DA
Diritto Artificiale n. 1, 2025

Giuseppe Ziino

Del1tt0 e cast1g0. Per un critica della applicabilità delle sanzioni penali ai sistemi di Intelligenza Artificiale

Abstract This article critically examines the potential application of criminal sanctions to artificial intelligence (AI) systems, particularly those categorized as "strong AI" or autonomous agents. These systems, capable of learning, reasoning, and acting independently, pose profound questions regarding accountability and the philosophical underpinnings of punishment.

Through an exploration of existing legal frameworks and ethical considerations, the study challenges the feasibility of applying traditional notions of punishment – such as retribution, deterrence, and rehabilitation– to non-human entities. It further evaluates whether AI systems can be analogized to legal persons or necessitate the establishment of a novel legal paradigm.

Ultimately, the paper advocates for the adoption of proactive legal approaches to address the unique challenges presented by increasingly sophisticated AI systems, while acknowledging the fundamental distinctions between human and machine agency.

Keywords
- Artificial Intelligence and Criminal Liability
- Autonomous Agents
- Punishment and Rehabilitation
- Legal Personhood for AI
- Ethics and AI Accountability.

Sommario

1. Introduzione

Un dato di comune esperienza insegna che l'essere umano, quando fa una scoperta, si pone contestualmente il problema di come controllarla; più precisamente, di come

indirizzarla agli scopi per i quali ha profuso il proprio impegno per giungere ad essa. È accaduto con il fuoco, con il motore a scoppio, con l'elettricità, con l'energia nucleare, con internet, e si sta riproponendo anche con riguardo ai si-stemi dotati di intelligenza artificiale (da ora in avanti I.A.).

Difatti, l'invenzione di algoritmi in grado di emulare, e in alcuni casi implementare, comportamenti e capacità umane, apportando indubbi vantaggi nei più disparati settori della vita quotidiana – quali l'assistenza sanitaria, i trasporti, l'informazione, l'istruzione, la difesa e la sicurezza, il sistema della giustizia, l'industria, ecc. – reca con sé l'inevitabile preoccupazione di come prevenire e reprimere quelle condotte illecite che tali sistemi potrebbero porre in essere, magari a seguito di un'anomalia o di un errato apprendimento delle informazioni di base necessarie per lo svolgimento delle attività cui sono destinati.

Tale questione si pone con particolare riguardo ai c.d. "sistemi sapienti" o sistemi di *strong A.I.*, ovverosia quegli agenti artificiali che sono in grado di comprendere, apprendere e ragionare in modo simile o superiore a quello umano in tutte le attività cognitive, assumendo quelle che a prima vista potremmo chiamare "decisioni" e, di conseguenza, realizzando comportamenti e attività in maniera autonoma.

Il concetto di I.A. "forte" suggerisce, dunque, che tali sistemi non solo eccellano in compiti specifici, ma dimostrino una forma più ampia di "intelligenza" generale, affrontando sfide complesse in modo simile a un essere umano. Per maggiore chiarezza, ci si si riferisce a quegli *autonomous agents* definiti di "terzo" e "quarto livello", in base ad una classificazione che tiene conto del grado di autonomia posseduto. I primi sono quelli orientati da algoritmi di apprendimento automatico e, pertanto, svincolati dal rispetto fisso e predeterminato di regole già stabilite dai programmatori, in quanto, grazie all'esperienza che acquisiscono nel tempo, riescono ad autocorreggere i propri comportamenti[1]; i secondi rappresentano il più elevato grado di autonomia, tanto che vengono anche definiti *multi-agent systems*, in quanto, grazie all'*Internet of Things*, riescono ad «interagire con altri agenti ed oggetti e sono capaci di adeguare autonomamente i loro comportamenti in base all'ambiente in cui

[1] Nello specifico, «gli algoritmi intelligenti, che poggiano sull'analisi dei dati, sul calcolo statistico ed il riconoscimento di modelli predeterminati (cartelli stradali, pedoni, auto ecc. ...) consentono agli agenti artificiali di svolgere in modo autonomo specifiche funzioni, di individuare problemi e definire soluzioni secondo modalità che neppure i loro programmatori possono prevedere a priori» (I. Salvadori, *Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale*, in *Rivista italiana di diritto e procedura penale*, 1/2021, p. 93).

si trovano ad operare»[2]. In questo quadro, sono sicuramente da annoverare i c.d. agenti artificiali *fully automated*, i quali, grazie all'alto grado di autonomia di cui godono, assumono comportamenti assolutamente imprevedibili da parte di un osservatore esterno[3].

2. *Machina delinquere non potest*

Prescindendo, adesso, da quelle disquisizioni in base alle quali l'I.A. forte sia un obiettivo futuro e attualmente privo di consistenza[4], bisogna, invece, porsi nell'ottica che siano reali e attuali quelle scene proposte dalla cinematografia hollywoodiana[5] di robot autonomi, che possiedono un'intelligenza del tutto identica, se non migliore, di quella umana, e che potenzialmente siano in grado di porre in essere condotte illecite dotate di rilevanza penale, al fine di comprendere se gli attuali strumenti giuridici che il diritto penale sostanziale offre siano idonei e configurabili in presenza di tali fattispecie[6].

In altri termini, si vuole ragionare sui seguenti quesiti: ma se un *robot* commette un reato, a esso perfettamente ed esclusivamente ascrivibile, come deve essere punito? In cosa consisterà la punizione per il male che esso ha procurato? Prima di individuare gli idonei meccanismi di imputazione della responsabilità penale per tali soggetti artificiali,

---

[2] Sulla loro struttura, come indicato da I.Salvadori, *op. cit.*, p. 93, v. G. Weiss (ed.), *Multiagent Systems*, Cambridge 2013; J Xie, C.C. Liu, *Multi Agent Systems and Their Application*, in *Journal of Internation Council on Electrical Engineering*, 2017, 188 ss. Più in generale, sull'interazione tra A.I. e IoT (c.d. AioT) si veda lo studio realizzato da Sas, *AioT. How IoT Leaders are Breaking Away*, 2019.

[3] I. Salvadori, *op. cit.*, pp. 90-94.

[4] Sul punto J.R. Searle, *Menti, Cervelli e programmi. Un dibattito sull'intelligenza artificiale*, a cura di G. Tofoni, Club Editore, 1 gennaio 1987, in base al quale «*secondo l'intelligenza artificiale forte, il computer non sarebbe soltanto, nello studio della mente, uno strumento; piuttosto un computer programmato opportunatamente è davvero una mente*»: in altri termini, John R. Searle, che si avvalse del famoso esperimento della Stanza Cinese, dimostrò che le macchine sono prive della consapevolezza, un peculiarità tipica dell'uomo; e, per tale motivo, è impossibile creare una macchina in grado di pensare, dal momento che dovrebbe essere un programma a ideare e creare un sistema di caratteri casuali identici al cervello. Una circostanza che al filosofo appare alquanto impossibile da realizzare. Gli stati mentali sono un prodotto dell'operazione del cervello, differentemente dai programmi che sono, al contrario, frutto di processi, elaborati dall'uomo, ma eseguiti da un computer.

[5] Si pensi ad esempio all'opera di cinematografia nota come "*Terminator*", più apertamente "*I, Robot*", ove sono raffigurati androidi intelligenti o supercomputer coscienti; o, ancora, tra i più recenti, "*Avengers: Age of Ultron*" che rappresenta un sistema di intelligenza artificiale creato per difendere il mondo e che in maniera del tutto autonoma e assolutamente imprevedibile sviluppa una sorta di "complesso di Dio" per la quale deve salvare l'umanità, distruggendola.

[6] La questione sulla *conditio autonoma* si pone il quesito "se e come le macchine intelligenti possono trovare la loro strada in una società umana". I. Mcewan, *Machines like Me*, Penguin 2019, che ne fornisce un tipico esempio letterario.

non sarebbe opportuno ragionare sulla minaccia giuridica che bisogna predisporre nei loro confronti? Vanno bene le sanzioni penali attualmente in vigore o conosciute?

Certamente, un argomento giuridico-filosofico di assoluta rilevanza che entra in gioco rispetto a simili quesiti è quello relativo alla pena e alle sue finalità, specie in un ordinamento costituzionale democratico come quello del nostro Paese, dove i diritti fondamentali dell'uomo assumono un valore preminente, sicché avrà senso irrogare una pena solo allorquando la stessa svolga una funzione costituzionalmente orientata.

Come noto, le funzioni della pena possono essere molteplici: da quella più antica, la retribuzione, a quella auspicata già da Cesare Beccaria, la prevenzione.

Nella società odierna – e segnatamente negli ordinamenti occidentali moderni – la pena assurge ad una dimensione che mira alla rieducazione del reo.

Invero, l'art. 27 della Carta Costituzionale[7] statuisce che la pena non può vertere in trattamenti inumani e degradanti e deve, viceversa, mirare alla rieducazione del condannato. Per cui, la funzione della rieducazione intende la pena non come una mera punizione per "neutralizzare" il delinquente, ma considera il reo quale agente che delinque «per difetto di educazione tale da non consentirgli di riconoscere i beni giuridici altrui e di riconoscere se stesso come parte del comune vivere sociale»[8]. Orbene, tale visione rappresenta il *punctum dolens* della funzione e finalità della sanzione penale irrogata a un agente artificiale autore di un evento delittuoso: proprio relativamente a tale questione si incentra la critica rivolta al c.d. dogma *machina delinquere non potest*.

Difatti, una pena comminata ad un software non potrebbe mai assolvere le tipiche funzioni che «la dottrina penalistica generalmente riconosce alla sanzione criminale»[9].

Tra i vari scopi assolti dalla pena, sicuramente in tale ambito quella che pone maggiori problematiche è la finalità rieducativa, in una prospettiva di "prevenzione generale speciale (special-preventiva)", posto che le macchine intelligenti, ad oggi, risultano ancora incapaci di comprendere il disvalore dell'azione e, dunque, appaiono impassibili a qualsivoglia rimprovero colpevole.

---

[7] Art. 27 Cost.: «La responsabilità penale è personale. L'imputato non è considerato colpevole sino alla condanna definitiva. Le pene non possono consistere in trattamenti contrari al senso di umanità e devono tendere alla rieducazione del condannato».

[8] T. Martines, *Diritto costituzionale*, Giuffrè, 2020, p. 80 ss.

[9] A. Cappellini, Machina delinquere non potest*? Brevi appunti su intelligenza artificiale e responsabilità penale*, in *Criminalia* 2019, pp. 15-16.

L'unica ipotesi in cui un algoritmo, fondato sulla tecnologia dell'I.A., sarebbe in grado di discernere la liceità o l'illiceità dell'azione compiuta è quella in cui appositamente il programmatore crei una macchina che possegga la specifica caratteristica di poter essere rieducata in seguito all'irrogazione di una sanzione.

Ma a questo punto, sorge spontaneo chiedersi fino a che punto sia efficace ed efficiente la rieducazione di uno strumento che ben potrebbe essere riprogrammato dall'operatore – attraverso nuove tecnologie di *machine learning* – che ripristini le capacità dell'agente artificiale, rielaborandole al perseguimento di fini lecitamente riconosciuti.

Pertanto, appare evidente, soprattutto in termini di economia processuale, come sia maggiormente proficuo correggere un comportamento criminoso del *robot* in modo diretto (mediante – come sopra detto – la formattazione del *software*), piuttosto che «uno strumento indiretto di pressione quale è la pena»[10].

3. Agenti artificiali e funzione della pena

Tuttavia, alcuni autori ritengono auspicabile il soddisfacimento della funzione rieducativa della pena anche qualora venga irrogata ad un agente artificiale: esemplificando, ciò potrebbe realizzarsi attraverso l'installazione, all'interno del *software*, di meccanismi di autoapprendimento o addirittura disattivando l'agente di I.A., realizzando così anche la funzione retributiva della sanzione penale[11].

Segnatamente, tale ultima funzione mira a far "pagare" il male recato dalla macchina attraverso l'inflizione di una pena.

Ebbene, anche se materialmente risulti possibile far scontare la pena (ad. esempio una misura detentiva) irrogata ad una macchina autrice di un reato, allo stesso tempo tale soluzione appare inefficace ed inefficiente in virtù del fatto che gli agenti artificiali, privi di sentimenti e pensieri umani, non carpirebbero quella privazione di libertà quale conseguenza dell'azione criminosa.

---

[10] Come osservato da A. Cappellini, *op. cit.*, per maggiore approfondimento sulla eventuale "rieducazione" dell'I.A. e i suoi rapporti con la riprogrammazione, da cfr. P. Asaro, *Determinism, machine agency, and responsibility*, in *Politica e società*, il Mulino, 2014, p. 282 ss.; sul punto v. anche Id., *A body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics*, in *Robot Ethics: the Ethical and Social Implications of Robotics*, Ed. Patrick Lin, Keith Abney and George A. Bekey (eds.), Cambridge: MIT Press 2012, pp. 169-186.

[11] Sul punto v. J. Kaplan, *Intelligenza artificiale. Guida al prossimo futuro*, Luiss University Press, 2018, p. 156.

Infine, in merito alla funzione di prevenzione generale (o general-preventiva) della pena, maggiori problemi sorgono circa l'esplicazione della stessa nei confronti di macchine intelligenti.

Preliminarmente, giova rammentare che tale funzione assolve ad una finalità deterrente alla spinta criminosa: distogliere i consociati dal commettere reati, in quanto "intimoriti" dall'irrogazione di una sanzione penale.

Sotto questo profilo ci si chiede come, un agente artificiale privo di sentimenti ed emozioni umane – e dunque di una coscienza – possa effettivamente paventare il timore o la "paura"[12] di dover subire una "punizione" qualora commetta azioni criminose.

Ed invero, l'effetto della deterrenza nei confronti della comunità robotica trova soluzioni solo su un piano meramente fantascientifico; alcuni autori, difatti, hanno teorizzato alcune soluzioni, come ad esempio l'elaborazione di precetti penali digitali affinché possano essere appresi dai consociati robotici, o ancora la diffusione, tramite apposite piattaforme network, di vicende "giudiziarie" intraprese da robot resisi responsabili di un delitto o in quanto potenziali delinquenti abituali[13].

Inoltre, nella fattispecie risulterebbe assente la funzione "comunicativa" della pena, posto che mancherebbe quel nesso di collegamento tale per cui la comminazione di una pena al robot *x* autore di un reato ponga un freno alla commissione di quel medesimo delitto da parte del robot z (ovverosia qualunque dei consociati).

Quanto sopra detto appare, ad oggi, impossibile nel mondo dell'I.A.: ed invero, l'irrogazione di una sanzione penale nei confronti di una singola macchina non può suscitare alcun «monito di deterrenza nei confronti dell'I.A. complessivamente considerata»[14], atteso che gli agenti artificiali risultano, allo stato, privi di qualsivoglia coscienza etica generalizzata.

A tal punto, è doveroso chiedersi come (e se) un sistema di I.A. possa – in un futuro – capire la portata dell'irrogazione di una sanzione finalizzata non alla punizione isolata di quel singolo evento, bensì alla veicolazione di un messaggio di dissuasione generalizzata.

---

[12] G. Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems*, Dordrecht, Springer, 2015.

[13] F. Mercorio, Big Data*, Intelligenza Artificiale, industria 4.0: un'*overview, in AA. VV., Metropolis*, dai social media all'intelligenza artificiale*, a cura di M. Carvelli, G. Sapelli, libreriauniversitaria.it, Padova 2019, p. 14.

[14] Studio Legale Fornari e Associati, *A.I. – ARTIFICIAL INTELLIGENCE: La frontiera del diritto,* in *www.fornarieassociati.com* (*https://fornarieassociati.com/focus-ai-artificial-intelligence-la-frontiera-del-diritto/*), Milano-Roma, Ottobre 2020.

4. Le macchine come "persone giuridiche"?

Apparirebbe, pertanto, necessario la possibilità di attribuire alle macchine una vera e propria capacità di intendere e volere, che rappresenta, tra l'altro, anche il fondamento affinché un soggetto sia imputabile, e quindi come tale, responsabile penalmente.

Ancora necessaria sarebbe la comprensione, da parte dell'agente, «di maturare un istante rappresentativo ed un istante volitivo da porre a sostegno delle proprie decisioni e conseguenti azioni»[15], in modo che si perfezionino tutti gli elementi indispensabili perché si possa parlare di responsabilità penale a titolo doloso.

In capo alle macchine, tuttavia – a differenza delle società o degli enti – manca la qualifica di "persona giuridica", con la conseguenza che l'irrogazione di una pena nei confronti di un agente artificiale comporterebbe la necessità di istituire una «penalità delle cose»[16], dove le macchine dovrebbero essere riconosciute come entità giuridiche che, sebbene non possano mai raggiungere la posizione dell'essere umano, siano capaci di agire in modo da incidere giuridicamente sul piano penale[17].

Sul piano fattuale, in ogni caso, l'inflizione di una pena ad un agente artificiale autore di un reato assolve all'unica finalità di soddisfare il sentimento di giustizia dell'uomo, che ne risulta offeso o danneggiato dalla condotta *criminis* commessa dalla macchina. Anche se, a parere di chi scrive, la distruzione o la ri-costruzione/riparazione di una "latta" non sarà mai in grado di appagare tale ineluttabile bisogno, soprattutto nelle ipotesi di crimini particolarmente efferati.

Ed in ogni caso, anche a voler ipotizzare l'applicazione di una sanzione nei confronti di un sistema intelligente che ha "voluto" la commissione di quel fatto illecito penalmente rilevante, difficile appare comunque considerare la macchina suddetta quale "colpevole in senso stretto", dal momento che mancherebbe la caratteristica dell'autoderminazione egoistica propria dell'essere umano[18].

Concludendo, solo se si addiverrà ad una concreta possibilità per gli agenti artificiali di introiettare la portata della pena e, dunque, consentire il reale espletamento delle funzio-

---

15 Studio Legale Fornari e Associati, *op. cit.*, pp. 4-6.

16 A. Cappellini, *op. cit.*, p. 20 ss.

17 Studio Legale Fornari e Associati, *op. cit.*, pp. 4-6 .

18 J. Haugeland, *Artificial intelligence: the very idea*, MIT Press Ltd, Cambridge (Massachusetts) 1985.

ni tipiche di questa – basate sul principio di colpevolezza – allora sarà possibile superare l'inutilità dell'irrogazione di una sanzione penale agli attanti robotici; allo stato, però, tale soluzione è pura fantascienza: la macchina, difatti, risulta priva di tutte quelle apprezzabili riflessioni etiche, morali, giuridiche, sociali e culturali proprie degli esseri umani[19].

Non per questo, tuttavia, il diritto penale può farsi cogliere impreparato e, quindi, attendere che il processo tecnologico giunga a tale punto di evoluzione. Il ragionamento sulle risposte giuridiche deve necessariamente precedere la realizzazione e proliferazione di androidi perfettamente sapienti e potenzialmente criminale.

[19] Studio Legale Fornari e Associati, *op. cit.*, pp. 4-6.