Si rende di seguito disponibile la versione accettata per la pubblicazione del saggio poi apparso come:

Fossa, F. (2020), Robot morali? Considerazioni filosofiche sulla *machine ethics*. Sistemi Intelligenti 2/2020, 425-444. https://www.rivisteweb.it/doi/10.1422/96334

Robot morali?

Considerazioni filosofiche sulla *machine ethics*

1. Introduzione

Il presente saggio propone una discussione critica dei presupposti metodologici ed epistemologici su cui si basa l'etica delle macchine—o *machine ethics* [ME]—allo scopo di determinare nel modo più chiaro possibile l'ambito di validità del sapere elaborato in tale sede. La prospettiva da cui si affronta la questione, dunque, non è tecnologica, ma filosofica: il problema sotto esame non concerne la traduzione algoritmica e l'implementazione dell'esperienza morale in sé, come compito robotico e computazionale, ma il modo in cui tale progetto è descritto e compreso, nonché le conclusioni che è lecito trarre alla luce dell'impostazione di un simile progetto di ricerca e delle sue ipotetiche o reali applicazioni.

L'idea di un'etica delle macchine, ovvero il tentativo di implementare una qualche forma di giudizio morale in robot autonomi o semi-autonomi, è trasmigrata nel giro di pochi decenni dai noti racconti di Isaac Asimov (1982) alla letteratura scientifica e ai laboratori specializzati (Wallach e Allen, 2009; Anderson e Anderson, 2011; Lin, Abney e Bekey, 2012). Non si è poi dovuto attendere altrettanto perché il tema invadesse la stampa, come dimostrano i numerosi articoli pubblicati su prestigiose riviste divulgative e rinomati quotidiani (Marcus, 2014; Deng, 2015; Parkin, 2017). L'esplosione di interesse circa la possibilità di progettare *robot morali* è certamente da inscrivere nella più ampia risonanza generata dalle recenti conquiste dell'automazione e dalle prospettive così aperte (Nourbakhsh, 2015; Cingolani e Metta, 2015; Carrozza, 2017; Kaplan, 2017). La vicenda della ME, però, è interessante non solo per il modo in cui incarna la vitalità della ricerca nei settori della robotica e dell'intelligenza artificiale o svela il fascino magnetico che simili iniziative esercitano sull'opinione pubblica. Da una prospettiva critica, la disciplina esemplifica anche un interessante fenomeno di ibridazione epistemologica tipico della scienza computazionale e robotica, per cui concetti e problemi tradizionalmente appartenenti alla filosofia assumono un'inedita formulazione tecnologica.

I costanti progressi nel campo dell'automazione determinano, tra le altre cose, una produttiva (ma rischiosa) contaminazione di due modalità del sapere solitamente e apparentemente disgiunte: il sapere, in senso lato, tecnologico e il sapere filosofico—morale, nel caso della ME. Ne risulta, in molti studi che affrontano il tema dei robot morali, una delicata commistione di registri che spesso trascura o sottovaluta la differenza dei due modi di sapere e le peculiari condizioni della loro mutua collaborazione al progetto di un'etica delle macchine. In sin troppi casi la confusione di dimensioni concettuali e linguistiche tra loro eterogenee si traduce in affermazioni sensazionalistiche circa le meraviglie che la robotica morale avrebbe in serbo per l'umanità futura; una retorica che, stuzzicando il voyeurismo fantascientifico che tanto sembra solleticare le voglie di futuro di autori, editori e lettori, è altrettanto ben attestata nella stampa divulgativa. Nondimeno, simili fraintendimenti offuscano l'intenzione reale per cui si persegue il progetto della ME e travisano ingenuamente (se non

furbescamente, per fare notizia) gli effettivi rapporti pratici ed epistemologici che determinano la relazione reciproca tra il sapere filosofico e tecnologico nell'ambito dell'etica delle macchine.
Particolare attenzione sarà dunque dedicata ai rapporti che si instaurano tra sapere tecnologico e filosofia morale nel contesto della ME, così da precisare i termini effettivi della loro collaborazione. Sulla base di ciò verrà poi chiarito in che misura il sapere ottenuto nel campo dell'imitazione tecnologica dell'esperienza morale possa retroagire sul sapere filosofico relativo all'etica umana. In altre parole, proverò a distinguere il problema tecnologico che sta alla base dell'etica delle macchine dalle speculazioni filosofiche (o pseudo-filosofiche) a cui è frequentemente abbinato, in modo da far emergere attraverso una critica di queste ultime la reale importanza del tentativo di implementare una qualche forma di algoritmica morale nei robot che siamo in procinto di costruire e usare. Nelle prossime pagine, quindi, si difenderà un'interpretazione dell'etica delle macchine in termini funzionali, che sottolinei cioè il carattere strumentale dei prodotti tecnologici e, conseguentemente, del loro possibile profilo morale, di contro ad una differente prospettiva che vede nella ME il mezzo conoscitivo più adeguato per lo studio degli oggetti morali. In definitiva, l'intenzione qui perseguita consiste nell'approntare gli strumenti critici necessari a riconoscere frettolose estensioni di concetti e nozioni propri di uno specifico progetto di ricerca ad ambiti del sapere ad esso non omogenei, così da apprezzare le potenzialità insite nella robotica morale senza cedere alla fallace attrattiva di utopiche aspettative di salvezza o distopiche profezie di sventura.

2. Abbiamo bisogno di robot morali?

Per prima cosa proviamo a chiarire il bisogno che sta alla base del progetto finalizzato alla costruzione di robot morali. Non diversamente da quanto intuito da Asimov (le famose tre leggi della robotica sono un'ipotesi discussa in ME; si veda esempio Clark 2011 e Anderson 2011c), la necessità di integrare al funzionamento dei robot il rispetto di determinati valori morali[1] deriva da ragioni interne all'avanzamento tecnologico. Un fattore trainante della ricerca robotica consiste nello sviluppo di sistemi *autonomi*, cioè—secondo un'accezione molto diffusa—capaci di svolgere le proprie funzioni indipendentemente dalla supervisione umana[2]. Le applicazioni robotiche autonome sono in grado di ricavare e processare da sé elevate quantità di informazione così da poter svolgere nel modo più efficiente anche compiti complessi e articolati. Diventa perciò possibile delegare ad automatismi

[1] Non è facile (o possibile?) proporre una definizione di "valore morale" che valga in ogni occasione e contesto. Tuttavia, per evitare fraintendimenti e dato il ruolo che la nozione giocherà nel presente saggio, è necessario provare almeno a delineare in generale il significato in cui sarà usata nelle pagine che seguono. Per "valore morale" intendo qui un contenuto di senso che orienta scelte pratiche verso la concretizzazione di un ideale del bene (anch'esso assai difficile da definire in modo assoluto e universale). Per quanto riguarda il presente discorso, si può meglio intuire il modo in cui si intende la nozione di valore morale se la si confronta con altri generi di rappresentazione che possono influenzare le scelte pratiche, come il criterio strumentale dell'efficienza. Mentre quest'ultimo influenza una scelta pratica determinandola verso la realizzazione dello scopo preposto nel "miglior" modo possibile (cioè massimizzando il risultato e minimizzando l'uso di risorse), il valore morale influenza una scelta pratica determinandola in base all'allineamento tra lo scopo perseguito e un ideale del bene, di ciò che ci sta a cuore e che riteniamo necessario rispettare e concretizzare in virtù della sua intrinseca desiderabilità. Per esempio, secondo molti la dignità umana (anch'essa, ancora, difficile da definire) è un valore da concretizzare nelle nostre scelte di vita e tale da mettere in secondo piano considerazioni determinate dal criterio dell'efficienza, in quanto il suo rispetto concorre a dare sostanza all'ideale morale del bene. Distinguere tra valore e bene è necessario perché valori diversi possono essere allineati al medesimo bene, cosa che sembrerebbe richiederne una mai facile gerarchizzazione.

[2] Cfr. per esempio Sullins (2011), p. 158: «I mean to use the term "autonomy" as engineers do, simply that the machine is not under the direct control of any other agent or user»; e Bekey (2012), p. 18: «What does it mean for machines to have autonomy? If we may simply stipulate it here, we define "autonomy" in robots as the capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time».

l'esecuzione di mansioni che, prima della loro introduzione, richiedevano lavoro umano per essere svolte. Alcune delle mansioni delegate, però, possono presupporre l'esercizio del giudizio morale da parte dell'agente umano coinvolto. Se un sistema autonomo è impiegato in un contesto pratico in cui certi valori morali sono in gioco, o se ad esso è delegato un compito che implica il rispetto o l'affermazione di valori morali, è consigliabile—se non necessario—che il robot in questione soddisfi le aspettative morali che derivano dal ruolo assegnatogli (Moor, 1995 e 2011)[3].

La condizione appena descritta, d'altronde, è una tappa inaggirabile dello sviluppo robotico (Wallach e Allen 2009, pp. 13-22). I vantaggi, in termini di risorse e accuratezza, che derivano dal ricorso a tecnologie autonome per la gestione di situazioni complesse sono difficili da negare e l'utilizzo di simili tecnologie in contesti dove sono in gioco anche valori morali è ormai imminente, quando non già reale. Tuttavia, la possibilità di un intervento diretto di un utente umano che prenda le decisioni morali al posto dell'automatismo è in molti casi da escludere: la lentezza e l'imprecisione che contraddistinguono l'azione umana vanificherebbero i vantaggi che giustificano il ricorso alle tecnologie autonome. L'impossibilità di esercitare un controllo in tempo reale sullo svolgimento di funzioni da parte dei sistemi autonomi, che si traduce in una loro certa imprevedibilità[4], è già di per sé motivo di preoccupazione da un punto di vista etico, dal momento che i passaggi intermedi possono generare effetti moralmente significativi almeno quanto il conseguimento finale dello scopo. Non resta dunque che affidarsi, anche da una prospettiva morale, a funzioni di controllo integrate che assicurino la coincidenza tra il funzionamento del robot e le nostre aspettative morali.

Per riassumere: a seconda delle situazioni in cui è calata, più una tecnologia si gestisce in autonomia, più è necessario che il suo funzionamento sia conforme non solo a criteri di efficienza, ma anche a valori che noi utenti umani consideriamo degni di rispetto e affermazione. Il concetto stesso di affidabilità (*reliability*) di una determinata tecnologia autonoma deve dunque essere ampliato tanto da includere la convergenza del funzionamento dell'agente artificiale non solo con il fine ad esso preposto, ma anche con i valori morali che sono in gioco nelle situazioni in cui si trova ad operare.

Come ci si può assicurare che simili automatismi non tengano conto solo dell'efficienza e della produttività, ma anche di valori, diritti e doveri? Ecco che emerge la necessità di dotare le tecnologie autonome di un intero nuovo insieme di capacità: rilevare quali valori siano in gioco in quali contesti pratici, risolvere contrasti tra i criteri strumentali della conformità a scopo e i criteri morali della conformità a valore, gerarchizzare i diversi valori morali di cui si ha contezza, e così via. L'esigenza è tanto più urgente quanto più si accresce la portata dell'autonomia robotica (Allen, Varner e Zinser, 2000; Torrance, 2011); un processo, questo, indissolubilmente legato all'attuale dinamica di sviluppo tecnologico. Un qualche analogo della facoltà umana di giudizio morale deve essere integrato ai programmi che regolano il funzionamento delle tecnologie autonome.

3. Costruire un robot morale

[3] L'esempio più citato di tecnologia autonoma che, dato il compito assegnatole, richiede l'implementazione di una qualche forma di giudizio morale è l'automobile a guida autonoma. La letteratura sul tema è già colossale: per una aggiornata introduzione cfr. la prima sezione di Lin, Abney e Jenkins (2017). Per una introduzione più "pratica" si visiti il sito http://moralmachine.mit.edu, dove ci si può misurare in prima persona con i problemi etici posti dalle *autonomous car*. La presentazione dello studio basato sui dati raccolti tramite il sito (Awad, Dsousa, Kim *et al.* 2018) ha ottenuto grande risonanza sulla stampa mondiale.

[4] L'imprevedibilità del comportamento delle tecnologie autonome è tra le ragioni che motivano i dubbi circa la fidatezza (*trustworthiness*) degli agenti artificiali, al di là della loro affidabilità (*reliability*). Su ciò si veda, ad esempio, Dzindolet, Peterson, Pomranky, Pierce e Beck (2003), Taddeo (2010), Coeckelbergh (2012), Tavani (2015).

Una volta chiarito il bisogno alla base della robotica morale, prendiamone in esame l'impianto epistemologico. Affinché gli automatismi manifestino un comportamento non solo conforme allo scopo loro preposto, ma anche ai valori socialmente condivisi, è necessario che siano progettati come se fossero *agenti morali* (Wallach e Allen 2009, p. 17). L'essere umano, essendo il solo ente che esibisce la piena capacità di agire in conformità a valori etici, è in questo frangente l'unico modello disponibile da imitare o riprodurre. Di conseguenza, è del tutto naturale che il punto di partenza per la progettazione di un *agente morale artificiale* coincida con la modellizzazione tecnologica dell'esperienza morale umana (Wallach, 2010). Per conseguire tale fine, è prima di tutto necessario rielaborare il sapere a nostra disposizione circa l'esperienza morale in un linguaggio omogeneo a quello che permette la costruzione di tecnologie autonome. Si potrà così pervenire a una struttura funzionale, deterministica e meccanicistica, che possa guidare i programmatori e gli ingegneri alla riproduzione imitativa del comportamento morale umano e alla sua integrazione in sistemi automatici.

Come suo primo postulato, dunque, la ME prevede che possa essere definita una base comune tra agente morale umano e artificiale tale da giustificare l'intero progetto di riproduzione tecnologica. Per determinare questa base comune, i caratteri definitori di ciò che è lecito ritenere un agente morale devono essere riformulati in senso astratto, in modo che possano essere ammessi nel novero degli agenti morali non solo gli esseri umani, ma tipologie multiple di enti (tra cui ovviamente i robot). Ciò implica che si debba procedere ad una riformulazione del concetto di agente morale che sia priva di ogni riferimento esclusivo alla classe degli esseri umani. L'uomo, in questo nuovo scenario, non rappresenterà più un oggetto qualitativamente unico, ma il picco di un continuo lungo la cui linea si possono situare anche agenti artificiali (Asaro, 2006). L'incontro tra le esigenze epistemologiche della scienza robotica e il sapere relativo all'esperienza morale umana deve generare, in altre parole, un modello non antropocentrico di agente morale che serva poi da guida alla progettazione di macchine etiche. Secondo quali direttive, però, può essere riformulata la categoria di agente morale in modo che includa i robot autonomi senza escludere l'uomo?[5]

Una simile riformulazione può giovarsi dei suggerimenti di Floridi e Sanders (2004), i quali hanno proposto uno schema interpretativo per l'estensione della categoria di agente morale a sistemi tecnologici. Secondo i due autori, i caratteri fondamentali di ogni agente—artificiale o meno—possono essere essenzialmente ricondotti a tre categorie formali:

- interattività: la capacità di reagire a stimoli ambientali;
- autonomia: la capacità di elaborare informazioni e dare avvio a processi in modo indipendente da stimoli ambientali;
- adattabilità: la capacità di modificare le leggi provvisorie che regolano il comportamento sulla base delle esperienze pregresse e della situazione particolare.

Se un agente, descritto da queste categorie, è inoltre in grado di causare effetti significativi da un punto di vista morale, allora si tratta di un agente morale. Diviene quindi possibile considerare un robot alla stregua di un agente morale qualora esso sia in grado di analizzare una situazione in conformità a valori, elaborare le informazioni così ottenute secondo istanze valoriali, prendere delle

[5] L'idea di includere agenti artificiali nel novero degli agenti morali, espressa con grande chiarezza da Anderson (2011b) e Sullins (2011), si basa sulla preliminare attribuzione di *agency* ad alcuni prodotti tecnologici tipica della *computer science* (Franklin & Gaesser 1996). Da un punto di vista filosofico, l'apertura verso forme tecnologiche di agenti morali è inoltre messa spesso in relazione ad analoghi progetti di estensione del dominio etico in modo da includere, ad esempio, animali e ecosistemi (Torrance, 2011; Gunkel, 2012). Per alcuni, tra cui Himma (2009), Johnson (2011), Bryson & Kime (2011), questo tipo di estensione è illegittima. A mio parere (cfr. *infra*) bisogna distinguere quanto vale nel dominio della ME e quanto invece vale al di là dei presupposti che è necessario assumere in questo ambito di ricerca.

decisioni che conducano ad atti effettivi e infine rielaborare queste stesse istanze alla luce delle sessioni passate[6].

Pur semplificando la complessità della questione, si può affermare che i compiti della ME siano quindi riconducibili alla riproduzione operativa della capacità di giudizio e delle facoltà umane in grado di presentare istanze al giudizio—immaginazione, emotività, empatia, e così via. La possibilità stessa della riproduzione, però, è assicurata dalla riformulazione del concetto di agente morale in un linguaggio omogeneo al sapere robotico. L'estensione del concetto di agente morale è funzionale al progetto dell'etica delle macchine in quanto prescinde dall'unicità delle condizioni umane dell'esperienza etica e, inquadrando il sapere relativo all'esperienza morale in una cornice epistemologica adeguata alla scienza tecnologica, getta le basi per un'approssimazione graduale degli automatismi alle prestazioni umane prese a modello.

L'autonomia nello svolgimento di funzioni morali e la sensibilità a valori non valgono però solo come caratteristiche del modello astratto, ma anche come criteri per la classificazione degli agenti morali; una classificazione che riunisce in un unico schema prodotti tecnologici e esseri umani, rendendo operativo il concetto di agente morale proprio dell'etica delle macchine. Secondo tale approccio, presentato nel modo più chiaro da Wallach e Allen (2009, pp. 25-37), si possono distinguere almeno tre tipologie di *moral agency: operational morality, functional morality,* e *full moral agency.*

A gradi nulli di autonomia e sensibilità a valori corrisponde la modalità più basilare dell'implementazione di istanze valoriali, che consiste nel ricorso a specifici accorgimenti (sicure, sistemi di sicurezza, ecc.) volti a prevenire, per quanto possibile, un uso o un funzionamento inappropriato della macchina. Simili congegni né soppesano alternative, né considerano diversi scenari: la conformazione dello strumento è ideata in modo da impedire usi pericolosi o funzioni dannose. Controllo, valutazione e decisione non sono gestiti in autonomia, ma imposti rigidamente dal costruttore, che in base alle proprie convinzioni, ai codici etici o alle normative vigenti impone dei limiti strutturali alle possibilità della macchina (*operational morality*). Ad un livello superiore si situano sistemi che adeguano in autonomia la propria funzione a istanze morali e che sono in grado di estrapolare informazioni di carattere morale sia dalla considerazione di una situazione o degli effetti di una decisione che sono chiamati a prendere, sia dalle sessioni passate. Simili sistemi formano una classe intermedia, nella quale i gradi di autonomia e sensibilità a valori possono variare molto e in modo indipendente (*functional morality*). Infine, alti gradi di autonomia e sensibilità, nonché la piena capacità di rendere ragione dei propri atti, definiscono l'agente morale nel senso più pieno della parola (*full moral agency*). L'unico oggetto di questa classe, allo stato attuale della ricerca, è l'uomo. Tuttavia, l'uomo non *definisce* che cosa sia un agente morale, ma semplicemente si trova ad esserne il caso più avanzato: i diversi agenti morali sono separati da niente più che una differenza di grado.

Ad oggi, la missione della ME consiste nello sviluppare automatismi le cui prestazioni soddisfino in buona parte il confronto con il comportamento di un agente morale umano[7]. Il settore di "agenti

[6] Considerare robot alla stregua di agenti pone anche la necessità di un adeguamento legislativo. Da alcuni, come Peter Asaro (2012), il taglio giuridico è considerato anzi una via d'accesso privilegiata a questioni che, da un punto di vista morale, risultano più difficili da dirimere. Ne è sorta una vivace discussione su molti temi (Calo, Froomkin e Kerr, 2016; Pagallo, 2013), tra cui la convenienza di attribuire ad agenti artificiali diritti (Gunkel 2018), responsabilità civile o penale (Bertolini, 2013; Hallevy, 2013), o personalità giuridica (Calverley, 2011; Solaiman 2017; Bryson, Diamantis, Grant 2017).

[7] In questi termini Allen, Varner e Zinser (2000) parlano di un *Moral Turing Test (MTT)*. Anche Moor (2006), com'è noto, ha proposto una classificazione degli agenti morali articolata in *ethical impact agents* (qualsiasi macchina che generi effetti valutabili da un punto di vista morale), *implicit ethical agents* (macchine progettate in modo tale che prevengano

morali" che l'etica delle macchine si sforza di popolare è quello mediano, situato tra una moralità puramente operazionale e lo sfaccettato fenomeno dell'esperienza morale umana. Con la tanto suggestiva quanto indeterminata espressione "robot morale" si intende per lo più un sistema artificiale che sia in grado di processare informazioni di carattere morale e produrre scelte conformi a valore senza bloccarsi, richiedere l'intervento di un utente umano o ignorare alcune istruzioni a vantaggio di altre. Per fare ciò, il sistema deve essere in grado di determinare quali siano e da dove provengano le informazioni significative per il giudizio morale che è chiamato a computare: dovrà quindi essere dotato di apparati percettivi in grado di catturare simili informazioni e inserirle nel processo decisionale, nonché di tutte le istruzioni relative a ciò che i suoi utenti ritengono, o in linea di massima dovrebbero ritenere, buono e giusto. I robot morali non sono altro che specchi delle convinzioni di chi li programma e—almeno lo si spera—di chi li usa; strumenti progettati in modo da svolgere funzioni utili e, allo stesso tempo, attenersi a specifiche indicazioni morali. Il tentativo di realizzare simili automatismi si basa sull'impianto epistemologico di cui è appena stata proposta una sintesi.

4. Etica delle macchine, etica degli umani

Come si è detto, il primo passo della ME consiste in una modellizzazione dell'esperienza morale umana, basata sul sapere disponibile, che sia omogenea al linguaggio e alle strutture della scienza robotica. Ma il modello scientifico-tecnologico così elaborato, di che cosa è modello? A prima vista, può sembrare che sia l'esperienza morale stessa, così com'è, ad essere qui colta e tradotta nel linguaggio univoco e verificabile della scienza. D'altra parte, il sapere sviluppato nel campo della ME si basa su un concetto di agente morale interamente definito secondo le norme del metodo scientifico. Potrebbe dunque sembrare lecito concludere che quanto si scopra e si conosca in questa sede non ammonti solo a un sapere circa la robotica morale, ma riguardi invece la conoscenza scientifica—cioè vera—dell'esperienza morale in quanto tale.

Nello scenario che ora si profila, l'etica delle macchine non sarebbe più solo un sapere volto alla realizzazione di robot morali, ma verrebbe promossa a conoscenza dell'agente morale in sé—in linea con l'adagio caro alla scienza tecnologica secondo cui si conosce davvero solo quanto si è in grado di costruire. Se le nozioni poste nell'ambito epistemologico dell'etica delle macchine riguardano l'esperienza morale in sé e se solo differenze di grado separano noi umani dai robot che siamo (o saremo) in grado di costruire, allora unicamente una questione di tempo separa agenti morali umani e artificiali. Ciò implica che il sapere nato dall'indagine della ME non coglierebbe tanto il modo in cui si possa imitare l'esperienza morale umana per mezzo della robotica, quanto la verità scientifica circa l'esperienza morale espressa nel linguaggio ad essa idoneo. Ma allora non è più l'esperienza morale umana il modello tramite cui comprendere la controparte robotica: al contrario, le categorie e i prodotti dell'etica delle macchine divengono, rispettivamente, la vera scienza dell'esperienza morale e il mezzo per risolverne i problemi pratici.

---

funzionamenti o usi eticamente inaccettabili), *explicit ethical agents* (macchine programmate in modo tale che sappiano svolgere ragionamenti e prendere decisioni conformi a istanze morali), e *full ethical agents* (sistemi in grado di prendere decisioni moralmente complesse e di giustificarle in autonomia). Come anche per Wallach e Allen, per Moor l'obiettivo dell'etica delle macchine non è tanto il livello dei *full ethical agents*, quanto quello intermedio rappresentato dagli *explicit ethical agents*. Asaro 2006, p. 15, aggiunge l'intelligente considerazione per cui rimane un obiettivo troppo ambizioso quello di costruire un *explicit ethical agent* capace di affrontare problemi morali generali: si deve concentrare l'attenzione sulla gestione autonoma di problemi morali specifici, relativi alla funzione svolta dal robot e dal suo ambito di applicazione. Con ciò concorda Anderson 2011a, p. 25, la quale aggiunge il suggerimento di cimentarsi con buoni *advisors* morali prima di impegnarsi nella costruzione di agenti morali artificiali.

Le tesi che sposano una simile prospettiva, non sempre però con eguale consapevolezza, sono ben attestate e di carattere molto vario, pur germogliando tutte lungo il medesimo ramo.
In alcuni casi si afferma che la robotica sia destinata a produrre agenti artificiali le cui capacità di risolvere problemi morali supereranno di gran lunga le prestazioni umane (Bostrom, 2003). Non solo noi umani abbiamo difficoltà a seguire la catena delle possibili conseguenze delle nostre decisioni, mancando spesso di realizzare le molteplici implicazioni delle nostre scelte. In più, gli attributi meno evidenti delle situazioni in cui ci troviamo coinvolti sfuggono alla grossolanità dei nostri apparati affettivi (percettivi e emotivi), impedendo un'analisi completa ed accurata del caso; per tacere poi delle innumerevoli interferenze a cui è esposta la nostra capacità di giudizio, in balìa com'è delle più spurie influenze[8]. Le suddette limitazioni non interessano invece il robot morale, le cui capacità di raccolta, analisi ed elaborazione dati dipendono unicamente dall'avanzamento della ricerca tecnologica. In futuro, si conclude, non sarà più l'essere umano a rappresentare la tipologia più avanzata di agente morale, ma il robot—che potrà così istruirci sul buono e sul giusto, se non realizzarli al nostro posto. Se c'è continuità tra agente morale artificiale e umano, è evidente che l'uomo sia destinato ad abdicare in favore dei suoi apparati tecnologici, con le conseguenze distopiche (Dietrich, 2007) o utopiche (Anderson, 2011d) che ognuno ritiene di trarne.
In altri luoghi si legge poi che la traduzione in linguaggio computazionale delle teorie e dei saperi morali non esibisce uno statuto puramente tecnologico, ma deve riflettersi anche sullo statuto filosofico delle teorie e dei saperi in questione (Wallach 2010; Gips, 1995; Anderson 2011a). Se, pur nella differenza concreta che distingue agenti artificiali e umani, non si dà una sostanziale rottura tra gli oggetti raccolti sotto il concetto di agente morale, si potrebbe pensare che una proposta inconsistente dal punto di vista dell'etica delle macchine rappresenti una teoria debole anche in relazione all'esperienza morale umana. Qui la ME diventa normativa nei confronti dell'etica filosofica: si sottintende infatti che il sapere ottenuto nel campo dell'etica delle macchine non informi solamente circa la robotica morale, ma sia invece relativo all'etica in quanto tale, e dunque anche all'etica umana. Resta solo un breve passo dall'affermare che l'etica delle macchine riuscirà laddove più di duemila anni di filosofia morale hanno ripetutamente fallito, non producendo altro che un sapere incerto, mutevole, esposto a continue diatribe e rifondazioni (Beavers 2012).
Difatti, la tesi più comune e di portata più generica consiste nell'affermare che lo studio tecnologico dei processi decisionali morali non conduca solo a migliori automatismi, ma accresca o corregga la conoscenza dell'uomo su se stesso e sulla sua esperienza morale: «programming or teaching a machine to act ethically will help us better understand ethics» (Moor, 2006, p. 21). L'ingegnere e il programmatore si troveranno a dover analizzare con grande precisione gli articolati processi alla base delle procedure decisionali umane; ne offriranno scomposizioni sempre più accurate e modelli sempre più fini. Nello sforzo di regolare la funzione degli automatismi, l'etica delle macchine può svelare come funziona l'esperienza morale umana e aprire nuove prospettive sull'oggetto classico dell'etica filosofica: come scrivono Wallach e Allen (2009, p. 56) «pressing ahead on the practical task of building AMAs will contribute to better understanding of the ontological and epistemological questions about the nature of ethics itself». Si presuppone così che ciò che avviene in modo confuso e oscuro nel caso dell'esperienza morale umana, e che ha occasionato infinite dispute filosofiche, sarà presto esibito nel modo più chiaro distinto dai robot morali. Apparentemente, l'etica delle macchine non assume solo il compito di adeguare il funzionamento degli automatismi ai valori che l'uomo

[8] Nadeau (2006) si spinge ad affermare che l'unico *vero* agente morale non può che essere artificiale, in quanto solo in questo caso è possibile separare il lato logico-razionale del giudizio dalle sue componenti emotive, passionali, e dalle altre influenze che ne offuscano la purezza.

reputa degni di essere perseguiti; allo stesso tempo, si propone anche come esercizio di autocomprensione filosofica: «the exercise of thinking through the way moral decisions are made with the granularity necessary to begin implementing similar faculties into (ro)bots is thus an exercise in self-understanding» (*ivi*, p. 11).
La circolarità di tali affermazioni, che saltano d'un balzo dalla modellizzazione tecnologica dell'esperienza morale umana alla sua retroflessione sull'esperienza morale umana stessa, deve insospettire circa la bontà dell'atto epistemologico che le motiva. Bisogna perciò chiedersi se la ME possa davvero dire qualcosa di positivo sull'esperienza morale umana, se lo specchio in cui uomo e automatismo si riflettono l'un l'altro possa veramente essere attraversato in entrambi i sensi. Com'è da intendere il rapporto tra i saperi dell'etica delle macchine e della filosofia morale?

5. La questione epistemologica

Per gettare luce sull'ambito di validità delle conoscenze prodotte nel contesto dell'etica delle macchine suggerisco di approfondire il rapporto epistemologico che si instaura tra quest'ultima e una delle discipline che tradizionalmente studiano e riflettono sull'esperienza etica, cioè la filosofia morale[9]. Il coinvolgimento del sapere filosofico nel progetto dell'etica delle macchine esemplifica l'interdisciplinarità caratteristica di molti indirizzi di ricerca nel campo della robotica e dell'intelligenza artificiale. In contesti interdisciplinari, ovviamente, la contaminazione di diversi registri epistemologici è inevitabile: il rischio di perdere consapevolezza delle differenze e dei confini che separano i diversi modi del sapere è il prezzo da pagare per la vivacità che lo scambio assicura. D'altronde, vale la pena di correre il rischio: ibridare il sapere robotico e filosofico è una delle vie che potrebbero condurre alla costruzione di robot morali.
L'operazione, da un punto di vista epistemologico, è sicuramente delicata e richiede cautela perché venga condotta nel rispetto delle specificità dei saperi coinvolti—cautela purtroppo ignorata da molta retorica. Le condizioni che fondano la cooperazione dei saperi filosofico e tecnologico al progetto della ME vengono spesso travisate, generando giochi di specchi che rendono concepibili le tesi più intriganti, sensazionalistiche, conturbanti, ma anche inconsistenti. Per evitare di perdere coscienza della linea che separa etica delle macchine e etica degli umani (Amigoni e Schiaffonati, 2005) è utile prendere le mosse da una riflessione circa i rapporti effettivi che legano tecnologi e filosofi nel progetto dell'etica delle macchine. Sarà così possibile chiarire quali siano i presupposti di una tanto inusuale cooperazione, il che coincide con la determinazione dei confini entro cui è lecito far valere le conoscenze elaborate in tale sede.
Se si eccettua il caso delle varie etiche applicate (*engineering ethics, computer ethics, roboethics, AI ethics* e così via), che esibiscono però un'ibridazione differente rispetto a quella ora in esame, i rapporti tra filosofia morale e mondo dell'ingegneria sono incentrati su una cauta separazione dei rispettivi saperi, secondo una tradizione millenaria ripresa volta per volta da entrambe le parti. Nel caso dell'etica delle macchine, invece, il filosofo può assistere costruttivamente il tecnologo nella sua opera di implementazione del comportamento morale. Qui il filosofo non ha tanto l'incarico critico

[9] La scelta di concentrarsi sulla filosofia morale è dovuta al fatto che il problema discusso in questo saggio riguarda la validità delle conoscenze sviluppate in ME, che in alcuni casi viene estesa tanto da soppiantare la prima. Bisogna però tenere a mente che l'interdisciplinarità della ME non coinvolge solamente la filosofia nel suo progetto, ma anche molti altri saperi che hanno per oggetto aspetti relativi alla dimensione morale umana quali, ad esempio, la psicologia, la biologia, l'antropologia, la sociologia, la teologia. Ancora, si tenga a mente che il discorso qui sviluppato circa i rapporti reciproci di filosofia morale e etica delle macchine non è generalizzabile ai rapporti interdisciplinari tra etica delle macchine e altri saperi: ogni caso deve essere attentamente valutato a sé.

di portare all'evidenza i valori impliciti nella costruzione di sistemi artificiali, cosicché ne venga messa in discussione la supposta neutralità morale (Nissembaum, 2001), né di tracciare le linee guida che assicurino l'adeguatezza etica del lavoro dei tecnologi (Veruggio, 2006). Al filosofo che collabora al progetto della ME pertiene soprattutto la missione positiva di affiancare ingegneri e programmatori impegnati nella riproduzione algoritmica dell'esperienza morale. In un senso quasi baconiano, il filosofo è ora chiamato a mettere "a frutto" o "in opera" la sua conoscenza del fenomeno etico.

Tecnologia e filosofia scoprono così una particolarissima intersezione dei loro interessi e dei loro saperi per molti versi così distanti. Da un canto, a partire dalla propria ricerca ingegneri e programmatori individuano nella filosofia uno dei saperi che possa guidarli nella determinazione degli elementi costitutivi dell'esperienza etica; dall'altro i filosofi, che intuiscono rischi e potenzialità dell'automazione, si trovano nelle migliori condizioni per rispondere all'appello e possono forse anche nutrire interesse nella prospettiva di una messa alla prova informatica o robotica delle proprie idee sulla morale (Beavers, 2011). Tuttavia, al di là del giustificabile entusiasmo che circonda questa sorta di rivalsa pratica dell'etica filosofica, se si vuole capire quali considerazioni sia lecito trarre dai risultati dell'etica delle macchine rimane imprescindibile una discussione del modo in cui le due discipline stringono il loro patto di collaborazione.

L'aspetto più importante da chiarire è che filosofo e tecnologo non celebrano, per così dire, delle nozze tra pari. Al contrario, il filosofo figura come consulente in un progetto a direzione tecnologica. L'etica delle macchine, d'altronde, è una sottosezione della ricerca robotica e informatica. Il fatto che sia il tecnologo a coinvolgere il filosofo morale nel suo progetto fa sì che quest'ultimo debba accettare una serie di presupposti, rifiutati i quali la collaborazione non può che rivelarsi insensata, paradossale. Il filosofo deve accettare i termini generali della sfida se vuole giocare un ruolo costruttivo nella progettazione dei robot morali; più precisamente, deve essere disposto a mettere da parte ogni contenuto che non si presti ad essere compreso nei termini del sapere che guida il progetto (Frank e Klincewicz, 2016). «A valid approach», scrive bene Beavers (2012, p. 340), «must also be an implementable one».

Non deve quindi stupire che l'etica delle macchine presenti un'interpretazione dell'esperienza morale dominata da quegli aspetti che più si prestano ad essere espressi in termini computazionali. Tantomeno deve stupire il fatto che l'etica delle macchine comprenda l'esperienza morale nel senso di un processo inferenziale o «a host of cognitive mechanisms» (Wallach, 2010, p. 249). Come l'automatismo è in grado di processare informazioni e, in base a queste, di dare corso ad eventi conformi allo scopo della propria funzione, così l'unico concetto di "esperienza morale" che abbia un qualche senso tecnologico non può che rimandare ad una serie di passaggi, esprimibile in un programma, che istruisca su come processare informazioni di carattere morale e, in base a queste, decidere tra alternative concorrenti. Ecco istituito il presupposto primario che il filosofo deve assumere come valido se vuole contribuire all'etica delle macchine—e su cui quindi, all'interno dell'ambito della ME, non ha senso sollevare obiezioni filosofiche: ogni sforzo dev'essere indirizzato ad un'*etica per sistemi deterministici*[10].

[10] Non sembra condivisibile l'obiezione di Floridi & Sanders (2004, p. 366), secondo cui gli automatismi possono essere considerati come agenti liberi, o sistemi non-deterministici, in quanto presentano i caratteri dell'autonomia, dell'adattabilità e della interattività: queste ultime categorie sono *già* definite secondo la logica deterministica di input e output. Una simile critica può essere mossa all'argomentazione di Moor (2006), per cui non ha senso sostenere che sistemi meccanicistici non possano esibire comportamenti morali, in quanto l'uomo stesso altro non è che una macchina biologica, ed è capace di agire moralmente. L'argomento è infatti condivisibile solo a patto che l'interpretazione tecnologica preceda lo studio dell'umano e lo determini (Sini, 2009). Ugualmente deboli mi sembrano le obiezioni di

In ME, dunque, la filosofia morale è interessante e utile non in quanto tale, ma nella misura in cui permetta una comprensione tecnologica dell'esperienza morale. Le nozioni sottoposte al tecnologo devono essere, almeno in linea di principio, esprimibili in termini funzionali e algoritmici. Di conseguenza, i contenuti del sapere filosofico che presuppongono una logica non formalizzabile e una causalità non deterministica rimangono inevitabilmente da parte. Se l'obiettivo è elaborare un programma e costruire un automatismo che imiti il comportamento morale umano, è già presupposta una presa di posizione sui temi più complessi e discussi della filosofia morale. Concetti come libertà, libero arbitrio, coscienza morale, autodeterminazione e interpretazione del valore restano inevitabilmente alla periferia dell'etica delle macchine. Ciò di cui il tecnologo ha bisogno è un modello computabile e riproducibile di esperienza morale: tale è la condizione di possibilità di ogni proficua collaborazione tra filosofi e tecnologi in ME. Una volta che sia posta la necessità dell'integrazione di capacità morali negli automatismi, il filosofo deve adeguare il proprio linguaggio al modo in cui il sapere tecnologico conosce i propri oggetti.

Il ruolo della filosofia morale nel progetto dell'etica delle macchine è dunque ausiliario. Le regole del gioco sono dettate dalla scienza tecnologica. Se il filosofo accetta la sfida, deve essere pronto a onorare i termini del patto. D'altra parte, l'imitazione tecnologica dell'esperienza morale non può che procedere lungo questa via e l'impostazione pare legittima a livello metodologico—a condizione, certamente, che filosofi e tecnologi non dimentichino i presupposti della loro collaborazione. Tali presupposti fondano un sapere che ha validità solo all'interno della collaborazione in questione: la legittimità della loro assunzione è correlata al progetto che rendono possibile. Si tratta, in definitiva, di ipotesi di lavoro, valide esclusivamente nell'ambito che dischiudono; e il sapere che si basa su di esse è valido in maniera ugualmente condizionale.

6. Giochi di specchi e cura del linguaggio

Alla luce di quanto detto, dovrebbe essere evidente che procedere all'interpretazione dell'esperienza morale umana attraverso le categorie e le strutture sviluppate in ME sia un atto epistemologico scorretto, preda di un circolo: l'oggetto iniziale viene ricondotto al modello nato da una selezione dei suoi contenuti. Le condizioni di possibilità della cooperazione tra tecnologi e filosofi al progetto della ME segnano così, allo stesso tempo, le ragioni per cui il sapere relativo ai robot morali non possa essere immediatamente esteso all'oggetto proprio della filosofia morale[11].

Abney (2012), pp. 48-50, che ignora completamente l'evidenza fenomenologica dell'umano essere esposti a situazioni pratiche in cui sono in gioco valori che si appellano direttamente a noi, e che dobbiamo fronteggiare senza poter semplicemente applicare istruzioni esterne o attenerci a standard da noi indipendenti.

[11] Questa conclusione non è da intendere in senso assoluto, come cifra di ogni rapporto tra filosofia e scienza tecnologica, ma riguarda solamente il tentativo di spiegare l'esperienza morale umana tramite la ME (e, semmai, analoghi tentativi di ricondurre un sapere più articolato alla sua riduzione tecnologica—cfr. XXX). Qui, cioè, non si vuole argomentare che teorie e concetti della filosofia non siano influenzabili dai rapporti che intrattengono con la tecnologia; al contrario, la filosofia ha tanto da apprendere dalle sue ibridazioni pratiche. La tesi avanzata sostiene solamente che la riduzione dell'etica umana all'etica delle macchine è basata su un atto epistemologico circolare e scorretto. Penso che sia invece corretto asserire che, in alcuni casi e secondo alcuni aspetti, l'implementazione di idee morali sia utile e produttiva—a condizione, però, che sia valutata con grande attenzione l'estensione di tale influenza. Ad esempio, mi sembra del tutto pacifico sostenere che l'implementazione di una dottrina morale possa produrre utili prove circa la consistenza logica e la coerenza interna della dottrina in questione. Tuttavia, per comprendere che cosa simili prove implichino dal punto di vista morale bisogna anche chiarire se a) la traduzione di una dottrina morale in termini computazionali non solo non ne rappresenti uno snaturamento, ma anzi ne colga l'essenza; e se b) coerenza interna e consistenza logica siano valori massimi e assoluti di una teoria morale; entrambe cose di cui è almeno lecito dubitare.

I problemi filosofici sollevati dall'etica delle macchine devono essere letti nel quadro della generale tendenza all'assottigliamento delle differenze percepite tra esseri umani e tecnologie autonome: più le potenzialità della tecnologia si accrescono, più la distanza dell'uomo dall'automatismo appare ridotta. A livello pratico, l'accostamento è forse ancora prematuro; a livello teorico, invece, la situazione è diversa. Come visto nel §4, già sono stati elaborati i termini per un discorso che dalla distinzione delle strutture e dei processi robotici muova verso la confusione esibita dai comportamenti umani. Esseri umani e robot si rispecchiano l'uno nell'altro (Fabris 2018). Con la ME, il rispecchiamento tocca l'esperienza morale.

Il rischioso gioco di specchi che ne deriva è mediato e allo stesso tempo celato dal linguaggio che abbiamo a disposizione per parlare degli agenti artificiali in generale e in particolare di quelli "morali". Essendo l'essere umano l'unico modello disponibile a cui ispirare la riproduzione tecnologica dell'esperienza morale, il linguaggio a cui si ricorre per presentare e descrivere l'agente morale artificiale non può che essere antropomorfo. La costituzione del lessico della ME attinge necessariamente alla dimensione di senso dell'agire umano. Per quanto ci si sforzi di concepire un paradigma non antropocentrico di agente morale, il linguaggio rimane antropocentrico—è l'unico disponibile.

È d'altronde del tutto naturale che le parole a cui si ricorre per descrivere l'imitazione robotica dell'umano siano le medesime che descrivono il modo d'essere del modello. Alle prese con i sistemi artificiali, il linguaggio si adatta da sé in modo che lo stesso termine possa significare sia il contenuto umano che l'analogo tecnologico. Si verifica così l'*estensione semantica* (XXXX) della parola antropomorfa all'oggetto robotico: un processo riscontrabile in modo trasversale negli studi di robotica e intelligenza artificiale. Sin da Wiener, McCarthy e Turing si legge di macchine che pensano, decidono, sentono, ricordano, imparano e agiscono quando esse calcolano, eseguono programmi, raccolgono dati, li immagazzinano, li riorganizzano e funzionano. È il linguaggio stesso, in un certo senso, ad allestire il gioco di specchi che induce poi ad impiegare l'accezione tecnologica di un termine per ridefinire l'analogo umano, appiattendolo su un sapere che è già dall'inizio una sua riduzione.

Considerare il fenomeno linguistico dell'estensione semantica è di massima importanza. È sulla sua base che vengono avanzate le affermazioni discusse nel §4. L'uso ordinario del linguaggio nei termini presentati alimenta l'impressione che robot e esseri umani siano una medesima cosa, e che quindi il sapere dell'etica delle macchine si applichi tanto agli uni quanto agli altri. Da qui deriva il rispecchiamento che porta ad assumere le ipotesi di lavoro della ME, in particolare modo la continuità tra agenti morali artificiali e umani, come tesi valide in assoluto. Il processo di estensione semantica induce a trascendere l'ambito di validità determinato dai presupposti operativi della ME. Allo stesso tempo, la naturalezza del gesto linguistico nasconde e rimuove la consapevolezza dell'infrazione epistemologica.

Con il tema del rispecchiamento si raggiunge il punto cruciale dell'intera questione. È stato mostrato come l'etica delle macchine non accolga il sapere della filosofia sull'esperienza morale umana in tutta la sua estensione, ma possa accedere solamente alla sua parte che si presta ad essere inquadrata in termini tecnologici. L'etica delle macchine lavora con un'immagine operativa dell'esperienza morale, adeguata al proprio discorso: «functional equivalence of behavior», dicono bene Wallach e Allen (2009, p. 68), «is all that can possibly matter for the practical issues of designing artificial moral agents». La copia algoritmica dell'esperienza morale umana rappresenta un'elezione della parte sul tutto giustificata esclusivamente da fattori interni al progetto a cui appartiene. Perciò, in qualsiasi luogo del suo percorso la ME non incontrerà mai l'umano, ma solo il robotico: la possibilità di accedere al fenomeno dell'umano e alla sua peculiare logica è stata accantonata in sede metodologica.

Di conseguenza, il ritorno della robotica morale all'esperienza morale umana non potrà che incontrare nuovamente e solamente se stessa.
L'impostazione metodologica dell'etica delle macchine offre tutti gli strumenti necessari a schermare il fenomeno del rispecchiamento. Se si assume una prospettiva interna al problema tecnologico dei robot morali, il coinvolgimento del sapere filosofico in posizione subordinata appare del tutto legittimo. La traslitterazione dei contenuti della filosofia morale nel linguaggio della scienza tecnologica, di cui conosciamo la logica, è anch'essa uno stadio necessario della collaborazione, giustificato dalle finalità e dalle esigenze del progetto. Da questa stessa impostazione metodologica deriva però anche il divieto di spiegare l'umano tramite il tecnologico.
Perdere di vista il confine che separa l'etica delle macchine dalla filosofia morale non è solo deleterio da un punto di vista filosofico, ma si ripercuote negativamente sulla stessa ME, in quanto distoglie l'attenzione dai suoi scopi reali e alimenta aspettative smisurate circa le sue possibilità. Il fine della ME non è né la riproduzione dell'agente morale umano, né tantomeno la scoperta di verità assolute in materia morale (della cui esistenza sarebbe più prudente dubitare), ma la costruzione di sistemi automatici che conformino in modo soddisfacente lo svolgimento della propria funzione a valori morali, quali che siano. Secondo un termine molto ben ideato, il vero obiettivo si può definire *functional morality*: la necessità che la tecnologia svolga funzioni nel rispetto di valori degni di essere affermati. L'ideale regolativo dell'agente artificiale perfetto, della copia tecnologica del soggetto morale umano, non è che un miraggio epistemologico. La missione dell'etica delle macchine è la costruzione di robot che siano in grado di funzionare in conformità a ciò che ci sta a cuore.
Per quanto possa essere suggestivo equiparare esseri umani e robot morali, immaginando macchine future capaci non solo di replicare l'esperienza morale umana, ma di perfezionarla, l'analisi del rapporto determinato dei saperi tecnologico e filosofico in ME mostra con chiarezza l'equivoco metodologico alla base di simili discorsi, cioè la confusione tra ipotesi di lavoro che valgono esclusivamente in un determinato ambito e tesi generali circa la natura del fenomeno morale umano. Mostrare l'inconsistenza della retorica che vede nella robotica morale la via per la costruzione di agenti morali artificiali tali e quali gli esseri umani non serve solo a richiamare l'attenzione su un uso scorretto del linguaggio e dell'argomentazione, ma anche—ed è assai più importante—a regolare le aspettative che è lecito nutrire nei confronti di un genere di tecnologie che potrebbe ben presto rivoluzionare la nostra quotidianità.

Riferimenti bibliografici

Allen, C., Varner, G., Zinser, J. (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence*, *12,* pp. 251-261.

Amigoni, F., Schiaffonati, V. (2005). Machine Ethics and Human Ethics: A Critical View. In *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*. Menlo Park: AAAI Press, pp. 103-104.

Anderson, S.L. (2011a). Machine Metaethics. In Anderson e Anderson (2011), pp. 21-27.

- (2011b). Philosophical Concerns with ME. In Anderson e Anderson (2011), pp. 162-167.
- (2011c). The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics. In Anderson e Anderson (2011), pp. 285-295.
- (2011d). How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better. In Anderson e Anderson (2011), pp. 524-530.

Anderson, M., e Anderson, S.L. (a cura di) (2011). *Machine Ethics.* Cambridge: Cambridge University Press.

Asaro, P. (2006). What Should We Want From a Robot Ethics? *IRIE – International Review of Information Ethics*, *6,* pp. 9-16.

- (2012). A Body to Kick, but Still no Soul to Damn: Legal Perspectives on Robotics. In Lin, Abney e Bekey (2012), pp. 169-186.

Asimov, I. (1982). *The Complete Robot*. New York: Doubleday.

Awad, E., Dsouza, S., Kim, R., Schultz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I. (2018). The Moral Machine Experiment. *Nature*, *563*, pp. 59-64.

Bekey, G.A. (2012). Current Trends in Robotics: Technology and Ethics. In Lin, Abney e Bekey (2012), pp. 18-34.

Bertolini, A. (2013). Robots as products: the case for a realistic analysis of robotic applications and liability rules. *Law, Innovation and Technology, 5(2)*, pp. 214-247.

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. https://nickbostrom.com/ethics/ai.html (ultima consultazione 8/11/2018)

Bryson, J.J., Diamantis, M., Grant, T.D. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*, pp. 273-291.

Bryson, J.J., Kime, P.K. (2011). Just an Artifact. Why Machines are Perceived as Moral Agents. In T. Walsh (a cura di), *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI 2011).* Menlo Park: AAAI Press.

Calo, R., Froomkin, A.M., Kerr, I. (a cura di) (2016). *Robot Law.* Chaltenham: Edward Elgar Publishing.

Carrozza, M. C. (2017). *I robot e noi.* Bologna: il Mulino.

Cingolani, R., Metta, G. (2015). *Umani e umanoidi. Vivere con i robot.* Bologna: il Mulino.

Clarke, R. (2011). Asimov's Laws of Robotics. Implications for Information Technology. In Anderson e Anderson (2011), pp. 254-284.

Coeckelbergh, M. (2012). Can We Trust Robots? *Ethics and Information Technology*, *14*, pp. 53-60.

Deng, B. (2015). Machine Ethics: the robot's dilemma. *Nature*, https://www.nature.com/news/machine-ethics-the-robot-s-dilemma-1.17881 (ultima consultazione 8/11/2018).

Dietrich, E. (2007). After Humans Are Gone. *Journal of Experimental and Theoretical Artificial Intelligence*, *19(1)*, pp. 55-67.

Dzindolet, M.T, Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58, pp. 697-718.

Fabris, A. (2018). La filosofia e lo specchio delle macchine. *InCircolo*, 6, pp. 28-38.

Floridi, L., Sanders, J.W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*, pp. 349-379.

Franklin, S., Graesser, A. (1996). Is it an agent, or just a program?: A taxonomy for autonomous agents. In J. Müller, M.J. Wooldridge, N.R. Jennings (a cura di), *Intelligent Agents III. Agent Theories, Architectures, and Languages.* Berlin-Heidelberg: Springer, pp. 21-35.

Gips, J. (1995). Towards the Ethical Robot. In G.K. Ford, C. Glymour, P.J. Hayes (a cura di), *Android Epistemology*. Cambridge: M.I.T. Press, pp. 243-252.

Gunkel, D.J. (2012). *The Machine Question. Critical Perspectives on AI, Robots and Ethics.* Cambridge: M.I.T. Press.

- (2018). *Robot Rights.* Cambridge: M.I.T. Press.

Hallevy, G. (2013). *When Robots kill. Artificial Intelligence under Criminal Law*. Boston: Northeastern University Press.

Himma, K.E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11(1),* pp. 19–29.

Johnson, D.G. (2011). Computer Systems. Moral Entities, but Not Moral Agents. In Anderson e Anderson (2011), pp. 168-183.

Kaplan, J. (2017). *Intelligenza artificiale. Guida al futuro prossimo.* Roma: Luiss University Press.

Lin, P., Abney, K., Bekey, G.A. (a cura di) (2012). *Robot Ethics. The Ethical and Social Implications of Robotics.* Cambridge: M.I.T. Press.

Lin, P., Abney, K., Jenkins, R. (a cura di) (2017). *Robot Ethics 2.0. From autonomous cars to Artificial Intelligence.* Oxford: Oxford University Press.

Marcus, G. (2014). Moral Machines. *The New Yorker*, https://www.newyorker.com/news/news-desk/moral-machines (ultima consultazione 8/11/2018).

Moor, J.H. (1995). Is Ethics Computable? *Metaphilosophy*, *26(1-2)*, pp. 1-21.

- (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *July/August 2006*, pp. 18-21.

Nadeau, J.E. (2006). Only androids can be ethical. In K.E. Ford, C. Glymour, P.J. Hayes, *Thinking about Android Epistemology.* Menlo Park: IAAA Press, pp. 241–248.

Nissenbaum, H. (2001). How Computer Systems Embody Values. *Computer*, *March 2001*, pp. 118-120.

Nourbakhsh, I. R. (2014). *Robot fra noi. Le creature intelligenti che stiamo per costruire.* Torino: Bollati Boringhieri.

Pagallo, U. (2013). *The Laws of Robots: Crimes, Contracts and Torts.* Dordrecht: Springer.

Parkin, S. (2017). Teaching Robots Right from Wrong. *The Economist 1843*, *June/July 2017*, pp. 68-73, https://www.1843magazine.com/features/teaching-robots-right-from-wrong (ultima consultazione 8/11/2018)

Sini, C. (2009). *L'uomo, la macchina, l'automa. Lavoro e conoscenza tra futuro prossimo e passato remoto.* Torino: Bollati Boringhieri.

Solaiman, S.M. (2017). Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial Intelligence and Law*, *25(2)*, pp. 155-179.

Sullins, J.P. (2011). When is a Robot a Moral Agent? In Anderson e Anderson (2011), pp. 151-161.

Taddeo, M. (2010). Modelling Trust in Artificial Agents, a First Step Toward the Analysis of E-Trust. *Minds and Machines, 20*, pp. 243-257.

Tavani, H.T. (2015). Levels of Trust in the Context of Machine Ethics. *Philosophy of Technology, 28(1)*, pp. 75-90.

Torrance, S. (2011). Machine Ethics and the Idea of a More-Than-Human Moral World. In Anderson e Anderson (2011), pp. 115-137.

Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of decision making. *Ethics and Information Technology*, *12(3)*, 243-250.

Wallach, W., Allen, C. (2009). *Moral Machines. Teaching Robots Right from Wrong.* Oxford: Oxford University Press.

Winfield, A.F.T., Blum, C., Liu, W. (2014). Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In M. Mistry, A. Leonardis, M. Witkowski, C.

Melhuish (a cura di), *Advances in Autonomous Robotics Systems*, Cham-Heidelberg-New York-Dordrecht: Springer, pp. 85-96.

Title:
Moral Machines? Philosophical Notes on Machines Ethics

Abstract:
The purpose of this essay is to determine the domain of validity of the notions developed in Machine Ethics [ME]. To this aim, I analyse the epistemological and methodological presuppositions that lie at the root of such technological project. On this basis, I then try and develop the theoretical means to identify and deconstruct improper applications of these notions to objects that do not belong to the same epistemic context, focusing in particular on the extent to which ME is supposed to feedback onto moral philosophy. By highlighting the inadequacy of many approaches to the supposed philosophical implications of ME, I wish to redirect attention to its actual scope and to stress its relevance for the social acceptance of autonomous technologies.
The essay is structured as follows. After a brief introduction (§1), in §2 I shed light upon the link between the current trend of robotic development toward greater degrees of autonomy and the corresponding need for artificial moral agents that fuels research in ME. I then present the epistemological profile of ME in §3, focusing on its main component, i.e., the modelling of human moral agency in the language of robotics and computer science. In §4 I deal with cases in which such model is brought to bear on human ethics and moral philosophy as well, whilst in §5 I develop a criticism of this extension based on an account of the actual epistemological relations that obtain between philosophical and technological knowledge in the context of the ME project. In §6 I discuss the role that the ordinary use of language plays in the process of extending machine-related notions to other domains of knowledge and, finally, I clarify what the appropriate scope of machine ethics is notwithstanding its many utopian or dystopian misinterpretations.

Keywords:
Machine Ethics, Moral Machines, Artificial Moral Agents