RIVISTA INTERNAZIONALE DI FILOSOFIA E PSICOLOGIA
DOI: 10.4453/rifp.2019.0006
ISSN 2039-4667; E-ISSN 2239-2629
Vol. 10 (2019), n. 1, pp. 78-93

TEMI ED EVENTI

Cervelli e carrelli: formalizzare la moralità nelle auto a guida autonoma

Eleonora Signorini[(α)]

Ricevuto: 30 novembre 2018; accettato: 11 aprile 2019

Riassunto L'indagine sui fondamenti neurali del giudizio morale è uno dei principali ed attuali temi di ricerca della Neuroscienza, il quale si intreccia inevitabilmente con tematiche relative all'Intelligenza Artificiale, al futuro dei trasporti e alla Filosofia della Mente. Gli esseri umani sono naturalmente dotati di un innato senso della morale, il quale è governato dalle intuizioni, ma sono anche provvisti di alcuni principi razionali. Il giudizio e il comportamento morale sono il risultato dell'integrazione fra le emozioni ("moralità veloce") e i processi razionali ("moralità lenta"), proprio come i processi cognitivi erano una combinazione di istinto ("pensiero veloce") e pura computazione ("pensiero lento"). Nella parte finale di questo lavoro, ho preso in considerazione i problemi di natura morale derivanti dall'introduzione dei *veicoli a guida autonoma*, i quali si presentano come un'applicazione diretta del *problema del carrello*: come dovrebbe essere programmato un veicolo per comportarsi nel caso di un incidente inevitabile, nel quale deve scegliere tra due mali?
PAROLE CHIAVE: Processi decisionali; Neuroscienza; Giudizi morali; Prospettiva comparata; Problema del carrello

Abstract *Brains and Trolleys: Formalizing Morality in Automated Driving Cars* - Inquiry into the neural bases of moral judgment is one of the current frontiers in neuroscientific research and is intertwined with issues in Artificial Intelligence, the future of transport, and Philosophy of Mind. Humans are naturally endowed with an innate moral sense, which is governed by intuition and informed by rational rules. Moral judgment and behavior are the byproduct of the integration of emotional (morality fast) and rational (morality slow) processes, just as cognitive processes are the result of a combination of emotional instinct (thinking fast) and computational rationality (thinking slow). In the latter part of this work, I also consider the ethical issues raised by the introduction of *Automated Driving Systems* by reexamining the *Trolley Problem*: how should an autonomous vehicle be programmed to behave in the event of an unavoidable accident, in which it has to choose between two harmful consequences?
KEYWORDS: Decision-making; Neuroscience; Moral Judgments; Comparative Perspective; Trolley Problem

IN QUESTO LAVORO SI INTENDE proporre una breve riflessione sui problemi aperti – di carattere tecnico e legislativo – riguardanti il funzionamento dei *veicoli a guida autonoma*. Negli anni, i veicoli *driverless* sono diventati sempre più efficienti e sicuri: la loro componente tecnologica è stata implementata e le loro prestazioni migliorate. E tuttavia persi-

[(α)]Dipartimento di Filosofia e Comunicazione, Università degli Studi di Bologna, via Azzo Gardino, 23 – 40122 Bologna (I)

E-mail: eleonora.signorini2@studio.unibo.it (✉)

stono ancor oggi alcuni problemi sia di natura tecnica (come, per esempio, la reazione che tali automi potrebbero avere a seguito di imprevisti) sia di carattere legislativo, che chiamano in causa decisioni etiche riguardo alla programmazione e alla fruibilità di tali veicoli. In queste pagine cercherò di mostrare come questi problemi siano una traduzione moderna di antichi dilemmi di natura etica, che per molto tempo hanno impegnato filosofi, psicologi e scienziati cognitivi. Queste riflessioni sui dilemmi morali toccheranno inevitabilmente il cosiddetto "*Problema del carrello*" nelle sue varianti e le principali indagini filosofiche e neuroscientifiche volte a comprendere come gli esseri umani implementano le loro scelte e i loro processi decisionali (sensazioni ed emozioni *vs* razionalità e computazione).

Il tentativo di programmare un comportamento morale negli automi *driverless* intreccia anche l'antico e ostico problema mente/corpo. La cognizione umana non è determinata unicamente dalla ragione e dalla computazione, ma è profondamente connessa con il corpo e con le sue sensazioni. La ragione inoltre non è l'unica legislatrice delle nostre decisioni in ambito morale. L'essere umano non è solo pura razionalità, ma è un organismo collocato in quella esperienza propriamente umana che è la vita: di conseguenza, l'esperienza, le intuizioni e le emozioni influenzano sia i processi cognitivi che quelli decisionali di carattere morale.

Esistono problemi morali irrisolvibili anche per l'essere umano, la cui moralità è caratterizzata da una componente razionale e una emotiva. Come potrebbero essere implementati e programmati aprioristicamente in un dispositivo automatico che, al massimo, potrebbe possedere (o simulare) solamente una componente, quella razionale, della moralità umana? Che tipo di moralità potrà essere implementata nei *veicoli a guida autonoma*? Che tipo di scelta vogliamo far compiere a questi veicoli? A quale tipo di etica dobbiamo dare la precedenza? Quale può essere il male minore o il bene maggiore a cui un automa deve obbedire? Come facciamo ad affidarci a questi veicoli, mettendo a repentaglio la nostra vita e quella dei nostri cari, se non sappiamo progettare a priori il loro comportamento?

Quando l'etica diventa un problema tecnico

Il progresso quantitativo e qualitativo degli ultimi decenni nel campo del *machine learning* e dei robot intelligenti ha posto l'umanità innanzi a nuovi interrogativi che spesso, come nel caso dei veicoli *driverless*, rievocavano lo spettro di antichi dilemmi. È opportuno precisare che la realtà fisica, specialmente quella di contesti complessi come la guida su strada, non è quasi mai schematizzabile nelle singole opzioni presentate generalmente dai dilemmi morali. Come vedremo, il *Problema del carrello* coinvolge unicamente due binari, i quali rappresentano un artefatto che semplifica molto la realtà della guida su strada, in cui esiste una pluralità e una continuità di alternative possibili.

Gli autonomi intelligenti, come definito dalla *Risoluzione* adottata nel 2017 dal Parlamento Europeo,[1] sono dotati della capacità di acquisire autonomia grazie ai sensori e allo scambio di dati con il proprio ambiente (interconnettività). Proprio come suggeriva Alan Turing tali robot autonomi possiedono capacità di apprendimento attraverso l'esperienza e l'interazione con il mondo esterno e sono inoltre dotati della capacità di adeguare il proprio comportamento e le proprie azioni all'ambiente. Tali automi che interagiscono con l'uomo e agiscono nella realtà esterna si troveranno quindi inevitabilmente di fronte a una serie di decisioni morali che imporranno al veicolo a guida autonoma la necessità di deliberare fra più alternative. L'introduzione di tali automi comporterà quindi una molteplicità di decisioni, che non saranno più solamente umane, ma proprie anche dei decisori autonomi che popoleranno i nostri ambienti naturali e i contesti virtuali.

Queste reificazioni di una macchina razionale e volontaria, che imiti quanto più

possibile una persona umana e che sia in grado di interagire con essa, non solo ripropongono questioni filosofiche relative ai dilemmi morali, ma introducono alcune questioni di *roboetica*,[2] connesse con la natura, la formalizzazione e i limiti della capacità decisionale dei sistemi di intelligenza artificiale. Con l'evolvere della tecnologia, i veicoli a guida autonoma potrebbero essere costretti a operare una scelta morale riguardo l'uccisione di una persona per salvarne un numero maggiore, oppure a dover scegliere se investire un bambino o una coppia di anziani, un motociclista con il casco o uno privo di protezione. Tali macchine dovranno quindi essere dotate o di una moralità di tipo operazionale (coincidente con le condotte morali stabilite aprioristicamente dal progettista), oppure funzionale (relativa alla capacità di valutare e rispondere alle sfide morali che si presentano mediante una sorta di esperienza pregressa), entrambi coerenti con i presupposti di autonomia e sensibilità etica. Non è chiaro tuttavia il prototipo di morale che dovrà essere inserito in un'agentività artificiale completa, dal momento che i filosofi morali si interrogano su situazioni analoghe (connesse ai dilemmi morali) da decenni, senza individuare risposte ultime e conclusive. Nonostante tale *vexata quaestio*, ai produttori di veicoli a guida autonoma spetta tuttavia il gravoso compito di definire gli algoritmi "morali" che guideranno tali veicoli in situazioni di danno inevitabile.

Una prima fase di questo imminente futuro della circolazione automobilistica (della cui evoluzione siamo testimoni[3]) impone la presenza del conducente, designato a prendere il controllo del veicolo qualora necessario, mentre in una seconda fase i veicoli saranno completamente *driverless* e interamente governati dalla tecnologia. È opportuno esplicitare come le tecnologie *driverless* siano state progettate per ridurre drasticamente il numero di decessi provocati da incidenti stradali,[4] per favorire la mobilità per individui disabili (fisicamente o cognitivamente), i quali non possono ottenere o mantenere l'idoneità alla guida, e infine per ridurre le emissioni inquinanti evitando numerose manovre inutili (come accelerate improvvise o inchiodate dettate da disattenzione). Benché la guida autonoma sia stata programmata per aumentare la sicurezza della circolazione, non potrà tuttavia eliminare totalmente gli incidenti stradali, soprattutto quelli imprevisti e causati da un fattore inaspettato (come un bambino che attraversa correndo la strada, sbucando all'improvviso da una posizione nascosta), il quale induce il veicolo a dover scegliere immediatamente fra due mali distinti e inevitabili.

È interessante notare come una situazione del genere sia concettualmente identica a quella che può verificarsi con conducenti umani; tuttavia sembra che la mentalità collettiva tolleri tranquillamente i numerosi e inesorabili errori umani, ma non quelli più sporadici e altrettanto inevitabili della tecnologia moderna. Un dilemma morale potrebbe presentarsi nella situazione generata da un attraversamento imprevisto e illecito: supponiamo di essere in auto quando improvvisamente alcuni pedoni, per esempio cinque, attraversano la carreggiata.

Le tre opzioni che si presentano sono: investire i pedoni che attraversano illecitamente fuori dalle opportune strisce (attenendoci rigorosamente alle norme della circolazione stradale); invadere la corsia opposta, collidendo con il veicolo che procede nel senso opposto, al cui interno vi è solo il conducente (rispondendo del bene maggiore), o, infine, deviare verso il marciapiede e investire un ignaro pedone che sta passeggiando (ubbidendo al medesimo principio). In un'ipotetica e analoga situazione, il comportamento del conducente umano, dettato dall'istinto, sarà generalmente tollerato dall'opinione comune e sollevato dalla responsabilità da un punto di vista giuridico, in virtù della presenza di una circostanza sopravvenuta e imprevista, sufficiente (secondo l'articolo 2054 del Codice della Strada)[5] ad abolire ogni nesso causale tra la condotta del conducente e il decesso dei soggetti coinvolti. Diversamente, il *veicolo a guida autonoma*, non disponendo indubbiamente di

istinto e sensibilità, dovrà necessariamente agire secondo un algoritmo. Di conseguenza il dilemma etico, legato a tali veicoli, sorge relativamente alla decisione preventiva del profilo etico, ovvero del comportamento e delle scelte che il *veicolo a guida autonoma* dovrà attuare in circostanze simili. Ma è possibile individuare e formalizzare un principio universale di scelta, in modo da programmarlo nei *veicoli a guida autonoma*?

Sebbene un automa (come, per esempio, i *veicoli a guida autonoma*) non possieda una coscienza morale, esso si troverà probabilmente a dover compiere delle azioni, le quali, se fossero compiute da un essere umano, sarebbero giudicate moralmente, proprio come, secondo Turing, i calcolatori digitali potevano svolgere compiti che sarebbero stati considerati intelligenti se eseguiti da un essere umano. È sensato ipotizzare che l'essere umano sia dotato di innati principi razionali che agiscono nei processi decisionali e morali. Proprio la componente razionale della giustificazione e dell'azione morale sembrerebbe garantire l'imparzialità, l'universalità e la legittimità di un comportamento. Il carattere razionale e computazionale dei processi che guidano le scelte etiche e le risposte morali degli esseri umani potrebbe essere in linea di principio formalizzabile e riducibile a determinati algoritmi decisionali, proprio come, secondo Alan Turing, il pensiero computazionale umano poteva essere riproducibile meccanicamente.

Questo porta a credere che i *veicoli a guida autonoma* possano essere programmati per agire moralmente, secondo determinati algoritmi "morali". Se il comportamento morale fosse regolato unicamente dalla ragione, allora esso potrebbe, in linea di principio, essere riducibile e formalizzabile in opportuni schemi comuni di origine algoritmica inseribili nel programma, che farebbero del dispositivo elettronico dotato di *software* un agente morale. Eppure, la moralità non è riducibile alla mera razionalità, ma comprende anche una componente emotiva ed esperienziale, patrimonio esclusivo degli esseri viventi. Di contro all'universalità dei processi decisionali razionali, quelli emotivi risultano parziali, inconsci ed effimeri; tuttavia sono il frutto dell'evoluzione e il risultato di un processo automatico e inconscio che permette agli esseri umani di agire immediatamente, senza dover riflettere. La componente emotiva dei processi decisionali umani è quindi fondamentale nei casi in cui la gestione del tempo è essenziale, ovvero nelle eventualità impreviste in cui è necessario agire in fretta. La moralità è una peculiarità umana e deve essere considerata nella molteplicità che contraddistingue anche l'intelligenza umana: quindi il tentativo di riprodurne una singola parte attraverso modelli computazionali, anche se molto evoluti, è destinato a fallire.

In ogni caso, l'automa agirà secondo un determinato algoritmo, il quale potrà essere programmato per rispettare un criterio puramente deontologico (per esempio, rispettando rigorosamente le regole della circolazione stradale), oppure sarà in grado di eseguire una valutazione potenziale e utilitaristica sul numero delle vittime (per esempio, sacrificando la vita di un innocente per salvarne altre) o infine potrà essere progettato seguendo la maggioranza delle decisioni di soggetti umani intervistati (magari degli utenti che eseguono il test online della *Moral Machine* ideato dal MIT).[6] Questa ultima possibilità, nonostante sia in contraddizione con il presupposto di sopperire all'inabilità degli uomini di scegliere in modo efficiente in situazioni critiche, necessita di un sondaggio riguardo al comportamento del veicolo che gli individui umani preferirebbero. *Prima facie* sembrerebbe la soluzione migliore, ma una ricerca condotta dallo psicologo cognitivo Jean-François Bonnefon[7] rileva come gli individui preferirebbero un veicolo utilitaristico, purché essi non ne siano i passeggeri. Questa situazione potrebbe generare un dilemma sociale, poiché la maggioranza delle persone desidererebbe vivere in una società in cui l'uso dei veicoli comporti meno incidenti possibili (e quindi meno vittime), tuttavia agendo tutti secondo il proprio egoisti-

co interesse, rischieremmo di creare condizioni meno sicure per tutti. Un ulteriore problema con i veicoli "utilitaristici" è che non attribuirebbero priorità ai loro passeggeri, sacrificandoli se necessario; questo principio, ovviamente, non incrementerebbe la vendita di tali mezzi.

Una prima possibilità per realizzare computer "morali" sembra consistere nell'adottare una legge etica specifica (indifferentemente deontologica o utilitaristica), formalizzarla in un programma e creare un automa che segua rigorosamente tale algoritmo; la difficoltà risiede tuttavia nel decidere quale sia la regola etica appropriata. La moralità umana infatti non sembra dar mostra di modelli generali universali o assiomi aprioristici (siano essi deontologici o consequenzialisti) che possano essere concepiti come strumenti esegetici delle azioni umane. Ogni legge morale, anche quella apparentemente più semplice come "non uccidere", può comportare un numero notevole di eccezioni e controesempi. Non è possibile infatti postulare aprioristicamente l'esistenza di un assioma morale indiscutibile e sempre vero, tale da essere formalizzato e inserito all'interno di un algoritmo: per esempio, il più chiaro ed evidente principio utilitarista secondo il quale «la morte di una persona rappresenta una tragedia minore di quella di cinque, indipendentemente da come sia stata causata»[8] non rappresenta una legge universale dal momento che può essere legittimamente discussa dalla ragione sulla base di pratici controesempi. Sebbene l'azione di sacrificare un uomo per salvarne cinque possa essere necessaria per massimizzare l'utilità, essa non è comunque sufficiente a giustificare tale azione. Un controesempio che mostra l'impossibilità di un principio universale utilitarista è il noto *dilemma del trapianto di organi*: cinque pazienti di un ospedale sono gravemente malati e necessitano urgentemente di un trapianto di reni, di polmoni e di cuore; il giorno stesso arriva in ospedale, per sottoporsi ad alcune analisi, un giovane uomo sano, i cui organi sarebbero compatibili con quelli necessari ai cinque pazienti. Se adottiamo come principio universale e indiscutibile l'assioma utilitarista, di conseguenza il chirurgo dovrebbe uccidere l'uomo sano al fine di impiantare i suoi organi nei cinque pazienti a rischio, andando contro ogni deontologia.

Patrick Lin[9] afferma che l'etica potrebbe non essere in sé internamente coerente (ovvero potrebbe non soddisfare il criterio di coerenza logica), il che renderebbe impossibile la sua riduzione e formalizzazione in programmi "morali" per i *software* digitali. Tale incoerenza logica della morale potrebbe anche essere giustificata da istinti egoistici tipicamente umani, dal momento che le nostre scelte morali potrebbero subire una decisiva modifica se la persona da sacrificare, per l'utilità maggiore, fosse un nostro caro: in questo caso la coerenza e la fedeltà al principio utilitarista sarebbe messa da parte. Il *dilemma del bambino che piange* rappresenta un caso prototipico del ruolo degli istinti egoistici nei processi decisionali morali: immaginiamo che dei soldati nemici abbiano invaso la nostra città per raderla al suolo e uccidere tutti i civili superstiti. Un gruppo di concittadini (di cui fai parte) si è nascosto in una cantina di un'abitazione che i soldati stanno in quel momento saccheggiando. Improvvisamente il tuo bambino comincia a piangere, quindi, per non farvi scoprire dai soldati, istintivamente cerchi di chiudergli la bocca con la mano. Se rimuovi la mano dalla sua bocca, il suo pianto richiamerà l'attenzione dei soldati e causerà la morte di tutto il gruppo nascosto nella cantina; se invece vuoi salvare il gruppo, sei costretto a soffocare il tuo bambino. Le nostre sensazioni ci renderanno inclini a non accettare il sacrificio della persona a noi più cara, mentre la nostra razionalità ci spingerà ad accettare la risposta utilitaristica del sacrificio di un'unica persona al fine di salvare un gruppo intero.

Dal momento che il fine ultimo è quello di implementare nella macchina un comportamento morale simile a quello umano, sembra che l'alternativa sia quella di valutare le singole circostanze piuttosto che limitarsi alla rigida applicazione di un assioma formale

universale. L'algoritmo "morale" dovrebbe probabilmente effettuare una valutazione delle possibili conseguenze, che potrebbero includere la perdita di vite umane o un eventuale danno a delle proprietà. Per esempio, il veicolo automatico dovrebbe confrontare l'opzione A (colpire cinque pedoni) che non comporterebbe alcun danno a proprietà e l'opzione B (colpire un muro e ferire forse mortalmente quattro passeggeri) che causerebbe un ingente danno a un monumento socialmente apprezzato (come un antico muro romano). Se l'algoritmo è programmato per valutare i danni ai beni (compresi i beni culturali e le risorse ambientali) in termini monetari e in seguito confrontarli con un'analoga stima monetaria equivalente alla perdita di una vita umana (usando, per esempio, la formula presentata nel *Value of a Statistical Life*),[10] allora molto probabilmente sarà indotto a scegliere l'opzione B, nonostante preveda la morte di un numero maggiore di persone. Nell'eventualità in cui l'algoritmo conferisca priorità alla vita umana e sia programmato per ridurre al minimo l'impatto sugli esseri umani, il veicolo potrebbe sacrificare beni culturali inestimabili al fine di salvare un essere umano. Tale alternativa appare paradossale, soprattutto nel caso di danno ambientale, in cui l'algoritmo dovrebbe inverosimilmente valutare tutte le possibili conseguenze dannose che succederanno all'azione e che potrebbero causare un danno maggiore per la salute e la vita di un numero più elevato di persone (per esempio, l'inquinamento di un fiume o l'esplosione di una fabbrica).

Un ultimo modo possibile per concretare automi morali potrebbe essere quello di creare un automa dotato di un apprendimento automatico (*machine learning*) e istruirlo a rispondere alle varie situazioni reali, al fine di ottenere un comportamento morale. Già nel 1950 Alan Turing sosteneva che invece di sforzarsi di realizzare un programma che simulasse la mente di un adulto, era più proficuo cercare di realizzarne uno che simulasse la mente di un bambino, in modo che esso potesse essere educato e istruito opportunamente. La geniale intuizione di Turing ha dato l'avvio a un'area di ricerca innovativa e interdisciplinare per lo studio degli automi artificiali *embodied*,[11] che prende ispirazione proprio dai meccanismi di sviluppo dei bambini, i quali sono capaci di acquisire competenze comportamentali, morali, cognitive, linguistiche e comunicative, attraverso una forma di apprendimento individuale e sociale (determinata da ambienti come la famiglia o la scuola). L'algoritmo di apprendimento della condotta, teorizzato già da Turing per addestrare la macchina a pensare, si rivela simile al modo in cui gli esseri umani imparano ad agire moralmente.

Lo scienziato cognitivo Gerd Gigerenzer sostiene come in numerose circostanze della vita quotidiana, gli esseri umani prendono decisioni sulla base di intuizioni o stratagemmi morali che hanno radici culturali o inerenti alla loro storia evolutiva. Per esempio, se un bambino viene rimproverato sin dall'infanzia per i suoi piccoli furti, allora esso proverà una sensazione sgradevole ogni volta che penserà all'azione di rubare, la quale verrà conseguentemente etichettata come negativa e immorale. Giulio Tononi, nel suo saggio *Un viaggio dal cervello all'anima*, ha invece proposto un'interessante teoria che afferma la dipendenza fra l'esperienza e l'integrazione dell'informazione. Secondo il neuroscienziato, la funzione primaria del cervello umano è l'integrazione di informazioni, un processo che, in linea di principio, può compiersi indifferentemente negli esseri umani o negli artefatti tecnologici. Tononi fornisce inoltre una formalizzazione rigorosa del calcolo dell'integrazione dell'informazione, apparentemente meccanizzabile, che gli permette di correggere il noto "*test di Turing*": un *software* artificiale deve dimostrarsi capace di emulare l'essere umano non riguardo alle sue capacità linguistiche e conversazionali (come avviene nel *test di Turing*), ma proprio nella capacità di integrare informazioni.

Il problema di definire un algoritmo morale nei *veicoli a guida autonoma* è molto complesso, dal momento che coinvolge tematiche e situazioni difficili anche per l'essere umano, l'unico

che detiene il primato (fino a questo momento) della moralità. Come facciamo a programmare un automa in modo che agisca moralmente in situazioni difficili, se nemmeno noi esseri umani siamo capaci di tanto?

Carrelli mentali

Il dilemma morale noto come *Problema del carrello* mette in scena quel "cortocircuito concettuale"[12] scaturito dalla compresenza e dalla indecidibilità di due alternative contrapposte, che generano un conflitto interno alla mente: grazie a tale esperimento mentale è possibile individuare un collegamento tra il piano puramente astratto dell'indagine filosofico-morale e quello meramente concreto della progettazione di *veicoli a guida autonoma*. Il primo dilemma morale venne presentato dalla filosofa Philippa Foot,[13] nella forma di un esperimento mentale noto come il *dilemma dello scambio*: immaginiamo di essere un passeggero di un vagone ferroviario[14] fuori controllo. Il conducente è svenuto e la locomotiva si sta dirigendo verso cinque persone ignare del pericolo che stanno passeggiando sui binari; queste ignare persone non riusciranno ad abbandonare i binari in tempo, poiché le banchine sono troppo scoscese. Tuttavia, il percorso dei binari presenta una deviazione a sinistra e quindi, azionando una leva, è possibile deviare la folle corsa della locomotiva; tuttavia, nel tracciato di sinistra, si trova un'altra persona che morirebbe inevitabilmente se azionassimo la leva dello scambio. A questo punto possiamo azionare la leva e deviare la locomotiva, uccidendo una persona ma salvandone comunque cinque; oppure, in alternativa, possiamo astenerci dal manovrare lo scambio, lasciando che la locomotiva investa le cinque persone.

La quasi totalità degli individui intervistati[15] risponde, con assoluta sicurezza, che sia moralmente giusto ed eticamente doveroso manovrare lo scambio per deviare il treno verso il binario secondario, al fine di salvare la vita dei cinque sfortunati. In questo caso, la morte della persona sul binario secondario rappresenta un accadimento collaterale, indesiderato (perché l'intenzione della nostra azione non era quella di provocare la morte di un uomo), ma comunque prevedibile e pronosticabile. Nell'esperimento mentale descritto da Philippa Foot, il soggetto esegue un'azione, quella di spingere una leva e, a seguito di questa azione e indipendente dalla sua volontà, un uomo viene ucciso: tale corso degli eventi non è stato determinato dal soggetto che ha azionato la leva ma è stato causato indirettamente e inconsapevolmente.

Judith Jarvis Thomson, sulla scia dell'esperimento mentale di Philippa Foot, introduce una variante del dilemma dello scambio, nota come *dilemma dell'uomo grasso*: immaginiamo di trovarci su un cavalcavia pedonale sovrastante un binario su cui viaggia un carrello fuori controllo.[16] Sul percorso del vagone ci sono cinque persone e, ancora una volta, le banchine sono troppo ripide perché essi possano abbandonare i binari in tempo. L'unico modo per fermare la locomotiva fuori controllo è lanciare sui binari un grosso peso, ma l'unico oggetto disponibile e di peso sufficiente è una persona molto robusta che si trova anch'essa sul cavalcavia: il peso dell'uomo sarebbe sufficiente a bloccare il vagone. Possiamo quindi spingere quest'uomo, facendolo cadere sul percorso della locomotiva e ucciderlo, oppure possiamo astenerci dal farlo, lasciando che muoiano cinque persone.

È interessante notare come i due casi presentano un bilancio costi/benefici identico (ovvero uccidere una persona per salvarne cinque) tuttavia, solo una minima percentuale degli individui[17] spingerebbe personalmente dal ponte l'uomo corpulento. Sembra infatti che la maggioranza degli intervistati consideri il primo scenario (il *dilemma dello scambio*) moralmente giusto, mentre il secondo scenario (il *dilemma dell'uomo grasso*) non eticamente sostenibile. Ciò nonostante, l'unica differenza che sussiste fra i due esperimenti mentali consiste nel fatto che nel primo caso la morte di un uomo è una conseguenza prevista, ma non desiderata, consciamente, mentre nel secondo caso la morte dell'uomo è una conseguenza intenzionalmente voluta. L'abbaglio morale, ovvero l'incoerenza

dei giudizi dei soggetti intervistati nei due diversi esperimenti mentali, è prova del fatto che occasionalmente le emozioni e i sentimenti tempestivi subentrano (e si sostituiscono) alla ragione, che altrimenti avrebbe indotto gli interrogati a considerare le due azioni moralmente equivalenti. Nel momento in cui riteniamo aprioristicamente che arrecare un danno a una persona sia moralmente sbagliato ed eticamente inaccettabile (indipendentemente dalle conseguenze dell'azione), effettuiamo una valutazione intuitiva, affidandoci ai giudizi deontologici. Diversamente, se giudichiamo a posteriori che l'aggravio provocato a una persona possa essere ammissibile in vista delle conseguenza positive che ne deriverebbero, ci affidiamo ai giudizi utilitaristici. La dottrina deontologica e quella utilitarista si presentano quindi come i corrispettivi filosofici di due distinte facoltà di decisione: le emozioni e le intuizioni da un lato, il calcolo razionale dall'altro.

C'è una sottile linea rossa che congiunge i due casi prototipici del *Problema del carrello* con la classica bipartizione etica fra utilitarismo e deontologia. Il giudizio utilitarista legittima il sacrificio dell'uomo corpulento come conseguenza logica e realizzazione pratica del principio del bene superiore, il quale induce a massimizzare il numero di vite salvate: le decisioni utilitariste vengono dedotte da una valutazione razionale delle conseguenze dell'azione morale, in base alla quale è lecito agire per il fine di un bene più grande.[18] Il giudizio deontologico invece ricusa tale azione, giudicandola come un atto immorale che viola i diritti e i doveri di ogni essere umano: le decisioni deontologiche si presentano come appannaggio esclusivo delle emozioni umane e delle intuizioni, fra cui la credenza che arrecare un danno intenzionalmente a un'altra persona sia moralmente sbagliato, indipendentemente dalle possibili conseguenze. Arrecare morte certa a un uomo, spingendolo in prima persona da un cavalcavia, comporta un impatto morale superiore rispetto all'azionare una leva che indirettamente sopprimerà un uomo.

Di conseguenza, il dilemma dell'uomo corpulento richiederà un maggior coinvolgimento emotivo, in quanto l'azione è *personale*[19], mentre il dilemma dello scambio comporterà un distacco emotivo, dal momento che l'azione è *impersonale*[20] e dettata una decisione automatica e razionale. In conclusione è opportuno quindi postulare come nelle decisioni utilitaristiche ci appelliamo a quella porzione di moralità umana deliberata e razionale, che necessita di uno sforzo cognitivo; inversamente, quando deliberiamo in maniera deontologica, ricorriamo alle valutazioni intuitive, all esperienze e alle emozioni.

La componente lenta e veloce dell'attività cognitiva e morale umana

L'uomo, come ci tramanda Omero, è dotato di una "mente colorata",[21] costituita dalla multiformità della ragione e su plurime dimensioni delle emozioni, oltre che di una moralità altrettanto variopinta e poliedrica. È curioso notare come l'uomo abbia da sempre ascritto alla morale una natura astratta e intangibile, incorporea e utopica, la stessa che riservava all'attività cognitiva.

Su questo tema i più importanti pensatori morali della tradizione filosofica, da Tommaso d'Aquino a Kant, da Hume a Bentham, si sono a lungo confrontati proponendo soluzioni diverse che sarebbe pleonastico affrontare in questa sede. È interessante invece considerare i recenti risultati ottenuti in ambito neuroscientifico, i quali hanno evidenziato come la moralità scaturisca (proprio come la cognizione) da una complessa interazione tra emozione e razionalità, tra istinto e computazione, rilevando quindi la sua dipendenza costitutiva dalla morfologia celebrale e neurale degli esseri umani.[22]

I meccanismi neuronali e le proprietà biologiche di particolari regioni del sistema nervoso umano hanno quindi contribuito a plasmare il dover essere degli individui, ovvero il loro comportamento morale. Queste numerose conquiste neuroscientifiche chiariscono e modificano la caratterizzazione dei giudizi morali e in particolare dei possibili fattori che contribuiscono a delinearne ambiti e portata,

ricadute utilitaristiche e deontologiche, e infine collegamenti con gli aspetti razionali ed emozionali della cognizione umana.[23]

È sensato ipotizzare che vi sia una notevole parte della cognizione umana che possa essere definita computazionale, ovvero riconducibile al modello generale della *macchina di Turing*. Tali fenomeni cognitivi computazionali umani comprendono la manipolazione e la trasformazione di simboli, lo svolgimento di un algoritmo più o meno esplicito e l'esecuzione di una procedura meccanica di calcolo. Così come Alan Turing (e più in generale i sostenitori della *Tesi computazionale della mente*)[24] sosteneva che tutto il pensiero umano fosse riconducibile ad attività computazionali e razionali, allo stesso modo gli psicologi Jean Piaget e Lawrence Kohlberg (e più in generale i sostenitori del *modello razionalista*) attribuivano alla ragione un ruolo fondamentale nei processi decisionali in contesti morali. Il modello cognitivo dello sviluppo morale di Kohlberg,[25] estensione di quello piagetiano, attribuisce quindi centralità nei processi morali alla deliberazione razionale e consapevole, trascurando le componenti più emotive e intuitive coinvolte nei processi decisionali. Nel secolo passato, l'indagine psicologica sulla moralità ha riconosciuto il primato del *modello razionalista*, collocandosi nella tradizione del pensiero di stampo kantiano. Kant sostiene infatti come solamente a partire dalla ragione si possa derivare l'*imperativo categorico*, una necessità morale assoluta e universale. L'indagine filosofica sulla morale, almeno fino alla modernità, ha quindi enfatizzato ed esaltato la componente della ragione nelle decisioni e nelle azioni in contesto morale e, di contro, ha sempre demonizzato e screditato la controparte emotiva: se la ragione consacrava l'uomo alle idee più alte e nobili, l'emozione gli rammentava costantemente la sua bruta animalità; se la ragione costituiva la massima esternazione dell'anima umana, l'emozione non era che la mera manifestazione bestiale del corpo; se la ragione promuoveva e valorizzava la nostra eccelsa vita morale, l'emozione era qualcosa di imbarazzante e impediva il suo corretto darsi.

La duratura e consolidata tradizione razionale e logicista ha promosso costantemente il primato del pensiero computazionale sulla sensibilità; eppure, intorno agli anni '80 e '90 del secolo passato, la controparte emotiva e istintiva irrompe in quella follia computazionale che sembrava dilagare in ogni settore della vita quotidiana (dall'intero universo al nostro cervello, dall'andamento dell'economia alla struttura interna del nostro DNA). Grazie soprattutto ai lavori dei filosofi Paul e Patricia Churchland,[26] John Searle[27] e Robert French[28] appare evidente come i fenomeni cognitivi computazionali non esauriscano affatto il campo, ben più vasto, di tutte le attività cognitive umane, in quanto la mente potrebbe essere composta da una componente non computazionale in senso stretto. In proposito Daniel Kahneman distingue due tipologie di attività mentale umana: il *pensiero veloce*, che comprende pensiero intuitivo, percezione e memoria, e il *pensiero lento*, più riflessivo e di natura computazionale.[29] Riportando la metafora delle forbici, la parte della cognizione umana emotiva e istintiva (*pensiero veloce*) e la parte razionale (*pensiero lento*) rappresentano le due lame delle forbici, necessarie per l'elaborazione del pensiero; le forbici tagliano solo quando le due lame operano congiuntamente, quindi, fuor di metafora, entrambe le tipologie di capacità intellettive sono indispensabili per comprendere il funzionamento del pensiero umano. Il cervello umano si scopre così essere caratterizzato dalla compresenza di più sistemi e da una complessa iterazione tra piani e strutture complementari.

Analogamente, sul finire del secolo scorso, l'approccio neuroscientifico e le nuove tecniche di ricerca sperimentale hanno portato un contributo decisivo nell'indagine relativa alle scelte morali, riconoscendo alle emozioni e alle intuizioni un ruolo significativo nei giudizi morali. Le emozioni rappresentano la realtà costitutiva dell'essere umano, il cui pensiero e comportamento morale non sembra più dipendere della ragione esplicita o delle fredde e lente computazioni mentali.[30] Tuttavia è opportuno specificare che le moderne indagini neurologiche non sono sufficienti a spiegare completamente gli aspetti

etici e morali dei processi decisionali umani, sebbene abbiano contribuito a negare l'dea che le nostre decisioni in ambito morale dipendano unicamente da fattori razionali e cognitivi. Questi contributi non hanno fatto altro che confermare, in via sperimentale, ciò che molti filosofi, come per esempio Martha Nussbaum, affermavano già da tempo: «le emozioni [...] non possono esser messe da parte facilmente nelle spiegazioni del giudizio etico, come tanto spesso è accaduto nella storia della filosofia. [...] le emozioni dovrebbero essere parte fondamentale dell'oggetto della filosofia morale [...] non possiamo ignorarle, come tanto spesso ha fatto la filosofia morale».[31]

Il nuovo secolo ha assistito a un notevole incremento dell'interesse sperimentale da parte di numerose discipline (come la psicologia sociale, le scienze cognitive, l'economia e la neuroscienza), riguardo al rapporto tra emozione e ragione in contesti morali. È stato quindi necessario, prima di poter comprendere la connessione nell'uomo tra queste istanze, assumere una concettualizzazione rigorosa (o comunque non ambigua) della nozione di emozione.

Un'interessante caratterizzazione delle emozioni umane è stata proposta dal neurologo Antonio Damasio,[32] il quale le ha definite come variazioni negli stati mentali del cervello e della costituzione fisica dell'essere umano nel suo complesso, innescate non solo da oggetti o eventi, ma semplicemente dal ricordo di essi. La controparte emotiva e istintiva della cognizione umana diventa in Damasio un'imprescindibile alleata della controparte razionale, necessariamente utile nello sviluppo del processo decisionale e nella formazione delle scelte morali. Analogamente, il *modello socio-intuizionista*, proposto dall'antropologo e psicologo cognitivo Jonathan Haidt, rileva come le capacità emotive degli individui (che includono le emozioni e i giudizi intuitivi) siano determinanti nella scelta morale, dal momento che le nostre decisioni etiche non hanno origine dalle attività lente, computazionali e *a priori* della ragione, ma da intuizioni lampo (*flash*).

I giudizi morali sono quindi deliberazioni che afferriamo con immediatezza, in base a processi automatici, impliciti e rapidi, che affiorano a posteriori con un'emozione e che risultano indipendenti e ingovernabili dalla ragione. Questi contributi di Damasio e di Haidt di fatto si inscrivono nella tradizione inaugurata dal filosofo David Hume, il quale riteneva come la sola ragione fosse impotente nell'orientare le azioni umane. Hume dichiara infatti che l'azione morale non si fonda a partire da una determinazione razionale finalizzata a raggiungere uno scopo (morale), ma si sviluppa unicamente da un'impressione istintiva, che ci induce a produrre movimenti del corpo e pensieri della mente.

Ritengo che sia nel campo della formulazione dei pensieri che nel campo del ragionamento morale sia necessario riconoscere una compartecipazione fra un elemento razionale e computazionale e uno emotivo e istintivo; quindi, sebbene sia sempre pericoloso ed azzardato attribuire etichette alla moralità, si potrebbe distinguere, sul modello kahnemaniano, una *moralità veloce* (indotta dalle emozioni e dall'istinto) e una *moralità lenta* (derivata dall'intelligenza razionale e computazionale). Joshua Greene, nel suo interessante volume *Moral Tribes*,[33] sembra implicitamente legittimare questa possibile distinzione mediante la teoria del giudizio morale *a doppio processo*, secondo la quale diverse regioni del cervello umano controllano le diverse risposte ai dilemmi morali: una inconscia, semiautomatica ed emotiva (*moralità veloce*), e l'altra cosciente, deliberata e razionale (*moralità lenta*). Greene afferma quindi come le neuroscienze siano riuscite a distinguere due diverse componenti della moralità, innescate da distinte reti cerebrali: una moralità cosciente, fondata sulla ragione e sul pensiero critico, che definisce *modalità manuale;* una *moralità automatica* basata sull'emozione e sull'intuizione (che, per esempio, ci proibisce di uccidere una persona, qualunque siano le conseguenze), che definisce automatica perché veloce, inconscia ed irrefrenabile. Nutrita nel corso dell'evoluzione umana, questa moralità *au-*

tomatica e *veloce*, sebbene ci conduca frequentemente alla decisione giusta, presenta un difetto terribile che consiste nella sua miopia, ovvero nell'incapacità di prevedere le conseguenze future delle nostre azioni.

Anche secondo lo psicologo Marc Hauser noi possediamo due modalità morali, dal momento che il nostro cervello è in grado di produrre giudizi razionali; inoltre l'essere umano è dotato di regole morali inconsce, intuitive e innate.[34] La *moralità lenta* potrebbe quindi identificarsi con l'insieme di principi razionali e di computazioni, indispensabile al fine di plasmare giudizi in merito a ciò che è oggettivamente giusto o sbagliato; invece la *moralità veloce* è stimolata e attivata dal sistema emotivo. L'atto morale nasce quindi da una tensione fra coscienza razionale e sfera emotiva, fra norma (il risultato universale della coscienza giudicante) e valore (che il cuore percepisce come tale), in assenza della quale non si supererebbe quella falsa dialettica tra l'universalismo formale di Kant e l'emotivismo di Hume.

È interessante notare come, non solo nella deliberazione dei *pensieri lenti* e dei *pensieri veloci*, ma anche nella formulazione della *moralità lenta* e della *moralità veloce* (quindi nella predilezione di decisioni utilitaristiche e deontologiche) entrino in funzione sezioni separate del cervello umano. Il neurologo Antonio Damasio aveva definito le emozioni come reazioni di tipo chimico e neurale, biologicamente determinate e prodotte da un sistema preciso e definito di regioni subcorticali come l'amigdala, la corteccia prefrontale, l'ippocampo, il sistema limbico.[35]

I progressi delle neuroscienze hanno reso possibile, mediante la PET (tomografia a emissione di positroni), la fMRI (risonanza magnetica funzionale) e le tecniche di *neuroimmagine*,[36] comparare l'organizzazione anatomica e gli stati di attività celebrale con i processi decisionali della *moralità lenta* e *veloce* (così come con l'attività cognitiva del *pensiero lento* e del *pensiero veloce*).

In particolare, le ricerche di Jonathan Haidt attestano che le zone del cervello si attivano diversamente nelle circostanze dei dilemmi personali e impersonali: le decisioni morali personali innescano le aree celebrali caratteristiche delle emozioni (come il giro frontale mediale, il giro cingolato posteriore e il giro angolare),[37] mentre per le decisioni morali impersonali si azionano le aree (prefrontali e parietali) coinvolte nel calcolo computazionale. Nell'esperimento mentale del carrello ferroviario, coloro che dovevano decidere se azionare la leva attivavano quell'area celebrale associata ai calcoli razionali (come la corteccia dorsolaterale prefrontale), mentre chi era chiamato a determinare se spingere l'uomo corpulento dal cavalcavia attivava altre aree celebrali (come l'amigdala e il solco temporale superiore, che raccoglie informazioni sulla persona in base a come muove le labbra, gli occhi o le mani).

La computazione biologica celebrale, associata alle zone celebrali tipiche del *pensiero lento*, è, in linea di principio, deterministica, prevedibile e riproducibile artificialmente. È opportuno ricordare che la risoluzione meccanica di un calcolo complicato, secondo Kahneman, era un esempio del funzionamento del *pensiero lento*, che si presentava come una lunga catena di stadi e azioni specifiche, avviata dal ricordo del programma cognitivo di calcolo imparato da bambini e terminante nella computazione finale del risultato. Di contro, il riconoscimento sicuro, veloce e intuitivo di un'immagine rappresentava una pratica dimostrazione del *pensiero veloce*, costituito dalle impressioni e dalle sensazioni che danno luogo alle nostre convinzioni e alle nostre credenze.

Il *pensiero lento* descritto da Daniel Kahneman presenta quindi una natura computazionale descrivibile formalmente, mentre il *pensiero veloce* comprende tutti gli altri fenomeni cognitivi che non si lasciano rappresentare da modelli computazionali standard. Alan Turing già nel 1950 sosteneva che la componente razionale e computazionale dell'intelligenza umana potesse essere riproducibile ed emulabile con precisione da un dispositivo computazionale: secondo Turing, i dispositivi di calcolo erano in possesso di

quei *pensieri lenti* descritti da Kahneman, dal momento che essi erano in grado di eseguire calcoli con una velocità e una precisione persino migliore di quella umana.

È opportuno inoltre ricordare come la brillante ideazione della *macchina di Turing* derivi proprio dall'osservazione di un individuo umano che esegue un calcolo avvalendosi dell'intelligenza computazionale *lenta*. Il tentativo, iniziato con Turing, di elevare il dispositivo computazionale al livello delle facoltà umane (sia cognitive che morali) si realizza attraverso operazioni di formalizzazione, i quali si traducono nella costruzione di algoritmi che governano la presunta intelligenza degli automi. Tuttavia, sebbene i dispositivi computazionali possiedano l'intelligenza *lenta*, non potranno mai essere in grado di riprodurre né tantomeno di formalizzare quel procedimento mentale che può essere definito come intuizione (inscritta da Kahneman nel *pensiero veloce*). L'intuizione è una facoltà esclusiva del pensiero umano, che consente di cogliere l'essenza di una cosa nella sua interezza e di riconoscere l'unità nel molteplice (come avrebbe detto Kant), la quale si manifesta indipendente da qualsiasi procedimento logico di tipo dimostrativo. Le tecnologie intelligenti moderne e future potranno forse essere in grado di utilizzare le due diverse funzioni che secondo Kahneman corrispondono alle nostre abilità cognitive, per acquisire la capacità squisitamente umana di connettere gli stimoli esterni, sociali e più generalmente ambientali con decisioni praticamente immediate (*pensiero veloce*), piuttosto che elaborare razionalmente il ragionamento computazionale e definire un percorso cognitivo più complesso e articolato (*pensiero lento*) in tempi decisamente più lenti. Finora, tuttavia, intuizioni ed emozioni restano patrimonio esclusivo degli esseri viventi.

A seguito della possibile riproducibilità meccanica delle due componenti del pensiero introdotte da Kahneman, è interessante prendere in esame la possibile formalizzazione di quelle che ho definito *moralità lenta* e *veloce*, di particolare interesse per la realizzazione di alcuni automi come i *veicoli a guida autonoma* o i droni militari automatici usati in combattimento. Pensare alla moralità umana come risultato complessivo dei contributi forniti da questi due processi (*lenti* e *veloci*) è un risultato importante per provare a risolvere i potenziali dilemmi morali dei veicoli *driverless*.[38]

Conclusioni

Molti dilemmi sono importanti non tanto per le risposte che essi forniscono, quanto per il dibattito sui presupposti che essi svelano. La realizzazione di automi chiamati a prendere decisioni morali ha sollevato notevoli dubbi riguardo alla definizione di moralità e ai processi decisionali che guidano le scelte umane. In questa sede non ho proposto un criterio definitorio e generale capace di esaurire la sconfinata natura della morale; mi sono invece limitata a indicare le due componenti chiamate in causa nei processi decisionali. La distinzione tra *moralità lenta* e *moralità veloce* si schiera contro ogni forma di riduzionismo della complessità e della varietà morale umana a un approccio puramente razionalistico (secondo il modello di Piaget e Kohlberg) o naturalistico (il rischio che la neuroscienza riduca i principi etici ai risultati dell'*imaging* celebrale e neurale).

È evidente come gli automi *driverless* potrebbero essere in grado di imitare (e quindi possedere) solamente la *moralità lenta* e computazionali e non le intuizioni inconsce e derivanti da quella esperienza esclusivamente umana che è la vita. I calcolatori digitali, artificiali e meccanici non possono competere con le emozioni e le sensazioni umane che caratterizzano la *moralità veloce*, propria di organismi dinamici, empatici, profondi, autentici, sensibili, vivi e naturali. Proprio come, secondo Alan Turing, i calcolatori digitali erano dotati di una specifica forma di intelligenza, quella computazionale e razionale dei *pensieri lenti* descritti da Kahneman, così i *veicoli a guida autonoma* saranno in grado di simulare unicamente la *moralità lenta*, che

comprende tutti quei processi decisionali per i quali la miglior descrizione e spiegazione è quella che fa ricorso a un sistema computazionale di tipo classico o standard, cioè assimilabile a una macchina di Turing. La *moralità veloce* infatti comprende tutti gli altri fenomeni decisionali che non si lasciano descrivere da adeguati modelli computazionali. Per questo tipo di fenomeni si potrebbe ricorrere ad altri tipi di modelli, quali i modelli confessionisti o altri tipi di modelli dinamici continui.

Tuttavia, è interessante notare come nel caso della simulazione delle capacità cognitive, i sistemi computazionali potrebbero possedere, in linea di principio, una "intelligenza incompleta", poiché priva della componente *veloce* e istintiva. Questa intelligenza, sebbene parziale, implementata con le tecniche del *machine learning*, consente ai dispositivi computazionali di gestire molti ambiti della nostra vita. Oggi i *software* guidano aerei, navigano nello spazio, riconoscono volti umani, individuano tumori nella diagnostica per immagini e possiedono anche una potenza predittiva con un alto livello di accuratezza. Un'intelligenza parziale, estranea alla componente istintiva ed emotiva, non risulta perciò meno precisa o potenzialmente pericolosa.

Nella distinzione fra *pensiero lento* e *pensiero veloce*, relativamente alle attività cognitive umane, era probabilmente più semplice riconoscere come l'intelligenza computazionale artificiale potesse essere considerata "reale", dal momento che i dispositivi computazionali sono effettivamente in grado di eseguire calcoli con una velocità e una precisione migliore di quella umana, mentre la componente emotiva poteva al massimo essere simulata (per un breve intervallo di tempo) dai dispositivi di calcolo. Gli automi tecnologici dispongono solamente di uno dei due sistemi cognitivi umani, quello *lento*, il quale non si traduce in una razionalità più debole e incompleta, ma in una razionalità altrettanto rigorosa e logicamente coerente. Allo stesso modo, potremmo sostenere che gli automi *driverless* siano al massimo dotati di una moralità imperfetta e parziale, poiché possiedono unicamente la componente *lenta* e razionale della moralità, diversamente dagli esseri umani, le cui risposte emotive guidano i processi decisionali cognitivi.

In proposito, uno studio condotto da Coricelli e Rustichini[39] ha rilevato come l'emozione del rimpianto (relativo alle scelte passate) abbia un ruolo decisivo nella valutazione umana delle conseguenze future delle decisioni. Tuttavia, la formalizzazione della componente razionale e computazionale negli automi privi di controllo umano (in linea di principio possibile) potrebbe implicare un potenziale danno per gli esseri umani. Tali automi si rivelano infatti un riflesso ottuso di una regolarità statistica che, trascendendo la componente *veloce* e istintiva della moralità umana, diventerebbero dei decisori troppo rigorosi. Gli esseri umani, in quanto prodotto della selezione naturale, hanno necessariamente incorporato nei loro giudizi morali tendenze egoistiche, le quali ne hanno permesso la sopravvivenza.

Le macchine, mediante il processo del *machine learning*, possono seguire rigidamente la logica del programma che hanno incorporato, mentre gli esseri umani, davanti ai dilemmi morali, spesso sono indirizzati ad agire correttamente dalle circostanze in cui si trovano oltre che da valori e da convincimenti personali. Molti soggetti umani interrogati si erano infatti rifiutati di spingere il noto uomo corpulento dal cavalcavia, contrariamente a un automa il quale, se programmato per massimizzare il benessere sociale, eseguirà il compito, logicamente coerente e consistente con il suo programma.

Una moralità parziale potrebbe comportare anche problemi giuridici e legislativi, riguardanti la sicurezza e la tranquillità della vita in una società in cui gli automi obbediscono a un imperativo supremo, che potrebbe richiedere il sacrificio di vite umane, spesso innocenti. Le tecnologie intelligenti e gli automi robotici potrebbero avviare una sorta di capolinea dell'azione morale e della decisione cognitiva umana, in cui l'intelligenza tecnologica si chiude e diventa indipendente

dal contributo umano. Una conseguenza rischiosa di tale chiusura alla cognizione e alla moralità umana e dell'implementazione di algoritmi morali tecnologici è che si assisterebbe al venir meno del pluralismo delle opinioni, tipicamente umano, affidandosi a tecnologie che rispondo a una logica conformista e aprioristica, che, quasi in maniera dittatoriale, viene imposta dall'esterno.

In conclusione, i nuovi quesiti nati dall'esigenza di attribuire una moralità ai *veicoli a guida automatica* chiamano in causa un numero molto elevato di problemi pratici oltre che dilemmi filosofici a cui difficilmente si potrà fornire una risposta esaustiva. L'essere umano è necessariamente considerato il modello di moralità a cui tali veicoli dovranno ispirarsi: ma il problema sorge nel momento in cui neanche il prototipo morale riesce a prendere decisioni in contesti particolari.

Inoltre, non solo non sappiamo come formalizzare le norme morali universali a cui la *moralità lenta* dell'automa dovrebbe appellarsi, ma anche se questo fosse possibile, il potenziale danno per l'umanità sarebbe inestimabile. Nel mio lavoro, ho cercato quindi definire le controparti istintive e computazionali della morale umana, attribuendo agli automi unicamente la *moralità lenta*, riconoscendo che non abbiamo ancora oggi delle risposte definitive ed esaurienti alla domanda "possono le macchine avere un comportamento morale?", una costante che chiama in causa annosi problemi. Restiamo tuttavia testimoni dell'entusiasmante e stupefacente divenire della tecnologia, che si avvicina sempre di più alla realizzazione di automi umani collocati in situazioni sempre più reali.

Note

[1] Proposta di risoluzione del Parlamento Europeo concernente norme di diritto civile sulla robotica del 27 gennaio 2017.

[2] La *roboetica* è una branca dell'etica applicata alla robotica che si occupa dell'applicazione alle macchine intelligenti di regole di comportamento fondate su valori morali, culturali ed etici universalmente condivisi e diretti alla tutela degli esseri umani.

[3] Oggigiorno vi sono numerosi dispositivi che permettono alla tecnologia di controllare il veicolo, escludendo il guidatore da ogni decisione o scelta: i dispositivi di frenata automatica, di antislittamento nella frenata e il parcheggio assistito rappresentano due esempi pratici del graduale sviluppo della guida *driverless*.

[4] Cfr. P. GAO, R. HENSELY, A. ZIELKE, *A Roadmap to the Future for the Auto Industry*, in: «McKinsey Quarterly», n. 4, 2004, pp. 42-53. Gli autori stimano che le *auto a guida autonoma* ridurranno circa il 90% degli incidenti stradali causati da errori umani.

[5] L'articolo sancisce l'assenza di colpa in sinistri stradali «dimostrando di avere tenuto un comportamento esente da colpa e perfettamente conforme alle regole del codice della strada, ma può risultare anche dall'accertamento che il comportamento della vittima sia stato il fattore causale esclusivo dell'evento dannoso, comunque non evitabile da parte del conducente».

[6] Il *Massachusets Institute of Technology* (MIT) ha escogitato la *Moral Machine*, ovvero una piattaforma social in cui viene richiesto agli utenti di esprimersi in merito a quali scelte dovrebbero fare i piloti automatici in casi altamente critici.

[7] J.-F. BONNEFON, A. SHARIFF, I. RAHWAN, *The Social Dilemma of Autonomous Vehicles*, in: «Science», vol. CCCLII, n. 6293, 2016, pp. 1573-1576.

[8] Cfr. P. AGNOLI, F. PICCOLO, *Probabilità e scelte razionali: una introduzione alla scienza delle decisioni*, Armando Editore, Roma 2008, p. 216.

[9] Patrick Lin dirige lo *Ethics + Emerging Sciences Group* presso il California Polytechnic State University.

[10] Cfr. W.K. VISCUSI, J.E. ALDY, *The Value of a Statistical Life: A Critical Review of Market Estimates Throughout the World*, in: «Journal of Risk and Uncertainty», vol. XXVII, n. 1, 2003, pp. 5-76.

[11] Cfr. A. MASTROGIORGIO, A. PETRACCA, *Razionalità incarnata*, in: «Sistemi intelligenti», vol. XXVII, n. 3, 2015, pp. 481-504.

[12] Cfr. M.G. ROSSI, *Dilemmi morali*, in: «Sistemi Intelligenti», vol. XXIII, n. 1, 2011, pp. 187-205, qui p. 187.

[13] Cfr. P. FOOT, *The Problem of Abortion and the Doctrine of the Double Effect*, in: «Oxford Review», vol. V, 1967, pp. 5-15; P. FOOT, *Moral Dilemmas*, Clarendon Press, Oxford 2002.

[14] I primi articoli di quella branca dell'etica chiamata ironicamente "*carrellologia*" facevano riferimento ai carrozze dei tram. Per una rassegna cfr. D.

EDMONDS, *Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us About Right and Wrong*, Princeton University Press, Princeton 2014 (trad. it. *Uccideresti l'uomo grasso? Il dilemma etico del male minore*, traduzione di G. GUERRIERO, Cortina, Milano 2014). Si veda anche A. MANFRINATI, L. LOTTO, M. SARLO, *Un nuovo set di 60 dilemmi morali; dati normativi italiani per giudizi di accettabilità morale, tempi di decisione e valutazioni emozionali*, in: «Giornale Italiano di Psicologia», vol. XXXV, n. 1, 2013, pp. 211-227.

[15] Hauser nel suo esperimento conclude che l'89% dei soggetti intervistati risponde in maniera affermativa al dilemma se sia lecito azionare lo scambio per deviare il percorso del treno (cfr. M. HAUSER, F. CUSHMAN, L. YOUNG, J.I.N. KANG-XING, J. MIKHAIL, *A Dissociation Between Moral Judgments and Justifications*, in: «Mind & Language», vol. XXII, n. 1, 2007, pp. 1-21).

[16] Cfr. J.J. THOMSON, *The Trolley Problem*, in: «The Yale Law Journal», vol. XCIV, n. 6, 1985, pp. 1395-1415.

[17] Hauser nel medesimo esperimento conclude che l'11% dei soggetti intervistati risponde in maniera affermativa al dilemma se sia lecito spingere l'uomo dal ponte (cfr. M. HAUSER, F. CUSHMAN, L. YOUNG, J.I.N. KANG-XING, J. MIKHAIL, *A Dissociation Between Moral Judgments and Justifications*, cit.).

[18] Cfr. M. DANIELE, M. BUCCIARELLI, *Le decisioni utilitariste nei dilemmi morali sacrificali possono basarsi su intuizioni*, in: «Sistemi Intelligenti», vol. XXX, n. 2, 2018, pp. 355-372; M. DI FRANCESCO, *Neurofilosofia, naturalismo e statuto dei giudizi morali*, in: «Etica e Politica», vol. IX, n. 2, 2007, pp. 126-143; P. SINGER, *Ethics and Intuitions*, in: «Journal of Ethics», vol. IX, n. 3-4, 2005, pp. 331-352.

[19] Joshua Greene definisce un'azione *personale* se comporta una violazione diretta, del soggetto protagonista del dilemma morale, che causa un grave danno fisico nei confronti di una determinata persona o di un gruppo di persone.

[20] Joshua Greene definisce un'azione *impersonale* se si limita a deviare una possibile minaccia preesistente che viene deviata in modo indiretto dal soggetto protagonista del dilemma morale.

[21] Cfr. P. CITATI, *La mente colorata*, Adelphi, Milano 2018.

[22] L'indagine filosofica sulla moralità è antica e complessa, dunque difficilmente riducibile a uno schema evolutivo secondo categorie dal potenziale dicotomico come quelle di ragione ed emozione. Questa apparente riduzione è tuttavia funzionale alla comprensione e all'identificazione di nuove connessioni con altre discipline (in questo caso le scienze cognitive e la neuroscienza), attraverso il darsi di nuove interpretazioni.

[23] Cfr. D. MARAZZITI, P. LANDI, S. BARONI, L. DELL'OSSO, *Esiste una neurobiologia del comportamento morale?*, in: «Giornale Italiano di Psicopatologia», vol. XVII, 2011, pp. 309-321; A. OLIVERIO, *Neuroscienza ed etica*, in: «Iride», vol. XXI, n. 52, 2008, pp. 193-215;

[24] La *Teoria Computazionale della Mente* afferma come tutto il pensiero umano sia riconducibile ad attività di tipo algoritmico, ovvero a trasformazioni di simboli in base a regole formali esplicitamente formulabili ed eseguibili in modo puramente meccanico (cfr. M. GIUNTI, R. GIUNTINI, *Macchine, calcolo e pensiero*, in: S. ANCINI (a cura di), *Sguardi sulla scienza nel giardino dei pensieri*, Mimesis, Milano 2007, pp. 39-67, qui p. 39).

[25] Cognitivo poiché fondato sul ragionamento razionale.

[26] Cfr. P.S. CHURCHLAND, P.M. CHURCHLAND, *Could a Machine Think?*, in: «Scientific American», vol. CCLXII, n. 1, 1990, pp. 32-37.

[27] Cfr. J.R. SEARLE, *Minds, Brains and Programs*, in: «Behavioral and Brain Sciences», vol. III, n. 3, 1980, pp. 417-424.

[28] R. FRENCH, *Subcognition and the Limits of the Turing Test*, in: «Mind», vol. XCIX, n. 393, 1990, pp. 53-65.

[29] Cfr. D. KAHNEMAN, *Thinking, Fast and Slow*, & Giroux, New York 2011 (trad. it. *Pensieri lenti e veloci*, traduzione di L. SERRA, Mondadori, Milano 2012).

[30] Cfr. A. MANFRINATI, *Il conflitto morale nella prospettiva delle neuroscienze*, in: A. DA RE, A. PONCHIO (a cura di), *Il conflitto morale*, Il Poligrafo, Padova 2011, pp. 69-82.

[31] M. NUSSBAUM, *Upheavals of Thought: The Intelligence of Emotions*, Cambridge University Press, Cambridge 2001, pp. 17-18.

[32] Antonio Damasio è un neurologo portoghese laureato in medicina all'Università di Lisbona. Fino al 2005 è stato direttore del Dipartimento di neurologia dell'*University of Iowa Hospitals and Clinics* negli Stati Uniti; oggi dirige il *Brain and Creativity Institute dell'University of South California* in cui insegna neurologia, psicologia e neuroscienze.

[33] Cfr. J.D. GREENE, *Moral Tribes. Emotion, Reason, and the Gap between Us and Them*, Atlantic Books, London 2015.

[34] È opportuno specificare che tali regole morali intuitive e innate non corrispondono a leggi morali universali e assolute, dal momento che le nostre emozioni e la nostra ragione sono sempre in grado di modificarle e adattarle all'esperienza: per esempio, uccidere un uomo è intuitivamente avvertito come sbagliato, in relazione alle nostre regole morali inconsce, ma la ragione ci mostra come, in alcuni possibili contesti (come nelle situazioni di legittima difesa), sia ammissibile uccidere intenzionalmente.

[35] J.D. GREENE, S.A. MORELLI, K. LOWEMBERG, L.E. NYSTROM, J.D. COHEN, *Cognitive Load Selectively Interface with Utilitarian Moral Judgment*, in: «Cognition», vol. CVII, n. 3, 2008, pp. 1144-1154.

[36] Le tecniche di *neuroimmagine* rilevano e misurano il metabolismo cerebrale nell'analisi e nello studio della relazione tra l'attività di determinate aree cerebrali e specifiche funzioni cerebrali.

[37] J. MOLL, R. DE OLIVEIRA-SOUZA, I. E. BRAMATI, J. GRAFMAN, *Functional Networks in Emotional Moral and Nonmoral Social Judgment*, in: «Neuroimage», vol. XVI, n. 3, Pt 1, 2002, pp. 696-703.

[38] Cfr. A. RENDA, *Ethics, Algorithms and Self-driving Cars – a CSI of the "Trolley Problem"*, in: «Policy Insight», n. 2, 2018 - URL: https://www.ceps.eu/system/files/PI%202018-02_Renda_TrolleyProblem.pdf; F. SANTONI DE SIO, *Ethics and Self-driving Cars: A White Paper on Responsible Innovation in Automated driving System*, Dutch Ministry of Infrastructure and the Environment Rijkswaterstaat, settembre 2016; F. SANTONI DE SIO, *Killing by Autonomous Vehicles and the Legal Doctrine of Necessity*, in: «Ethical Theory and Moral Practice», vol. XX, n. 2, 2017, pp. 411-429; G. TAMBURRINI, *Autonomia delle macchine e filosofia dell'intelligenza artificiale*, in: «Rivista di Filosofia», vol. CVIII, n. 2, 2017, pp. 263-275; L. BUTTI, *Driverless: sicurezza, responsabilità legali ed algoritmi eticamente complessi nella circolazione di auto senza conducente*, in: B&P AVVOCATI, n. 4, 2016.

[39] G. CORICELLI, A. RUSTICHINI, *Counterfactual Thinking and Emotions: Regret and Envy Learning*, in: «Philosophical Transactions of the Royal Society – B: Biological Science», vol. CCCLXV, n. 1538, 2010, pp. 241-247.