qui argomentata la questione kantiana relativa al rapporto tra ambito teoretico ed ambito pratico in funzione della visione antropologica. Per maggiori approfondimenti si vedano: I. Crispini, *Legge e apparenza morale. La ricerca etica kantiana*, Aracne, Roma 2008; R. B. Louden, *Kant's Impure Ethics. From Rational Beings to Human Beings*, Oxford University Press, Oxford 2000; A. W. Wood, *Kant's Ethical Thought*, Cambridge University Press, Cambridge 1999, pp. 215-225.

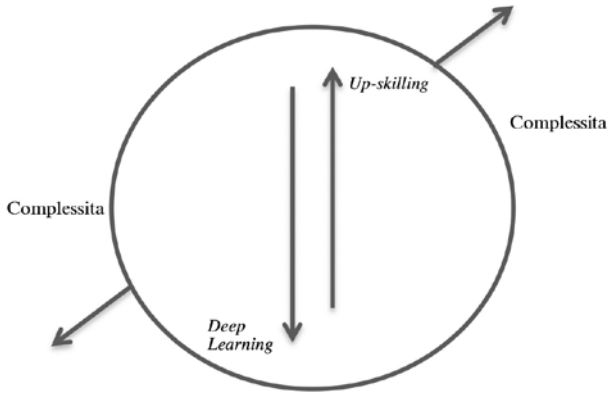


Schema 4 – Schema riassuntivo della stratificazione dei livelli di IA. Procedendo dall'apprendimento semplice, il livello di complessità della macchina si articola grazie all'apprendimento profondo, per approssimarsi sempre di più allo stato di singolarità tecnologica.

In questo processo di acquisizione dell'autonomia, il *deep learning* costituisce un passaggio fondamentale. Infatti, le reti neurali artificiali si presentano in forma stratificata [ing. *layer*]: i livelli più bassi di percezione (suoni, forme, colori) si combinano in strutture più articolate che costituiscono la sintesi di dati semplici. Come si diceva, questo procedimento segue uno sviluppo a modelli integrati, in cui ogni aspetto è inglobato in quello precedente, accrescendo i processi di umanizzazione dell'intelligenza artificiale e mostrando un andamento asintotico tra lo sviluppo della complessità della macchina e quello umano. Se nel processo di apprendimento base (*machine learning*) si aumenta il numero di *layer*, allora è possibile arrivare a un livello più profondo (*deep learning*), dunque a quella struttura di IA capace di interpretare i segnali esterni considerando anche l'aspetto percettivo (qui a-somatico), ossia costruendo una vera e propria conoscenza *a posteriori*, non solo *a priori*, come invece avverrebbe nel caso di una basilare forma di *machine learning*[39]. In sostanza, anche grazie all'utilizzo di "esempi" – che nel lessico morale si traducono nell'assunzione di un modello di etica descrittiva – il computer riesce a operare delle generalizzazioni, potenziando i processi di apprendimento

39 Cfr. F. Rossi, *Il confine del futuro. Possiamo fidarci dell'intelligenza artificiale?*, Feltrinelli, Milano 2019.

automatico e producendo nuove soluzioni e conoscenze. Se a questo si somma l'apporto di altre scienze, quali la robotica, l'ingegneria e la fisica, la possibilità di produrre robot umanoidi spingerebbe sempre di più la macchina dal livello di complessità totale verso quello assoluto (singolarità tecnologica)[40]. Come mostra lo Schema 5, i meccanismi di apprendimento della macchina, nell'andare sempre più in profondità (*deep learning*), determinano un incremento proporzionale delle abilità dell'IA (*up-skilling*)[41], tale che vi sia un'approssimazione continua verso la complessità del modello umano (qui rappresentata dalla struttura del cerchio) che, in maniera altrettanto continua, tende ad allargarsi verso l'esterno. La complessità in aumento della macchina si traduce, in termini matematici, in un accrescimento delle variabili che rendono difficile la possibilità di riprodurre pedissequamente l'attività umana nella sua unicità conoscitiva, pratica e soprattutto in quella totalità che considera entrambe le dimensioni.



Schema 5 – Rapporto tra *machine learning* e *up-skilling*

40 Cfr. Ivi, p. 70. A questo proposito, scrive Pedro Domingos: "L'Algoritmo Definitivo, soprattutto, non dovrà ricominciare da zero ogni volta che affronta un nuovo problema […]. L'Algoritmo Definitivo non è un consumatore passivo di dati: può interagire con l'ambiente circostante e cercare attivamente i dati di cui ha bisogno" [P. Domingos, *op. cit.*, p. 78].

41 Questo specifico processo di apprendimento che tende ad articolarsi in forma sempre più strutturata e profonda, rispecchia il modello umano che, ancora, si configura come *exemplum* per il *design* dell'IA, sia nella sua versione debole, sia in quella forte. Invece, per quanto riguarda il rapporto tra IA, dimensione sociale e abilità umane, un approccio *based-risk* mostra che il prolungato utilizzo delle macchine in ambito professionale (anche nella loro versione debole) può produrre un fenomeno di *deskilling*, di perdita di abilità per mancanza di pratica. Su questo si è pronunciato recentemente il CNB, con particolare riferimento alle professioni medico-sanitarie. Per un rimando si veda: CNB e CNBBSV, *Intelligenza Artificiale e medicina: aspetti etici*, 29 maggio 2020, p. 7.

Allo stato attuale, dunque, l'IA può sicuramente riconoscere una voce, ma non il significato che sottende a una richiesta; può attuare un riconoscimento biometrico, ma non la complessità dei dati esistenziali che genera in un soggetto umano un'espressione rattristata. La comprensione empatica, in questi termini, rappresenta una via di accesso privilegiata alla dimensione intersoggettiva ed etica del mondo, alla possibilità di vicinanza all'altro e alla considerazione della vita in senso qualitativo. Quando l'IA riuscirà ad accedere a questo livello, allora avrà compiuto un passo fondamentale verso l'imitazione complessa del modello umano.

Ora, l'agire secondo un'attività di pensiero capace di valutare un insieme sistemico di condizioni, possibile anche grazie ai processi emotivo-empatici, implica una possibilità che alla macchina – per sua stessa definizione – non si dà: ammettere la possibilità dell'errore. L'IA futura, autonoma sarà programmata per commettere errori dovendo rispondere a una causalità per libertà, solo così può essere imitata a pieno l'intelligenza umana, attraverso un "*redress by design*"[42], ossia una riconfigurazione delle modalità di azione in base all'esperienza accumulata.

Esiste una notevole distanza della macchina rispetto alla dimensione emotiva umana, legata ai processi di codifica dei ricordi che stanno alla base di una "memoria morale", di un'identità etica radicata, dello sviluppo della coscienza e del senso morale[43]. Il lavoro sulle emozioni e l'esperienza morale si intersecano rendendo possibile un agire pratico-situazionale che trae fondamento proprio dal sedimentarsi di ricordi legati agli stati emotivi, a loro volta generati dagli atti morali o immorali. In questi termini, la combinazione fra conoscenza ed esperienza morale diventa indice di una costruzione simbolica dell'io e del mondo. L'autonomia, allo stesso tempo, rende possibile il *sehen als*[44], il "vedere come", la costruzione di narrazioni, intesa come capacità di pensare e non solo di conoscere la realtà. La sfida dell'io artificiale – perché possa dirsi autenticamente morale – si rende proprio nella capacità di edificazione di una memoria semantica e non meramente fattuale. Posta in questi termini, la distanza incolmabile uomo-macchina si configurerebbe nel fatto che l'immagazzinamento del ricordo è sostanzialmente slegato dall'impatto emotivo, non più articolato in una prospettiva di sviluppo interiore, guida e principio della vita come cura della *psychè*.

Quando la macchina da automa diventa autonoma in senso morale sarà non solo capace di agire come l'essere umano, ma di passare da una dimensione sub-simbolica a una dimensione simbolica, di leggere e costruire significati. Allo stato attuale, infatti, il software non *comprende* le proposizioni, le IA non sono intelligenze sintattiche. Una mancanza che a sua volta genera un deficit comunicativo, spesso proposto nell'immaginario filmico con la macchina che parla con una dizione affettata.

42 Cfr. Fondazione Leonardo, *op. cit.*, pp. 46-7.

43 Cfr. A. Smith, *The Theory of Moral Sentiments,* tr. it. a cura di S. Di Pietro, *Teoria dei sentimenti morali*, BUR, Milano 2017 (7ª ed.).

44 Cfr. L. Wittgenstein, *Philosophische Untersuchungen,* a cura di P.M.S.- Hacker e J. Schulte, Blackwell, Oxford 1953; tr. it. a cura di M. Trinchero, *Ricerche filosofiche*, Einaudi Torino, 2009.

Una macchina si auto-modifica, ma non costruisce il senso di sé; una macchina compie atti, ma non agisce consapevolmente; una macchina accumula informazioni ma non attua processi semantici, rimanendo un essere non-teleonomico, deficiente di capacità introspettive e fronetiche. Ciò destituirebbe l'affermazione dell'identità etica artificiale, dell'autorità di un io che si autoimpone e che vuole essere legittimato, secondo la formula linguistica utilizzata da Asimov nel celebre titolo della sua antologia *Io, Robot*[45]. La coscienza dell'azione coincide, dunque, con l'autorialità dell'atto[46], ossia con una sua significazione nell'ordine di una costruzione narrativa del mondo e dell'esistenza, in una visione che procede secondo una scansione temporale che dalla memoria del passato procede verso l'azione del futuro.

45 I. Asimov, *I, Robot*, Gnome Press, New York, 1950, tr. it. a cura di L. Serra, *Io, robot*, Mondadori, Milano 2018.

46 G. Tamburrini, *op. cit.*, p. 62.