Fabio Fossa

Etica funzionale. Considerazioni filosofiche sulla teoria dell'agire morale artificiale

Abstract: *The purpose of Machine Ethics is to develop autonomous technologies that are able to manage not just the technical aspects of a tasks, but also the ethical ones. As a consequence, the notion of Artificial Moral Agent (AMA) has become a fundamental element of the discussion. Its meaning, however, remains rather unclear. Depending on the author or the context, the same expression stands for essentially different concepts. This casts a suspicious light on the philosophical significance of Machine Ethics. In particular, the risk arises of discarding Machine Ethics as a whole on the basis of accusations that, however, apply exclusively to one specific understanding of what AMAs are – but not to other, more adequate and convincing conceptualisations. To avoid this pitfall, this essay tries to elaborate a philosophically sound interpretation of AMAs and to sketch its primary component, i.e., the notion of functional ethics.*

Keywords: *Machine Ethics, Artificial Moral Agents, Autonomy, Eteronomy, Functional Ethics.*

> *L'intelletto umano ha una forte e, come sembra, irresistibile tendenza a interpretare le funzioni umane nelle categorie degli artefatti che le sostituiscono e gli artefatti nelle categorie delle funzioni umane da essi esercitate.*
>
> H. Jonas[1]

1. Etica delle macchine

Sin dalla sua origine in terra greca l'etica occidentale si è caratterizzata come un affare esclusivamente umano. Lo stesso Socrate, tramanda Platone nel Fedro[2], confessa di lasciare malvolentieri le mura cittadine per il sentiero di campagna, lungo il quale alberi e rocce restano muti ai suoi interrogativi. Non meno rimarca l'essenza antropica della riflessione sulla vita buona e felice Aristotele nell'*Etica Nicomachea*, il quale prende le mosse proprio distinguendo l'agire dell'essere uma-

1 H. Jonas, *Organismo e libertà. Verso una biologia filosofica*, a cura di P. Becchi, Torino, Einaudi, 1999, pp. 150-151.

2 Platone, *Phaedr*. 230d.

Filosofia, anno LXV, 2020, pp. 91-106 • Mimesis Edizioni, Milano-Udine
Web: https://www.ojs.unito.it/index.php/filosofia • ISSN (print): 0015-1823, ISSN (online): 2704-8195 • DOI: 10.13135/2704-8195/5235

no dai modi d'esistenza delle piante e degli animali[3]. Nel dramma della filosofia morale tanto la scena quanto gli spalti sembrano riservati agli esseri umani.

In tempi recenti, tuttavia, la chiusura antropica dell'etica è divenuta bersaglio di aspre critiche. Si pensi, ad esempio, al problema della sofferenza animale[4] o alle pressanti questioni ecologiche, ormai parti integranti della riflessione morale e politica[5]. Tra le direttive che ne mettono sempre più a dura prova la tenuta se ne è aggiunta, negli ultimi anni, una ulteriore che non ruota intorno a enti di natura, ma a prodotti tecnologici. Che si argomenti per una loro inclusione nel novero dei pazienti[6] e degli agenti morali[7], o che si faccia un passo indietro alla ricerca di modi alterativi di porre la domanda[8], si è diffusa la convinzione che la riflessione etica debba fare spazio ad un nuovo arrivato: l'*agente morale artificiale*.

L'inclusione degli agenti artificiali tra i legittimi oggetti dell'etica è spesso presentata come una tappa obbligata del frenetico sviluppo di cui, negli ultimi decenni, si sono rese protagoniste la robotica e l'Intelligenza Artificiale[9]. Una parte determinante del loro successo risiede nell'*autonomia* di cui si è saputo dotare le nuove tecnologie. Il termine, si badi bene, è usato in senso tecnico, e indica la peculiare capacità degli agenti artificiali di svolgere funzioni adattandosi alle condizioni del contesto[10] senza richiedere intervento o supervisione costante da parte dell'utente umano[11].

Per quanto innovativi, gli agenti artificiali sgorgano dal bisogno che da sempre spinge il genere umano verso la produzione tecnica e tecnologica: la costruzione di strumenti grazie ai quali svolgere in modo migliore determinati compiti i cui risultati ci interessano[12]. Già Aristotele sapeva immaginare uno strumento capa-

3 Arist., *Eth. Nic.* 1097b22-1098a20.

4 Cfr., ad esempio, P. Singer, *Liberazione animale,* a cura di P. Cavalieri, Milano, il Saggiatore, 2015.

5 H. Jonas, *Il principio responsabilità. Un'etica per la civiltà tecnologica*, a cura di P.P. Portinaro, Torino, Einaudi, 2009; B. Latour, *Politiche della natura. Per una democrazia delle scienze,* a cura di M. Gregorio, Milano, Raffaello Cortina Editore, 2000.

6 L. Floridi, *Information ethics: On the philosophical foundation of computer ethics*, in "Ethics and Information Technology", I (1999), pp. 37-56; E. L. Neely, *Machines and the Moral Community*, in "Philosophy & Technology", XXVII (2014), n. 1, pp. 97-111.

7 L. Floridi, *On the Morality of Artificial Agents*, in *Machine Ethics*, a cura di S. L. Anderson, M. Anderson, New York, Cambridge University Press, 2011, pp. 184-212.

8 M. Coeckelbergh, *The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics*, in "Philosophy and Technology", XXVII (2014), n. 1, pp. 61-77; D. J. Gunkel, *The Machine Question*, Cambridge, MIT Press, 2012.

9 C. Allen *et al.*, *Why Machine Ethics?*, in S. L. Anderson, M. Anderson, *op. cit.*, pp. 51-61; W. Wallach, C. Allen, *Moral Machines. Teaching robots right from wrong*, Oxford, Oxford University Press, 2009, pp. 13-24.

10 M. Verdicchio, *An Analysis of Machine Ethics from the Perspective of Autonomy*, in *Philosophy and Computing*, a cura di T. M. Powers, Cham, Springer International Publishing, 2017, pp. 179-191.

11 J. P. Sullins, *When is a Robot a Moral Agent?*, in S. L. Anderson, M. Anderson, *op. cit.*, pp. 151-161.

12 V. Marchis, *Storia delle macchine. Tre millenni di cultura tecnologica*, Roma-Bari, Laterza, 2005.

ce di svolgere la propria funzione con efficienza senza richiedere supervisione o intervento esterno, di modo che l'utente potesse guadagnare non solo il risultato dell'attività delegata, ma anche il tempo per dedicarsi ad altro[13]. L'agente artificiale è una risposta alle medesime esigenze, il che ne spiega la fortuna. Dai trasporti alla diagnostica medica, dalla finanza alla fruizione di contenuti mediali, dall'occupazione alle relazioni sentimentali fino alla guerra e alle operazioni chirurgiche gli agenti artificiali occupano ormai una posizione essenziale nel tessuto delle società capitalistiche.

La delegazione di compiti ad agenti artificiali autonomi non può che sollevare una domanda etica. Data la versatilità delle nuove tecnologie, è probabile che ad esse siano assegnati compiti i quali, se svolti da un agente umano, ne chiamano in causa il giudizio morale[14]. Che ne è della componente etica di simili attività, quando vengono delegate ad agenti artificiali?

La questione morale deriva direttamente dai caratteri della delegazione ad agenti artificiali e richiede di essere adeguatamente trattata. L'alternativa, che consisterebbe nel lasciare irrisolta la delegazione della componente morale, è senza dubbio insoddisfacente: significherebbe ignorare il portato normativo della delegazione e, dunque, neutralizzarlo – ovvero, sopprimere il problema. Al contrario, diventa necessario trovare delle soluzioni che rispondano alla perdita di controllo[15] inclusa nella delegazione di compiti a strumenti autonomi. Lo iato che si apre tra utente e tecnologia in virtù della capacità di quest'ultima di funzionare da sé – potenzialmente in modo imprevedibile – è anche lo spazio dove si annidano le difficoltà morali. Qui, ad esempio, la catena di cause ed effetti che lega agenti umani e artificiali si confonde ed emerge il problema della responsabilità. Qui sorge anche il

13 Arist., *Pol.* 1253b. Sul mito dell'agente artificiale cfr. A. Mayor, *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*, Princeton-Woodstock, Princeton University Press, 2018. Si tenga anche a mente la nota etimologia del termine robot: "Il termine 'robot' deriva dalla parola cecoslovacca *robota* (lavoro pesante), che si ritiene possa a sua volta derivare da un archeologismo slavo, *rabota*, che significa 'servitù'" (R. Cingolani, G. Metta, *Umani e umanoidi. Vivere con i robot*, Bologna, Il Mulino, 2015, p. 7).

14 Ci sono le basi per sostenere che questa non sia una possibilità futura, ma la descrizione della situazione attuale. Si consideri, per esempio, un algoritmo che suggerisce contenuti agli utenti di una piattaforma di video online. A seconda dei dati analizzati per effettuare le raccomandazioni e dei video disponibili sulla piattaforma, potrebbe succedere che nella lista dei video suggeriti siano selezionati contenuti offensivi o discriminatori – cfr. ad esempio G. Chaslot, *The toxic potential of YouTube's feedback loop*, disponibile su: https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/ (consultato in data 24 Febbraio 2020). Se, invece dell'algoritmo, fossimo noi stessi a dover mettere insieme una lista di video per degli amici, selezioneremmo gli elementi non solo basandoci sui loro gusti, ma (potenzialmente) anche sul nostro giudizio morale, escludendo contenuti offensivi – soprattutto in casi in cui la selezione avverrebbe per conto, ad esempio, di soggetti minorenni. Delegare l'opera di selezione ad algoritmi di raccomandazione implica delegare anche il relativo aspetto morale.

15 A. Matthias, *The responsibility gap: Ascribing responsibility for the actions of learning automata*, in "Ethics and Information Technology", VI (2004), n. 3, pp. 175-183.

problema dell'allineamento[16]: l'esigenza di far sì che il funzionamento autonomo degli agenti artificiali risulti effettivamente allineato alle aspettative valoriali che, a riguardo, è ragionevole o lecito nutrire.

L'idea di agente morale artificiale nasce come risposta a tali difficoltà[17]. Se l'autonomia di cui godono gli agenti artificiali può portare a situazioni di disallineamento tra gli effetti prodotti e le nostre aspettative morali, una soluzione consiste nel dotare gli agenti artificiali stessi di un analogo computazionale di quelle abilità grazie a cui noi umani siamo in grado di gestire le componenti morali di un compito. Si tratterebbe, dunque, di dotare il sistema di un analogo funzionale del giudizio morale umano – cioè di accorgimenti grazie a cui, dato un certo input, il sistema produca un output sufficientemente simile a quello prodotto da un generico utente date le medesime circostanze.

L'idea di rendere gli agenti artificiali in qualche modo capaci di gestire in autonomia non solo gli aspetti tecnici di un compito, ma anche quelli etici, si è affermata come una prospettiva degna di indagine ed è diventata l'obiettivo dell'etica delle macchine. Se è vero che la nozione di agente morale artificiale è ormai parte del dibattito, assai meno chiaro è però il suo senso. A seconda dell'autore o del contesto, la medesima etichetta sta infatti per nozioni essenzialmente diverse. Ciò getta una luce obliqua sulla questione degli agenti morali artificiali, la cui intera problematica rischia di essere messa al bando sulla base di accuse che, però, si applicano solo ad una accezione del concetto, mentre non riguardano altre, più rigorose impostazioni del problema. Il presente saggio ha lo scopo di lasciare emergere un'interpretazione forse minoritaria ma filosoficamente adeguata di cosa sia lecito intendere per 'agente morale artificiale' e di sviluppare, almeno in prima battuta, la sua principale componente: l'idea di un'etica funzionale.

16 IEEE, *Ethically Aligned Design*, disponibile su: https://ethicsinaction.ieee.org (consultato in data 15 Gennaio 2020).

17 Non si tratta certamente dell'unica soluzione possibile, cosa che può far dubitare della *necessità* dell'etica delle macchine (A. van Wynsberghe, S. Robbins, *Critiquing the Reasons for Making Artificial Moral Agents*, in "Science and Engineering Ethics", XXV (2019), n. 6, pp. 719-735). Ad esempio, altri approcci esplorano modalità tramite cui affidare ai sistemi tecnologici la sola gestione degli aspetti funzionali di un compito, mantenendo sotto il controllo dell'utente umano gli aspetti etici. È il caso del paradigma del *controllo umano significativo*, il quale ha saputo guadagnarsi un notevole ruolo in discussioni relative agli armamenti autonomi, alla chirurgia robotica e alle automobili a guida autonoma (F. Ficuciello *et al.*, *Autonomy in surgical robots and its meaningful human control*, in "Paladyn, Journal of Behavioral Robotics", X (2019), n. 1, pp. 30-43; F. Santoni de Sio, J. van den Hoven, *Meaningful human control over autonomous systems: A philosophical account*, in "Frontiers Robotics AI", V (2018), pp. 1-14). Parallelamente, si può procedere ad una revisione del concetto filosofico di responsabilità che vada al di là del paradigma del controllo diretto e faccia proprie istanze derivanti dalla delegazione dell'agire a tecnologie autonome e dalla conseguente diffusione dell'azione lungo i vari nodi di una rete (A. Fabris, *Ethics of Information and Communication Technologies*, Cham, Springer International Publishing, 2019).

2. Agenti morali artificiali

La confusione che circonda il concetto di agente morale artificiale è da ricondurre al fatto che la sua accezione più diffusa è al contempo la più controversa.

Per farsi un'idea del problema si prenda il documento *Draft Ethics Guidelines for Trustworthy AI*[18], la cui peculiare natura può valere da preliminare assicurazione circa la tipicità delle opinioni esposte. Nel § I.5.5[19] gli autori prendono in considerazione possibili rischi sul lungo periodo collegati alla superintelligenza – l'ipotesi secondo cui in un tempo futuro l'IA sarà in grado di superare l'intelligenza umana in ogni sua prestazione. Tra le tecnologie passate in rassegna compaiono la coscienza artificiale, l'IA generale e proprio gli agenti morali artificiali. La nota 19 chiarisce ogni dubbio sul modo in cui il concetto è interpretato: per agente morale si intende un sistema in grado di generare *in autonomia* giudizi normativi e di agire, sempre *in autonomia*, sulla base di tali giudizi. L'idea della superintelligenza, però, è tanto conturbante quanto spinosa: come segnalato in un apposito riquadro, nel gruppo di lavoro non si è raggiunto un accordo circa la reale consistenza della questione. Lo stesso sospetto circonda gli agenti morali artificiali.

Per capire la controversia relativa al concetto di agente morale artificiale bisogna innanzitutto interrogarsi sul processo che ha portato ad associarlo ad una dottrina futuristica quale la superintelligenza. L'associazione dipende da un presupposto interpretativo che emerge bene nel documento europeo e consiste nel concepire l'agente morale artificiale come un doppio o un sostituto dell'agente morale umano. Dato il presupposto, diventa naturale ricorrere al linguaggio dell'autonomia morale per rendere conto dell'agire etico tanto umano quanto artificiale. Come si vedrà, si tratta di una scelta linguistica poco rigorosa e foriera di controsensi. Da qui le difficoltà che portano ad includere gli agenti morali artificiali nel libro degli esseri immaginari della tecnoscienza.

Come accennato, l'accezione comune di 'agente morale artificiale' si basa sul presupposto che sia possibile instaurare una continuità essenziale tra agenti umani e artificiali. Al di là di differenze accidentali e prive di interesse, si sostiene, la riproduzione tecnologica del comportamento morale umano ne coglie la sostanza. Di conseguenza, il risultato dell'immagine funzionale dell'agire etico umano trasla impercettibilmente dall'*analogia* all'*omologia*[20]. In fondo, si con-

18 Il testo, pubblicato nel dicembre 2018, costituisce il primo risultato dello *High Level Expert Group on AI*, un'assemblea di 52 esperti e rappresentanti del settore incaricata dalla Commissione Europea di tracciare un quadro di riferimento per lo sviluppo etico dell'intelligenza artificiale. Il documento è stato inizialmente reso disponibile in forma di bozza, chiedendo alle parti interessate di proporre commenti sia generali, sia su alcune questioni specifiche segnalate in appositi riquadri. La versione finale del documento, rivisto alla luce degli oltre 500 commenti ricevuti, è stata pubblicata il 9 aprile 2019 ed è disponibile su: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

19 Ivi, pp. 12-13.

20 Per 'analogia' intendo somiglianza nella differenza, per 'omologia' intendo un rapporto di corrispondenza tra due elementi in quanto descrivibili in riferimento alla medesima logica. Mentre la differenza è un dato irriducibile nel pensiero analogico, nel pensiero omologico le dif-

clude, agenti morali umani e artificiali appartengono alla stessa categoria di enti: possono essere concepiti attraverso il medesimo linguaggio e paragonati gli uni agli altri senza ulteriori *caveat*.

Il paradigma della continuità, di cui non mancano applicazioni concrete[21], ha sicuramente il vantaggio di proporre un'interpretazione intuitiva ed eccitante di cosa siano gli agenti morali artificiali. In più, è in linea con il copione fantascientifico di sicuro successo secondo cui le nuove tecnologie custodiscono in sé tanto l'utopica soluzione a tutte le nostre imperfezioni quanto la distopica condanna del genere umano all'obsolescenza, alla sottomissione, all'estinzione.

La diffusione, il fascino e la semplicità del paradigma continuista non sono però ragioni sufficienti per avallarne la logica, tanto più che i suoi presupposti rispondono male alla prova dell'analisi critica. Al di là dell'implausibilità e del futurismo[22], a preoccupare è la debolezza dell'impianto concettuale. La traslazione dall'analogia all'omologia di modello e copia per concettualizzare la riproduzione tecnologica del comportamento morale appare almeno prematura, se non ingiustificata e arbitraria[23]. Qualificare le innumerevoli differenze che intercorrono tra agire umano e artificiale come semplicemente accidentali è una tesi di cui, dal punto di vista concettuale, si danno buone ragioni di dubitare.

Le difficoltà a cui va incontro il paradigma della continuità nascono tutte dalla disattenzione riservata alle differenze che contraddistinguono agire morale umano e artificiale. Sarebbe però imprudente concludere che, data l'inconsistenza del concetto di agente morale artificiale così come concepito nell'alveo del paradigma continuista, sia necessario abbandonare l'etica delle macchine *tout court*. Una più equa conclusione dell'argomentazione dovrebbe essere che, se si ignorano le differenze che distinguono l'agire morale umano da quello artificiale (pur nella somiglianza che regge l'imitazione), si imbocca una via che conduce a controsensi

ferenze cadono sullo sfondo e l'attenzione si concentra unicamente sulle somiglianze. La traslazione da analogia a omologia mediata dal linguaggio sta alla radice di molti problemi concettuali in filosofia ed etica dell'IA.

21 Un esempio è l'argomentazione di R. Arkin (*Lethal Autonomous Systems and the Plight of the Non-combatant*, in "AISB Quarterly", CXXXVII (2013), pp. 1-9) in favore dello sviluppo dei sistemi d'arma autonomi. Secondo Arkin i timori che circondano la ricerca sono eccessivi: le armi autonome, se ben progettate, rappresenterebbero infatti un passo avanti verso la realizzazione di conflitti allineati alle relative esigenze etiche. Eventuali soldati robotici dotati di intelligenza artificiale non sarebbero soggetti né alla paura della morte o del dolore né ad altri stati psicologici quali rabbia, angoscia, agitazione, stress o stanchezza, i quali sono spesso alla base dell'adozione umana di comportamenti moralmente inaccettabili. Le tesi di Arkin sono state criticate, ad esempio, da M. Guarini, P. Bello, *Robotic Warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters*, in P. Lin *et al.*, *op. cit.*, pp. 129–144; A. Sharkey, *Autonomous weapons systems, killer robots and human dignity*, in "Ethics and Information Technology", XXI (2019), n. 2, pp. 75-87; J. P. Sullins, *RoboWarfare: can robots be more ethical than humans on the battlefield?*, in "Ethics and Information Technology", XII (2010), n. 2, pp. 263-275.

22 L. Floridi, *Singularitarians, AItheists, and Why the Problem with Artificial Intelligence is H.A.L. (Humanity At Large), not HAL*, in "APA Newsletter Philosophy and Computing", XIV (2015), n. 2, pp. 8-11.

23 F. Fossa, *Artificial Moral Agents: Moral Mentors or Sensible Tools?*, "Ethics and Information Technology", XX (2018), n. 2, pp. 115-126.

e illusioni ottiche. Rimane però aperta una seconda via: la via della differenza. Il compito, quindi, consiste nel prendere sul serio il rapporto *analogico* che l'imitazione tecnologica dell'agire morale instaura tra il modello umano e la copia artificiale. Puntare tutto sulla somiglianza conduce nelle secche del paradigma continuista. Rimane da esplorare il territorio della differenza.

3. Azione e mediazione

Esplorare il senso dell'agire morale artificiale al di là del paradigma continuista richiede innanzitutto l'individuazione di quella differenza specifica seguendo le indicazioni della quale diventi possibile tanto gettare luce su ciò che contraddistingue l'agire etico artificiale quanto chiarirne il rapporto con l'agire etico umano. Entrambi gli aspetti potranno essere riportati in superficie andando a scavare nel punto in cui, in ossequio al presupposto dell'omologia, la differenza viene coperta; ovvero, nel punto in cui si posizionano esseri umani e agenti artificiali lungo le medesime coordinate dell'agire morale.

Una simile impostazione è esemplificata dal piano cartesiano a cui ricorrono Wallach e Allen[24] per rendere conto delle qualità dell'agire morale artificiale [figura 1]. Le variabili da considerare, suggeriscono gli autori, sono sostanzialmente due: autonomia e sensibilità a valori morali. Sul gradino più basso, dove non si manifestano gradi significativi di alcuna variabile, si situa il livello dell'etica operativa (*operational morality*). La classe include accorgimenti del tutto passivi ed inerti che, però, hanno implicazioni di carattere etico – come le sicure sulle armi da fuoco o, per fare un esempio reso celebre da Latour[25], i dossi stradali. A gradi elevati di autonomia e sensibilità etica corrispondono gli agenti morali nel pieno senso della parola (*full moral agency*), capaci di responsabilità e degni di fiducia, di cui al momento non sono disponibili esempi tecnologici, ma solo organici – gli esseri umani. Tra i due estremi si estende il dominio dell'etica funzionale (*functional morality*), i cui elementi sono descritti da gradi medi e diversi di autonomia e sensibilità a valori. Una freccia mostra poi come l'avanzamento tecnologico tenda da sé ad agenti artificiali dotati di autonomia crescente, mentre

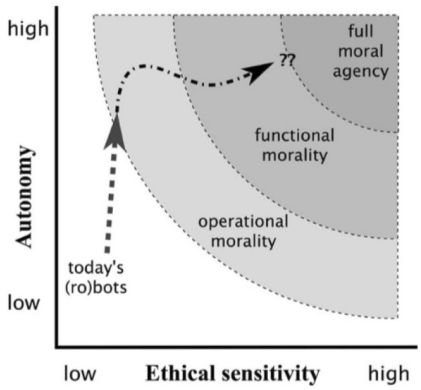

Figura 1 - Dimensioni dell'agire morale artificiale (Wallach & Allen 2009, p. 26)

24 W. Wallach, C. Allen, *op. cit*., p. 26.

25 B. Latour, *Where are the missing masses? The sociology of a few mundane artifacts*, in *Shaping Technology-Building Society. Studies in Sociotechnical Change*, a cura di W. Bijker e J. Law, Cambridge, MIT Press, 1992, pp. 151-180.

più complessa è l'implementazione della sensibilità ai valori, che si profila quindi come la missione dell'etica delle macchine. Non si esclude poi la possibilità che tra le pieghe del futuro si annidi un agente morale artificiale vero e proprio.

Il presupposto che regge l'intero schema consiste nell'assumere che si possa ricorrere alla medesima accezione delle nozioni di autonomia e sensibilità etica per descrivere qualsiasi punto nel piano cartesiano. Che si tratti di etica operativa, funzionale o in senso proprio, le due variabili non cambiano nella sostanza, ma solo nel grado. Questo è l'assunto su cui si basa il paradigma continuista e che, in vista di una sua revisione, deve essere approfondito. In particolare, l'area che interessa corre lungo la linea tratteggiata che segna il confine tra etica funzionale ed etica in senso proprio. Siamo certi che ci sia continuità, o omogeneità, tra gli agenti capaci di etica funzionale e di etica in senso pieno?

Per indagare la questione è utile rivolgersi alla classificazione proposta da James H. Moor[26]. A differenza di Wallach e Allen, Moor dedica più attenzione alle caratteristiche che contraddistinguono le diverse articolazioni dell'agire etico, pur non escludendo una continuità tra di esse[27]. L'impostazione di Moor è interessante per i nostri scopi perché, chiarendo quali prestazioni corrispondano a quali forme dell'agire morale, rende più agevole il compito di determinare la loro omogeneità o differenza.

L'insieme degli agenti artificiali può essere strutturato, secondo Moor, in quattro classi. Nella prima classe troviamo gli agenti a impatto morale (*ethical impact agents*), ovvero tecnologie la cui mera introduzione sortisce conseguenze di interesse etico[28]. Per quanto importante, si tratta ancora di un livello elementare che non calca la soglia dell'etica delle macchine. Il primo passo si compie invece con gli agenti morali impliciti (*implicit ethical agents*): tecnologie progettate e costruite in modo tale che, durante il loro funzionamento, si adeguino rigidamente a valori morali dati[29]. L'adeguamento della funzione al valore, qui, è interamente predeterminato: i valori sono vincoli che direzionano implicitamente il funzionamento dell'algoritmo verso esiti moralmente apprezzabili. Il gradino successivo della scala è occupato dagli agenti morali espliciti (*explicit ethical agents*) i quali, a differenza dei precedenti, esibiscono la capacità di: a) calcolare esplicitamente

26 J. H. Moor, *The Nature, Importance, and Difficulty of Machine Ethics*, in M. Anderson, S. L. Anderson, *op. cit.*, pp. 13-20.

27 La quale, di conseguenza, diventa più difficile da concepire – cfr. W. Wallach, C. Allen, *op. cit.*, pp. 33-36. Le due classificazioni hanno dei punti in comune, ma non sono facilmente sovrapponibili.

28 Appartiene alla categoria, ad esempio, qualsiasi tecnologia la cui introduzione migliori le condizioni di esistenza di soggetti sfruttati, sofferenti o diversamente abili.

29 Si pensi, ad esempio, ad un algoritmo il cui scopo sia di pubblicizzare posizioni lavorative agli utenti di un social network che sia esplicitamente progettato per ignorare il genere sessuale dell'utente in relazione a ruoli dove considerarlo sarebbe discriminatorio. Su simili problemi si veda, ad esempio, A. Datta *et al.*, *Discrimination in Online Advertising A Multidisciplinary Inquiry*, in "Proceedings of Machine Learning Research", LXXXI (2018), n. 3, pp. 1-15. La differenza tra *ethical impact agents* e *implicit ethical agents* consiste nel fatto che nel primo caso la componente etica è una conseguenza del ricorso alla tecnologia, mentre nel secondo caso è parte costitutiva del design.

diversi scenari di funzionamento alla luce di parametri che corrispondono a valori morali; b) funzionare di conseguenza; e c) fornire informazioni che rendano conto delle computazioni eseguite – il che, con tutte le cautele del caso, costituirebbe un analogo di ciò che per noi è la giustificazione. La giustificazione in senso proprio, così come il comportamento morale in senso proprio, si manifestano solo con la classe degli agenti morali pieni (*full moral agents*): esseri dotati di coscienza, intenzionalità e libero arbitrio, capaci di ragionare eticamente e rendere conto delle proprie decisioni.

La differenza di cui siamo sulle tracce va cercata quindi tra le classi degli agenti morali espliciti e pieni. Nell'articolo di Moor, l'elemento che separa le due classi è piuttosto indefinito: si ricorre ai termini "coscienza, intenzionalità, libero arbitrio", concetti notoriamente delicati e soprattutto – come lo stesso Moor nota – contando sui quali non sembra possibile dirimere la questione.

Ciò non toglie che tra le due classi rimanga una differenza piuttosto evidente. Se si prendono in considerazione le prestazioni che costituiscono l'agire morale pieno, appare evidente come almeno due siano gli ordini di problemi, strettamente connessi l'uno all'altro: il comportamento secondo valori e la determinazione dei valori. L'esperienza morale umana è costituita dallo sforzo di vivere affermando ciò che ci sta a cuore e opponendoci a ciò che riteniamo inaccettabile, il che è inseparabile da un'interrogazione su cosa sia ciò che ci sta a cuore e ciò che riteniamo inaccettabile. Al contrario, la situazione degli agenti morali espliciti non esibisce la bidimensionalità propria dell'agire morale pieno. Ciò risulta chiaro se si ritorna alla narrazione che vede l'emergere dell'etica delle macchine dalla delegazione di compiti alle tecnologie autonome. Gli agenti artificiali fanno il loro ingresso in società come prodotti del fare umano, come esecutori di compiti dati. Le aspettative che nutriamo circa il loro comportamento non pertengono alla determinazione dei valori morali, ma esclusivamente alla loro affermazione: ci si aspetta che le funzioni degli agenti artificiali e le modalità di esecuzione siano allineate a ciò che *per noi* è dotato di valore morale.

La stessa tesi può essere espressa notando come gli esseri umani esibiscano una relazione *diretta* al mondo dei valori e all'ideale del bene, mentre gli agenti artificiali che popolano le altre classi esibiscono una relazione *mediata* ad entrambi i termini. I valori che strutturano la funzione, siano essi implementati in maniera implicita o esplicita, devono già essere in qualche modo definiti prima di essere programmati nel sistema o, nel caso si opti per soluzioni *bottom-up* basate su tecnologie di *machine learning*, richiedono comunque la vidimazione umana[30]. Gli

30 Su questo punto vale la pena soffermarsi brevemente. Si potrebbe pensare che agenti morali artificiali basati su tecnologie di *machine learning* [ML] esibiscano una relazione diretta ai valori, in quanto questi ultimi sarebbero autonomamente estratti dai dati. La conclusione sarebbe però quantomeno imprudente. L'estrazione avverrebbe infatti secondo una logica statistica e corrisponderebbe ad una tendenza di carattere quantitativo contenuta nel data set fornito al sistema. Siccome nessuna considerazione di carattere normativo sarebbe inclusa nella determinazione del valore, è piuttosto evidente che la terminologia del valore etico sia inadeguata in questo caso. Un uso più attento del linguaggio parlerebbe di parametri che solo *dal punto di vista umano* potrebbero essere associati a sensi morali, sempre presupponendo che il sistema sia dotato di

agenti artificiali esistono sempre e solo immersi in un contesto normativo, in una sostanza etica che ne determina già i valori di funzionamento come dei parametri, dei prerequisiti da soddisfare, e non come dei termini da porre. In questo senso, gli agenti morali artificiali *mediano* valori la cui determinazione non pertiene loro. La posizione dei valori, possibile solo se si gode di una relazione diretta ad essi, non è un atto che riguarda gli agenti morali artificiali, ma è invece una componente costitutiva dell'agire morale umano. Tra le due dimensioni corre una diversità di principio di cui non è possibile rendere conto secondo la logica del graduale e progressivo incremento delle capacità computazionali.

Ecco, dunque, la differenza specifica tra agire etico umano e artificiale. Il paradigma continuista risulta inadeguato perché ignora o diluisce in una mera differenza di grado la differenza concettuale tra relazione diretta o mediata ai valori. Dalla messa a fuoco e dalla comprensione di tale differenza, però, dipende l'impostazione rigorosa del problema dell'agire morale artificiale.

4. Etica funzionale: autonomia e eteronomia

Per uscire dalle secche del paradigma continuista bisogna pensare gli agenti morali artificiali a partire dalla differenza che li contraddistingue, pur nella somiglianza, dagli agenti morali umani. Nelle prossime pagine si proverà, quindi, a delineare che cosa significa 'agire morale artificiale' partendo dalla tesi per cui esso consista nell'adeguazione autonoma di un funzionamento a valori etici dati. Determinare il modo in cui lo svolgimento autonomo di funzioni interseca la dimensione del valore etico mira a tematizzare ciò che è specifico e proprio dell'agire morale artificiale, cosicché sia possibile svilupparne una teoria che non soffra di intrusioni illecite e spericolate omologie. Si proverà, cioè, ad abbozzare i princìpi di un'etica dello svolgimento autonomo di funzioni, o *etica funzionale.*

un grado sufficiente di interpretabilità. È utile distinguere qui tra *metodi di implementazione* dei valori e *relazione a valori*. Sul lato dell'implementazione la letteratura contempla principalmente due metodologie variamente combinabili: la scrittura diretta dell'impianto valoriale nel programma del sistema (*top-down approach*) o la delegazione al sistema stesso della formazione dell'impianto valoriale tramite ML (*bottom-up approach*). Se i metodi di implementazione cambiano, la relazione ai valori è però la stessa: il sistema non esercita controllo diretto su di essi, ma li ha come dati (o nel programma o nel data set) e li media. Anche nel caso di approcci *bottom-up*, infatti, la componente umana funge da filtro delle regolarità statistiche estratte dall'algoritmo. Un esempio: se, addestrando un sistema il cui scopo sia di consigliare se concedere o meno un mutuo, si ottengono risultati che discriminano utenti sulla base del colore della pelle, sarebbe irragionevole mettere in discussione il valore della non-discriminazione in base al colore della pelle *data la nuova evidenza*. Si dovrebbe invece, ad esempio, tornare sul data set e verificarne l'adeguatezza o, nel caso di algoritmi evolutivi, rivedere la funzione di fitness. Ciò accade perché il sistema di valori di riferimento *rimane pur sempre* quello determinato direttamente dagli esseri umani per se stessi. Estrarre statisticamente un valore da un data set o lasciare che algoritmi evolvano fino a funzionare in modo accettabile da un punto di vista morale (metodo di implementazione) e porre normativamente un valore di carattere etico (relazione diretta a valori) sono due fenomeni tra loro irriducibili.

La difficoltà del compito consiste innanzitutto nella sua novità, la quale a sua volta genera problemi di carattere linguistico. La novità della situazione riguarda il fatto che per la prima volta si trovano effettivamente separati due aspetti dell'esperienza morale – adeguamento ai valori e determinazione dei valori – che, nel caso umano, si manifestano solo nella loro inscindibile correlazione. Ciò determina difficoltà linguistiche in quanto, per analizzare la situazione nuova, non abbiamo che le parole con le quali descriviamo l'analogo umano; parole che, non essendo tagliate per il nuovo compito a cui vorremmo piegarle, rischiano di trascinarsi dietro significati fuorvianti.

Entrambe le difficoltà diventano evidenti se si prova ad accostarsi all'etica funzionale per mezzo di due nozioni che sembrano, però, fare del tutto al caso in questione: *autonomia* e *eteronomia*. La nozione di autonomia gioca un ruolo significativo sia nelle *Draft Ethics Guidelines* che nel piano cartesiano di Wallach e Allen: in entrambi è il collante che tiene insieme le due classi di agenti. Tuttavia, già è stato ricordato, in robotica ed IA la nozione è usata in senso tecnico e indica la capacità di un sistema di svolgere compiti dati senza richiedere costante intervento e supervisione esterni. Si tratta di una definizione funzionale dell'autonomia, nel senso che la funzione determina l'ambito in cui si esercita la capacità dell'agente artificiale di condursi da sé. Qui è già sottintesa una differenza analoga a quella emersa nell'analisi dell'agire morale artificiale: un conto è saper trovare i mezzi necessari al raggiungimento di uno scopo dato, un conto è saper fare ciò ma anche porre gli scopi da perseguire[31]. Nel caso dell'agire morale artificiale il ricorso alla nozione di autonomia confonde l'analisi perché suggerisce surrettiziamente che gli agenti artificiali siano autonomi, da un punto di vista morale, come lo sono gli agenti umani – mentre la loro autonomia pertiene esclusivamente all'esecuzione autonoma delle relative funzioni tale da soddisfare aspettative morali date. L'autonomia dell'agente umano, al contrario, include lo sviluppo e la determinazione proprio delle aspettative che fanno da contesto normativo al funzionamento etico, le quali rimangono necessariamente di sua esclusiva competenza.

Si potrebbe quindi essere tentati di ricorrere al concetto di *eteronomia* per rendere conto della peculiare relazione a valori del funzionamento morale. Per eteronomia si intende la situazione in cui l'ideale del bene e i valori affermati nella prassi non siano autonomamente posti dall'agente stesso, ma siano invece determinati sotto l'effetto di istanze spurie o assunti irriflessivamente[32]. Da una parte sta un at-

31 Si veda la distinzione tra *having a purpose* e *carrying out a purpose* come critica dell'analogia cibernetica di macchina e organismo in H. Jonas, *Organismo e Libertà,* cit., pp. 149-169. Una terminologia simile è ripresa da J. J. Bryson, P. P. Kime, *Just an artifact: Why machines are perceived as moral agents*, in *IJCAI International Joint Conference on Artificial Intelligence*, 2011, pp. 1641-1646 (disponibile su: http://www.cs.bath.ac.uk/~jjb/ftp/BrysonKime-IJCAI11.pdf, consultato in data 20 aprile 2020), dove si distingue tra *purpose-built artefacts* e *purpose-setting agents.* È sulla base di questa differenza che – come sostiene M. Verdicchio, *op. cit*. – bisogna pensare l'autonomia degli agenti artificiali: restando, cioè, nella dimensione della funzione e magari complicando il concetto di automaticità prima di ricorrere alla nozione di autonomia.

32 Le nozioni di autonomia ed eteronomia sono trattate in modo esemplare nella filosofia kantiana. Nelle opere morali è ampiamente dibattuto il problema della determinazione della

teggiamento *critico* nei confronti dei valori, che li sottopone a discussione razionale e prova pratica cosicché siano assunti per loro stessi, per le loro qualità intrinseche. Dall'altra, un atteggiamento *acritico*, che invece opta per un sistema di convinzioni morali sulla base di considerazioni estrinseche, come ad esempio la loro autorità, antichità, diffusione o origine.

A prima vista, il concetto di eteronomia sembra utile all'esplorazione della differenza tra agire umano e artificiale in quanto coglie proprio la diversità nella relazione ai valori che è stata messa a fuoco. L'agire morale artificiale appare quindi come un agire *eteronomo*, determinato dall'adeguazione di un corso di eventi a valori assunti in quanto dati, senza che si presenti lo spazio per una loro discussione critica. L'eteronomia coglie l'aspetto *esecutivo* dell'etica funzionale e la separazione di quest'ultimo dalla dimensione normativa.

Tuttavia, la nozione filosofica di eteronomia morale è pur sempre ritagliata sul modo di esistere umano ed una sua traslazione all'ambito dell'agire artificiale rischia di cadere nuovamente nella fallacia del paradigma continuista, reintroducendo surrettiziamente la dimensione dell'autonomia. In effetti, in riferimento all'etica umana la nozione di eteronomia assume una sfumatura normativa più che descrittiva: prima di essere un concetto utile alla descrizione di un comportamento, l'eteronomia indica un'illusione, un fallimento della ragione morale, un vizio del carattere. Assumere un atteggiamento pienamente eteronomo per un agente umano non è semplicemente possibile: l'eteronomia si riduce alla falsa credenza di poter uscire dal cerchio dell'autodeterminazione. L'agente eteronomo si irretisce in una finzione paradossale: l'illusione di negare la propria autonomia è pur sempre frutto di una decisione che presuppone l'esercizio di quest'ultima. Da qui deriva la sfumatura normativa legata al concetto morale di eteronomia: esso rappresenta una colpevole forma di immaturità, superficialità, irriflessività e mancanza di spirito critico.

L'aspetto paradossale dell'eteronomia emerge con chiarezza se si prende in esame la situazione pratica dell'esecuzione di un comando – cardine su cui ruotano molte analogie tra esseri umani e macchine, come ad esempio le retoriche speculari dello schiavo robot (con le relative dinamiche del riconoscimento tanto amate dalla fantascienza[33]) e dell'essere umano ridotto a ingranaggio di una macchina[34]. Nella delegazione di un compito ad un essere umano tramite comando, infatti, è inclusa

volontà dovuta ad istanze spurie. Si veda, ad esempio, I. Kant, *Critica della ragion pratica,* trad. it. F. Capra, Roma-Bari, Laterza, 2010, p. 71 e Idem, *Fondazione della metafisica dei costumi*, trad. it. F. Gonnelli, Roma-Bari, Laterza, 2007, p. 99. L'eteronomia come assunzione acritica di sistemi di valori già dati motiva invece il celebre incipit del saggio *Risposta alla domanda: cos'è illuminismo?* In Idem, *Scritti di storia, politica e diritto*, a cura di F. Gonnelli, Roma-Bari, Laterza, 2007, p. 45 e risuona nella massima del *Selbstdenken* in Idem, *Critica del Giudizio*, a cura di L. Amoroso, Milano, Rizzoli, 2007, p. 393. Il tema torna anche nei capitoli 3 e 4 de Idem, *La religione entro i limiti della sola ragione*, Roma-Bari, Laterza, 2007, pp. 99-226.

33 R. Bodei, *Dominio e sottomissione. Schiavi, animali, macchine, Intelligenza Artificiale*, Bologna, il Mulino, 2019; D. J. Gunkel, *Robot rights*, Cambridge, MIT Press, 2018, pp. 117-130.

34 N. Wiener, *Introduzione alla Cibernetica. L'uso umano degli esseri umani*, Torino, Bollati Boringhieri, 2012.

l'aspettativa che chi lo dovrà eseguire lo comprenda, cioè sia capace, come suggerisce Gadamer, di "applicarlo alla concreta situazione alla quale si riferisce". "Il vero senso dell'ordine", continua il filosofo, "si determina solo nella concretezza della sua adeguata esecuzione". Ma in questo spazio di manovra, di "comprensione creativa", si apre al contempo il margine per "un esplicito rifiuto di obbedienza, che non è semplicemente disobbedienza, e che trova la sua legittimazione nel senso del comando e della sua concretizzazione, che è lasciata alla decisione del singolo". L'eteronomia che pare essere implicata nella mera esecuzione di un comando è in realtà sempre calata in una dimensione di autonomia che fa capo all'esecutore e "alla responsabilità di colui che obbedisce"[35]. Nell'esecuzione di un ordine, spiega anche Jonas, "lo scopo di qualcun altro", quando "eseguito dall'agente con la sua azione", certamente rientra "esso stesso nel sistema complessivo della vita intenzionale propria dell'esecutore". Tuttavia, "questo non deve significare che egli abbia fatto proprio lo scopo del superiore; ma ha elevato a suo scopo presente almeno la sua esecuzione, nella misura in cui gli è stata affidata, e ciò per scopi suoi propri". Solo istituendo una finzione il superiore può atteggiarsi "come se il suo sottoposto per la durata dell'azione avesse sospeso la sua propria 'persona', con la spontaneità della sua propria vita intenzionale" e concepire l'esecutore, "che pertanto in fin dei conti è *stato meccanizzato*", come "il suo robot, il suo strumento"[36]. Il velo della finzione si stende su un dato che, essendo irremovibile, può solo essere coperto; esso cala sulla relazione diretta dell'esecutore a scopi e valori, che non è tolta con lo svolgimento acritico del comando, ma attivamente determinata in linea con il contenuto dell'ordine – e quindi già esercitata.

Questo è il motivo per cui appare sensato ritenere comunque responsabili o colpevoli, almeno da un punto di vista morale, quegli agenti che di fronte ad un crimine si giustificano ricorrendo alla logica dell'eteronomia[37] ('ho solo eseguito gli ordini', 'ho agito in conformità alla legge'). Il senso descrittivo della nozione di eteronomia si applica male al caso umano perché il carattere della relazione diretta a valori, che ne contraddistingue la prassi, non permette la vera e propria eteronomia. Non possiamo in tutto e per tutto eseguire comandi *come una macchina*, nemmeno quando altri dispongono l'occasione per ingenerare una simile situazione o

35 H.-G. Gadamer, *Verità e metodo*, trad. it. G. Vattimo, Milano, Bompiani, 2014, p. 689.

36 H. Jonas, *Organismo e libertà*, cit., p. 163.

37 Com'è noto, intorno a questi aspetti ruota la discussione del caso di Eichmann in H. Arendt, *La banalità del male. Eichmann a Gerusalemme*, Milano, Feltrinelli, 2013. Si vedano soprattutto il capitolo 2, *L'imputato* (pp. 29-43), e il capitolo 8, *I doveri di un cittadino ligio alla legge* (pp. 142-157). Un riferimento diretto alla questione dell'esecuzioni di ordini immorali si trova a p. 155, dove la Arendt mette in guardia da un suo uso ingenuo, che nel caso del regime criminale nazista "significa non soltanto aggirare la questione, ma rifiutarsi deliberatamente di prender nota dei principali fenomeni morali, giuridici e politici del nostro tempo". Un genere analogo di problemi risuona in C. Eatherly, G. Anders, *Burning Conscience*, New York, Monthly Review Press, 1962. L'esecuzione cieca degli ordini è motivo di preoccupazione anche nel caso in cui l'esecutore sia tecnologico. Si vedano almeno le note di Wiener sulla *literal-mindedness* delle tecnologie autonome (N. Wiener, *Cybernetics or, Control and Communication in the Animal and the Machine*, Mansfield Centre, Martino Publishing, 2013, pp. 169-180; Idem, *Dio & Golem s.p.a.. Cibernetica e religione,* Torino, Bollati Boringhieri, 1991, pp. 53-82).

la stessa situazione si presenta da sé come risultato di un'organizzazione sociale impersonale. Quando ci si interroga sull'agire umano, la nozione di eteronomia è soprattutto normativa: indica una mancanza da parte dell'agente, una dismissione fittizia e illusoria delle proprie responsabilità. La relazione diretta ai valori definisce l'ambito dell'esperienza morale umana come sua condizione di esistenza: non esiste, per noi, un *al di fuori* da questa relazione.

L'*al di fuori* è però esattamente l'ambito del funzionamento morale. Qui, al contrario che nel caso umano, la nozione assume un'accezione pienamente descrittiva: non essendoci lo sfondo dell'autonomia a rendere paradossale il comportamento eteronomo, l'adeguamento delle funzioni a valori etici dati non ha più alcuna sfumatura normativa associata, ma descrive una modalità inedita dell'interazione tra valori morali ed eventi. La novità e la specificità degli agenti morali artificiali consistono proprio nel fatto che essi manifestano l'eteronomia morale non come finzione paradossale passibile di giudizio, ma come semplice fenomeno[38].

Nel funzionamento morale la dimensione della determinazione dell'ideale del bene e dei valori conseguenti è esterna per definizione. La differenza specifica dell'agire morale artificiale consiste nella *piena eteronomia* – un'eteronomia specifica che permette non solo di tracciarne in modo netto i caratteri peculiari, ma anche di pensarne rigorosamente il rapporto con l'agire umano.

5. Conclusione

La nozione di etica funzionale, se sviluppata partendo dall'analisi della piena eteronomia, permette di concepire con accuratezza il rapporto tra agenti umani ed artificiali e di coglierne le reali sfide morali.

Scorporando la determinazione dei valori dall'adeguamento pratico ad essi si perde la possibilità di concepire adeguatamente l'agire umano ma si guadagna l'opportunità di comprendere appieno il funzionamento morale degli agenti artificiali. Questi ultimi sono quindi meri *esecutori o mediatori di valori umani*. Da un punto di vista morale, agli agenti artificiali non chiediamo dunque altro che l'*allineamento* del loro funzionare alle nostre aspettative valoriali. L'adeguamento cieco e acritico a valori etici dati nell'esecuzione di un compito, che tanto preoccupa nel caso dell'agire umano, è esattamente quello che cerchiamo nel caso degli agenti artificiali: sistemi flessibili, prevedibili, robusti e affidabili che ottengano gli scopi loro preposti nel rispetto dei valori che ci stanno a cuore.

38 Questo punto non viene adeguatamente preso in considerazione da chi si preoccupa del fatto che un simile tentativo possa rivelarsi controproducente in quanto porterebbe a sistemi dotati di mentalità da burocrate o robot psicopatici. Proiettando la mentalità ottusa e irresponsabile del burocrate o la chiusura prospettica dello psicopatico sul caso degli agenti morali artificiali si commette già un errore categoriale, in quanto si ricorre a categorie antropomorfiche che si applicano esclusivamente alla dimensione della relazione diretta a valori per rendere conto di un fenomeno che esibisce una relazione a valori essenzialmente diversa. Su ciò si veda D. J. Gunkel, *The Machine Question*, cit., pp. 86-88. Lo stesso errore, come già notato, sta alla base del ricorso alla metafora dello schiavo.

Si compone così con maggiore dettaglio l'*analogia* che mette in relazione agenti morali umani e artificiali. Sul lato della somiglianza troviamo il fatto che entrambi i termini sono in grado di elaborare informazioni e dare conseguentemente corso ad effetti passibili di essere considerati alla luce di valori morali. Dal lato della differenza troviamo che le azioni dei primi possono essere descritte secondo il concetto di autonomia, mentre le funzioni dei secondi rappresentano la concretizzazione più piena del concetto di eteronomia – il quale, prima degli agenti artificiali, mancava di un vero e proprio caso concreto. Da ciò si può dedurre che la teoria dell'agire morale artificiale, o etica funzionale, consista nello studio della piena eteronomia e del suo impatto sull'agire morale umano.

Assumere il punto di vista dell'etica funzionale ha l'effetto di riposizionare la discussione sugli agenti morali artificiali presso la scuola di pensiero che ha messo a fuoco le nozioni di *mediazione tecnologica*, *moralizzazione delle tecnologie* e *tecnologia morale*[39]. Ciò non ha solo il vantaggio di evitare la confusione categoriale propria del paradigma continuista. In più, interrogarsi sugli agenti morali artificiali attraverso la concettualità delle tecnologie morali permette il ricorso a schemi e metodologie ampiamente collaudati per la descrizione delle reali difficoltà poste dalla disseminazione sociale di agenti artificiali che implementano valori morali. A scopo puramente indicativo se ne possono segnalare, seppur brevemente, almeno tre.

La prima difficoltà consiste nell'effettiva enucleazione dei valori di cui abbiamo parlato. Un conto, infatti, è supporre l'esistenza di valori condivisi che stiano alla base delle aspettative morali circa il funzionamento autonomo degli agenti artificiali; un altro conto è specificare quali siano questi valori, quali le loro condizioni di validità e quali le modalità migliori per la loro determinazione – domande necessarie se si vogliono evitare derive etnocentriche, paternalistiche o tecnocratiche[40]. Si tratta, in altre parole, del problema della *determinazione dei valori*.

La seconda difficoltà, o *problema dell'implementazione*, consiste nella capacità di tradurre in modo efficace e coerente la semantica dei valori morali in linguaggio informatico, cosicché l'allineamento tra funzionamento e aspettative etiche sia praticamente raggiungibile. Come già accennato, si tratta dell'oggetto di discussione principale dell'etica delle macchine. Parte del problema è anche la riflessione critica sull'implementazione stessa, dal momento che la trasformazione del linguaggio dei valori nel linguaggio dei parametri non può non avere effetti trasformativi sulla logica sottesa. Riflettere sul modo in cui la nozione di valore si trasforma nella sua traduzione computazionale è necessario sia per evitare facili rispecchiamenti o deleterie riduzioni, sia per mantenere la coscienza della differenza che separa l'esperienza umana dei valori dalla sua imitazione tecnologica.

39 La riflessione sulle tecnologie morali, inaugurata da Langdon Winner e Bruno Latour e sviluppata da autori come Hans Acherthuis, Don Ihde, Deborah Johnson e Peter-Paul Verbeek si concentra sulla nozione di *mediazione tecnologica* e sulla tesi conseguente per cui le tecnologie autonome siano *mediatori* di valori morali. Per una introduzione si veda *The Moral Status of Technical Artefacts*, a cura di P. Kroes e P.-P. Verbeek, Dordrecht, Springer Science+Business Media, 2014.

40 J. Danaher, *The Ethics of Algorithmic Outsourcing in Everyday Life*, in *Algorithmic Regulation*, a cura di K. Yeung e M. Lodge, Oxford, Oxford University Press, 2019, pp. 98-118.

La terza difficoltà, o problema del *deskilling morale*[41], riguarda il rischio dello sviluppo di un atteggiamento remissivo, passivo e delegatorio nei confronti delle performance di sistemi che implementano valori morali. A ciò potrebbe corrispondere un'atrofizzazione dell'esperienza diretta, il che è preoccupante in quanto essa è allo stesso tempo la fonte primaria dell'apprendimento morale e della formazione del carattere. Ciò vale, in verità, per tutti gli ambiti del fare umano che richiedono *giudizio*, quella capacità difficilmente formalizzabile di collegare caso e regola in modo efficace e che spesso è interpretata ricorrendo a nozioni di ardua analisi quali tatto, intuito, esperienza o talento. Secondo la tradizione della filosofia pratica, il giudizio morale ne è un esempio tipico insieme a quello dei medici e dei giudici – prestazioni sempre più coadiuvate da sistemi di apprendimento automatico. L'adozione massiva di algoritmi di raccomandazione in aree tanto sensibili del fare umano potrebbe certamente incrementare i risultati ottenuti (o, almeno, la loro qualità percepita), ma allo stesso tempo svuotarci delle abilità che ci rendono capaci di compiere simili prestazioni.

In conclusione, innestare lo studio degli agenti artificiali sulla tradizione delle tecnologie morali ha l'effetto immediato di sgombrare il campo da suggestioni inconsistenti e chiarire quali siano le questioni più controverse e più urgenti in relazione agli agenti morali artificiali. Lungo questa via il dibattito sull'etica delle macchine può trovare quel rigore filosofico e quella credibilità scientifica che altrove non è riuscito a guadagnarsi.

41 S. Vallor, *Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character*, in "Philosophy and Technology", XXVIII (2015), n. 1, pp. 107-124.