Intelligenza Artificiale e responsabilità penale: prime considerazioni*

Riccardo Borsari

Sommario

1. Evoluzione dell'Intelligenza Artificiale e diritto penale

Secondo una definizione accettata a livello internazionale, l'"Intelligenza Artificiale" è quella disciplina, appartenente all'informatica, che studia i fondamenti teorici, le metodologie e le tecniche che consentono di progettare sistemi *hardware* e sistemi di programmi *software* capaci di fornire all'elaboratore elettronico delle prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana[1].

La rapida evoluzione dell'Intelligenza Artificiale e le sempre più numerose applicazioni che essa trova nei diversi settori della vita quotidiana impongono una profonda riflessione sulle sue implicazioni in ambito giuridico. Qui interessano, in particolare, quelle riguardanti il diritto penale (sostanziale), dove lo straordinario sviluppo dell'Intelligenza Artificiale dell'ultimo decennio ha sollevato questioni assai delicate.

La più delicata – sul piano filosofico ed etico, prima ancora che giuridico – è senz'altro quella concernente la possibilità di concepire le entità intelligenti come autori di reato. I sistemi di Intelligenza Artificiale di ultima generazione sono infatti dotati di un grado di autonomia dall'uomo tale da mettere in crisi il modello tradizionale della responsabilità indiretta di quest'ultimo per i fatti di reato verificatisi a causa del comportamento dell'entità di Intelligenza Artificiale.

Meno problematica è invece la questione, postasi soprattutto in previsione di un ulteriore progresso tecnologico, se sistemi di Intelligenza Artificiale possano essere assunti

* Il presente contributo corrisponde, con alcune integrazioni, al testo dell'intervento svolto in occasione del convegno JusTech e Industry 4.0, i cambiamenti indotti dalle nuove tecnologie nel diritto delle imprese, Università degli Studi di Padova, 14 febbraio 2019.

1 M. Somalvico, *Intelligenza artificiale*, Milano, 1987.

ad autonomi beni giuridici meritevoli di tutela penale e, per questa via, a vere e proprie vittime del reato, secondo una prospettiva sperimentata di recente con riferimento a forme di vita "non umane", come gli animali. La qualità di vittima del reato, in effetti, non presuppone gli stessi requisiti psicologici richiesti ai fini della responsabilità penale, che – come si avrà modo di vedere – potrebbero rendere difficoltoso il riconoscimento delle entità intelligenti come autori del reato[2].

2. Intelligenza Artificiale e criminalità

Recenti studi[3] hanno documentato l'impatto straordinario dell'Intelligenza Artificiale in diverse aree criminali.

Ad esempio, in ambito economico (soprattutto nel settore dei mercati finanziari[4]), si è evidenziato come i *social bot* (*software* che, automatizzando *account* di *social media*, simulano di essere utenti umani) – sono stati impiegati per il *pump and dump*, ossia per quella particolare tipologia di frode che consiste nel fare lievitare artificialmente il prezzo di un titolo, mediante dichiarazioni false, fuorvianti o esagerate, con l'obiettivo di vendere titoli acquistati a buon mercato ad un prezzo superiore. Modelli di simulazione di mercati hanno inoltre dimostrato che un agente commerciale artificiale – attraverso l'apprendimento di rinforzo (tecnica di apprendimento automatico basata sull'assegnazione alla macchina di una "ricompensa" a fronte di una scelta corretta) – può imparare la pratica dello *spoofing* finanziario, cioè piazzare ordini, in modo continuativo per un certo periodo di tempo, senza avere l'intenzione di eseguirli, al fine di manipolare i prezzi di mercato.

Un'altra area criminale in cui viene ampiamente sfruttata l'intelligenza artificiale è il traffico di droga detto *business to business*, che si avvale di droni e sottomarini controllati a distanza. I sottomarini senza equipaggio offrono un chiaro esempio del potenziale duplice utilizzo, positivo e negativo, dell'Intelligenza Artificiale: sono stati ideati per scopi legittimi (difesa, protezione delle frontiere, pattugliamento delle acque), rivelandosi tuttavia funzionali anche ad attività illegali.

Ma è nell'area dei reati contro la persona che si riscontrano i più comuni impieghi dell'Intelligenza Artificiale. In particolare, i *social bot* possono essere utilizzati come strumenti di molestie, dirette o indirette (come il *retweeting* o il gradimento di *tweet* negativi allo scopo di creare una falsa impressione di animosità su larga scala nei confronti di una persona) – esemplare la vicenda del *social twitter* "Tay" di Microsoft, che ha rapidamente imparato dall'interazione con gli altri utenti a dirigere *twitter* osceni ad un'attivista femminista.

---

[2] S. Riondato, *Robot: talune implicazioni di diritto penale,* in P. Moro, C. Sarra (a cura di), *Tecnodiritto. Temi e problemi di informatica e robotica giuridica,* Milano, 2017, 85 ss.

[3] T. C. King, N. Aggarwal, M. Taddeo, L. Floridi, *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*, in *Science and engineering ethics*, 14 febbraio 2019, 1 ss.

[4] *Ibidem*. Con riferimento ai sistemi di *common law*, la letteratura analizzata dagli autori ha segnalato preoccupazioni in relazione all'utilizzo di sistemi di AI nelle c.d. *cartel offences*, e in particolare nei casi di manipolazioni del mercato e collusioni poste in essere tramite il mantenimento di prezzi concordati artificialmente e tacitamente.

3. Le entità intelligenti come strumento del reato commesso dall'uomo

Fino al recente passato, oltre alle macchine automatiche di vecchia generazione esistevano soltanto sistemi di Intelligenza Artificiale che operavano mediante algoritmi integralmente preimpostati dal programmatore, il cui comportamento, quindi, è del tutto predeterminato.
Rispetto a questi sistemi di Intelligenza Artificiale (si pensi ai *software* utilizzati per disattivare il sistema di sicurezza informatica di una banca o a quelli per distruggere o danneggiare i dati di un computer), il modello di imputazione della responsabilità indiretta dell'uomo ha tenuto e tiene senza particolari problemi. Infatti, non solo, per quanto complessa sia, l'azione dell'entità intelligente (ad esempio, il movimento di un braccio robotico) può essere imputata, sul piano sia oggettivo (causale) che soggettivo (coscienza e volontà), a chi l'abbia utilizzata, programmata e, per le entità dotate di corpo fisico, costruita; ma soprattutto, poiché il comportamento di siffatti sistemi di Intelligenza Artificiale è predeterminato e perciò prevedibile, ad una di queste persone potrà sempre muoversi un rimprovero, almeno a titolo di colpa, per il fatto di reato verificatosi a causa di detto comportamento. In tale prospettiva, l'entità intelligente si atteggia a mero strumento del reato commesso dall'uomo[5].
Per esempio, se un comandante militare imposta intenzionalmente un drone automatico affinché uccida dei civili, risponderà senz'altro, per dolo, della loro morte; se, invece, il drone spara a dei civili per un difetto di programmazione o di costruzione, a risponderne, per colpa, saranno, rispettivamente, il programmatore o il costruttore, che avrebbero potuto prevedere ed evitare l'evento.
La peculiare natura dello strumento può comunque avere un particolare rilievo sul piano penale – tra l'altro, in sede di commisurazione della pena, quale fattore aggravante o attenuante (v. l'art. 133 c.p., secondo cui il giudice deve tener conto, in primo luogo, dei mezzi dell'azione).
Inoltre, siccome cose che furono destinate alla commissione del reato, le macchine intelligenti possono o devono essere confiscate, in funzione preventiva, anche a prescindere da una sentenza di condanna (v. l'art. 240 c.p.)[6].

4. Le entità intelligenti di ultima generazione e la crisi del modello di imputazione della responsabilità indiretta dell'uomo

Come anticipato, il modello di imputazione della responsabilità indiretta dell'uomo entra in crisi di fronte ai sistemi di Intelligenza Artificiale di ultima generazione, cioè quelle che operano in base ad algoritmi aperti ad automodifiche strutturali, determinate dall'esperienza del sistema stesso.
Grazie ai cosiddetti meccanismi di *machine learning* (apprendimento automatico), un si-

[5] S. Riondato, *op. cit.,* cit., 85 ss.

[6] *Ibidem*.

stema di Intelligenza Artificiale è infatti capace di imparare dall'esperienza e di modificare di conseguenza il proprio comportamento, adattandolo agli stimoli nel frattempo ricevuti – uno degli esempi più rilevanti è quello delle automobili a guida autonoma. In molti casi, peraltro, mediante il ricorso alle tecnologie di *cloud computing*, un'entità intelligente, scambiandosi informazioni con altre, anche operanti in ambienti diversi, può incrementare esponenzialmente il proprio apprendimento. Senza considerare – soprattutto in chiave futura – la possibile interazione in *cloud* di sistemi di Intelligenza Artificiale e uomini nel vasto mondo dell'*Internet of Things*.

È evidente, allora, che il comportamento di tali sistemi in Intelligenza Artificiale non è interamente predeterminato, e perciò prevedibile, con la conseguenza che può divenire problematico individuare una persona umana cui muovere un rimprovero per il fatto di reato si sia verificato a causa di tale comportamento.

Il problema si pone in termini sicuramente meno gravi quando, a causa del comportamento imprevedibile dell'entità intelligente, si sia verificato un evento penalmente sanzionato che comunque l'uomo – di regola, l'utilizzatore – si sia prefisso e abbia voluto. Si ritiene, infatti, che, in caso di divergenza tra il decorso causale prefigurato e quello effettivo, sia comunque ravvisabile il dolo in capo all'agente se l'evento verificatosi è pur sempre realizzazione dello specifico rischio insito nella sua azione iniziale. Resta, tuttavia, che, secondo l'interpretazione conforme al principio costituzionale di personalità della responsabilità penale (art. 27, c. 1, Cost.), l'autore non potrà essere chiamato a rispondere dell'evento penalmente sanzionato (verificatosi a causa del comportamento imprevedibile dell'entità intelligente) più grave (delitto preterintenzionale – per es. omicidio preterintenzionale – art. 584 c.p.) o comunque diverso da quello da lui voluto (*aberratio delicti* – art. 83 c.p.) oppure nei confronti di una persona diversa da quella cui l'offesa era diretta (*aberratio ictus* – art. 82 c.p.).

Ma è soprattutto fuori dalle ipotesi in cui il comportamento imprevedibile dell'entità intelligente si innesti su di una condotta dolosa della persona umana che il modello di imputazione indiretta della responsabilità dell'uomo si rivela inadeguato. Perché, se l'utilizzatore non può essere chiamato a rispondere di un evento verificatosi a causa del comportamento imprevedibile dell'entità intelligente, almeno finché non sia comunque in grado evitarlo (si pensi ad un'automobile a guida semi-autonomatica, che sia cioè dotata di comandi che consentano all'utente di intervenire in caso di emergenza), non potrà nemmeno risalirsi sempre al programmatore o al produttore.

Il funzionamento degli agenti intelligenti di ultima generazione si articola infatti sulle metodologie di *deep learning,* che sono fondate su quelle che vengono definite le *black box algorithms*. In queste tecnologie il processo che dagli *input* conduce agli *output* rimane avvolto da un inevitabile grado di opacità, per cui non si riesce a comprendere come il *software* abbia posto in essere il risultato finale[7], il quale rimane al di fuori delle capacità previsionali dei programmatori[8].

Il rischio, pertanto, è che interessi anche primari rimangano sguarniti di tutela penale

[7] S. Doncieux, J. Mouret, *Beyond black-box optimization: a review of selective pressures for evolutionary robotics*, in *Evolutionary Intelligence*, 7, 2014, 71 ss.

[8] S. Beck, *Google cars, software agents, autonomous weapons systems – New challenges for criminal law?*, in E. Hilgendorf, U. Seidel (eds.), *Robotics, Autonomics, and the Law*, Baden-Baden, 2017, 227 ss.

di fronte a pericoli di aggressione che evidentemente diventeranno sempre più gravi e diffusi.

5. Superamento dell'assioma del *machina delinquere (et puniri) non potest*?

Di fronte a siffatto scenario, ci si chiede se le macchine intelligenti non possano essere concepite come soggetti attivi del reato, superandosi l'assioma del *machina delinquere* (*et puniri*) *non potest*.

5.1. La tesi positiva di Gabriel Hallevy

Una risposta positiva a tale quesito proviene da una parte della letteratura straniera di *common law* e, in particolare, da Gabriel Hallevy[9].
Secondo l'Autore, non ci sono ragioni valide per negare la punibilità dei sistemi di Intelligenza Artificiale.
In particolare, quanto, anzitutto, all'elemento oggettivo del reato, l'*actus reus* – inteso, com'è negli ordinamenti di *common law*, in termini meramente materialistici – potrebbe essere ricondotto direttamente al sistema di intelligenza artificiale, sia che si tratti di una condotta attiva (integrata da un movimento fisicamente apprezzabile della macchina – ad esempio, il movimento di un braccio robotico) sia che si tratti di un'omissione (integrata dall'inerzia della macchina).
In secondo luogo, per quanto concerne il profilo psicologico, in capo all'entità intelligente potrebbero benissimo configurarsi alcune forme di *mens rea* e, in particolare, la *negligence* e addirittura il *general intent* (categoria dogmatica che ricomprende *intention*, *knowledge* e *recklessness*). Infatti, guardando alle modalità tecniche di funzionamento delle macchine più avanzate, esse, acquisendo i dati dal mondo esterno e rielaborandoli, "si rappresentano" la realtà, potendo definirsi quali *aware* in virtù delle capacità, da una parte, di assorbire dati fattuali attraverso i molteplici sensori di cui dispongono, dall'altra, di creare una rilevante generale immagine della realtà dall'analisi dei dati raccolti.
Inoltre, secondo l'autore, i sistemi di IA di ultima generazione possono "prevedere" e "volere" un certo risultato come conseguenza della propria azione, grazie ai processi di *decion-making* che, dalla valutazione delle probabilità che un determinato evento possa verificarsi, permettono alla macchina di direzionare il proprio operato conseguentemente.
Infine, l'agire dell'agente intelligente può qualificarsi anche quale *reckless* (imprudente), nei casi in cui il sistema non abbia preso in considerazione una probabilità che avrebbe dovuto essere compresa sulla base degli *inputs* inseriti, o nei casi di *miscalculation*, quando si verifica un errore di calcolo nei processi di apprendimento. È irrilevante, peraltro, che la macchina non provi sentimenti, in quanto estranei di regola al dolo.
In definitiva, per Hallevy, alla base dell'assioma del *machina delinquere* (*et puniri*) *non potest*

[9] G. Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems*, Berlino, 2015, 47 ss.

sta lo stesso pregiudizio antropocentrico che per molto tempo si è opposto al riconoscimento della responsabilità penale delle persone giuridiche, finché non è giustamente prevalsa l'esigenza di regolazione sociale. D'altra parte, a differenza delle persone giuridiche, i sistemi di intelligenza artificiale sono dotati sempre – in quanto *software* pur sempre funzionanti tramite *hardware* – di un corpo fisico su cui può incidere la pena (ad esempio, distruzione).
Hallevy ha quindi teorizzato tre paradigmi di responsabilità, tutti fondati sul presupposto necessario del riconoscimento della personalità giuridica alle entità intelligenti.
Il primo, definito *perpetration through another*, rappresenta l'aggiornamento di quello, tradizionale, di responsabilità indiretta dell'uomo: in base ad esso, i sistemi di intelligenza artificiale sono ricondotti alla categoria degli "agenti innocenti", che vengono strumentalizzati per la commissione del reato da una persona umana, che potrà individuarsi nel programmatore del *software* o nell'utente finale, e che ne risponderà in via esclusiva.
Il secondo (*natural probable consequence*) e il terzo (*direct liability*) paradigma prevedono, invece, la possibilità di individuare una responsabilità dell'entità intelligente, in via cumulativa o autonoma rispetto alla responsabilità del programmatore e/o dell'utente.

5.2. Le obiezioni mosse alla teoria di Hallevy

Alla tesi di Hallevy viene obiettato essenzialmente che al riconoscimento dei sistemi di intelligenza artificiale come soggetti attivi del reato si oppone il principio di colpevolezza che informa di sé gli ordinamenti penali moderni. In quello italiano, tale principio è sancito dal già citato art. 27, co. 1, cost., secondo cui «la responsabilità penale è personale».
Infatti, nell'ambito delle tradizionali teorie sulla pena, la colpevolezza, che ne rappresenta il presupposto logico necessario, postula a sua volta la libertà del volere, per il semplice fatto che, in tanto ha senso infliggere una pena che compensi il male arrecato dal reato, in quanto il suo autore avesse la possibilità di agire diversamente. Una tale possibilità non apparterrebbe ai sistemi di intelligenza artificiale, perché anche quelli capaci di agire indipendentemente e in modo imprevedibile sono programmati per farlo; né, comunque, la responsabilità di tali sistemi è ancora un'idea acquisita nella nostra coscienza morale e nella nostra vita sociale, a differenza di quella dell'uomo, rispetto al quale pure la libertà del volere non è stata ancora dimostrata e, anzi, è stata di recente messa in discussione dalle neuroscienze.
D'altra parte, nei confronti di un sistema di intelligenza artificiale, la pena non potrebbe svolgere nemmeno una delle diverse funzioni generalmente attribuite alla pena. In ordine a quella di prevenzione generale, basti considerare che i sistemi di intelligenza artificiale non sono capaci di provare timore e sono quindi immuni dall'effetto dissuasivo della minaccia della pena e, tantomeno sono in grado di cogliere l'effetto pedagogico connesso alla comminatoria legislativa della sanzione più grave, di accreditamento sociale dei valori tutelati. Per quanto concerne, invece, la funzione di prevenzione speciale – intesa, alla stregua del principio rieducativo enunciato dall'art. 27, co. 3, Cost. – come risocializzazione e, quindi, non come mera neutralizzazione (ad esempio,

disattivazione della macchina) –, i sistemi di intelligenza artificiale non sono in grado di apprendere dalla sanzione irrogata, a meno che non siano a ciò programmati; mentre un trattamento per così dire terapeutico, mediante l'implementazione di meccanismi di *machine learning*, non potrebbe essere imposta, richiedendo il consenso del destinatario. All'argomento dell'assimilazione dei sistemi di intelligenza artificiale agli enti, si è peraltro replicato che, nell'ambito della *corporate liability*, sia il comando sotteso alla norma incriminatrice che la pena si rivolge in definitiva a uomini.

5.3. Alternative

Di fronte al vuoto di tutela penale connesso ai nuovi sistemi di Intelligenza Artificiale, a parte i possibili rimedi civilistici o amministrativi (si potrebbero ipotizzare misure corrispondenti a quelle sanitarie previste per gli animali pericolosi), le strade sono due: o si vieta radicalmente la realizzazione di tali sistemi, in base al principio di precauzione, con la conseguente rinuncia ai benefici sociali apportati dagli stessi; oppure si individua un'area di rischio consentito, attraverso complessi bilanciamenti tra l'utilità collettiva e i rischi imponderabili dei vari sistemi. Ad esempio, potrebbe essere consentito l'ingresso sul mercato di auto senza conducente, nonostante il rischio di cagionare lesioni, se non la morte, ad altri utenti della strada, perché tali auto sono comunque migliore soluzione per i problemi del traffico; e vietare senz'altro la realizzazione di droni armati.
Non manca nemmeno, tuttavia, chi, in funzione provocatoria, ipotizza pratiche afflittive *lato sensu* penali nei confronti dei sistemi di intelligenza artificiale, che, in quanto *res*, potrebbe essere sacrificata per apportare un beneficio psicologico della vittima, assecondando la sua irrazionale sete di vendetta.