DECISIONE UMANA E DECISIONE ROBOTICA
UN'IPOTESI DI RESPONSABILITÀ DA PROCREAZIONE ROBOTICA

di Maria Beatrice Magro
(Professore ordinario di Diritto penale
presso l'Università G. Marconi di Roma)

La mente intuitiva è un dono sacro e la mente razionale è un servo fedele. Abbiamo creato una società che onora il servo e ha dimenticato il dono".
(Albert Einstein)

SOMMARIO: 1. L'automazione della tecnologia di apprendimento automatico o *Auto Machine Learning*. 2. Il problema dell'agire imprevedibile degli agenti artificiali: *the black box alghoritms*. 3. Gli agenti intelligenti agiscono autonomamente, e sono perciò agenti liberi? La tesi funzionalista. 4. La punizione penale degli agenti non umani come la responsabilità degli enti giuridici. 5. Il problema della colpevolezza degli agenti artificiali intelligenti. 6. La questione dello statuto morale delle macchine umanizzate: il comportamentismo metodologico e la computazione emotiva. 7. La tesi strutturalista- ontologica. Le proprietà dell'intelligenza umana: pensiero logico e pensiero creativo intuitivo. 8. La coscienza artificiale e i limiti della computazione emotiva degli agenti artificiali umanizzati. 9. La responsabilità a titolo di colpa (eventuale) dello sviluppatore da procreazione di agenti artificiali. Il principio della produzione robotica responsabile e benefica.

1. Gli agenti artificiali intelligenti (IA) sono tra noi, appartengono al nostro ambiente, un ambiente antropocentrico, concepito a misura e a beneficio dell'uomo, ed ormai popolato da artefatti artificiali creati a suo servizio. È opportuno interrogarsi su quali potranno essere gli sconvolgimenti che, in un prossimo futuro, l'immissione massiccia di IA in ambienti aperti e dinamici produrrà sul sistema della responsabilità giuridica per eventuali eventi dannosi causati dal loro agire.

Per rispondere alla questione occorre prima chiarire quali sono le proprietà degli agenti intelligenti di ultima generazione e accennare alla loro ambizione di replicare, anzi, di superare, ogni facoltà umana, tanto da poter aspirare di averne il medesimo *status* di agente morale.

Innanzitutto, sappiamo che tutti gli algoritmi delle moderne IA accedono e archiviano enormi quantità di dati e informazioni, assai superiore a quella che può conservare la mente umana; sappiamo inoltre che sono dotati di una maggiore e più

veloce capacità logico-computazionale rispetto quella umana, accompagnata dalla capacità di riconoscere l'ambiente e di interagire con la realtà circostante (reattività) e persino di comunicare, coordinarsi, cooperare, negoziare con altri agenti o con gli esseri umani (interattività).

Sappiamo inoltre che le moderne ed evolute forme di Intelligenza Artificiale esibiscono una sofisticatissima capacità di *problem solving*, utilizzando algoritmi che consentono di implementare automaticamente la propria banca dati mediante la codifica di nuove informazioni (c.d. *Auto Machine Learning*)[1]. *Auto ML* consiste nell'automatizzazione della tecnologia di apprendimento automatico, attraverso l'applicazione ricorsiva dell'apprendimento automatico a se stesso. In tal modo le IA utilizzano processi di automazione in grado di inglobare dati non elaborati dal modello di apprendimento iniziale, in grado di fornire una risposta automatizzata all'*input* esterno a prescindere dalle regole logiche impostate dallo stesso programmatore, e persino a prescindere dallo stesso programmatore umano.

L'apprendimento automatico è un nuovo approccio all'automazione e costituisce una svolta epocale nell'evoluzione delle IA. Mentre nella programmazione classica si costruisce un programma per risolvere un problema sulla base di una pregressa ed esatta conoscenza della soluzione del problema (o delle informazioni da cui ottenere la conoscenza necessaria per risolvere il problema), con l'apprendimento automatico, invece, il programmatore costruisce un modello che "trova" e "impara" la soluzione a quel problema stesso, senza che lo stesso programmatore conosca la soluzione o persino abbia previsto il problema in termini matematici. Questa nuova tecnologia supera o elude la mancanza di conoscenza dell'uomo in due sensi: in termini di ideazione del problema e della sua soluzione e in termini di conoscenza degli strumenti con cui la soluzione potrebbe essere prodotta automaticamente. In parole semplici, l'alto grado di automazione di *Auto ML* consente ai "non esperti" di utilizzare modelli e tecniche di apprendimento automatico senza essere competenti in quel campo di conoscenza, e senza neppure fornire alla macchina i dati necessari per elaborare il modello di automatizzazione di un processo.

[1] Alla fine degli anni '80 e '90, la prima forma di intelligenza artificiale è stata costruita come un "sistema esperto" la cui conoscenza di base veniva codificata sotto forma di regole logiche da ingegneri e sviluppatori del *software*. Tuttavia, si rivelava impossibile codificare manualmente tutte le conoscenze necessarie per affrontare situazioni complesse, e perciò l'evoluzione delle intelligenze artificiali è andata verso sistemi di autoapprendimento che consentissero alle macchine di accedere ai dati conoscitivi da sole. E' nato così *Auto ML* che costituisce l'automazione del processo di applicazione della tecnologia di apprendimento automatico; J. Steinhoff, *The Automation of Automating Automation: Automation in the AI Industry*, Comunicazione al *Canadian Communication Association Annual Conference*, in *UBC* 2-6 giugno 2019; U. Pagallo, *From Automation to Autonomous Systems: a legal phenomenology with problems of accountability*, in *Proceedings of IJCAI-17* 2017, 17.

Dunque, *Auto ML* aggiunge alla macchina un *quid pluris* rispetto la iniziale programmazione (c.d. *deep learning*): perciò questi sistemi informatici sono in grado di reagire (cioè elaborare modelli decisionali) anche a quelle situazioni che non sono neppure previste dal programmatore, né interpretabili o spiegabili sulla base delle regole con cui l'essere umano ha costruito il modello originario di operatività. Tutto ciò consente ai sistemi di *cognitive computing* di risolvere problemi di grande complessità con risultati ottimali che migliorano sempre più nel tempo anche di fronte a dati informativi non noti[2]. Grazie agli algoritmi appartenenti al sistema di *deep learning* e di apprendimento automatico i robot intelligenti possono prendere "decisioni" individuali del tutto autonome, sottratte al controllo umano che, retrospettivamente, neanche il sistema o il programmatore comprende più completamente.

Questi sistemi informatizzati esperti, se immessi nell'ambiente aperto e dinamico, sono in grado di reagire a situazioni non previste dal programma in modo difforme alle regole con cui l'essere umano li ha programmati.

2. Uno dei problemi connessi alla "responsabilità connessa all' uso di IA" è che di solito sono numerosi gli utenti umani, le organizzazioni, le componenti meccaniche e gli algoritmi coinvolti nella produzione dell'evento dannoso, con conseguente grandi difficoltà nella individuazione dei soggetti responsabili[3].

In risposta al problema, si è elaborato il concetto di "responsabilità distribuita", proprio con riferimento a queste ed altre forme di agentività distribuita e frammentata, che tuttavia non risolve il problema pratico sul come distribuire la responsabilità su tutti questi soggetti, in relazione a ciascun contributo e alle interazioni con il sistema informatico automatizzato[4].

Gli esempi sono innumerevoli. Consideriamo ad esempio l'incidente d'auto Uber a guida autonoma avvenuto nel marzo 2018 in Arizona che ha causato la morte di un

[2] Le moderne IA sono macchine interattive, reattive e proattive, non solo in grado di accedere alle informazioni memorizzate e di interpretare l'informazione, ma anche di conservare le informazioni al fine di utilizzarle nei successivi processi decisionali, le cui regole decisionali, necessariamente abbozzate dal programmatore, vengono elaborate solo dopo una rilevazione statistica da parte del robot. Inoltre la tecnologia di automazione del processo di apprendimento automatico in macchine interattive e reattive fa sì che anche l'obiettivo o il fine cui sono parametrate le *performance* non sia predeterminato dai programmatori originari, o che non sia da loro interpretabile, consentendo quindi un'evoluzione della macchina che va ben oltre i confini dell'originaria programmazione.

[3] Questo aspetto della attribuzione e della distribuzione di responsabilità viene definito come "*the problem of many hands*", in *Responsability and IA*, *Council of Europe Study* 2019.

[4] M. Coeckelbergh, Arti*ficial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability* in *SEE* 2019, disponibile su https://doi.org/10.1007/s11948-019-00146-8; J. P. Günther, *Roboter und rechtliche Verantwortung – eine Untersuchung der Benutzer- und Herstellerhaftung*, Monaco 2016.

pedone indentificato come un "falso positivo"[5]. Chi è il responsabile? Potrebbero esserlo gli sviluppatori del *software*, la casa automobilistica che ha prodotto la macchina, Uber proprietaria della macchina, l'utente dell'auto, il pedone incauto e persino il soggetto pubblico regolatore che ha consentito di testare l' auto a guida autonoma senza adottare le dovute cautele (cioè lo stato dell'Arizona)[6]. Oppure consideriamo il caso di un *chatbot* (*software* che dialoga con umani) di Twitter che, dopo l'interazione con gli utenti, ha iniziato a produrre commenti razzisti e misogini. Anche in questo caso sono coinvolti molti soggetti: gli sviluppatori del processo, il *designer* del *software*, ma anche l'azienda e persino coloro che interagivano con il *chatbot*. Si pensi ancora al sistema automatizzato e velocissimo delle transazioni finanziarie, al *trading* di alta frequenza, o al pilota automatico di grandi velivoli commerciali, i cui i piloti umani non sono in grado di partecipare o inibire tempestivamente la guida, o ancora al sistema antimissile velocissimo e automatizzato che consente, grazie alla sua velocità di reazione, di intercettare l'attacco nemico che l'uomo, con i suoi limiti cognitivi, non avrebbe mai il tempo di ostacolare[7].

Tanti sono i soggetti e le componenti che interagiscono nella produzione dell'evento dannoso finale. Chi di loro può essere considerato responsabile, se come abbiamo detto, gli agenti artificiali, sebbene inizialmente programmati da agenti umani, non sono costituiti da agenti umani, né agiscono attraverso agenti umani? Una volta programmato, l'IA non fa più affidamento sul programmatore, interagisce con il mondo senza la necessità che il programmatore funga da burattinaio. Per le ragioni descritte nel paragrafo precedente, il comportamento di *Intelligent Agent* è imprevedibile non solo quando si trova in una situazione per la quale non è stata

[5] Dopo oltre un anno di indagini, la Commissione della *National Trasportation Safety Board* (NTSB) ha individuato la colpa sia dell'agente umano che della macchina. È emerso infatti che il pilota-tester fosse stato distratto dal suo cellulare durante il viaggio (guardava video in *streaming*), e che non erano state predisposte adeguate procedure di valutazione del rischio per la sicurezza da parte del gruppo *Advanced Technologies* di Uber. Mancava, in sostanza, un responsabile della sicurezza incaricato dall'azienda per la valutazione e l'attenuazione dei rischi. Sotto accusa, nel report del NTSB, anche le telecamere dell'auto che tutto controllavano fuorché la possibile disattenzione del co-pilota impedendo una supervisione che vigilasse alla violazione delle regole imposte dalla stessa azienda. Guidare - ovvero prestare attenzione - con questi veicoli può essere molto noioso, e Uber si è resa responsabile della mancanza di un sistema che impedisse la disattenzione dell'agente umano addetto al controllo. D'altra parte Tesla, il produttore delle auto a guida autonoma si è difesa dicendo che la sua tecnologia non rappresenta un sistema di guida autonoma completo, ma di sola "assistenza" alla guida e dunque il guidatore deve sempre prestare attenzione alla strada ed essere pronto ad intervenire ogni volta che il sistema lo richieda, *La repubblica*, 21 novembre 2019.

[6] E. Hilgendorf, *Automatisiertes Fahren. Und Recht. Ein Ueberblick*, in *JA* 2018, 801 ss.; C. Staub, *Strafrechtliche Fragen zum Automatisierten Fahren*, in *NZV* 2019, 320 ss.

[7] Il fenomeno è stato descritto come un incolmabile "divario di responsabilità" da A. Matthias, *The responsibility gap: Ascribing responsibility for the actions of learning automata*, in *EIT,* 2004,175.

programma una risposta adeguata, ma anche quando, a causa di una nuova "esperienza", inizi a modellare i dati acquisiti autonomamente.

Qui sta l'origine di una grande inquietudine: l'uomo non può né prevedere, né controllare totalmente il comportamento dell'agente artificiale in situazioni non pianificate. A differenza delle macchine intelligenti tradizionali, questi agenti di nuova generazione sono sottratti al controllo umano e prendono decisioni autonome, flessibili e non predeterminate nella misura in cui i loro sensori consentono loro di analizzare e di elaborare i dati provenienti dall'ambiente. Quanto più i loro algoritmi sono duttili e capaci di migliorarsi interagendo con il mondo esterno, quanto più è grande la quantità d'informazioni che elaborano, tanto maggiore è la loro capacità computazionale e la capacità di modellare i dati su cui operano autonomamente. In poche parole, il loro comportamento può essere *ex ante* imprevedibile da un punto di vista che non è solo soggettivo (cioè del programmatore costruttore che non lo ha previsto), ma – verrebbe da dire – ma da un punto di vista oggettivo- tecnologico. Si parla perciò di *black box algorithms* per indicare che tra i dati di *input* e i comportamenti tenuti come *output* vi sia un'opacità, un vuoto di comprensione da parte dell'osservatore umano esterno, che fa sì appunto che le condotte di queste IA siano gravate, in un'ottica *a priori*, da un ineliminabile margine di imponderabilità[8]. Anzi, potremmo dire che, in una certa misura, l'imprevedibilità dell'agente intelligente è "pre- programmata", e con essa i rischi associati a terzi.

La fabbricazione e l'immissione nell'ambiente di agenti intelligenti pone quindi seriamente in crisi il sistema della responsabilità penale dell'uomo a titolo di colpa a causa dell'evoluzione non programmata e non prevedibile del loro comportamento. Dal momento che gli agenti intelligenti utilizzano informazioni provenienti dall'ambiente in modo indipendente e agiscono attenendosi ai risultati di questa

[8] U. Pagallo, *Saggio sui robot e il diritto penale*, in *Scritti in memoria di Giuliano Marini*, a cura di Vinciguerra e Dassano, Napoli 2010, 595 s.; S. Riondato, *Robot: talune implicazioni di diritto penale*, in *Tecnodiritto. Temi e problemi di informatica e robotica giuridica*, a cura di P. Moro e C. Sarra, Milano 2017, 85 ss.; F. Basile, *Intelligenza artificiale e diritto penale: quattro possibili percorsi di indagine*, in www.*dirittopenaleuomo.org*, 29 settembre 2019; A. Perin, *Standardizzazione, automazione e responsabilità medica. Dalle recenti riforme alla definizione di un modello d'imputazione solidaristico e liberale*, in *BLJ* 2019, 207 ss; S. Beck, *Google Cars, Software Agents, Autonomous Weapons Systems – New Challanges for Criminal Law*, in *Robotics, Autonomics and the Law*, a cura di Hilgendorf e Seidel, 2017, 217 ss.; H. Surden e M.A. Williams, *Technological Opacity, Predictability, and Self-Driving Cars*, in *CLR,* 2016, 157 ss.; S. Doncieux e J.B. Mouret, *Beyond black-box optimization: a review of selective pressures for evolutionary robotics*, in *EI* 2014, 71 ss.; S. Gless, E. Silverman, e T. Weigend, *If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability*, in *NCLR* 2016, disponibile su https://doi.org/10.1525/sp.2007.54.1.23.
Inoltre mi si permetta il rinvio M. B. Magro, *Robot, cyborg e intelligenze artificiali*, in *Cybercrime*, a cura di A. Cadoppi, M. Canestrari, A. Manna, M. Papa, Torino 2019, 1080 ss.; Id., *Biorobotics, robotics and criminal law: some hints and reflections,* in *PC* 2016, 235 ss.; Id., *Biorobots, robotics and criminal law,* in *GRLP*, a cura di S. Riondato, D. Provolo, F. Yenisey, Padova 2014, 477 ss.

valutazione senza subire un'ulteriore interferenza da parte dell'uomo, l'operatore umano (programmatore o utilizzatore) che lo monitora può non specificamente prevedere (se non in termini del tutto generici) quali saranno i modelli di comportamento che il robot sceglierà nel riconoscere e interpretare la realtà, né potrà attivare un sistema di controllo manuale da remoto, che giungerebbe inevitabilmente troppo tardi. Come possono gli esseri umani essere ritenuti responsabili se non possono esercitare un completo controllo *ex ante* su questi agenti veloci e automatizzati?

3. La diffusione e immissione generale nell'ambiente antropocentrico di agenti intelligenti pone una domanda di fondo: l'agire autonomo delle IA intelligenti è solo una deviazione aleatoria del progetto iniziale dovuta a cause accidentali o è ispirato ad una sorta di "*libero arbitrio artificiale*"?[9]

Fino a poco tempo fa, lo studio della mente umana e della robotica appartenevano a campi del tutto estranei per competenze e metodologie. Oggigiorno, la distanza tra i due campi di ricerca è invece ridotta, fino quasi a lambirsi reciprocamente, e ciò a causa del progresso della modellazione computazionale artificiale delle moderne ed evolute IA[10]. Per migliorare l'apprendimento artificiale, la ricerca scientifica si è infatti ispirata ai processi di apprendimento neurologico tipico dei bambini, e all'idea secondo cui l'intelligenza umana non è solo cervello, ma è legata al corpo e all'esperienza corporea con il mondo esterno. Gli evoluti robot artificiali dotati di *AutoML* percepiscono l'ambiente esterno mediante sensori, e agiscono in base alle loro percezioni, quindi esibiscono un tipo di intelligenza simile a quella umana.

Il paradigma teorico di base di questo indirizzo suppone una perfetta - o quasi-analogia tra mente umana e computer, un isomorfismo mente-computer, non in senso *ontologico-strutturale*, ma in senso *funzionale*. Non vi sarebbe alcuna differenza funzionale tra intelligenza umana e intelligenza artificiale (il *software*), perché ciò che

[9] Franklin Foer in *World Without Mind*, si chiede: *algorithms are meant to erode free will, to relieve humans of the burden of choosing* ponendo la questione del "libero arbitrio" degli agenti artificiali, F. Foer, *World Without Mind. The existential threat of big tech,* 2017. Cfr. W. Wallach e C. Allen, *Moral machines – teaching robots right from wrong*, Oxford 2009.

[10] L'intersezione tra robotica, neuroscienze e biologia della mente si manifesta maggiormente nel tentativo di imitare, con la massima precisione possibile, il cervello e in particolare quei meccanismi che sottendono la pianificazione e l'esecuzione di azioni negli esseri umani e altri primati qualificati, in modo da giungere alla definizione e all'implementazione di un modello funzionale artificiale delle aree cerebrali capace di replicare, anzi, di sorpassare la mente umana. La biologia sintetica costituisce un pioneristico tentativo di riprodurre artificialmente l'architettura neuronale del sistema nervoso umano così da progettare agenti artificiali totalmente intercambiabili con l'uomo.

cambia è la sede o il supporto fisico (l'*hardware*): il cervello e i suoi circuiti neurologici nell'uomo e le parti metalliche ed elettriche nel computer. Se l'intelligenza è espressione dell'attività neuronale, allora anche la riproduzione artificiale di tale attività deve riprodurre, per ciò stesso, l'intelligenza umana; se gli stati mentali umani sono qualificati dalla loro funzione, indipendentemente dalla specificità materiale del sistema neurale organico, è possibile supporre che sia riproducibile un sistema neurale artificiale ottimale, una mente artificiale intelligente o super-intelligente (c.d. IA forte)!

Questa tesi, denominata funzionalista, assume che le funzioni cognitive di un essere umano siano indistinguibili da quelle della macchina, anzi, che i sistemi artificiali di *cognitive computing* godano di una maggiore intelligenza, in quanto esibiscono una logica computazionale- conseguenziale e capacità di modellazione superiore a quella degli esseri umani, che invece sono afflitti da una "razionalità limitata". Le IA non solo possono riprodurre l'attività mentale dell'uomo, ma possono persino migliorare le prestazioni umane, correggere ed evitare i loro errori, le loro distorsioni e limitatezze cognitive, e prendere quindi decisioni assai più veloci, più efficaci ed utili[11].

Certamente gli agenti artificiali sono dotati della libertà di evolversi e cambiare i loro programmi e i loro modelli comportamentali; quindi esercitano una loro libertà, quella libertà che è propria dei robot intelligenti. Ma non è la stessa *libertà cosciente* che esercita l'uomo, la cui architettura cerebrale costituisce un mistero inspiegabile. Casomai, il vero problema non è se le IA siano libere o meno, ma se possano essere considerate responsabili dal punto di vista penalistico. Il problema della responsabilità non è vincolato alla definizione di quale misura di libertà goda l'agente artificiale. Si può essere responsabili senza assumere un concetto per tutti gli agenti (artificiali o no) univoco e incontestato di libertà[12]. Quindi il problema è chiarire se l'agente artificiale, così come gli enti artificiali, possa essere considerato in un prossimo futuro, un'entità responsabile anche penalmente. Secondo una estremista concezione normativa della responsabilità, la libertà di azione non è un fenomeno naturale di cui occorre individuare i referenti ontologici, ma un mero attributo normativo dell'agire del responsabile autore del reato, funzionale agli scopi dell'organizzazione sociale.

Questo concetto di libertà e di colpevolezza potrebbe fornire una buona base di partenza per progettare sistemi normativi di attribuzione della responsabilità giuridica delle IA.

[11] C.R. Sunstein, *Algorithms, correcting biases. Forthcoming*, in *SR* 2018, disponibile su https://ssrn.com/abstract=3300171.

[12] In proposito, M. *Simmler* e *N. Markwalder, Roboter in der Verantwortung? – Zur Neuauflage der Debatte um den funktionalen Schuldbegriff*, in *ZSTW* 2017, 20 ss.

4. L'idea della responsabilità penale delle IA forti come entità autonome raccoglie ormai diversi proseliti. Oltre al primo sponsor, Gabriel Hallevy[13], si annoverano altri sostenitori convinti della opportunità di esportare il modello della responsabilità degli enti giuridici anche ad altri enti legali non umani, adducendo molti di quegli argomenti che hanno animato il dibattito sulla responsabilità penale dell'ente giuridico[14]. Uno di questi sostiene che il *deficit* di responsabilità- qualora non vi fossero spazi per la responsabilità umana – produrrebbe inevitabili e scivolose conseguenze incentivando prassi perverse[15].

Il tema quindi presenta forti analogie con quello dell'affermazione di un'etica aziendale e di una conseguente responsabilità da reato degli enti collettivi, sistema con cui potremmo confrontarci nell'ipotesi di danni arrecati dall'agire autonomo di IA. Le analogie sono evidenti: gli enti collettivi non hanno né corpo né anima, ma sono comunque "soggetti giuridici", cioè autori di reati (per il tramite delle persone fisiche incardinate in essi) per la legge penale; i robot invece hanno un "corpo" fisico che interagisce con l'ambiente tramite sensori, una materia su cui far ricadere la sanzione penale (ad esempio, la disattivazione o riprogrammazione della macchina o la sua distruzione) e sono dotati di autonomia decisionale. Ciò consente di ipotizzare un sistema di responsabilità dell'agente artificiale che, sulla falsariga della responsabilità amministrativa da reato degli enti, nel capovolgerne i presupposti, renderebbe responsabile l'agente artificiale soggettivizzato per il suo operare e per quello dell'uomo *frontman*, ovvero l'utilizzatore, il programmatore, il designer, il produttore, etc.

Vi sono certamente molte somiglianze tra ente giuridico collettivo e ente artificiale ma, come Peter Asaro ha osservato giustamente, anche molte differenze[16]. Queste differenze impongono di pensare che la logica sottostante alla scelta di punizione

---

[13] G. Hallevy, *The Criminal Liability of Artificial Intelligence Entities– from Science Fiction to Legal Social Control*, in *AIPJ* 2010, 271 ss.; Id., *When Robots Kill: Artificial Intelligence Under Criminal Law*, London, *UPNE*, 2013.

[14] Lawrence Solum scrive: "*The problem of punishment is not unique to artificial intelligences, however. Corporations are recognized as legal persons and are subject to criminal liability despite the fact that they are not human beings*", L. Solum, *Legal Personhood for Artificial Intelligences*, in *NCLR* 1992, 1248, disponibile su https://doi.org/10.3868/s050- 004-015-0003-8.

[15] S. Chopra e L. White, *A Legal Theory for Autonomous Artificial Agents*, in *A Legal Theory For Autonomous Artificial Agents* 2011, 169, disponibile su https://doi.org/10.3998/mpub.356801.

[16] *A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics*, in *Robot Ethics: The Ethical and Social Implications of Robotics* 2011, 170, 182. In proposito, A. Cappellini, *Machina delinquere non potest. Brevi appunti su intelligenza artificiale e responsabilità penale,* in *Criminalia*, 2018, 499 ss.; M. Bassini, L. Liguori, O. Pollicino, *Sistemi di Intelligenza Artificiale, responsabilità e* accountability. *Verso nuovi paradigmi?*, in Intelligenza artificiale, protezione dei dati personali e regolazione, a cura di F. Pizzetti, Torino 2018, 334.

dell'ente giuridico non sia così proficua se riportata nel contesto degli agenti di intelligenza artificiale per una serie di considerazioni che partono tutte da una constatazione: a differenza delle società, gli agenti non sono costituiti da agenti umani.

Innanzitutto, punire un agente artificiale non avrebbe alcun effetto dissuasivo e deterrente sugli umani che lo utilizzano, programmano o immettono nell'ambiente: mentre le persone umane svolgono un ruolo fondamentale e costitutivo nel processo decisionale aziendale e societario, non svolgono un ruolo simile nell'agire dell'agente artificiale, anzi possono essere del tutto estranei al processo decisionale robotico.

Inoltre, la punizione penale del robot intelligente (ovvero la sua distruzione o riprogrammazione) sarebbe inutile sotto il profilo dell'efficacia deterrente anche se rivolta all'ente artificiale in sè considerato. Chi è privo di una "buona o cattiva coscienza", anche intesa come entità legale, in quanto privo di un sistema etico di riferimento, difficilmente può essere considerato soggetto responsabile, perché non è possibile con lui instaurare alcun dialogo etico, non è possibile muovere alcun rimprovero né avviare alcun percorso di recupero o di riparazione[17].

Oltre alla difficoltà di immaginare una reale possibilità di instaurare un dialogo comunicativo finalizzato ad obiettivi di prevenzione speciale, anche la possibilità che un agente intelligente sia sottoposto ad una sanzione economica è al momento difficile da immaginare, visto che un agente intelligente – di per sé – non ha risorse proprie.

In tal caso la misura sanzionatoria della distruzione fisica o della riprogrammazione dell'agente artificiale verrebbe, in definitiva, a ricadere indirettamente sul proprietario o utilizzatore umano o ente giuridico (magari incolpevole) e non sull'agente intelligente.

Certamente la previsione di una diretta responsabilità a carico di in robot intelligente può indirettamente motivare il suo creatore a modulare diversamente il suo *design* in modo da prevedere meccanismi di controllo umano, o apponendo limiti all'apprendimento evolutivo, o sollecitare il proprietario a usare ancora maggiori cautele. Ma giudicare o inibire la produzione di un dispositivo pericoloso o dannoso è piuttosto diverso dal considerare moralmente responsabile il dispositivo[18].

5. Inoltre, sotto il profilo penalistico, la questione della responsabilità giuridica dei robot intelligenti deve passare attraverso le forche caudine del principio di

[17] A proposito della impossibilità di ipotizzare una giustizia riparativa robotica, Cfr. H. Adzi e D. Roio, *Restorative Justice in Artificial Intelligence Crimes*, in *JDC*, novembre 2019; si veda anche M.B. Magro, sul fondamento neuroscientifico della giustizia riparativa, *Neurosciences and Restorative Justice*, in *Restorative Approach and Social Innovation: from theoretical Grounds to sustainsable Practices*, a cura di G. Grandi e S. Grigoletto, Padova 2019, 131 ss..

[18] V. R. Bhargava, M. Velasquez, *Is Corporate Responsibility Relevant to Artificial Intelligence Responsibility?,* in *GJL&PP* 2019, 829 ss..

colpevolezza e delle sue interpretazioni. Il problema della responsabilità penale delle IA si scontra inevitabilmente con quello della individuazione di stati mentali o soggettivi artificiali dei robot soprattutto se assumiamo una concezione della colpevolezza che rivendichi un fondamento ontologico di tipo psichico. Perciò, per concepire una sanzione punitiva "personalmente" a carico di un agente artificiale, dovremmo immaginare che questo sia in grado di percepire o comprendere la sua condotta; dovremmo cioè ravvisare un requisito minimo ontologico che consenta di attribuire la soggettività giuridica (compresa la titolarità di diritti) e quindi anche la responsabilità penale per il suo comportamento[19].

Il punto è che ritenere responsabile un agente artificiale suppone, oltre al convincimento sul suo *status* morale e giuridico, che questo sia capace di stati mentali come quelli umani, che sia capace di una risposta emotiva negativa, come il risentimento, l'indignazione, la vergogna, la sofferenza. Insomma, è possibile che, in un ipotetico futuro, gli agenti intelligenti riescano a sviluppare coscienza, sentimenti, empatia e moralità, così da poter infliggere loro una punizione?

6. Le forme più raffinate e evolute di IA tendono a replicare anche gli aspetti più intimi e essenziali della intelligenza umana, sviluppando robot antropomorfici in grado di percepire e interagire alle emozioni umane, dotati di una sorta di emozioni artificiali e coscienza artificiale che li rende più simili a noi (c.d. computazione emotiva)[20]. Questi raffinati robot intelligenti hanno rinvigorito il dibattito sul loro stato morale di soggetti.

La teoria che fornisce risposta positiva alla questione è denominata "*comportamentismo metodologico*", o della "*equivalenza performativa*", secondo la quale i robot possono avere uno *status* morale significativo se sono approssimativamente equivalenti dal punto di vista delle prestazioni ad altre entità a cui si riconosce comunemente quello stato morale[21]. Ciò significa che se un robot si

[19] D. Gunkel, *The other question: Can and should robots have rights? Ethics and Information Technology,* 2018, 87 ss.

[20] La computazione emotiva è un sistema che ha accesso allo stato emotivo degli utenti, sulla base di dati biometrici correlabili a risposte emotive (sudorazione, dilatazione delle pupille e così via), che di dati comportamentali (quali gesti, postura e altri comportamenti manifesti dei soggetti) e che consente all'agente intelligente di interagire con l'uomo nel modo più appropriato: in proposito, E.A. Wilson, *Affect and Artificial Intelligence*, in *UWP* 2011; C. Pelauchaud, *Modelling Multimodal Expression of Emotion in a Virtual Agent*, in *PTBSCEMM* 2009, 3539 ss.; F. Fallon, *Integrated Information Theory (IIT) and Artificial Consciousness*, in AA.VV., *Advanced research on biologically- Robotic,* Pennsylvania 2017 23 ss.; J. Kaplan, *Intelligenza artificiale. Guida al futuro prossimo,* Roma 2018.

[21] J. Danaher, *Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism*, in *SEE*, disponibile su https://doi.org/10.1007/s11948-019-00119-x; D. Gunkel, *Robot Rights*, Cambridge, *MITPress* 2018; R. Sparrow,

comporta come un'altra entità a cui attribuiamo uno *status* morale, allora gli dovrebbe essere riconosciuto lo stesso *status* morale.

Il problema è che il "*comportamentismo metodologico*" ben potrebbe accontentarsi di una soglia performativa piuttosto bassa: certamente non è sufficiente vestire una macchina come un essere umano, farlo apparire come un essere umano, per riconoscerle lo *status* di soggetto autonomo titolare di diritti e doveri[22]. Perciò l'artificio performativo deve spingersi a chiedere qualcosa di più, cioè un "apparato mentale interno" che consenta al robot di sentire, pensare e vedere il mondo in un modo simile al nostro[23]. Se riteniamo che solo la capacità di comprensione simbolica ed emotiva e la consapevolezza di un Se pensante possano fornire *standard* abbastanza elevati per il riconoscimento della soggettività, occorre – quantomeno - che queste caratteristiche siano riscontrabili in termini comportamentali.

Gli psicologi comportamentisti pensano sia scientificamente scorretto postulare l'esistenza di stati mentali interni non osservabili che si ritiene siano costitutivi dell'intelligenza umana come la sensibilità, la consapevolezza, la coscienza fenomenica. Costoro ritengono che tutto ciò che possiamo fare è dedurre tali stati mentali – la cui esistenza è solo ipotizzata metafisicamente perché non possono essere conosciuti e osservati direttamente - per il tramite di comportamenti esterni[24]. Solo la presenza (o l'assenza) di prestazioni comportamentali coerenti con l'esistenza di una coscienza fenomenica di livello superiore potrebbe qualificare i robot antropomorfici come soggetti, non la presenza di qualche essenza metafisica interiore, in sé e per sé e considerata, che non è qualcosa che può essere osservato e misurato dal punto di vista scientifico.

Can machines be people? Reflections on the turing triage test, in *RETESIR*, a cura di P. Lin, K. Abney e G. A. Bekey, Cambridge 2012, 301 ss.

[22] Si pensi al robot Sophia e alla sua raffinata capacità di riprodurre l'essere umano nella locomozione, gestualità e nell'interazione, tale da farle guadagnare – provocatoriamente - la cittadinanza saudita, peraltro paradossalmente proprio in un paese in cui non tutti i esseri umani ne godono (lavoratori e donne). In proposito, J. Vincent, *Pretending to give robots citizenship helps no one*, in *TVerge* 2017, disponibile su https://www.theverge.com/2017/10/30/16552006/robot-rights-citizenship-saudi- arabia-sophia; Z. Stone, *Everything You Need To Know About Sophia, The World's First Robot Citizen*, in *Forbes* 7 novembre 2017, disponibile su https://www.forbes.com/sites/zarastone/2017/11/07/everything-you-need-to-know-about-sophia-the- worlds-first-robot-citizen/#4e76f02b46fa.

[23] Quindi, ad esempio, se i robot intelligenti si dovessero comportare come se provassero dolore, e se la capacità di sentire dolore fosse un presupposto dello *status* morale di persona, allora dovremmo riconoscere loro soggettività morale. Si consideri che il comportamentismo metodologico è alla base del classico test di Turing; A. Turing, *Computing Machinery and Intelligence*, in *Mind* 1950, 433 ss.

[24] Per comprendere questa teoria, è importante chiarire che il concetto di "comportamento" va interpretato in senso lato, come comprensivo non solo di comportamenti fisici esterni (cioè il movimento di arti e labbra) ma inclusivo anche delle attività cerebrali direttamente osservabili e registrabili.

I moderni neuroscienziati cognitivi dubitano che le osservazioni dei circuiti neurologici non siano direttamente corrispondenti a stati mentali, e perciò ricorrono al comportamentismo per fornire una prova dell'esistenza degli stati mentali altrimenti imperscrutabili. Infatti, dall'osservazione dell'attività neurologica del cervello, possono ben dedurre solo delle *correlazioni* tra quei modelli cerebrali e gli stati mentali, che dovranno essere verificate attraverso test comportamentali[25].

Perché si dovrebbe accogliere il comportamentismo etico o metodologico? La ragione è ovvia: esso rispetta i nostri limiti epistemici, ovvero l'incapacità della nostra conoscenza di accedere alle proprietà metafisiche delle cose in sé, di cui abbiamo accesso sempre e solo attraverso alle sue rappresentazioni.

Il comportamentismo vuole fornire una risposta alle posizioni degli strutturalisti secondo cui la capacità dei robot intelligenti di interagire anche a livello emotivo non significa affatto che queste macchine provino "vere" emozioni e siano dotati di un pensiero cosciente, cioè non vadano oltre una mera *simulazione* di emozioni, sentimenti, stati mentali umani coscienti, sostenuta dalla nostra proiezione. Il *computer* può anche essere un ottimo manipolatore formale di simboli, ma questa attività non corrisponde ad una reale "*comprensione*" dei significati di tali simboli. Il sistema simbolico-computazionale si comporta come "se" capisse la realtà circostante, ma non riproduce l'attività umana. La sua dimensione soggettiva di tipo qualitativo è simulata: i robot possono simulare compassione, paura, rabbia, pazienza, ma non provare compassione, paura, rabbia, pazienza[26].

Un'obiezione correlata - e forse quella che cattura il disagio di molte persone sul robot Sophia e i suoi presunti diritti di cittadinanza - è che qualsiasi equivalenza performativa tra robot e umani possa essere raggiunta, essa è frutto di simulazioni, artifici, inganni, ma non è espressione del processo decisionale umano o di stati mentali. I produttori di robot faranno di tutto affinchè le loro creazioni imitino alcuni segni comportamentali che associamo ad esseri con un significativo stato morale, ma a causa della natura interna dei robot, questi segnali comportamentali non saranno

[25] Ad esempio, un neuroscienziato comportamentista può affermare che una particolare attività cerebrale visibile attraverso la tecnologia di *neuroimaging* sia correlata ad uno stato mentale, solo dopo aver chiesto all'individuo sottoposto alla prova neuroscientifica cosa ha provato mentre la neuro-immagine registra quell' attività neurologica.

[26] La posizione comportamentista risponde all'obiezione ribadendo la sua metodologia: se si pensa che certi stati metafisici siano le "vere" basi per l'attribuzione dello *status* morale e che tali stati mentali siano accessibili solo attraverso lo *standard* performativo comportamentale, significa che l'entità artificiale che si comporta *come se* provi dolore o altri stati mentali o proprietà mentali associate allo *status* morale di soggetto, non si potrà certo dire che tali esibizioni siano "false" o "ingannevoli" perché si sospetta che l'entità artificiale manchi di qualche essenza metafisica interiore.

mai correlati alle proprietà metafisiche che noi attribuiamo all'uomo, come la coscienza. È la *performance* a testare l'essenza metafisica.

7. La tesi strutturalista-ontologica sostiene al contrario la non autenticità del pensiero meccanico e la diversità ontologica-qualitativa tra intelligenza artificiale e intelligenza naturale: se è differente l'*hardware*, è diverso anche il *software*! I suoi sostenitori evidenziano che l'elaborazione automatica di modelli decisionali di cui sono capaci gli agenti artificiali è basata solo su procedure formali di calcolo (comunemente considerata un'elaborazione razionale), ed è improntata ad una logica consequenziale: in questo ambito è innegabile che essi esibiscano un elevato livello di razionalità, ma non di intelligenza umana!

L'obiezione strutturalista mette in risalto la diversità "ontologica" tra decisione umana (prodotto dalla materia biologica) e decisione robotica (prodotto della materia inorganica), a prescindere dall'equivalenza delle proprietà funzionali o dalla capacità della macchina di replicare il comportamento umano. Ma come funziona il cervello umano e quale è l'essenza della intelligenza umana, che le IA moderne ed evolute vorrebbero replicare?

Invero, recenti studi di psicologia cognitiva hanno dimostrato che gli uomini non utilizzano affatto modelli logico-formali di tipo computazionale nell'elaborazione delle loro decisioni e che le risposte che danno ad un *input* esterno dipende principalmente dal modo in cui esso viene recepito ed elaborato emotivamente, piuttosto che da una chiara, completa e razionale rappresentazione cosciente. Ciò accade a causa di strutturali "*limiti cognitivi*" della mente umana, in quanto l'individuo è dotato di una limitata capacità nel ricevere, ricordare, elaborare e valutare tutte le informazioni di cui dispone[27]. Peraltro, questi approcci affermano che gli individui, per lo più, non hanno neppure bisogno di una completa conoscenza della realtà, in quanto essi raramente assumono le decisioni soppesando razionalmente tutte le informazioni ottenibili, ma piuttosto fanno uso di procedure o modelli mentali di tipo intuitivo, inconscio, automatico ed emotivo.

Si è scoperto che gli aspetti emozionali canalizzati nel processo decisionale non sono da considerare un'interferenza fuorviante, originata da una "*razionalità limitata*". Al contrario, essi sono quel potenziale neurologico in grado di offrire un *trade off* che colma il *gap* tra i limiti cognitivi della mente umana e la complessità della situazione

[27] D. Kahneman, P. Slovic, A. Tversky, *Judgement under uncertainty: Heuristics and biases*, Cambridge 1982; G. Gigerenzer, *Decisioni intuitive*, Cortina 2009. Su questi temi, applicati alla c.d. neuroeconomia o economia comportamentale, M.B. Magro, *Manipolazioni di mercato e diritto penale. Una critica al modello di razionalità economica*, Milano 2012.

reale da gestire; e ciò avviene per qualunque tipo di scelta, da quella impulsiva - passionale a quella più fredda e pianificata. Di fronte ad una scelta, quando la parte emozionale è in conflitto con quella cognitiva, a meno che la prima non dia risposte di scarsa intensità, tende a vincere o imporsi quella inconscia emotiva e non quella cognitiva cosciente.

Le emozioni sono, di fatto, informazioni, cioè concorrono, insieme ad altre informazioni, a determinare il nostro comportamento e costituiscono un tratto costitutivo fondamentale dell'intelligenza umana (e anche degli animali mammiferi), inseparabile dalle altre nostre caratteristiche. Le emozioni sono profondamente pervasive e innestate nel corpo, inteso sia come insieme di organi sia come depositario della nostra identità e della nostra storia.

Da questi studi sul ruolo delle emozioni come informazioni utili alla decisione, si è arrivati a postulare che il nostro pensiero, il nostro ragionamento, e dunque la nostra decisione, funzioni secondo una doppia modalità: un sistema o pensiero logico-analitico, ed un sistema o pensiero intuitivo-esperienziale (teorie del doppio processo). In breve, la sfera emotiva, nella sua funzione di linea-guida delle decisioni, è un indispensabile complemento della razionalità e svolge un decisivo ruolo epistemico funzionale (e non disfunzionale, come si credeva) alla conoscenza, alla comprensione e alla decisione, tanto che, paradossalmente, quando i sistemi affettivi e emotivi sono danneggiati, anche il sistema deliberativo entra in crisi[28].

Sembra, insomma, che sia impossibile riprodurre l'intelligenza umana, che appare un concetto vuoto senza che il suo possessore abbia un corpo biologico, fatto di carne e ossa. L'assenza di un corpo impedisce che i robot abbiano una dimensione emotiva, e ciò fa sì che il loro sistema cognitivo si atteggi in modo del tutto particolare rispetto a quello di un uomo, ma anche rispetto quello di un gatto o di un cane[29].

---

[28] Il sistema analitico -razionale usa inferenze, è governato da regole controllabili e logiche; è principalmente verbale e consapevole; lavora in sequenza, ed è abitualmente lento, richiede un controllo consapevole delle operazioni di ragionamento ed è tendenzialmente faticoso. Il sistema intuitivo- emotivo-esperenziale costituisce invece la forma privilegiata di conoscenza che consente di superare le rigidità e i limiti del pensiero logico, accedendo ad una più profonda comprensione che attiene ad una dimensione innata e inconscia. Il sistema intuitivo è preconscio, veloce, prevalentemente automatico, lavora in parallelo, non in sequenza, è poco accessibile alla consapevolezza e richiede poco sforzo cognitivo. E' implicito e difficilmente controllabile. Ha permesso agli esseri umani di sopravvivere nel loro lungo cammino evolutivo e rimane ancora oggi il modo più naturale e più frequente con cui, per esempio, valutiamo il rischio. Il pensiero intuitivo permette di comprendere subito la realtà, senza la mediazione della logica o dell'analisi, senza l'impiego del linguaggio verbale, bensì si fonda sulla base di indizi e sensazioni (c.d. euristica dell'emozione o *affect*). Così M. C. Nussbaum, *L'intelligenza delle emozioni*, Bologna 2011 (ed. or., *Upheavals of Thought. The Intelligence of Emotions*, Cambridge 2001).

[29] Ad esempio C. R. Kaczor, *Abortion and Unborn Human Life*, 2011, sostiene che l'appartenenza alla specie umana sotto il profilo ontologico costituisce una prerogativa fondamentale ed esclusiva per il riconoscimento dello *status* morale di soggetto.

I robot non sono fragili, mortali, moralmente bisognosi come gli umani o gli animali. Le loro parti possono essere facilmente sostituite in caso di lesioni o incidenti; la loro "mente" e i loro ricordi possono essere sottoposti a *backup* e ripristinati. Proprio questa solidità esistenziale li allontana da tutto ciò che è vivente, a prescindere dalle loro *performance* più o meno equivalenti a quelle umane o animali. In quanto capaci di sola razionalità computazionale, le macchine intelligenti, anche le più raffinate ed evolute non possono essere paragonate alle straordinarie potenzialità della mente umana, che funziona sulla base di circuiti neurologici stabilizzati, ma non ne è affatto schiava.

Si è infatti evidenziato che il pensiero umano non è solo razionale e conseguenziale, ma anche a carattere ipotetico-intuitivo: gli uomini sono in grado di avanzare ipotesi esplicative sulla base di inferenze idonee a stabilire nessi basati su generalizzazioni nuove che non si basano sulla logica conseguenziale (in altri termini, spiegazioni consentite da "abduzioni straordinarie" o "creative"). Ciò avviene perché la mente (e i suoi innumerevoli processi inconsci) non svolge solo una funzione di custodia degli archivi accurati di memoria passata, ma presenta anche una struttura intuitiva, proattiva, proiettata verso il futuro, che consente di immaginare in modo creativo situazioni future del tutto nuove, esplorare possibilità di vita futura[30]; il cervello umano, più è libero e sano, più riesce a sopprimere un circuito neurologico comunemente associato ad una certa rappresentazione di informazioni e sostituirlo con un altro[31]. In tal modo riesce a inibire e correggere gli ingranaggi automatici riconsegnandoli a nuove reti neurali, alla ricerca di nuovi percorsi, alla creatività, all'attenzione e alla volontà cosciente.

Questo è il libero arbitrio dell'essere umano; non una proprietà metafisica, ma una proprietà del carattere, empiricamente fondata. La libertà umana non è uno *status,* una condizione, ma un obiettivo cui tendere, un privilegio dell'uomo, un tratto del carattere che si conquista. La libertà non è un tutt'uno con lo stato delle cose spontaneo ma è capacità di trasformazione, di uscire dagli schemi ripetitivi e inconsapevoli stratificati nel sistema limbico.

Per questi aspetti l'intelligenza umana non è modellabile e replicabile in quanto la mutabilità degli orientamenti che una persona assume in tempi, luoghi e contesti

---

[30] J. Hawkins e S. Blakeslee, *On Intelligence. How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*, Times Books 2004.

[31] Infatti, sappiamo che le memorie del passato non sono esatte, e, per di più, non sempre replicare il passato si rivela la strategia migliore. Il che dimostra che la mente non è solo un archivio di ricordi più o meno passati, non è solo il luogo della memoria implicita o esplicita (cioè cosciente o non cosciente); in proposito, M.B. Magro, *Mente sana e mente alterata. Uno studio neuroscientifico sul vizio di mente*, in *A* 2019 e bibliografia ivi indicata; in particolare D. L. Schacter, D. R. Addis, K. K. Szpunar, *Escaping the past: Contributions of the hippocampus to future thinking and imagination*, in *The hippocampus from cells to systems: Structure, connectivity, and functional contributions to memory and flexible cognitio, a cura di* D.E. Hannula e M.C. Duff, New York 2017, 439 ss..

diversi ne vanifica la significatività. Un agente intelligente, anche molto sofisticato, è intelligente solo nella misura in cui gli è possibile automatizzare, e cioè modellare, un processo decisionale: le sue regole di comportamento sono il risultato di ripetute simulazioni di comportamenti e di migliaia di rilevazioni statistiche che costituiscono la base per elaborare modelli. Gli agenti intelligenti – diversamente dagli umani – non agiscono "per analogia" o per "intuito"; da qui il paradosso che le macchine riescono a riprodurre il più alto rigore logico ma non riescono a essere programmate su un livello minimo di intelligenza comune. Le IA non possono essere paragonate all'uomo, senza ridurre quest'ultimo ad un profilo comportamentale monodimensionale e rigido.

8. Ai robot intelligenti manca qualunque proprietà della coscienza, anche di livello inferiore. Il termine "coscienza" va chiarito. Studiosi cognitivi moderni distinguono la *coscienza cognitiva* o *di accesso* dalla *coscienza fenomenica*. Con il concetto di *coscienza di accesso* o *cognitiva* si indica che una qualche informazione esterna è sottoposta al controllo del comportamento; essa rende cioè "*accessibile*" a noi stessi i nostri stati mentali e i nostri comportamenti, in quanto presiede ai processi di formazione di ragionamenti pratici, di credenze, di riflessioni razionali, alla formazione di stati intenzionali intesi nel solo aspetto funzionale e rappresentazionale[32]. In poche parole, la "*coscienza di accesso*" è la capacità di un sistema di avere accesso ai propri stati interni, ai fini della verbalizzazione, della organizzazione dell'azione e anche della costruzione di modelli utilizzabili nell'interazione sociale. Della coscienza cognitiva è possibile valutare i correlati neuronali, determinanti per la pianificazione dell'atto o per il controllo degli impulsi, ma anche, attraverso le tecniche di *neuroimaging* che misurano l'attività metabolico-funzionale, è possibile individuarne quei segni spontanei (cioè non attivati da stimolo elettrico) non accessibili a test comportamentali e del tutto inconsapevoli.

La *coscienza fenomenica* inerisce invece all'aspetto assolutamente *qualitativo* delle esperienze e delle sensazioni soggettive perché indica l'esperienza percettiva "*in prima persona*", cioè le *rappresentazioni mentali delle esperienze percettive*. Essa costituisce il vero problema sotto il profilo naturalistico, in quanto sfugge a una precisa localizzazione neurologica e ad un'indagine oggettiva: le sensazioni soggettive non sono accessibili a un osservatore terzo e hanno un'esistenza solo in prima persona, in quanto le percezioni sono sempre le sensazioni di qualcuno e sono sempre raccontate

[32] A. Paternoster, *Introduzione alla filosofia della mente*, Roma 2014, 170; M. Di Francesco, *Introduzione alla filosofia della mente,* Roma 2000, 43 s..

da qualcuno, e non possono essere osservate né oggettivate da un terzo all'esterno[33]. La coscienza fenomenica sfugge a qualsiasi spiegazione fisicalistica perché non corrisponde a correlati neurologici ben localizzati [34] ; perciò sfida ogni ipotesi funzionalista a fornire una esauriente spiegazione dei processi mentali.

La coscienza cognitiva o di accesso dell'uomo non è sempre consapevole (coscienza e consapevolezza sono concetti diversi). Al contrario, studi neuroscientifici hanno evidenziato come l'uomo disponga di una architettura della mente assai complessa, che si esprime anche attraverso rappresentazioni mentali *implicite non consapevoli* che riflettono un sistema cognitivo, che sono responsabili di stati mentali non accessibili alla coscienza. Anzi, il fatto che la maggior parte della corteccia non produce sensazioni coscienti e, se interessata da *deficit* o lesioni, non produce menomazioni a livello cognitivo, ha fatto desumere che la maggiore parte della nostra attività neurologica sia inconscia. Si è quindi concettualizzato il c.d. "*inconscio cognitivo*", ovvero quella parte del funzionamento mentale che è inconscia e che mai potrà emergere a livello di coscienza. La conoscenza che si sviluppa a partire da tali disposizioni innate è di tipo implicito e non richiede, per il suo funzionamento, né la coscienza né l'autocoscienza.

Questo sistema cognitivo, sebbene sommerso, non esprime un processo meccanico di tipo logico -computazionale, ma è sempre intimamente legato alla dimensione emotiva, fisica ed esperenziale umana. Anche i processi neurologi importanti come quelli motivazionali, emotivi ed affettivi, possono verificarsi al di fuori della consapevolezza, evidenziando che l'attività mentale è radicata in sistemi motivazionali ed emozionali filogeneticamente antichi, capaci di influenzare la sviluppo della mente e di operare totalmente al di fuori della piena consapevolezza[35]. Gli uomini, in poche parole, possono sperimentare stati emotivi e agire di conseguenza senza esserne consapevoli né coscienti (possono quindi provare qualcosa senza sapere che la stanno provando), in quanto il processamento emotivo ha inizio al di fuori della consapevolezza (c.d. *unconscious will*).

Se ci riferiamo alla coscienza cognitiva, ovvero come capacità di avere accesso alle informazioni, possiamo ritenere che i robot godano in buona misura di questa dimensione, in quanto accedono e memorizzano una quantità straordinaria di

[33] N. Block, J. Fodor, *What psychological states are not,* in *PR* 1972, 159 ss; N. Block, *Comparing the major theories of consciousness*, in *The Cognitive Neurosciences* a cura di Gazzaniga, *MIT Press* 2009, 1111 ss. disponibile in https://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Theories_of_Consciousness.pdf. V. inoltre D.J. Chalmers, *Facing up to the problem of consciousness*, in *JCStud* 1995, 200 ss. disponibile in http://consc.net/papers/facing.html; ID., *Moving forward on the problem of consciousness*, in *JCStud* 1997, 3 ss., disponibile in http://consc.net/papers/moving.html.

[34] J. Smart, *Sensations and Brain Processes*, in *PR* 1959, 141ss.

[35] J. Panksepp, *On the Embodied Neural Nature of Core Emotional Affects*, in *JCS* 2005, 158 ss..

informazioni; difficile però qualificare questa dimensione come sistema cognitivo o come dimensione soggettiva, in quanto l'accesso alle informazioni avviene in modo del tutto automatico e meccanico, senza coscienza neppure potenziale, in quanto privo della dimensione emotiva-simbolica. Gli agenti artificiali intelligenti godono di un elaborato sistema cognitivo di accesso, ma assolutamente sganciato da ogni dimensione emotiva e, soprattutto, che non riflette una architettura, sia pure inconscia, della mente. A differenza dell'agire robotico, uno stato mentale è rinvenibile sempre in un'azione umana, anche in quella coartata, anche nell'atteggiamento di chi perde il controllo delle proprie azioni per un blocco emotivo, una condizione di panico, un difetto di attenzione, un *lapsus*, un *dejavu*, un meccanismo di rimozione inconscio, o ancora a causa di un autentico automatismo inconsapevole che risponde a una reazione di tipo modulare non cognitiva, dovuto ai limiti (o alle potenzialità inconsce) della mente umana. E del resto, a riprova della dimensione soggettiva della colpa dell'uomo, l'unico limite che pone l'ordinamento è che questo stato mentale non sia determinato da caso fortuito o forza maggiore[36].

Ma la questione ancor più dibattuta concerne la c.d. coscienza fenomenica, poiché essa sembrerebbe essere un salto epistemologico, sembrerebbe porre un *gap* esplicativo tra i fenomeni naturali e il fenomeno mentale, in quanto fenomeno esclusivamente in possesso di chi lo ha vissuto[37]. Poiché l'ontologia della coscienza fenomenica è essenzialmente soggettiva, le sue proprietà non possono essere analizzate adoperando l'epistemologia delle scienze naturali, per definizione oggettiva. Per questo la coscienza fenomenica non può essere studiata con la stessa epistemologia con la quale è studiato il mondo naturale oggettivo[38].

Ebbene, ciò che manca agli agenti intelligenti è anche questa dimensione fenomenica o qualitativa della coscienza, la *coscienza* dei significati dell'esperienza. I robot intelligenti non hanno coscienza di secondo livello, cioè non sono né possono essere coscienti di essere coscienti. La macchina computazionale non è in grado di manifestare la coscienza fenomenica che caratterizza gli esseri umani e, sia pure in forme diverse, gli animali. Le macchine intelligenti sono psicologicamente isomorfe all'uomo, ossia hanno una identità di organizzazione funzionale, ma certamente non hanno consapevolezza del proprio agire. Ciò che caratterizza l'agire robotico è il suo

[36] A proposito della colpa cosciente, G. A. De Francesco, *Dolo eventuale e dintorni, tra riflessioni teoriche e problematiche applicative*, in *CP* 2015, 4624.

[37] C. Umiltà, *Consciousness and control of action*, in *The Cambridge handbook of consciousness,* a cura di P.D. Zelazo, M. Moscovitch, E. Thompson, Cambridge 2007, 327 ss.; S. Gozzano, *La coscienza*, Roma 2009; P. Pecere, *La coscienza fenomenica tra neuroscienze e metafisica. L'Ignorabimus di du Bois-Reymond e il futuro della "science of consciousness"*, in *RF* 2018, 215 ss.

[38] T. Nagel, *What Is it Like to Be a Bat?*, in *PR* 1974, 435 ss..

carattere assolutamente non consapevole (nel senso privo di coscienza sia di primo che di secondo livello) e, quindi, semanticamente vuoto, dei simboli elaborati da un sistema artificiale e, in conclusione si nega che i processi mentali possano essere ridotti a processi meccanici di tipo logico-computazionale.

Dal punto di vista ontologico, ammesso che i robot possano esibire un livello performativo pari o superiore a quello degli umani, comunque il loro agire è assai diverso da quello umano perché manca la loro emotività, cioè quella emotività che connota simbolicamente ogni informazione esterna. Avere stati psicologici rilevabili *ab extra* non significa possedere autentica consapevolezza dei comportamenti!

9. L'impossibilità di interpretare o di spiegare il modello decisionale degli agenti artificiali pone sul tavolo il tema della progettazione responsabile delle tecnologie di automazione di apprendimento automatico sottratte al controllo umano, in aggiunta a quello della responsabilità colposa, spesso marginale, a carico dell'utilizzatore umano di un robot che una volta immesso nell'ambiente aperto, sfugge ad un tempestivo controllo umano. Fino a che punto gli sviluppatori di macchine intelligenti ed evolute possono essere considerati responsabili per il "duplice e malefico" uso di una tecnologia concepita a scopo benefico ma asservita a scopi bellici o persino criminali? Si potrebbe affermare una responsabilità a carico degli sviluppatori per non aver previsto i potenziali usi alternativi dannosi e persino gli illeciti commessi dalle loro creazioni? E' possibile punire penalmente questa cecità, più o meno cosciente, connessa alle conseguenze più remote delle proprie azioni[39]?

Va messo in chiaro che l'argomento diffusamente esposto della mancanza di prevedibilità (e spiegabilità scientifica) dell'agire robotico non solleva il progettatore dalla responsabilità penale colposa poiché, così come la categoria della colpa è elaborata nel diritto vivente, la prevedibilità astratta non richiede alcuna previsione dettagliata e specifica dell'eventuale evento dannoso. Come è a noi noto, la costruzione positiva della c.d. "*colpa eventuale*" connessa ad accadimenti catastrofici, spesso dislocati spazialmente e temporalmente al di là di ogni relazione di contiguità con l'agente, non richiede affatto la prevedibilità dell'evento *hic et nunc*, essendo sufficiente la prevedibilità del rischio o anzi l'impossibilità di escludere la verificazione di un qualsiasi evento di danno, anche concretamente imprevedibile, quale ipotesi

[39] È il tema dei c.d. *IA crimes* connessi alle frodi, ai reati finanziari, all'*e-commerce*, ma anche alla vita umana e all'integrità fisica. In proposito, T. King, N. Aggarwal, M. Taddeo e L. Floridi, *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*, in *SSRN* 2018, disponibile su: https://papers.ssrn.com/abstract=3183238.

alternativa e virtuale che oltrepassa la misura del rischio consentito[40]. Ebbene, è davvero difficile negare che il ricercatore (programmatore o sviluppatore del *design*) non sia a conoscenza (o abbia astrattamente previsto) dei potenziali e futuribili rischi connessi all'suo delle tecnologie di *autoML*, se si ritiene sufficiente una previsione così generica e astratta.

Il progressivo depauperamento del contenuto della colpa soggettiva, in particolare dei suoi profili più propriamente personali, a favore di una eccessiva normativizzazione in termini di riduzione a pura inottemperanza cautelare, non può che sfociare in modelli di allocazione della responsabilità di tipo oggettivo-assicurativo-indennitario [41] . Questi sarebbero gli esiti di una eccessiva marginalizzazione della dimensione soggettiva (ed umana) della colpa, in cui la dimensione soggettiva si riduce a mera connessione pseudo-intellettiva con un dato stocastico priva di qualunque dimensione ontologica, volta piuttosto a colmare vuoti di tutela.

La questione quindi tira in ballo scelte che spettano al legislatore, avendo preso atto che le nuove sfide per l'umanità non dovrebbero dipendere esclusivamente da strategie imprenditoriali (si pensi che Google è uno dei più grandi investitori sulla ricerca robotica) ma riflettere principi e decisioni democratiche. Perciò, nella consapevolezza della necessità di una prevenzione efficace del rischio per i diritti umani e per le libertà fondamentali che queste tecnologie potrebbero minacciare, si profilano all'orizzonte forme diverse di responsabilità e di intervento penale[42].

In quest'ordine di idee qualche studioso si chiede se lo sviluppo di futuri robot intelligenti non possa essere anche parzialmente contrastato da barriere di diritto penale che condizionino la libertà di ricerca scientifica ponendo alcuni limiti alla progettazione di sistemi informatici intelligenti dotati di tecnologie di apprendimento automatizzato senza controllo umano che sostituiscano in toto la decisione umana[43]. La norma penale potrebbe vietare quella ricerca che sfoci nella creazione di intelligenze artificiali sottratte al controllo umano tali da mettere in pericolo la collettività, circoscrivendo il rischio consentito, o quanto meno penalizzando

[40] G. Civello, *La "colpa eventuale" nella società del rischio. Epistemologia dell'incertezza e "verità soggettiva" della colpa"*, Torino 2013; D. Notaro, *"A ciascuno il suo": nesso di causalità (e colpa) in materia penale fra scienza, ragione ed emozione*, in *CM* 2013, 531 ss.; C. Perini, *Il concetto di rischio nel diritto penale moderno*, Milano 2010.

[41] In una prospettiva civilistica o assicurativa, E. Macrì, A. Furlanetto, *I robot tra mito e realtà nelle interazioni con le persone, negli ambienti sociali e negli ospedali. Un approccio tra risk managment e diritto robots*, in *RIML* 2017, 1045 ss.

[42] In proposito, F. Raso, H. Hillgoss, V. Krishnamurthy, C. Bavitz, L. Kim, *Artificial Intelligence and Human Rights: opportunities and risk*, in *BKC* 2018.

[43] Così, K. Gaede, *Künstliche Intelligenz –Rechte und Strafen für Roboter? Plädoyer für eine Regulierung künstlicher Intelligenz jenseits ihrer reinen Anwendung*, in *NOMOS* 2019, 76 e 81 s.

l'inosservanza di speciali obblighi di comunicazione, di notifica, di richiesta di autorizzazione. In particolare, mettendo in luce anche le potenzialità della ricerca scientifica in tema di maternità surrogata che potrebbe condurre a creazione di ibridi, si suggerisce l'adozione di norme penali che limitino l'evoluzione tecnologia assumendo a modello l'*Embryo Protection Act*, il quale fornisce un esempio di norma penale che vieta la creazione di entità che possano mettere in pericolo la nostra convivenza.

Piuttosto che ipotizzare una responsabilità diretta a carico dei robot intelligenti, si fa strada l'ipotesi di tratteggiare una regolamentazione che, nel rispetto della libertà di ricerca scientifica, faccia propri alcuni fondamentali principi etici. Così come regoliamo la ricerca su virus e batteri o lo sviluppo e la fabbricazione di armi, allo stesso modo anche la ricerca scientifica su IA dovrebbe essere contenuta da una regolamentazione che prevenga questi rischi futuri per i diritti e le libertà fondamentali dell'uomo, che limiti la tecnologia di autoapprendimento e che ponga la ricerca robotica a servizio dell'umanità[44].

In tal senso, allo scopo di preservare i diritti umani, il Parlamento europeo, il Consiglio d'Europa e la Commissione europea sulle IA hanno espresso indicazioni e linee guida affinchè i robot intelligenti vengano progettati in modo tale da poter essere controllati dagli umani, lasciando anche intravedere la possibilità che siano introdotti a livello nazionale divieti penali alla progettazione e diffusione di intelligenze artificiali che mettano in pericolo l'incolumità pubblica in ragione del loro uso[45]. Altrettanto, a livello internazionale sono diffusi studi come il *Future of Life Institut*[46] e il dossier The *Malicious Use of Artificial Intelligence*: *Forecasting, mitigation and prevention* che ci mettono in guardia dalla produzione di entità artificiali dannose per l'uomo[47].

A queste considerazioni aggiungo le mie personali osservazioni.

È ben vero che la funzione della tecnologia è quella di realizzare l'obiettivo tecnico. È anche vero che, al di là di quella funzione, dello scopo che è proprio dei progettisti e

[44] S. Beck, *Jenseits von Mensch und Maschine, Ethische und rechtliche Fragen zum Umgang mit Robotern, Künstlicher Intelligenz und Cyborgs*, Baden-Baden 2012; S. *Gleß, K. Seelmann, Intelligente Agenten und das Recht*, Baden-Baden 2016; Min Kyn Kim, *Roboterrecht in der modernen Gesellscjaft. Vorschlaege zur Gesetzgebung und Reform*, 2018; Y. Wang, M. Xiong, H. Olya, *Toward an Understanding of Responsible Artificial Intelligence Practices*, conference paper ottobre 2019.

[45] Il tema dei limiti etici alla libertà di ricerca nella fase della progettazione e all'uso dei robot ha ispirato il Parlamento europeo nella *Risoluzione sulle norme di diritto civile nel settore della robotica* del 2017. Cfr. inoltre *Responsability and IA*, *Council of Europe Study* 2019 e le linee guida della commissione europea, *Ethics Guidelines for Trustworthy AI* 2018, 12.

[46] *Future of Life Institute*: *Autonomous Weapons: an Open Letter from AI & Robotics Researchers*, disponibile su https://futureoflife.org/open-letter-autonomous-wea pons.

[47] *The Malicious Use of Artificial Intelligence*: *Forecasting, mitigation and prevention,* febbraio 2018, Cornell University.

degli utenti della tecnologia, occorre tenere presente le implicazioni etiche relative al quanto e al come quel sistema tecnologico potrà incidere sulla vita degli agenti umani e all'impatto sul loro sistema di responsabilità. Ritengo perciò che debba essere valorizzata quella tecnologia che costruisce e sviluppa sistemi di IA che contribuiscono all'esercizio della responsabilità umana e dei suoi standards. Ciò significa che la ricerca scientifica dovrebbe far propri i principi etici che consentano un utilizzo di agenti artificiali che non espropriino, ma al contrario sollecitino, l'attenzione umana e quindi la responsabilità umana. Ciò non significa intralciare o ostacolare la libertà di ricerca, ma solo che dovremmo astenerci dal creare entità artificiali autonome che sostituiscano totalmente la decisione umana, facendo così a meno delle potenzialità straordinarie della mente umana; significa attivare e predisporre meccanismi che non affievoliscano la capacità di controllo (e autocontrollo) umano a causa del supporto robotico, ma, al contrario, che la sollecitino, la supportino e la potenzino.