DANIELA TAFANI

Sulla moralità artificiale.
Le decisioni delle macchine tra etica e diritto

Centrale, per il progetto dell'etica delle macchine, è la convinzione (o la speranza) che l'etica possa essere resa computabile, che possa essere affinata abbastanza da poter essere programmata in una macchina.
Susan Leigh Anderson, 2011

qui è all'opera una caratteristica molto umana del singolo scienziato. La chiamo legge dello strumento, e può essere formulata come segue: dai un martello a un bambino e scoprirà che tutto ciò che incontra ha bisogno di qualche colpo. Non sorprende affatto scoprire che uno scienziato formula i problemi in un modo che richiede, per la loro soluzione, solo quelle tecniche in cui egli stesso è particolarmente abile.
Abraham Kaplan, 1964

1. Il dilemma del carrello ferroviario

Gli esperimenti mentali costituiscono, per la riflessione filosofica, uno strumento concettuale utile e potente: in virtù del loro carattere narrativo, consentono di mettere alla prova una dottrina, verificando se i suoi esiti più estremi appaiano intuitivamente accettabili e sfidando chi la sostenga a sottoscriverli esplicitamente[1]. Qualora gli esperimenti di pensiero siano mal formulati, tuttavia, o prelevati dal loro contesto originario e traslati in luoghi impropri, si trasformano in strampalati dispositivi di persuasione o in strumenti fuorvianti e pericolosi[2]: è questo il caso, come si cercherà di mostrare nelle pagine che seguono, del dilemma del carrello ferroviario (*trolley problem*), quale problema che si imporrebbe, ineludibile, alla neonata disciplina dell'etica delle macchine[3].

[1] Per un quadro delle tesi alternative su cosa sia, e a cosa possa servire, un esperimento di pensiero, v. *Routledge Companion to Thought Experiments*, a cura di J.R. Brown, Y. Fehige, M. Stuart, Abingdon/New York, Routledge, 2018, e in particolare il contributo di G. Brun, *Thought Experiments in Ethics*, pp. 195-210.
[2] Cfr. D.C. Dennett, *Intuition pumps and other tools for thinking*, New York, London, Norton & Company, 2013, pp. 2-3; trad. it. a cura di S. Frediani con il titolo *Strumenti per pensare*, Milano, Cortina, 2014, p. 3.
[3] L'etica delle macchine «si occupa di dare alle *macchine* principi etici o una procedura per scoprire un modo di risolvere i dilemmi etici in cui potrebbero imbattersi, consentendo loro di operare in modo eticamente responsabile attraverso il proprio processo decisionale etico» (*Machine Ethics*, ed. by M. Anderson, S.L. Anderson, Cambridge, Cambridge University Press, 2011, p. 1). V. anche W. Wallach, C. Allen, *Moral machines: Teaching Robots Right from Wrong*, Oxford, Oxford University Press, 2009.

Nel 1967, in un articolo sull'aborto[4], Philippa Foot esaminava la «dottrina del doppio effetto»[5], ossia la tesi per cui sarebbe talora lecito causare in virtù di un'intenzione obliqua – vale a dire, meramente prevedendolo – ciò a cui non è lecito mirare direttamente. Foot presentava, tra gli altri, i tre esempi che seguono:

> Supponiamo che un giudice o un magistrato si trovi di fronte a rivoltosi che chiedono che venga trovato un colpevole per un certo crimine e che minaccino di compiere altrimenti la loro sanguinosa vendetta su una particolare sezione della comunità. Essendo sconosciuto il vero colpevole, il giudice ritiene di poter prevenire lo spargimento di sangue solo incastrando una persona innocente e facendola giustiziare. Accanto a questo esempio, ne è posto un altro in cui un pilota, il cui aereo sta per schiantarsi, stia decidendo se deviare da un'area più abitata a una meno abitata. Per rendere il parallelo il più vicino possibile, si può supporre piuttosto che egli sia l'autista di un tram fuori controllo, che può solo deviare da uno stretto binario a un altro; cinque uomini stanno lavorando su un binario e un uomo sull'altro; chiunque sia sul binario in cui entra è destinato ad essere ucciso.[6]

Da un'analisi di questi e di molti altri esempi, Foot traeva la conclusione che la distinzione tra intenzione diretta e intenzione obliqua non fosse decisiva. A consentire di discernere le differenze moralmente rilevanti nei diversi casi era piuttosto la distinzione tra doveri negativi e doveri positivi, ossia tra l'obbligo di astenersi da certe azioni, quali l'omicidio e il furto, e i doveri di cura o di aiuto[7], nonché la priorità dei primi sui secondi: nel caso del giudice, ad esempio, il dovere di non nuocere a un innocente avrebbe dovuto prevalere sul dovere di recare aiuto.
Il dilemma del tram fuori controllo sarebbe stato discusso, pochi anni dopo, da Judith Jarvis Johnson, e riformulato come «il problema del carrello ferroviario [trolley], in onore dell'esempio della signora Foot»[8]. Come tale ha avuto, nei decenni successivi, una straordinaria fortuna: insieme agli altri esperimenti mentali elencati da Foot, è infatti oggetto di continua discussione e rielaborazione, nell'ambito della filosofia morale e delle neuroscienze dell'etica[9].

---

[4] P. Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, «Oxford Review», V, 1967, pp. 5-15.

[5] La formulazione originaria della dottrina del doppio effetto è fatta risalire in genere a Tommaso d'Aquino, il quale a sua volta rinviava a Agostino: «S. Agostino afferma: "Non sia mai che ci venga imputato quel male occasionale con cui possiamo colpire qualcuno mentre noi facciamo in vista del bene delle azioni lecite" […] Ecco perché secondo il diritto, se uno nel compiere una cosa lecita con le debite precauzioni provoca l'uccisione di un uomo, non incorre nel reato di omicidio; se invece egli provoca la morte di un uomo nel compiere una cosa illecita, oppure nel compiere cose lecite non prende le dovute precauzioni, non può sfuggire il reato di omicidio» (Tommaso d'Aquino, *La Somma teologica*, trad. it. a cura dei Frati Domenicani, Bologna, Edizioni Studio Domenicano, 2014, vol. 3, IIa IIæ, Q. 64, A. 8, pp. 650-51).

[6] P. Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, cit., p. 8.

[7] Foot rimandava alla formulazione di J. Salmond in *Jurisprudence*, London, Sweet & Maxwell, 11a ed. 1957, senza ricordare che la tesi della priorità dei doveri negativi su quelli positivi era stata sostenuta, ad esempio, da Kant, quale priorità dei doveri di giustizia, detti anche perfetti o essenziali, sui doveri di bontà, imperfetti o accidentali; v. W. Kersting, voce «Pflichten, unvollkommene/vollkommene», in *Historisches Wörterbuch der Philosophie*, vol. 7, hrsg. von K. Gründer, Basel/Stuttgart, 1989, coll. 433-39.

[8] J. J. Thomson, *Killing, letting die and the trolley problem*, «The Monist», LIX, 1976, pp. 204–17.

[9] Cfr. F.M. Kamm, *The Trolley Problem Mysteries*, con commenti di J.J. Thomson, T. Hurka, S. Kagan; a cura di E. Rakowski, New York, Oxford University Press, 2016; un'introduzione divulgativa in D. Edmonds, *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*, Princeton, Princeton University Press, 2013; trad. it. a cura di G. Guerrerio con il titolo *Uccideresti l'uomo grasso?Il dilemma etico del male minore*, Milano, Cortina, 2014. Cfr. J.D. Greene, *Solving the trolley problem*, in *A Companion to Experimental Philosophy*, a cura di J. Sytsma and W. Buckwalter, Chichester, Wiley-Blackwell,

Quello che Foot non avrebbe immaginato, nel 1967, è che il suo dilemma del tram sarebbe stato ritenuto, dopo qualche decennio, oltre che un dilemma che ha ad oggetto un veicolo, altresì un dilemma per il veicolo stesso, nei casi in cui questo debba concretamente prendere decisioni, da solo, in situazioni di emergenza.
Il dilemma del trolley, nelle sue molte varianti, è presentato, nel dibattito contemporaneo sulle auto a guida autonoma, come un problema morale irrisolto e come una fattispecie giuridica ancora non contemplata da alcuna legge, che occorre dunque affrontare e risolvere anzitutto dal punto di vista morale, per colmare poi il relativo vuoto legislativo, con norme che regolino di conseguenza il traffico automatizzato e connesso. In realtà – come si cercherà di sostenere nel paragrafo che segue - negli Stati le cui carte costituzionali sanciscano la tutela dei diritti individuali fondamentali, il diritto contiene vincoli precisi alla soluzione dei dilemmi delle auto a guida autonoma, sebbene non li preveda specificamente. Occorre dunque tener presente che, in tali Stati, soluzioni lesive dei diritti fondamentali non possono trovare - salvo modifiche costituzionali – alcun recepimento giuridico.
Un veicolo a guida autonoma non è un agente morale in senso proprio e dunque, nei casi in cui si trovi di fronte alla prospettiva di un incidente inevitabile, non è - come si cercherà di argomentare nel terzo paragrafo - un soggetto che si trovi di fronte a un dilemma morale. E' una macchina, che reagirà sulla base di istruzioni programmate e, in quanto dotata di intelligenza artificiale, sulla base degli esiti, solo in parte prevedibili, del suo apprendimento automatico. L'idea che gli eventuali incidenti delle auto a guida autonoma pongano questioni straordinarie - quali le scelte che le macchine stesse dovrebbero compiere - anziché, come il funzionamento di molte altre macchine, mere questioni di sicurezza, di trasparenza, di controllo e di cautela, sorge forse dall'inclinazione umana a considerare gli oggetti antropomorfi quali agenti o, addirittura, quali agenti morali.

2. Auto che «decidono» chi uccidere

Il mito della sicurezza delle automobili a guida autonoma è nato insieme alla diffusione di massa delle autovetture: nel 1935, in *The safest place*[10] – un cortometraggio di educazione stradale commissionato dalla Divisione Chevrolet della General Motors – un tranquillo signore che canticchia per casa rischia, più volte, di rompersi l'osso del collo, in una serie di incidenti domestici ai quali sfugge per miracolo, finché finalmente non esce di casa («usciamo di qui» – recita la voce fuori campo – «prima che qualcuno si faccia male»). La stessa voce fuori campo accompagna poi un'elegante vettura – mentre percorre sicura le strade, senza conducente e senza passeggeri – ricordando allo spettatore che quel «salotto su ruote» sarebbe il luogo più sicuro al mondo, se solo potesse essere dotato di un «meccanismo di guida automatico»: gli incidenti non derivano infatti che dalle imprudenze dei guidatori umani, che tengono per sé, come fosse un segreto, la direzione in cui intendono svoltare o che scambiano la partenza a un semaforo per lo scatto alla partenza di una gara automobilistica.
L'incondizionata fiducia – esibita nel cortometraggio del 1935 – nell'infallibilità, e dunque nella sicurezza, della tecnologia, era finalizzata, oltre che a educare gli automobilisti,

2016, pp. 175-89; Idem, *Beyond Point-and-Shoot Morality: Why Cognitive (Neuro) Science Matters for Ethics*, «Ethics», CXXIV, 4, 2014, pp. 695–726.
[10] https://archive.org/details/SafestPl1935 (ultimo accesso, a questo e agli altri indirizzi web citati nel presente articolo, il 13 dicembre 2019).

soprattutto a rassicurarli, considerato che gli incidenti stradali avevano ormai assunto, negli Stati Uniti, la rilevanza di un problema sociale[11]. Un'analoga, trionfale fiducia (chissà se fondata su analoghe ragioni commerciali) accompagna oggi l'introduzione delle prime vetture a guida parzialmente autonoma[12] – prospettando una diminuzione del 90% degli incidenti stradali[13] – senza che sia in genere considerata l'eventualità di una mera sostituzione degli errori umani con gli errori delle macchine[14].
Guardando con ideologico ottimismo[15] ben oltre il presente stadio di sviluppo tecnologico dei veicoli autonomi - nel quale hanno luogo incidenti mortali perché, ad esempio, la vettura scambia un camion bianco per una porzione di cielo e vi si schianta contro[16] – molti ritengono che tali veicoli saranno in grado di affrontare le situazioni di emergenza, grazie alla loro potenza e rapidità di calcolo, sulla base di analisi raffinate e istantanee della situazione, anziché, come gli esseri umani, improvvisando in modo istintivo. Per far fronte ai rari incidenti inevitabili, sarà dunque opportuno – si sostiene – programmare anticipatamente i veicoli a guida autonoma affinché scelgano chi uccidere, nei casi in cui un danno mortale sia inevitabile e in cui sia certo che ciascuna delle manovre alternative intraprese dal veicolo mieterà una diversa vittima. Con ciò, il dilemma del carrello ferroviario ha apparentemente trovato una nuova, concreta urgenza e addirittura, nell'ambito del fitto dibattito sulla moralità artificiale, un suo specifico dominio: l'«etica degli incidenti stradali» dei veicoli a guida autonoma[17].
Oltre ai casi – analoghi alla versione originaria del dilemma – in cui l'alternativa sia tra sacrificare un solo individuo o, invece, più individui,[18] si ritiene che debbano essere considerati anche i casi in cui le vittime alternative siano singole persone:

> Immagina che, in un lontano futuro, la tua macchina autonoma fronteggi questa terribile scelta: deve o sterzare a sinistra e colpire una bambina di 8 anni, o sterzare a destra e colpire una nonna di 80 anni. Data la velocità della macchina, ciascuna vittima sarebbero certamente uccisa nell'impatto. Se non sterzi, entrambe le vittime verranno colpite e uccise; quindi c'è un buon motivo per pensare che dovresti sterzare in una direzione o

---

[11] F. Kröger, *Automated Driving in Its Social, Historical and Cultural Contexts*, in *Autonomous Driving. Technical, Legal and Social Aspects*, a cura di M. Maurer, J. Gerdes, B. Lenz, H. Winner, Berlin/Heidelberg, Springer, 2016, pp. 41-68.
[12] L'autonomia dei veicoli è stata classificata dalla Society of Automotive Engineers in sei livelli, dal livello 0, in cui è il conducente umano a svolgere tutte le operazioni di guida, al livello 5, in cui la guida è interamente automatizzata e l'occupante umano della vettura è un mero passeggero (https://www.sae.org/standards/content/j3016_201806/).
[13] Cfr. la National Highway Traffic Safety Administration, agenzia del Dipartimento dei trasporti statunitense: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety; v. anche M. Bertoncello, D. Wee, *Ten ways autonomous driving could redefine the automotive world*, McKinsey&Company, https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world.
[14] Tematizza invece la questione C. Misselhorn, *Grundfragen der Machinenethik*, Ditzingen, Reclam, 2018, pp. 198-200.
[15] V. F. Operto, *Elementi di roboetica. Analisi di alcune metaetiche nella progettazione e programmazione di veicoli autonomi*, «InCircolo», VI, 2018, pp. 89-108: 98.
[16] Cfr. il resoconto di Tesla, l'azienda produttrice del veicolo coinvolto nell'incidente: https://www.tesla.com/it_IT/blog/tragic-loss.
[17] V. ad es., tra gli innumerevoli articoli comparsi in meno di un decennio, P. Lin, *Why Ethics Matters for Autonomous Cars*, in *Autonomous Driving*, cit, pp. 69-85. G. Keeling, *Why Trolley Problems Matter for the Ethics of Automated Vehicles*, «Science and Engineering Ethics», 2019, pp. 1-15. Un'utile ricognizione in S. Nyholm, *The Ethics of Crashes with Self-Driving Cars: A Roadmap*, I e II, «Philosophy Compass», XIII, 7, 2018, https://onlinelibrary.wiley.com/doi/epdf/10.1111/phc3.12507 e /phc3.12506.
[18] J.-F. Bonnefon, A. Shariff, I. Rahwan, *The social dilemma of autonomous vehicles*, «Science», CCCLII, 6293, 2016, pp. 1573-76.

nell'altra. Ma quale sarebbe la decisione eticamente corretta? Se tu stessi programmando la macchina a guida autonoma, come la istruiresti a comportarsi se mai dovesse imbattersi in un caso del genere, per quanto raro?[19]

Programmare un'auto affinché selezioni, tra due persone, quella con cui collidere, equivale a introdurre nei veicoli civili algoritmi che prendano di mira, quando non siano in grado di considerare alternative ulteriori, un individuo anziché un altro, sulla base delle sue caratteristiche rilevabili, in modo analogo agli algoritmi di selezione degli obiettivi in ambito militare[20]. Ma come stabilire, per ciascuno degli scenari possibili, quale tra le due potenziali vittime debba essere colpita?
Alcuni propongono che i processi decisionali dei veicoli autonomi siano automatizzati, nei casi in cui non siano già disponibili principi etici condivisi, sulla base di un modello computazionale della scelta sociale, in grado di aggregare le preferenze espresse dalle persone quanto alle soluzioni dei «dilemmi morali nel contesto di incidenti inevitabili». Per raccogliere tali preferenze, è stata costituita, presso il Media Lab del Massachusetts Institute of Technology, una piattaforma di voto *online* denominata «Moral Machine»[21], che ha raccolto quasi 40 milioni di decisioni, da 233 paesi o territori[22].
La piattaforma si presenta con le sembianze di un videogioco («un "gioco serio" multilingue online»[23]), che genera scenari di incidenti sulla base di nove fattori:

risparmiare gli esseri umani (rispetto agli animali domestici), restare nella propria corsia (rispetto a sterzare), risparmiare i passeggeri (rispetto ai pedoni), risparmiare più vite (rispetto a un numero minore di vite), risparmiare gli uomini (rispetto alle donne), risparmiare i giovani (rispetto agli anziani), risparmiare i pedoni che attraversano legalmente (rispetto a coloro che attraversano illegalmente), risparmiare chi sia in forma (rispetto a chi sia meno in forma) e risparmiare coloro che hanno uno status sociale più elevato (rispetto a quelli con uno status sociale inferiore).[24]

Le preferenze globali espresse dalle persone che hanno risposto sono nettamente a favore dei bambini rispetto agli anziani e, seppure più debolmente, delle donne rispetto agli uomini, degli uomini d'affari rispetto ai senza tetto e degli atleti rispetto alle persone sovrappeso. Ciò non implica – concedono gli autori – che i decisori politici debbano recepire necessariamente tali opinioni; dovranno tuttavia essere preparati a giustificare decisioni che si discostino dall'opinione pubblica quanto alle preferenze più forti, quali quella a favore dei bambini[25].
Il ricorso alla «moralità pubblica»[26] – ossia a un'aggregazione, secondo un modello di scelta sociale, delle opinioni individuali raccolte – è presentato come strumento di ultima istanza, da utilizzarsi nei casi in cui non vi siano soluzioni tecniche, per evitare il dilemma, né vincoli

[19] P. Lin, *Why Ethics Matters for Autonomous Cars*, cit., pp. 69-70.
[20] *Ivi*, p. 72.
[21] http://moralmachine.mit.edu.
[22] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, *The Moral Machine experiment*, «Nature», DLXIII, 2018, pp. 59–64; R. Noothigattu, S.S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, A.D. Procaccia, *A Voting-Based System for Ethical Decision Making*, 2017, https://arxiv.org/pdf/1709.06692.pdf . L'idea che la scelta sociale computazionale, ossia l'utilizzo di algoritmi per aggregare le preferenze individuali, possa fornire gli strumenti per giungere a decisioni etiche collettive, è tratta da J. Greene, F. Rossi, J. Tasioulas, K. B. Venable, B.C. Williams, *Embedding Ethical Principles in Collective Decision Support Systems*, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence and the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference*, a cura di D. Schuurmans, M. Wellmam, Phoenix, AAAI, 2016, vol. VI, pp. 4147-4151.
[23] E. Awad et al., *The Moral Machine experiment*, cit., p. 59.
[24] *Ivi*, p. 60.
[25] *Ibidem*.
[26] *Ivi*, p. 59.

giuridici o verità morali accertate, in grado di dettare il da farsi. La tesi è che i risultati, opportunamente ricomposti, dell'osservazione empirica dei comportamenti morali simulati in una sorta di videogioco possano supplire al vuoto legislativo e alla mancanza di una condivisione e della necessaria formalizzazione dei principî – malgrado secoli di discussioni, si osserva – in materia di etica normativa[27].

In effetti, né il diritto contemporaneo né alcuna dottrina morale universalmente condivisa contengono una scala dei valori delle vite umane, sulla base della quale predisporre gli opportuni algoritmi sacrificali, che traducano in termini computazionali la gerarchia dei bersagli dei veicoli a guida autonoma, sulla base del loro genere, dell'età, della forma fisica o della ricchezza. Non si tratta, tuttavia, di una lacuna giuridica: la risposta alla domanda su chi uccidere, tra un vecchio o un bambino, o tra un uomo d'affari e un senza tetto, non si trova, nel diritto, perché la domanda stessa è vietata. Una scala dei valori delle vite consentirebbe senz'altro di dare a un sistema di guida autonoma istruzioni chiare e formalizzabili, ma sarebbe in contrasto con i diritti fondamentali dell'uomo alla vita e alla non discriminazione, quali sono sanciti, tra gli altri, nella Dichiarazione universale dei diritti dell'uomo, nella Carta dei diritti fondamentali dell'Unione Europea e, in molti paesi, nelle carte costituzionali[28].

Nel trattare il dilemma relativo all'impatto con una bambina di 8 anni o con una nonna di 80 anni, Patrick Lin osserva che la discriminazione sulla base dell'età è vietata dal codice etico dell'Institute of Electrical and Electronics Engineers, nonché dalla Costituzione tedesca e dal XIV Emendamento della Costituzione degli Stati Uniti (esaminati in quest'ordine), ma ritiene comunque – avendo formulato il dilemma in modo da prevedere che entrambe le vittime saranno colpite, qualora si decida di non sterzare in direzione di alcuna delle due – che sia meglio uccidere una sola persona, pur sulla base di una discriminazione, piuttosto che ucciderne due[29].

Qualora si accetti l'assunto di Lin - per cui tutti i possibili dilemmi devono essere anticipatamente considerati e risolti, senza alcun vaglio sulla legittimità dei dilemmi medesimi – occorrerà inserire tra le alternative anche quella tra uccidere un bambino bianco e un bambino nero, oppure tra uccidere un cattolico e un ebreo, e poi pretendere che si tratti di un dilemma giuridicamente lecito e attendere gli esiti del prossimo sondaggio planetario effettuato, eventualmente in 3D[30], tramite un videogioco della moralità[31].

A una conclusione rispettosa della Legge fondamentale tedesca è pervenuta invece la Commissione etica incaricata, nel 2016, dal Ministero federale tedesco dei trasporti, di redigere un codice etico per il traffico automobilistico automatizzato e connesso. Il codice, pubblicato nel 2017, prevede che «in situazioni di incidente inevitabili, qualsiasi distinzione basata sulle caratteristiche personali (età, sesso, costituzione fisica o mentale)» sia

---

[27] R. Noothigattu et al., *A Voting-Based System for Ethical Decision Making*, cit. Cfr. L.R. Sütfeld, R. Gast, P. König, G. Pipa, *Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios: Applicability of Value-of-Life-Based Models and Influences of Time Pressure*, «Frontiers in behavioral neuroscience», XI, 2017, https://www.frontiersin.org/articles/10.3389/fnbeh.2017.00122/full.

[28] Basti, qui, la Dichiarazione universale dei diritti dell'uomo, Artt. 2 e 3; la Carta dei diritti fondamentali dell'Unione Europea, all'Art. 21 «Non discriminazione», nomina esplicitamente anche l'età, tra i fattori su cui è vietato fondare discriminazioni.

[29] P. Lin, *Why Ethics Matters for Autonomous Cars*, cit., p. 70.

[30] Lo propongono L.R.Sütfeld et al., *Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios*, cit., p. 5.

[31] Derek Leben osserva che, utilizzando proprietà sociali come criteri per i giudizi morali, l'esperimento del MIT esamina i pregiudizi discriminatori delle persone, più che il loro giudizio morale (*Ethics for Robots. How to Design a moral Algorithm*, London and New York, Routledge, 2019, p. 112).

«severamente proibita». E' inoltre vietata «la compensazione» tra le vittime: è cioè inammissibile, in situazioni di emergenza, sacrificare una persona per salvarne altre[32]. La Commissione richiama, su questo punto, la celebre sentenza della Corte costituzionale federale del 2006, che ha dichiarato nulla – in quanto incompatibile con la tutela della dignità umana e il diritto alla vita sanciti costituzionalmente – la previsione della Legge sulla sicurezza aerea del 2005[33], che autorizzava l'uso delle armi, da parte dell'aviazione militare, contro aeromobili dirottati: l'uccisione dei passeggeri tenuti in ostaggio, allo scopo di proteggere una popolazione più ampia, avrebbe infatti degradato i passeggeri a mere cose, trattandoli come mezzi e negando loro il valore che spetta all'uomo in se stesso[34].
La citazione kantiana[35] contenuta in quest'ultima parte della motivazione della sentenza mostra che la mancanza di principi condivisi è meno drammatica di quanto la dipingano coloro che, nell'interrogarsi su come programmare i veicoli a guida autonoma, passano sommariamente in rassegna le principali alternative teoriche in materia di etica normativa, se ne ritraggono, sconcertati dal secolare disaccordo, e si volgono a consultare *online* la popolazione mondiale, onde ricavarne, per via algoritmica, una soluzione del *trolley dilemma* che possa successivamente essere recepita nelle leggi che dovranno regolare il traffico automatizzato e connesso. Gli Stati che, dopo la seconda guerra mondiale, hanno introdotto nei propri sistemi giuridici costituzioni rigide, sovraordinate alla legislazione ordinaria, nelle quali sono tutelati alcuni diritti umani, hanno infatti compiuto con ciò una scelta deontologica di fondo, alla quale devono conformarsi - pena la loro nullità - tutte le leggi ordinarie.
La Commissione etica tedesca ha respinto anche il principio che sta a fondamento dei cosiddetti «dispositivi etici» personalizzati, che sono stati proposti – a partire dalla consueta constatazione sulla mancanza di principi morali condivisi – quale alternativa alle impostazioni predefinite dalle case automobilistiche. Simili dispositivi consentirebbero al proprietario o al conducente del veicolo autonomo di definire preliminarmente la configurazione «etica» del veicolo – ed eventualmente salvarla in un dispositivo elettronico personale[36], attraverso un menu a tendina, come accade per parametri moralmente indifferenti, o attraverso una «manopola etica»[37], in grado di recepire la scelta (altruistica, egoistica o imparziale) sulla distribuzione dei rischi, in caso di incidenti. La scelta egoistica, eventuale, di sacrificare sempre i pedoni o i passanti, piuttosto che i passeggeri del veicolo

---

[32] Ethik-Kommission, *Automatisiertes und Vernetztes Fahren*. Eingesetzt durch den Bundesminister für Verkehr und digitale Infrastruktur, Bericht, Juni 2017, p. 11, https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.html.

[33] Luftsicherheitsgesetz, § 14 Abs. 3.

[34] Bundesverfassungsgericht, Urteil des Ersten Senats vom 15. Februar 2006 - 1 BvR 357/05 - Rn. (1-156), http://www.bverfg.de/e/rs20060215_1bvr035705.html.

[35] *Kant's gesammelte Schriften*, a cura dell'Accademia Prussiana delle Scienze, Berlin, de Gruyter, 1900 ss., 4, p. 429; trad. it. a cura di P. Chiodi con il titolo *Fondazione della metafisica dei costumi*, in I. Kant, *Scritti morali*, Torino, Utet, 1970, p. 88: «*Agisci in modo da trattare l'umanità, sia nella tua persona sia in quella di ogni altro, sempre anche come fine e mai semplicemente come mezzo*».

[36] W. Loh, J. Loh, *Autonomy and Responsibility in Hybrid Systems. The Example of Autonomous Cars*, in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, a cura di P. Lin, K. Abney and R. Jenkins, New York, Oxford University Press, 2017, pp. 35-50: 46; v. anche. J. Millar, *Ethics Settings for Autonomous Vehicles*, in *Robot Ethics 2.0*, cit., pp. 20-34. Critico P. Lin, *Here's a Terrible Idea: Robot Cars With Adjustable Ethics Settings*, «Wired», 2014, https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings.

[37] G. Contissa, F. Lagioia, G. Sartor, *The Ethical Knob: Ethically-customisable automated vehicles and the law*, «Artificial Intelligence and Law», XXV,3, 2017, pp. 365-378; Iidem, *La manopola etica: i veicoli autonomi eticamente personalizzabili e il diritto*, «Sistemi intelligenti», III, 2017, pp. 601-14.

autonomo, sarebbe resa lecita – si sostiene – dallo stato di necessità assunto in ipotesi[38]. La Commissione – valutando che una simile scelta si configurerebbe comunque come una violazione intenzionale e sistematica dei diritti di tali malcapitati passanti – ritiene che non sia «lecito che coloro che sono coinvolti nella produzione dei rischi della mobilità sacrifichino coloro che non sono coinvolti»[39]. Del resto, come è stato osservato, la tesi dell'opportunità di predisporre dispositivi che raccolgano le preferenze morali individuali, per applicarle poi nel traffico automobilistico, si fonda sulla convinzione che siano lasciate alla scelta etica individuale comportamenti in realtà già regolati da norme giuridiche, o comunque di pertinenza di queste[40].

Anche prescindendo dalle questioni affrontate dalla Commissione etica, e in particolare dalle caratteristiche individuali delle potenziali vittime, la struttura concettuale del dilemma del carrello ferroviario – per il suo carattere semplificato, dagli esiti netti e certi, proprio degli esperimenti mentali, e astratto dalle conseguenze sociali di una sua eventuale reiterazione – lo rende inadatto a rappresentare le questioni etiche che si presentano nel traffico stradale reale, giacché queste richiedono piuttosto, data la costitutiva incertezza riguardo alle reazioni dei soggetti coinvolti, una stima delle probabilità e dei rischi[41]. Sarebbe più appropriato discutere, anziché di dilemmi morali, degli standard di sicurezza funzionale – in relazione agli obiettivi pertinenti per la sicurezza dei veicoli autonomi, quali quelli di evitare le collisioni e di restare nella carreggiata – e della valutazione e gestione dei rischi[42]. C'è anche chi argomenta, sulla base di una considerazione degli aspetti tecnici, che sia opportuno prevedere invariabilmente, nei vari casi usualmente classificati come dilemmi del carrello ferroviario, che il veicolo, semplicemente, freni[43].

E' possibile che la tentazione di concettualizzare le «decisioni» dei veicoli autonomi in termini di soluzioni a dilemmi etici derivi dall'assimilazione di tali veicoli ai soggetti umani ch'essi sostituiscono nel ruolo di conducenti: il dilemma del *trolley* è dunque forse fuori luogo, nell'ambito del traffico automatizzato, non solo quanto all'oggetto del dilemma, ma altresì quanto al soggetto a cui si ritiene spetti, in senso proprio, decidere.

---

[38] V., sul tema, F. Santoni De Sio, *Killing by Autonomous Vehicles and the Legal Doctrine of Necessity*, «Ethical Theory and Moral Practice», XX, 2, 2017, pp. 411–29; sulla differenza radicale tra una decisione assunta da un singolo individuo in una situazione d'emergenza e una decisione assunta da vari portatori di interessi su come programmare un certo tipo di tecnologia a rispondere alle situazioni che potrebbero verificarsi, v. S. Nyholm, J. Smids, *The ethics of accident-algorithms for self-driving cars: An applied trolley problem?*, «Ethical Theory and Moral Practice», XIX, 5, 2016, pp. 1275–89.

[39] Ethik-Kommission, *Automatisiertes und Vernetztes Fahren*, cit., p. 11; sulla rilevanza della distinzione tra soggetti coinvolti e non, cfr. D. Hübner, L. White, *Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation*, «Ethical Theory and Moral Practice», XXI, 3, 2018, pp. 685-98.

[40] Cfr. A. Etzioni, O. Etzioni, *Incorporating Ethics into Artificial Intelligence*, «The Journal of Ethics», XXI, 4, 2017, pp. 403-18.

[41] V. S. Nyholm, J. Smids, *The Ethics of Accident-Algorithms for Self-Driving Cars: An applied trolley problem?*, cit.; A. Etzioni, O. Etzioni, *Incorporating Ethics into Artificial Intelligence*, «The Journal of Ethics», XXI, 4, 2017, pp. 403-18; N.J. Gogoll, & J. F. Müller, Autonomous cars: In favor of a mandatory ethics setting, «Science and Engineering Ethics», XXIII, 3, 2017, 681-700.

[42] V. R. Johansson, J. Nilsson. *Disarming the Trolley Problem – Why Self-driving Cars do not Need to Choose Whom to Kill*, in *Workshop CARS 2016 - Critical Automotive applications : Robustness & Safety*, 2016, https://hal.archives-ouvertes.fr/hal-01375606/document; N.J. Goodall, *Away from trolleys and toward risk-management*, «Applied Artificial Intelligence», XXX, 8, 2016, pp. 810-21. Propone di prevedere una circolazione separata dei veicoli autonomi, G. Tamburrini, *Autonomia delle macchine e filosofia dell'intelligenza artificiale*, «Rivista di filosofia», LVIII, 2, 2017, pp. 263-75.

[43] R. Davnall, *Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics*, in uscita in «Science and Engineering Ethics», 2019, https://doi.org/10.1007/s11948-019-00102-6.

3. La moralità del termostato

Lo sviluppo di sistemi di intelligenza artificiale in grado di selezionare e intraprendere, senza un diretto intervento umano, uno tra più corsi di azione alternativi, in contesti in cui tali azioni avrebbero – se compiute da esseri umani – una rilevanza morale, ha dato origine, da circa un decennio, a un dibattito sulla moralità artificiale, quale oggetto di un specifico ambito di ricerca interdisciplinare, che coinvolge informatica, filosofia e robotica: l'etica delle macchine o, più propriamente, l'etica *per* le macchine[44].
Il riconoscimento della sensatezza di un simile campo di indagine si fonda sull'assunzione metaetica che sia possibile dotare le macchine della capacità di decidere e di agire moralmente, ossia che l'etica sia traducibile in termini computazionali e che le macchine possano essere agenti morali.
L'opportunità di qualificare quali agenti morali le entità artificiali, seppur prive di coscienza, di emozioni e sentimenti, di stati mentali e di intenzionalità - e dunque non responsabili, e neppure consapevoli, delle azioni che compiono - è sostenuta da Luciano Floridi e Jeff W. Sanders, quale superamento di un concezione antropocentrica e antropomorfica dell'azione morale, in grado di aprire nuove prospettive di indagine sulla «moralità distribuita»: a un determinato livello di astrazione, un'entità che soddisfi i tre requisiti dell'interattività, dell'autonomia (la capacità di effettuare transizioni interne per mutare il proprio stato) e dell'adattabilità (la capacità di modificare le proprie stesse regole di transizione) può essere considerata un agente e, nel caso in cui stia giocando «il gioco morale», un agente morale.
Anche un termostato ospedaliero, dunque, purché in grado di monitorare, oltre alla temperatura dell'ambiente, anche il benessere dei pazienti, e di regolare di conseguenza la temperatura, potrebbe essere qualificato come agente morale (moralmente buono nel caso in cui i suoi *output* mantengano il benessere dei pazienti entro una soglia prefissata)[45].
L'approccio di Floridi e Sanders è dichiaratamente «fenomenologico»: il livello di astrazione coincide con una collezione di proprietà osservabili, ciascuna con un *set* ben definito di valori possibili; l'attribuzione a un sistema di intelligenza artificiale della proprietà dell'adattabilità dipende quindi dall'ignoranza dell'osservatore, al livello di astrazione prescelto, riguardo al codice o all'algoritmo che ne regola le transizioni.
Il livello di astrazione ritenuto appropriato da Floridi e Sanders per la definizione del concetto di agente morale non include alcuno degli «speciali stati interni», quali gli stati intenzionali – ossia «l'intendere, il desiderare o il voler agire in un certo modo, e l'essere epistemicamente consapevoli del proprio comportamento» – che costituiscono «una condizione carina, ma non necessaria» dell'essere agenti morali. Con ciò, Floridi e Sanders propongono un concetto di «moralità aresponsabile», che pare loro più adeguato al cyberspazio di una concezione della moralità ancorata agli esseri umani (*human-based*).
In una fase di transizione tecnologica e giuridica, in cui la mancanza di controllo e di prevedibilità dei sistemi di intelligenza artificiali richiede che la questione della responsabilità (morale, ma soprattutto giuridica) delle aziende costruttrici di tali sistemi sia affrontata in termini nuovi, la tesi di Floridi e Sanders non può che essere accolta con

[44] Cfr., oltre ai volumi citati più indietro, alla nota 3, l'ottimo contributo di C. Misselhorn, *Grundfragen der Machinenethik*, cit.
[45] L. Floridi, J.W. Sanders, *On the morality of artificial agents*, «Minds and Machines», XIV, 3, 2004, pp. 349–79.

entusiasmo, dai rappresentanti di quelle stesse aziende, giacché difende un «ampliamento della classe degli agenti morali» in virtù del quale «possiamo fermare il regresso del cercare l'individuo *responsabile* quando succede qualcosa di male».
L'operazione di considerare l'agire morale a un determinato livello di astrazione è senz'altro lecita; occorre tuttavia considerare quali conseguenze teoriche e pratiche una simile operazione comporti. Dal punto di vista sociale e giuridico, occorre tener presente il rischio che un simile concetto di moralità aresponsabile e distribuita sollevi gli esseri umani dalle loro responsabilità, generali e particolari, quanto alla direzione dello sviluppo tecnologico e ai suoi specifici prodotti, così che «alla fine nessuno si assuma la responsabilità del loro agire»[46] e il comportamento delle macchine sia disconnesso da quello degli uomini che le disegnano, le costruiscono e le utilizzano[47].
Dal punto di vista teorico, la concezione di una «moralità senza mente» («mind-less morality»), tale da includere gli agenti artificiali nella classe degli agenti morali, coincide con una forma estrema di riduzionismo, che interpreta in senso oggettivo, anzichè meramente esplicativo, l'analogia tra il comportamento conforme a scopi delle macchine e quello umano[48]. Per qualificare i sistemi artificiali quali agenti morali, si sottrae legittimità concettuale alle caratteristiche più propriamente umane, quali la coscienza e l'intenzionalità[49] – degradandone il concetto stesso a residuo di una mentalità magica – in virtù della loro irriducibilità a una descrizione in termini meramente funzionali[50].
Sostenere che un livello di astrazione che prescinde dalla coscienza di sé, dall'intenzionalità, nonché da volizioni e sentimenti, sia appropriato a descrivere un agente morale[51], equivale, paradossalmente – proprio mentre si ritiene di liquidare le teorie dei «fantasmi nella macchina»[52] – a sottoscrivere una precisa tesi metafisica, ossia una forma di dualismo cartesiano che concepisce in modo disincarnato la razionalità, la conoscenza e anche la moralità umane[53] e assimila la mente a un *software*, riproducibile, perciò, per via algoritmica. Una simile concezione si trova in contrasto con le più recenti acquisizioni scientifiche sulla

[46] C. Misselhorn, *Grundfragen der Machinenethik*, cit., pp. 14, 133.
[47] D.G. Johnson, *Computer Systems: Moral Entities, but not Moral Agents*, in *Machine Ethics*, cit, pp. 168-83: 183.
[48] V. F. Fossa, *Fare e funzionare. Sull'analogia di robot e organismo*, «InCircolo», VI, 2018, pp. 73-88. Cfr. P. Moro, *Libertà del* robot*? Sull'etica delle machine intelligenti*, in *Filosofia del diritto e nuove tecnologie. Prospettive di ricerca tra teoria e pratica*, a cura di R. Brighi, S. Zullo, Roma, Aracne, 2015, pp. 525-44.
[49] Sulla questione, non si può, qui, che limitarsi a rinviare a due testi classici: T. Nagel, *What Is It Like to Be a Bat?*, «The Philosophical Review», LVIII, 4, 1974, pp. 435–50; trad. it. a cura di T. Falchi con il titolo *Che cosa si prova ad essere un pipistrello*, Roma, Castelvecchi, 2013; J.R. Searle, *Minds, brains, and programs*, «The Behavioral and Brain Sciences», III, 1980, pp. 417–24; trad. it. a cura di G. Tonfoni con il titolo *Menti, cervelli e programmi: un dibattito sull'intelligenza artificiale*, Milano, CLUP - CLUED, 1984.
[50] W. Wallach, C. Allen, *Moral machines: Teaching Robots Right from Wrong*, cit., p. 53: «la comprensione è talvolta equiparata alla coscienza – un altro termine con connotazioni magiche».
[51] Per una discussione a più voci sulla proposta di Floridi e Sanders, cfr. *The Machine Question: AI, Ethics and Moral Responsibility*, a cura di D.J. Gunkel, J.J. Bryson and S. Torrance, Birmingham, 2012, http://events.cs.bham.ac.uk/turing12/proceedings/14.pdf. Tra i contributi più recenti, offrono una sintetica ricognizione delle alternative teoriche D. Brhdadi, C. Munthe, *Artificial Moral Agency: Philosophical Assumptions, Methodological Challenges, and Normative Solutions*, 2019, https://www.researchgate.net/publication/311196481.
[52] L. Floridi, J.W. Sanders, *On the morality of artificial agents*, cit..
[53] J. Bryson, P. Kime, *Just an Artifact: Why Machines are Perceived as Moral Agents*, in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, a cura di T. Walsh, Menlo Park, AAAI Press, 2011, pp. 1641-46: 1642: «la tendenza alla sovraidentificazione con le macchine deriva da una concezione errata della vita umana, che pone come sue caratteristiche chiave le facoltà del linguaggio, della matematica e della ragione».

natura incarnata della mente e sul ruolo, anche cognitivo, delle emozioni e dei sentimenti[54]. Si fondano su una tale concezione – e ne testimoniano, con i loro scarsi progressi, l'insufficienza – i tentativi compiuti fin qui di dotare del senso comune i sistemi di intelligenza artificiale[55].

Il livello di astrazione identificato da Floridi e Sanders come appropriato a descrivere un agente morale misconosce inoltre la funzione essenziale, in ambito morale, della capacità – attribuita dalle neuroscienze contemporanee al «circuito dell'empatia»[56] e su cui vertono le riflessioni filosofiche sul concetto di riconoscimento[57] – di distinguere le persone dalle cose e di attribuire alle prime un valore incondizionato, una dignità e dei diritti. La scelta di attribuire lo statuto di agenti morali ai soli esseri umani – mantenendo ferma la distinzione tra persone e cose – non si fonda su una mera «speculazione psicologica»[58], incapace di restare paga delle proprietà esteriormente osservabili: nelle carte costituzionali contemporanee, tale scelta è stata sancita giuridicamente in esito a un'esperienza tragica e immane, su che cosa possa comportare, concretamente, la riduzione concettuale delle persone a cose[59].

Non è detto, naturalmente, che una simile riduzione sia sottoscritta da tutti coloro che intendono in senso letterale e non figurato i «dilemmi etici delle macchine». Opera piuttosto nell'immaginario collettivo[60] la tendenza opposta, ossia l'inclinazione antropomorfica ad attribuire alla macchina la facoltà di riflettere, giudicare e compiere in senso proprio una scelta morale, anziché la mera capacità di seguire regole programmate o apprese.[61] Spesso lo slittamento dal senso figurato a un senso letterale del linguaggio – in virtù del quale il lessico tratto dalla biologia è utilizzato per descrivere il funzionamento di macchine, che dunque

[54] A. Damasio, *Descartes' Error: Emotion, Reason and the Human Brain*, New York, Grosset/Putnam, 1994; trad. it. a cura di F. Macaluso con il titolo *L'errore di Cartesio. Emozione, ragione e cervello umano*, Milano, Adelphi, 1995; Idem, *The Strange Order of Things: Life, Feeling, and the Making of Cultures*, New York, Pantheon, 2018; trad. it. a cura di S. Ferraresi con il titolo *Lo strano ordine delle cose*, Milano, Adelphi, 2018.

[55]Cfr. Y. Shoham, R. Perrault, E. Brynjolfsson, J. Clark, J. Manyika, J.C. Niebles, T. Lyons, J. Etchemendy, B. Grosz and Z. Bauer, *The AI Index 2018 Annual Report*, Stanford, Stanford University, 2018, p. 64.

[56] S. Baron-Cohen, *The Science of Evil: On Empathy and the Origins of Cruelty*, New York, Basic Books, 2011; trad. it. a cura di G. Guerrerio con il titolo *La scienza del male: l'empatia e le origini della crudeltà*, Milano, Cortina, 2012.

[57] Cfr. ad es. E. Nowak-Juchacz, *Das Anerkennungsprinzip bei Kant, Fichte und Hegel*, «Fichte-Studien», XXIII, 2003, pp. 75-84.

[58] L. Floridi, J.W. Sanders, *On the morality of artificial agents*, cit..

[59] Simon Baron-Cohen (*La scienza del male*, cit., pp. 1-5) collega la sua volontà di capire, da scienziato, «i fattori che inducono le persone a trattare gli altri come oggetti» a un episodio biografico: il racconto di una reale trasformazione di persone in oggetti (quando aveva sette anni, il padre gli disse che i nazisti avevano trasformato gli ebrei in saponette).

[60] Cfr. B. Henry, *Dal Golem ai cyborgs: trasmigrazioni nell'immaginario*, Livorno, Belforte, 2013; *Roboethics in film*, a cura di F. Battaglia, N. Weidenfeld, Pisa, Pisa University Press, 2015; F. Battaglia, *Macchine morali*, «Scienza e Filosofia», XIII, 2015, pp. 193-202.

[61] K. Weber, *Autonomie und Moralität als Zuschreibung: Über die begriffliche und inhaltliche Sinnlosigkeit einer Maschinenethik*, in *Maschinenethik. Normative Grenzen autonomer Systeme*, a cura di M. Rath, F. Krotz, M. Karmasin, Wiesbaden, Springer, 2019, pp. 193-208.

«imparano», «agiscono» e «decidono» – non è tematizzato, né consapevole[62] e reca in sé le caratteristiche del pensiero magico[63].
Come ha osservato Fabio Fossa, si tratta di un processo di estensione semantica del linguaggio ordinario al quale non è possibile resistere, entro l'uso del linguaggio medesimo. Occorre tuttavia tener presente che utilizzare una medesima terminologia per gli uomini e per le macchine «suggerisce che non vi sia alcuna differenza significativa» tra l'archetipo e la sua imitazione tecnologica e induce a «trattare le macchine come organismi», generando, nei confronti della tecnologia, aspettative irrealistiche, e «gli organismi come macchine», legittimando l'applicazione, agli uomini, di schemi concettuali impropri, di derivazione tecnologica[64].
Il pericolo maggiore, a parere di chi scrive, deriva dall'assimilazione dell'opacità, ossia dell'inspiegabilità del comportamento dei sistemi di intelligenza artificiale, a una loro autonomia morale: l'imprevedibilità che caratterizza le reti neurali artificiali – in grado di apprendere in modo automatico e non sorvegliato, ricavando modelli e strutture a partire da enormi quantità di dati e di assumere decisioni secondo criteri che restano opachi all'osservatore umano – è solo esteriormente assimilabile a quella del comportamento umano. La macchina è una «scatola nera» non nel senso che ci siano sconosciuti i suoi desideri, le sue intenzioni e la sua volontà, ma solo in quanto ha elaborato modelli, a partire dai dati, che non ci è in genere possibile spiegare e sulla cui base la macchina agisce e reagisce: «piccoli cambiamenti nel mondo percepito potrebbero tramutare, in modo imprevedibile, una risposta appropriata in una del tutto inappropriata, con conseguenze potenzialmente letali»[65].
La trasparenza costituisce perciò un requisito cruciale - al momento non soddisfatto, seppure sempre incluso tra i *desiderata* o le raccomandazioni[66] - dei sistemi di intelligenza artificiale: la decisione, umana, di lasciare che macchine o algoritmi producano effetti su altri esseri umani dovrebbe essere condizionata all'intelligibilità del modo in cui agiscono e alla possibilità di ascrivere tali effetti alla responsabilità di esseri umani. L'alternativa, di automatizzare alcuni processi decisionali, delegandoli a sistemi di intelligenza artificiale che

---

[62] Una raccomandazione a includere nella discussione sull'etica dei sistemi autonomi e intelligenti una valutazione critica delle assunzioni antropomorfiche è contenuta nella seconda versione del documento sottoposto alla discussione pubblica dall'Institute of Electrical and Electronics Engineers (IEEE), nell'ambito della «Global Initiative on Ethics of Autonomous and Intelligent Systems» (*Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 2018, pp. 194-95., https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

[63] Cfr. M. Musiał, *Enchanting Robots. Intimacy, Magic, and Technology*, Cham, Palgrave Macmillan, 2019, pp. 63-113.

[64] F. Fossa, *Creativity and the Machine: How Technology Reshapes Language*, «ODRADEK. Studies in Philosophy of Literature, Aesthetics and New Media Theories», III, 1-2, 2017, pp. 177-208.

[65] *When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making*, Informatics Europe & EUACM, 2018, p. 11, https://www.acm.org/binaries/content/assets/public-policy/ie-euacm-adm-report-2018.pdf. Cfr. E. Sirgiovanni, G. Corbellini, *IA e neuroetica: evoluzione spontanea e valori morali*, «Giornale italiano di psicologia»», Fascicolo 1, marzo 2018, pp. 147-52:149.

[66] V. ad es. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, «Minds and Machines», XXVIII, 4, 2018, pp. 689–707; European Commission's High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, 2019, p. 13, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60430; Commission National Informatique & Libertés (CNIL), *Comment permettre à l'homme de garder la main? Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, 2017, p. 51, https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf.

si siano costruiti modelli che ci sono sconosciuti e che non includono le più elementari regole del buon senso umano, comporta la conseguenza di una perdita di prevedibilità e di controllo su quei processi[67], i cui costi e i cui benefici dovranno essere oggetto, di volta in volta, di una attenta valutazione comparativa.

4. Conclusione

Lo sviluppo di macchine in grado di assumere decisioni in modo autonomo – ossia senza il controllo o l'intervento umani – può indurre a ritenere possibili, e lecite, tre operazioni di surroga, variamente intrecciate: del diritto con l'etica, in primo luogo, con la formulazione sul piano morale di dilemmi che implicano discriminazioni vietate, a livello costituzionale, in molti ordinamenti giuridici; della valutazione morale, in secondo luogo, con un modello computazionale delle preferenze socialmente prevalenti e, infine, dell'agire morale umano con il funzionamento autonomo, imprevedibile e opaco, delle macchine[68].
Un dilemma, quale il *trolley problem*, che richieda di decidere quale essere umano sacrificare, nel traffico automobilistico, attiene all'ambito giuridico, non a quello delle scelte morali individuali, e non può essere affrontato assumendo come risolutivo l'esito di un sondaggio planetario. Il diritto circoscrive, nelle carte costituzionali, le versioni in cui sia lecito porre un tale dilemma, escludendo, come si è visto, quelle che comportino violazioni dei diritti umani fondamentali alla vita e alla non discriminazione. Il diritto, che disciplina minuzionamente, nel codice stradale, comportamenti ben meno rilevanti, non può che avocare a sé, nei casi, eventuali, in cui sia davvero necessaria (e non sia più opportuno affrontare la questione in termini di standard di sicurezza o di valutazione e gestione dei rischi) la scelta su chi uccidere in modo sistematico, nel traffico automobilistico.
Quanto ai dilemmi morali, essi non possono porsi a un veicolo a guida autonoma, né a nessuna altra macchina, poiché – nell'attuale fase di sviluppo tecnologico dei sistemi di intelligenza artificiale – una macchina non è un agente morale in senso proprio[69]: non è consapevole di cosa fa, non ha inclinazioni, né desideri ed è priva di empatia, oltre che di senso comune. Che la macchina, anziché eseguire le istruzioni contenute in un programma, applichi modelli elaborati in modo automatico e non sorvegliato, rende il suo comportamento opaco alle spiegazioni umane, ma non, per questo, intenzionale, né responsabile.
La questione cruciale non è dunque se sia o no opportuno costruire agenti morali artificiali[70], giacché questo non rientra, al momento, tra le possibilità tecnologiche date, ma ricordare che le macchine che mimano, in diversi contesti, il comportamento umano, pongono questioni morali agli uomini che le progettano, le costruiscono, le utilizzano e ne regolano giuridicamente ruoli e funzioni[71]: incombe agli uomini l'onere di trovare soluzioni

[67] Cfr. A. Fabris, *La filosofia e lo specchio delle macchine*, «InCircolo», VI, 2018, pp. 28-38: 33.
[68] Sono grata, per l'invito a considerare questo importante aspetto, a uno dei revisori anonimi di questo articolo.
[69] Sull'alternativa tra la tesi della continuità e quella della discontinuità, tra agenti morali umani e artificiali, v. F. Fossa, *Artificial Moral Agents: Moral Mentors or Sensible Tools?* «Ethics and Information Technologies», XX, 2, 2018, pp. 115-26.
[70] Cfr. A. Van Wynsberghe, S. Robbins, *Critiquing the Reasons for Making Artificial Moral Agents*, «Science and Engineering Ethics», 2018, https://doi.org/10.1007/s11948-018-0030-8; A. Poulsen, M. Anderson, S.L. Anderson, B. Byford, F. Fossa, E.L. Neely, A. Rosas, A. Winfield, *Responses to a Critique of Artificial Moral Agents*, 2019, https://arxiv.org/ftp/arxiv/papers/1903/1903.07021.pdf.
[71] Cfr. A. Fabris, *Etica delle macchine*, in «Teoria», XXXVI, 2, 2016, pp. 119-136; M. Verdicchio, *An Analysis of Machine Ethics from the Perspective of Autonomy*, in *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*, ed. by T. Powers, Cham, Springer, 2017, pp. 179-91.

tecnologiche tali da garantire che, nell'elaborare dati relativi a comportamenti umani, le macchine non costruiscano modelli che ne replichino – amplificandone a dismisura gli effetti – i pregiudizi e le discriminazioni[72], e che funzionino invece in modo che non siano lesi i diritti di alcuno. Solo in questo senso, traslato, improprio, metaforico e antropomorfo, può darsi, oggi, un'etica delle macchine[73].

[72] Cfr. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, Crown, 2016; trad. it. a cura di D. Cavallini con il titolo *Armi di distruzione matematica: come i Big data aumentano la disuguaglianza e minacciano la democrazia*, Milano, Bompiani, 2017.

[73] E' corretta la definizione formita da Oliver Bendel: «"Morale delle macchine" è un'espressione tecnica, come "intelligenza artificiale". Allude a un *setting* che gli uomini hanno, e si vogliono imitare o simulare componenti di esso. […] Le macchine morali e immorali non sono buone o cattive, non hanno libero arbitrio né coscienza, né intuizione né empatia» (voce «Moralische Maschinen», in *Gabler Wirtschaftslexikon*, Wiesbaden, Springer Gabler, 2019, https://wirtschaftslexikon.gabler.de/definition/moralische-maschinen-119940). V. anche M.S. Vaccarezza, *Macchine morali: responsabilità e saggezza pratica degli agenti artificiali*, in *Etica e responsabilità*, a cura di F. Miano, Napoli, Orthotes, 2018, pp. 309-18.