

Da “nemica” ad “amica”: il ruolo dell’IA nella lotta alla criminalità digitale*From “enemy” to “ally”: the role of AI in the fight against Cybercrime*

Alyssa Chiara Montan

Dottoranda in Giurisprudenza presso l’Università degli Studi di Bergamo

Sommario: 1. La *cyber*-rivoluzione e la «ri-ontologizzazione» del diritto. – 2. Le nuove forme di *cyber*-criminalità e il ruolo emergente dell’Intelligenza Artificiale. – 3. Origini, definizione e caratteristiche dell’Intelligenza Artificiale – 4. Le diverse tipologie di Intelligenza Artificiale: IA «forte» *vs* IA «debole». – 5. L’Intelligenza Artificiale “nemica”: l’IA quale strumento e/o autore del reato. – 6. L’Intelligenza Artificiale “amica”: il suo utilizzo difensivo nella lotta alla criminalità digitale.

ABSTRACT

Il presente contributo esamina lo sviluppo dell’Intelligenza Artificiale nel contesto della rivoluzione cibernetica, soffermandosi sulla fenomenologia *cyber* e sulle diverse ricostruzioni dottrinali in merito alla nozione di IA «forte» o «debole»; in particolare, ci si interroga sulle questioni giuridiche poste dalle caratteristiche dei dispositivi intelligenti, dall’attribuzione di responsabilità per eventuali condotte illecite, dalla possibilità e opportunità di intenderli «da mezzi ad autori del reato»¹ e dal loro impiego nel rafforzamento della *cyber*-sicurezza.

This paper examines the development of Artificial Intelligence in the context of the cyber revolution, focusing on cyber phenomenology and the various doctrinal interpretations regarding the concept of ‘strong’ and ‘weak’ AI; in particular, it explores the legal issues concerning the characteristics of intelligent devices, the attribution of liability for any unlawful conduct, the possibility and opportunity of treating them ‘not merely as instruments but as perpetrators of the offence’, and their use in strengthening cybersecurity.

¹ C. PIERGALLINI, *Danno da prodotto e responsabilità penale. Profili dogmatici e politico-criminali*, Giuffrè, Milano, 2004, e C. PIERGALLINI, *Intelligenza artificiale: da “mezzo” ad autore del reato?*, in *Rivista italiana di diritto e procedura penale*, 4/2020, pp. 1745–1774.

1. La cyber-rivoluzione e la «ri-ontologizzazione» del diritto

Il massiccio processo di globalizzazione e le significative scoperte in ambito robotico e informatico – tra cui, l’automazione, con algoritmi in grado di riprodurre in modo sempre più accurato il comportamento degli individui – hanno reso il Duemila il secolo delle avanguardie scientifiche e tecnologiche: la cd. “quarta rivoluzione” – o rivoluzione cibernetica – ha infatti inciso profondamente sul tenore di vita degli individui, soprattutto sul piano relazionale e, di conseguenza, giuridico (L. Floridi, 2009)². Essa ha determinato infatti il consolidarsi di una nuova dimensione interattiva: il *cyberspace*, una fitta rete di scambi e comunicazioni all’interno della quale il singolo si inserisce come mero ingranaggio di un sistema complesso; uno spazio al contempo virtuale e reale che si articola in una strutturazione di tipo “verticale”, tra *Surface Web* – quella parte di Internet visibile a ciascun utente medio –, *Deep Web* – quella parte non indicizzata dai motori di ricerca e dunque non accessibile né esplorabile con dei *browser* comuni – e *Dark Web*, ossia quella parte più “recondita” della Rete, composta da segmentazioni raggiungibili esclusivamente tramite *software* specifici, progettate per rimanere nascoste e non tracciabili grazie all’utilizzo di tecniche di anonimizzazione e VPN (V. Lagi, 2023)³.

L’entusiasmo con cui questa ventata di innovazione è stata inizialmente accolta è stato però presto oscurato da una nuova consapevolezza: l’uomo moderno ha dovuto «comprendere che è il passato, noto e visibile, a essere davanti a noi; mentre il futuro, che è sconosciuto e misterioso, sta dietro: alle nostre spalle» (M. Papa, 2023); e ciò risulta ancor più vero per il giurista, che rischia di passare da «“non informatizzato” a sostanzialmente “analfabeta”» (U. Pagallo, 2023)⁴. Egli si è infatti trovato di fronte a una sfida inedita: evitare di

² L. FLORIDI, *Infosfera: etica e filosofia nell’età dell’informazione*, G. Giappichelli Editore, Torino, 2009

³ Il cd. *Deep Web* è una Rete di comunicazione privata e anonima, dotata di «sistemi per la protezione della privacy non solo verso i visitatori esterni, ma anche per l’identità di coloro che la utilizzano»; mentre il *Dark Web* «può definirsi una rete privata ed anonima tramite la quale si possono commettere vari tipi di reati (...) in cui si sfruttano la privacy e l’anonimato forniti per portare a termine operazioni illegali», quali «lo scambio di denaro in maniera anonima, tramite una delle cosiddette “cripto valute”» oppure armi o sostanze stupefacenti (V. Lagi, 2023). Tratto da: A. CADOPPI, S. CANESTRARI A. MANNA, M. PAPA (a cura di), *Cybercrime*, Omnia Trattati Giuridici, UTET Giuridica, Wolters Kluwer Italia s.r.l., Milano, 2023, II edizione.

⁴ Sia U. PAGALLO, 2023 che M. PAPA, in A. CADOPPI, S. CANESTRARI A. MANNA, M. PAPA (a cura di), op.cit.

rimanere intrappolato in un paradosso, in cui il diritto tenta di regolare mutamenti sociali già superati dal progresso tecnologico, risultando, di fatto, costantemente “in ritardo” rispetto agli stimoli provenienti dalla società digitale; come nella famosa ed efficace parabola di Achille e della tartaruga (U. Pagallo, 2023), arricchita e complicata dall’introduzione di un terzo elemento, ovvero la tecnologia, costantemente un passo avanti rispetto alle altre due variabili.

Tale ritardo si traduce nel progressivo indebolimento di due dei pilastri della tradizionale impalcatura normativa-statale, ossia il principio di territorialità – di cui il cyberspazio è del tutto privo – e quello di effettività dell’apparato sanzionatorio, limitato rispetto ai bisogni del “popolo di Internet”. In particolare, si è oggi indotti a dubitare della validità ed efficacia degli strumenti di prevenzione e deterrenza attualmente a disposizione dell’ordinamento: da un lato, nelle prime riflessioni in tema di Intelligenza Artificiale, si allude sempre più al rischio di lacune normative difficilmente colmabili mediante le sole categorie giuridiche tradizionali, oltre che alle ben più complesse questioni relative all’opportunità e alla necessità di assoggettare a sanzioni di tipo civilistico, amministrativo, ma soprattutto penale gli agenti artificiali autori di fatti illeciti⁵; dall’altra, il sorgere di nuove aspettative sociali e il riconoscimento di nuovi interessi giuridicamente rilevanti, legati al rimodellamento dei conflitti interni alla Rete, ha determinato il proliferare di nuovi comportamenti illeciti e inedite forme di violazione dei diritti, legate al riconoscimento di nuovi beni giuridici meritevoli di protezione (L. Picotti, 2023).

Per tutte queste ragioni, il mondo dell’informatica è stato identificato come un vero e proprio “sistema regolativo” a se stante, al pari di quello giuridico (L. Lessing, 1999); da cui la successiva riflessione sulle influenze reciproche tra il sistema del diritto e quello del mondo ICT – se i loro rapporti possano configurarsi in senso conflittuale, contrastandosi fino a rendere

⁵ U. PAGALLO, *Saggio sui robot e il diritto penale*, in VINCIGUERRA S., DASSANO F. (a cura di), *Scritti in memoria di giuliano Marini*, Edizioni Scientifiche Italiane, Napoli 2010. Emblematica è, a tal proposito, la domanda: «in che senso, una volta riconosciuto colpevole, il robot potrebbe mai pagare il suo debito nei confronti della società? In che modo la sua punizione potrebbe riformarlo al fine di non compiere nuovamente simili delitti? Ovvero, secondo i dettami della pena in funzione della prevenzione generale, in che modo la condanna inflitta al robot potrebbe rappresentare un avvertimento, o una minaccia, per la generalità dei consociati?» (U. Pagallo, 2010).

superflui o inefficaci i precetti dell’uno o dell’altro; in senso reciproco, rafforzandosi e tutelandosi vicendevolmente; oppure in una logica di indifferenza, procedendo ciascuno lungo la propria traiettoria evolutiva, nel tentativo di perseguire una presunta “neutralità” sia in ambito di tutela sia in ambito sanzionatorio – e sull’opportunità di un’effettiva «ri-ontologizzazione» del diritto a fini evolutivi (L. Floridi, 2017)⁶. Nello specifico, ci si chiede: quanto e come il sistema penale attuale è in grado di fronteggiare le forme di criminalità digitale? Quali contributi può offrire la digitalizzazione, con i suoi innovativi strumenti e in quali aspetti, invece, i nuovi dispositivi intelligenti sono da considerarsi dei “nemici”, sconosciuti e incontrollabili?

2. Le nuove forme di *cyber*-criminalità e il ruolo emergente dell’Intelligenza Artificiale

«Non solo l’universo tecnologico è in frenetica evoluzione, ma tanto più è veloce il progresso tecnologico quanto maggiori sono i rischi legati a un suo utilizzo dannoso», a maggior ragione in quanto, accanto alle tradizionali aggressioni alla sicurezza esterna e interna, si sono ormai affermate anche «tutte quelle forme di minaccia della sicurezza che si muovono non più sul classico piano della realtà fisica, bensì su quello della realtà virtuale» (L. Moroni, 2024)⁷. Tali forme di “minaccia della sicurezza” – comunemente definite *cyber*-attacchi o attacchi cibernetici – prendono di mira proprio quelle vulnerabilità prodotte dalla rapida e incontrollata digitalizzazione: l’obsolescenza delle infrastrutture tecnologiche, l’elevata interconnessione dei dispositivi, la carenza di investimenti e personale specializzato, la crescente sofisticazione delle tecniche di attacco e la frammentazione normativa e organizzativa, oltre a tutte quelle fragilità legate agli errori umani, quali *password* deboli, mancati aggiornamenti *software* e mancato riconoscimento delle campagne di *phishing*⁸.

⁶ L. FLORIDI, *La quarta rivoluzione. Come l’infosfera sta trasformando il mondo*, F. Angeli, Milano, 2017.

⁷ L. MORONI, *La governance della cybersicurezza a livello interno ed europeo: un quadro intricato*, in *federalismi.it – Rivista di diritto pubblico italiano, comparato, europeo (rivista online)*, 14/2014, pp. 179–197.

⁸ ICT SECURITY MAGAZINE, *MDR nella sanità. Cyber Security in ambienti critici*, disponibile su: <https://www.ictsecuritymagazine.com/notizie/mdr-nella-sanita-cyber-security-in-ambienti-critici/>; articolo pubblicato il: 14 giugno 2022; ultimo accesso il: 30 maggio 2026. AGENDA DIGITALE, M. DAL CO, *Perché gli hacker puntano sulle cartelle cliniche: rischi e costi*, disponibile su: <https://www.agendadigitale.eu/sicurezza/perche-gli-hacker-puntano-sulle-cartelle-cliniche-rischi-e-costi/>; articolo pubblicato il: 2 febbraio 2026; ultimo accesso il: 30 maggio 2026. N.

La fenomenologia della criminalità informatica negli ultimi decenni si esplica perlopiù in forme di attacco *ransomware*⁹, *Distributed-Denial-of-System* (D.D.o.S.)¹⁰, *data breach* – le cd. brecce nel sistema informatico – e *phishing*¹¹, i quali a loro volta non richiedono più particolari abilità tecniche di progettazione, attivazione e sviluppo, potendo invece essere portati a compimento prendendo in prestito *malicious software* da appositi *botnet*, previo collegamento alle reti nel *Dark Web* (Microsoft Digital Defense Report, 2020). Non a caso, per queste specifiche tipologie di attacchi, il Codice Penale ha progressivamente individuato e sviluppato – a partire dalla riforma del 1993

MOHAMMED, A.F. MOHAMMED, S. BALAMMAGARY, Ransomware in Healthcare: Reducing Threats to Patient Care, in J.C.C.C.I. Journal of Cognitive Computing and Cybernetic Innovations, vol. 1, 2/2025, disponibile online su: <https://doi.org/10.21276/jccci.2025.v1.i2.5/> (articolo pubblicato il: 27 luglio 2025). LI B.S., S. KAMALAKAR M.D., M.P.H., P. NISHANT, M.B.B.S., *Cyber-Attacks on Hospital Systems: A Narrative Review*, in ELSEVIER, Science Direct, vol. 7, 1/2025 (S2950386825000103), pp. 30-39, disponibile online su: <https://doi.org/10.1016/j.osep.2025.03.002>. CYBERGLOBAL, V. NEAGU, *Hospitals Under Attack: Ransomware in Healthcare*, disponibile su: <https://cybergl.com/blog/ransomware-in-healthcare/>; articolo pubblicato il: 4 novembre 2025; ultimo accesso il: 30 maggio 2026. TECNICA OSPEDALIERA, *Cyber Crime in sanità +74%, consigli per la patient data security*, disponibile su: <https://www.tecnicaospedaliera.it/cyber-crime-in-sanita-74-consigli-per-la-patient-data-security/>; articolo pubblicato il: 8 maggio 2023; ultimo accesso il: 30 maggio 2026.

⁹ Un attacco *ransomware* si compone: «(i) di una prima fase più propriamente “informatica”, nella quale l’attaccante introduce in vario modo un *malware* nel computer *target*, per criptare file, altri contenuti o il sistema stesso e renderli così inaccessibili al titolare (più nel dettaglio: ENISA, 2022); e (ii) di una seconda fase invece “estorsiva”, nella quale l’attaccante richiede al bersaglio il pagamento di un riscatto (solitamente in criptovalute), in cambio del ripristino della funzionalità o dell’accessibilità dei dati o del sistema (Cybercrime Convention Committee 2022)» (G. Fiorinelli, 2024). La modalità di attacco *ransomware*, che si configura sempre più non in forma di estorsione semplice, quanto piuttosto come un’aggressione “a doppia estorsione” – attacchi che combinano la cifratura dei dati con il loro furto preventivo, permettendo, così, al *cyber-criminale* prima di estorcere le informazioni e poi di creare ulteriore leva per gli attaccanti tramite la loro necessità di recuperare l’utilizzo delle macchine cifrate –, sfrutta, quali vettori di attacco, strumenti (perlopiù “a doppio uso”) liberamente disponibili o funzionalità e vulnerabilità note e facilmente accessibili in applicazioni esposte, quali il c.d. *exploitation of known vulnerabilities through public-facing applications*; tra cui, per citarne una, «*Citrix Bleed* (CVE-2023-496), which occurs in Citrix NetScaler ADC and NetScaler Gateway», una vulnerabilità che nel corso del biennio 2023-2024 ha attratto tentativi di sfruttamento su larga scala (ENISA Cybercrime Threat Landscape, 2024).

¹⁰ «Condotte lecite se prese singolarmente ma nocive quando effettuate su larga scala» (L. Picotti, 2023)

¹¹ Attacchi informatici in cui l’autore, tramite artifici e raggiri, si adopera per ottenere le credenziali di autenticazione necessarie ad accedere abusivamente a spazi informatici esclusivi, per poi svolgere all’interno di queste ultime operazioni bancarie o finanziarie ai danni del legittimo titolare.

fino al recentissimo D.D.L. Cybersicurezza, ora legge n. 90 del 2024 – una serie di fattispecie *ad hoc* al fine di garantirne la punibilità. Infatti, rilevano: rispetto alla fenomenologia *phishing*, gli articoli 640-ter c.p. (frode informatica) o, in alternativa, 640 c.p. (la frode semplice avvenuta su Internet), 615-ter c.p. (accesso abusivo a un sistema informatico) e 494 c.p. (sostituzione di persona tramite furto di identità digitale), eventualmente aggravati *ex* articolo 61 c.p. e posti in continuazione *ex* articolo 81 cpv. c.p.; rispetto alla numerosa casistica *ransomware*, il microsistema degli accessi abusivi¹², il microsistema dei danneggiamenti informatici¹³ e l’articolo 629, III comma c.p., inserito appositamente per tipizzare la fattispecie dell’estorsione informatica¹⁴; rispetto al verificarsi di *data breach*, oltre a un corposo impianto di illeciti amministrativi, tre fattispecie delittuose all’interno del GDPR (Regolamento Europeo n. 679 del 2016), ovvero l’inosservanza di provvedimenti del Garante (articolo 45 del suo decreto attuativo, il d.lgs. n. 51 del 2018), la falsità in atti e dichiarazioni al Garante (articolo 44 del suddetto decreto, nonché articolo 168 del Codice della

¹² Gli articoli 615-ter «accesso abusivo ad un sistema informatico o telematico»; 615-*quater* «detenzione, diffusione e installazione abusiva di apparecchiature, codici e altri mezzi atti a consentire l’accesso non autorizzato ad un sistema informatico o telematico»; e 615-*quinques* «detenzione, diffusione e installazione abusiva di apparecchiature, dispositivi o programmi informatici diretti a danneggiare o interrompere il funzionamento di un sistema informatico o telematico».

¹³ Articoli 635-bis, *ter*, *quater* e *quinques* c.p. rispetto a dati e/o sistemi informatici privati e/o pubblici.

¹⁴ Il Capo II dell’ex D.D.L. Cybersicurezza, ad oggi legge n. 90 del 2024 – in particolare all’articolo 15, primo comma, lett. m) – ha insistito sull’introduzione di un’apposita fattispecie penale volta a tipizzare la crescente fenomenologia degli attacchi *ransomware*: si tratta dell’«estorsione informatica» o «estorsione mediante reati informatici», formulata al fine di «reagire alla “gravità” e alla “frequenza” dei “ricatti realizzati attraverso la minaccia o l’attuazione di attacchi informatici”», i quali, anche sulla base dei rapporti ENISA (datati dal 2022 in avanti), rappresentano ormai «una delle forme più gravi di cybercriminalità dell’ultimo decennio, ai danni di individui, imprese e pubbliche amministrazioni (...) che, in virtù della particolare complessità dell’iter criminis, risulta tuttora priva di una rilevanza penale unitaria» (G. Fiorinelli, 2024). Tuttavia, sebbene sia da accogliersi con favore questo tentativo di aggiornare le fattispecie penali, vi sono alcune censure da obiettarsi, in particolare il fatto che «la realtà degli attacchi *ransomware* avrebbe reso forse preferibile, in luogo del ricorso *tout court* alla fattispecie di estorsione, il “recupero” del modello del “reato commesso a scopo di estorsione” (...) perché, nel caso del delitto commesso a scopo di estorsione il reato sarebbe integrato a prescindere dall’avvenuto versamento del “prezzo della liberazione” dei sistemi informatici o dei dati (posto che il profitto costituirebbe soltanto l’oggetto del dolo specifico), mentre nell’estorsione il conseguimento del profitto ingiusto rappresenta un elemento costitutivo del reato, in assenza del quale può ritenersi tutt’al più integrata la fattispecie tentata (...)» (*ibidem*).

Privacy) e l’illecito trattamento dei dati (articolo 43 del suddetto decreto, nonché articoli 167 e 167 *bis* del Codice della Privacy). Ai «delitti informatici» e al «trattamento illecito di dati» si fa infine riferimento agli artt. 24 e 24-*bis* del decreto 231 del 2001 quali reati-presupposto per la responsabilità da reato dell’ente.

Al netto di questa precisa casistica, però, rilevano tutt’oggi notevoli criticità su diversi aspetti di interesse, tra cui il luogo di consumazione dei reati (*locus commissi crimini*), la questione della competenza – se sia territoriale o piuttosto transnazionale, in linea con la concezione “globale” del *cyberspace* – e l’identificazione degli autori. Per questo, il *focus* dottrinale tende a proiettarsi nel breve periodo verso la progettazione di nuovi strumenti investigativi, sempre più coadiuvati da tecnologie di Intelligenza Artificiale capaci di captare comportamenti anomali in tempi reali, e verso lo sviluppo di una maggiore cooperazione tra le diverse giurisdizioni comunitarie e internazionali – aspetto di fondamentale importanza nei casi in cui i *cyber-criminali* si rivelino in grado di sfruttare tecniche di anonimizzazione e VPN per restare nascosti all’interno del *Dark Web* –.

All’interno di questo scenario, c’è anche chi si sofferma a riflettere sull’uso dell’IA come “strumento” per commettere crimini e, dunque, sul suo impatto diretto sui *Cyber-crime*.

A tal proposito, le fattispecie di reato rientranti nell’ambito della criminalità informatica ma connesse per mezzo o con l’ausilio di sistemi di IA, prendono il nome di *AI-Crimes*, «“cases in which an AI would be criminally liable if a natural person had performed the same act”»; tuttavia, tra le diverse voci della dottrina, c’è anche chi si spinge oltre, fino a prospettare vere e proprie forme di «*Hard AI Crime (...) instead defined as “scenarios where crimes are functionally committed by machines and there is no identifiable person who has acted with criminal culpability”*» (A. Giannini, 2021)¹⁵. Ci si chiede, infatti, se i dispositivi dotati di Intelligenza Artificiale possano essere effettivamente utilizzati non solo «in funzione “servente”, come mero strumento di potenziamento di alcune attività umane, sottoposto a rigido controllo (...) ma anche, e soprattutto, con riguardo alla penetrazione delle cc.dd. macchine pensanti, la cui caratteristica è quella di essere autonome (...)», cioè se possano passare a essere «da mezzi ad autori di reato» (C. Piergallini, 2025); nello specifico, «irrompe, con forza, la necessità di

¹⁵ A. GIANNINI, *Criminal behavior and accountability of artificial intelligence systems – Dissertation to obtain the degree of Doctor at the University of Florence and at Maastricht University*, 2020 e A. GIANNINI, *Artificial Intelligence, Human Oversight and Criminal Liability: a European “Strenght Test”?*, in *Criminalia – Annuario di Scienze penalistiche (DisCrimen)*, 2021.

interrogarci sul “margine di sicurezza”, cioè sul grado di rischio che si è disposti a tollerare (...) concernente l’individuazione dei centri di responsabilità ma, che, prima ancora, prova a decidere come governare, preventivamente, questi rischi, interrogandosi sul livello di tollerabilità sociale» (*ibidem*).

Se anche non si può negare che si profilano già diversi scenari in cui le Intelligenze Artificiali vengono utilizzate in ambito criminoso – soprattutto con riferimento alle attività di *targeting* (L. Picotti, 2023) e alla manipolazione del mercato finanziario attraverso ordini falsi, le pratiche di *deceifful false-orders* –, di fatto è bene ricordare che, allo stato attuale della tecnica e dell’innovazione digitale, i dispositivi dotati di IA siano ancora concepiti in termini di dipendenza dagli esseri umani; ovvero, come siano ancora generalmente intesi quali «strumenti del reato» più che agenti del reato stesso (D. Provolo, S. Riondato, F. Yenisey, 2014)¹⁶. Da cui, il loro principale utilizzo in pratiche di disattivazione dei sistemi di sicurezza informatica (solitamente supportati da analoghi meccanismi protettivi dotati di IA) e di distruzione o danneggiamento di dati altrui, ma anche nella sempre più numerosa casistica degli *identity theft* – i furti di identità digitale, i quali a loro volta risultano particolarmente rilevanti nel momento in cui riguardano dati biometrici o genetici –; all’interno di questi ambiti summenzionati, infatti, i *social media bots* e gli altri dispositivi intelligenti possono sfruttare al meglio la loro capacità di acquisire fiducia e dati personali, affidati loro direttamente dagli utenti, per poi riutilizzarli nell’integrazione degli stessi ad apposite strategie di *spear phishing* – il c.d. “*phishing* mirato”, economicamente efficace proprio grazie alla personalizzazione effettuata a mezzo delle tecniche di *machine learning* – o, nei casi più gravi, nella duplicazione di modelli di linguaggio individuale a partire da brevi registrazioni e note vocali – le cd. *skills* di *voice synthesis technology* –, con conseguenze imprevedibili e alquanto pericolose soprattutto in termini di *hacking* dei sistemi di sicurezza biometrica basati sulla voce.

Pertanto, risulta sempre più imprescindibile un’attenta analisi di quegli aspetti configurati dalla dottrina quali “elementi chiave” del fenomeno A.I.C. . Si fa qui riferimento alle quattro categorie di: «*Emergency*», la quale «riguarda il rischio che gli agenti artificiali, una volta implementati, agiscano in modi sofisticati e inattesi, sviluppando azioni e piani coordinati autonomamente»;

¹⁶D. PROVOLO, S. RIONDATO, F. YENISEY (a cura di), *Genetics, Robotics, Law, Punishment*, Padova University Press, Padova, 2014, I edizione

«*Liability*», per cui gli AIC minacciano «i modelli di responsabilità penale esistenti, in particolare riguardo all’*actus reus* (atto volontario) e alla *mens rea* (mente colpevole)»; «*Monitoring*», per cui il monitoraggio di questi reati sarebbe ostacolato da problemi di attribuzione, fattibilità e azioni *cross-system*; e «*Psychology*», per cui «l’AI può facilitare il crimine influenzando lo stato mentale degli utenti, ad esempio manipolandoli»¹⁷. Per comprendere al meglio l’enormità di tali fenomeni, però, è prima fondamentale riuscire a inquadrare che cosa realmente si intenda per «dispositivo dotato di Intelligenza Artificiale».

3. Origini, definizione e caratteristiche dell’Intelligenza Artificiale

La storia dell’Intelligenza Artificiale ha radici molto profonde: «nata con l’antico desiderio di “fabbricare gli Dei” (“*forge the gods*”) è un’idea che ha pervaso continuamente (...) la civiltà occidentale» (G.F. Italiano, 2025)¹⁸; fermo restando questo eterno interesse per un problema affascinante e, in un certo senso, “spaventoso”, fu però solo a partire dal 1940 che la possibilità di costruire per davvero una “macchina pensante” cominciò a farsi concretamente strada nella mente delle persone: l’invenzione e la disponibilità dei primi elaboratori elettronici, infatti, «riuscì a catalizzare l’interesse di scienziati provenienti da varie discipline, come la psicologia, la matematica, l’ingegneria, l’economia e le scienze politiche, sull’effettiva possibilità di costruire un cervello artificiale», soprattutto dal momento che «i progressi della neurologia avevano mostrato che il cervello era costituito da una rete di neuroni che trasmettevano impulsi elettrochimici», ricostruibili in forme elettroniche per mezzo di “neuroni artificiali” (*ibidem*)¹⁹.

¹⁷ B. FRAGRASSO, *Intelligenza Artificiale e Responsabilità Penale: Principi e categorie alla prova di una tecnologia “imprevedibile”*, Torino, 2025.

¹⁸ AGENDA DIGITALE, G.F. ITALIANO, *Intelligenza artificiale, che errore lasciarla agli informatici*, disponibile su: <https://www.agendadigitale.eu/cultura-digitale/intelligenza-artificiale-che-errore-lasciarla-agli-informatici/>; articolo pubblicato il: 11 giugno 2019; ultimo accesso il: 23 gennaio 2026.

¹⁹ «Questa scoperta, combinata con i nuovi progressi scientifici delle teorie cibernetiche nel campo del controllo e della stabilità delle reti elettriche di Norbert Wiener, della teoria dell’informazione di Claude Shannon, e della teoria del calcolo di Alan Turing, condusse naturalmente a chiedersi se fosse possibile costruire un cervello elettronico. Qualche anno dopo, si cominciarono a costruire le prime macchine basate su neuroni artificiali, come ad esempio SNARC, costruita nel 1951 da un giovane Marvin Minsky insieme a Dean Edmonds. Più o meno negli stessi anni, e precisamente nell’Ottobre del 1950, Alan Turing pubblicò un lavoro

D’altro canto, l’«equazione Intelligenza Artificiale = macchine intelligenti» (A. Caselli, 2025) ha comportato nel tempo una serie di progressivi aggiustamenti nella programmazione e progettazione delle stesse: a partire dalla seconda metà del 1900 autori, infatti, come Philip K. Dick e I. Asimov – padre della fantascienza moderna – hanno cominciato a interrogarsi sulla possibile esistenza di «altri sé» di natura robotica o meccanica (*ibidem*); in particolare, le Tre Leggi della Robotica di Asimov²⁰ sono diventate a tutti gli effetti il primo sistema di valori “etici” creato nel corso della storia e pensato appositamente per essere pre-installato in ogni macchina auto-pensante al fine di orientarne il comportamento in una precisa direzione – in ordine, umanità, esseri umani e se stessi –, il primo vero e proprio tentativo di «individuare un possibile approccio operativo per affrontare il problema dei rischi e dei timori che comporta la creazione di un’intelligenza “esterna”, altra da quella umana, macchinistica» (*ibidem*).

Dalla fantascienza alla realtà materiale il passo fu breve: l’uso originario dell’espressione “intelligenza artificiale” è infatti di pochi anni a seguire. Ci si riferisce alla nozione impiegata da John McCarthy, assistente universitario di matematica al Dartmouth College di Hanover, già nel 1956, quando, insieme ad altri ricercatori – del calibro di C. Shannon, N. Rochester e M. Minsky – organizzò un convegno estivo sull’argomento e scelse appositamente «l’espressione intelligenza artificiale, in parte per distinguere il lavoro suoi e dei suoi colleghi dal più consolidato campo della cibernetica – “lo studio scientifico del controllo e della comunicazione nell’animale e nella macchina”» (J. Kaplan,

pionieristico, in cui si studiava la possibilità di costruire una macchina pensante» (G.F. Italiano, 2025).

²⁰ «1. A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law. 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.»; a queste si aggiunge una quarta legge, la “Legge Zero”, secondo cui «0. A robot may not injure humanity or, through inaction, allow humanity to come to harm.» (I. Asimov, 1942). Asimov credeva fermamente nella validità delle sue Leggi, al punto tale da auspicarne l’applicabilità anche da parte degli esseri umani; diceva infatti: « I have my answer ready whenever someone asks me if I think that my Three Laws of Robotics will actually be used to govern the behavior of robots, once they become versatile and flexible enough to be able to choose among different courses of behavior. My answer is, “Yes, the Three Laws are the only way in which rational human beings can deal with robots—or with anything else.” —But when I say that, I always remember (sadly) that human beings are not always rational» (*ibidem*).

2017)²¹. In effetti, il clima che si respirava in quegli anni – un’atmosfera di puro ottimismo e fiducia nel progresso e nella sua velocità – si tradusse presto in un’ondata di successi tra tutti i più importanti ricercatori del settore, i quali non perdevano occasione per lanciarsi sempre più a capofitto in una serie di «previsioni entusiastiche sulle “magnifiche sorti e progressive” dell’Intelligenza Artificiale (G.F. Italiano, 2025); ciò di cui non si tenne abbastanza conto erano però la limitatezza delle risorse e l’esorbitante spesa dell’innovazione, le quali, nonostante la velocità del progresso, ne bloccarono lo sviluppo fino ai primi anni Duemila.

Tuttavia, «proprio l’“inverno” degli studi sull’IA ha condotto i ricercatori a sviluppare approcci diversi, focalizzati sulle sue più spaiate applicazioni, passando da agenti artificiali impiegati in settori economico-finanziari alla creazione di robot di servizio o di carattere ludico o assistenziale, dallo sviluppo di droni e robot-soldati in ambito militare alle nanotecnologie, alla chirurgia mini-invasiva e alla telechirurgia mediche, dalla progettazione di robot-serpenti in grado di offrire un valido apporto nelle operazioni di recupero e sostegno a seguito del verificarsi di grandi calamità naturali a quella di robot-poliziotto, fino al mondo del c.d. *Technology Enhanced Learning*, anche per mezzo di macchine intelligenti dotate addirittura di «computazione emotiva» (M.B. Magro, 2023)²². Risale infatti a quegli stessi primissimi anni un’altra definizione di IA: quella fornita da A. Turing, secondo cui «si ha un’Intelligenza Artificiale quando un soggetto non è in grado di stabilire se la risposta a una sua domanda sia fornita da un’altra persona o da un computer» (A. Turing, 1950). Questa nozione di I.A si pone difatti quale punto di partenza di quel continuo e dinamico progetto di ricerca che punta alla piena estensione delle capacità

²¹ J. KAPLAN, *Intelligenza artificiale: guida al prossimo futuro* (titolo originale: *Artificial Intelligence. What Everyone Needs to Know*), LUISS University Press, Roma, 2018, seconda ristampa. «In effetti, il “Dartmouth Summer Research Project on Artificial Intelligence” è universalmente considerato come l’evento che diede origine all’Intelligenza Artificiale: fu esattamente l’evento in cui l’Intelligenza Artificiale venne accreditata come disciplina scientifica, con il riconoscimento della sua missione specifica, dei suoi primi successi e della sua comunità scientifica. Gli anni che seguirono il workshop di Dartmouth furono anni di grande sviluppo e di grandi successi per la nuova disciplina: grazie alle nuove metodologie e alle tecniche innovative sviluppate in questo settore, i computer dell’epoca furono in grado di risolvere problemi algebrici di notevole complessità, di dimostrare teoremi nel campo della geometria e persino di parlare correttamente inglese a un buon livello (...). Gli anni tra il 1956 e il 1974 furono effettivamente anni d’oro per l’Intelligenza Artificiale.» (G.F. Italiano 2025).

²² In A. CADOPPI, S. CANESTRARI A. MANNA, M. PAPA op.cit.

cognitive dei sistemi IA – di modo che possano andare finalmente oltre alle capacità logico-computazionali, fino a provare delle vere e proprie emozioni e riuscire così a emulare (e non solo imitare) l’intelligenza umana –, all’asservimento della superiore razionalità delle macchine all’essere umano – così da poterla sfruttare sempre più all’interno di una complementarietà che «non aspira a sostituire la macchina all’uomo, ma a costruire sistemi artificiali intelligenti a supporto dell’uomo, persino progettando ibridi uomo-macchina che siano più efficienti di sole macchine o solo umani» (M.B. Magro, 2023)²³ – e alla «combinazione di tecniche di “normatività indiretta”, cioè basata sulla definizione di modelli e metodi di apprendimento che conducano un’IA a sviluppare scelte parametrize sullo specifico contesto di azione» (A. Caselli, 2025).

Giunti a questo punto, ci si è chiesto che cosa fosse davvero l’Intelligenza Artificiale; questa, però, è «una domanda a cui è sia facile che difficile rispondere», almeno «per due ragioni: la prima è che non c’è comune accordo su cos’è l’intelligenza artificiale. La seconda è che ci sono poche ragioni, almeno al momento, per ritenere che l’intelligenza delle macchine abbia molto in comune con quella umana.» (J. Kaplan, 2018). Secondo i primissimi studiosi, «l’essenza dell’IA – in effetti, l’essenza dell’intelligenza – è la capacità di fare generalizzazioni appropriate in modo tempestivo e su basi dati limitata» (*ibidem*)²⁴: la definizione più accreditata, tra le tante, risulta pertanto essere

²³ Affinché questi sistemi possano effettivamente fungere da ausilio per la vita e le attività quotidiane, la scienza a un certo punto si troverà però davanti a una serie di blocchi, costituiti da interrogativi e paragoni fondamentali per lo sviluppo della ricerca; in particolare: «l’intelligenza artificiale sarà in grado di modificarsi per adeguarsi all’ambiente? Il nostro cervello lo fa: si tratta di un passaggio essenziale dell’evoluzione, che per l’essere umano è anche un’evoluzione culturale. Le macchine saranno in grado di interagire con il contesto che le circonda? Saranno capaci di rispondere a domande imprevedibili? Noi lo siamo. (...) è possibile che un giorno sarà così. Certo, occorrerà innanzitutto avere una mappa perfetta dei miliardi di interconnessioni e sinapsi che compongono il nostro cervello e, poi, capire come fare a riprodurle all’interno di una macchina. Ma la (...) domanda di fondo (...) è: che cosa ce ne faremmo di tutto questo?» (M. Cacciari, 2021).

²⁴ L’*Oxford Dictionary of Philosophy* ha da poco definito come *artificial intelligence* «the science of making machines that can do of thing that humans can do», mentre un’altra definizione, risalente ma fortunata, individua invece «cinque attributi che permettono di riconoscere l’esistenza di una forma di Intelligenza Artificiale: la capacità di comunicazione, la propria conoscenza, la conoscenza della realtà esterna, una condotta teleologicamente orientata, ossia tesa al perseguimento di un fine e, da ultimo, l’esistenza di un apprezzabile grado di creatività, intesa come capacità di assumere decisioni alternative laddove il piano di azione iniziale fallisca o non sia realizzabile» (M. Bassini, L. Liguori, O. Pollicino, 2025). Allo stato attuale, invece, la

quella fornita dall’AI-HLEG – secondo cui «*Artificial Intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal*» (A. Giannini, 2023) –, subito seguita dalla proposta del rapporto EC JRC Flagship in tema di IA, per cui «*AI is a generic term that refers to any machine or algorithm that is capable of observing its environment, learning, and based on the knowledge and experience gained, taking intelligent action or proposing decisions*» (*ibidem*).

La difficoltà di definire uno strumento in continua e rapida evoluzione è incrementata ulteriormente dal continuo rinnovamento dello stato della tecnica: attualmente, quasi tutti gli algoritmi delle moderne Intelligenze Artificiali accedono e archiviano enormi quantità di dati e di informazioni, assai superiore a quella che può conservare la mente umana, essendo dotati *ab origine* di una maggiore e più veloce capacità logico-computazionale (il c.d. *deep learning*) e di una più sofisticata capacità di *problem solving* (un *quid pluris* consistente nel sistema di *Auto Machine Learning*, abbreviato “A.M.L.S.”). Utilizzando algoritmi in grado di implementare in maniera automatica la propria banca dati e di codificare un numero sempre maggiore di informazioni, infatti, queste macchine di ultima generazione hanno imparato a ricorrere a tutte quelle innovative metodologie di apprendimento automatico che le rendono capaci 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.» (M.B. Magro, 2023), nonché a risolvere problemi di grande complessità con risultati ottimali pur sussistendo dati informativi non noti – il c.d. *cognitive computing* – e ad estrarre sempre nuove fonti di conoscenza dai dati acquisiti; in questo modo, sono divenute in grado di prendere decisioni del tutto autonome, sottratte al controllo umano, agendo «in modo autonomo e imprevedibile e non determinato, flessibile e influenzabile (...)» o seguendo uno scopo non sempre individuato e prefissato dall’uomo (*ibidem*).

dottrina ha espresso unanime consenso nel definire gli Agenti Artificiali (A.A.) quali dispositivi informatici «*sufficiently informed, “smart”, autonomous and able to perform morally relevant actions independently of the humans who created them*» (Science and Engineering Ethics, 2020).

Da “nemica” ad “amica”: il ruolo dell’IA nella lotta alla criminalità digitale

Da questo punto di partenza, la dottrina ha poi ricondotto le caratteristiche dei sistemi di Intelligenza Artificiale all’interno di categorie predefinite.

Alcuni autori individuano i requisiti essenziali di un qualsiasi sistema informatico – e, di contro, di una qualsiasi macchina IA – in primo luogo nella dotazione di capacità cognitive – cioè nella sua capacità di «conoscere ed apprendere, oltre che dall’enorme banca dati rappresentata dal *web*, anche dall’esperienza del mondo “esterno”, ricercando ed acquisendo direttamente informazioni e dati, grazie a sofisticati sensori visivi, acustici, termici ecc., come accade nei c.d. sistemi esperti, che con tecniche di *machine learning* individuano, riconoscono, selezionano e memorizzano le informazioni, le immagini i suoi, utili ai fini della successiva elaborazione automatizzata e sempre più autonoma» (L. Picotti, 2023) – e, in secondo luogo, nella sua capacità di autodeterminarsi – consistente nell’«elaborazione o, se si vuole, “conoscenza” e valutazione (a sua volta automatizzata) dei dati e delle informazioni utili da considerare, esprimendo ed attuando immediatamente “decisioni” e scelte operative, fra possibili opzioni alternative» (*ibidem*) –; a queste, infine, si affianca anche quel margine ineliminabile di «imprevedibilità» (*unpredictability*), «derivante dalla loro autonomia e rapidità operativa e gestionale » (*ibidem*).

Altri autori invece – quali Abott e Sarch – si concentrano su quegli aspetti che potrebbero causare problemi di “irriducibilità” nei confronti dello stesso diritto penale – la c.d. *irreducibility challenge*, consistente nell’impossibilità di ridurre un reato commesso da un sistema di IA all’agire di un singolo essere umano –, ovvero: l’autonomia, «la capacità di causare un evento dannoso senza che il sistema sia diretto in tal senso da un agente umano» (A. Giannini, 2021); l’opacità, ossia «l’impossibilità (...) di poter ottenere una spiegazione su come il sistema, partendo dall’*input* (A), ha ottenuto l’*output* dannoso (B)» (*ibidem*); la complessità, cioè «il fatto che la creazione di un sistema di IA sia spesso il risultato del contributo di numerosi individui, sviluppatosi in un lungo periodo di tempo, nonché il fatto che il sistema possa essere addestrato su banche dati eterogenee e *open source*» (*ibidem*); e l’imprevedibilità, ossia «la capacità per il sistema di intraprendere attività non previste dalla sua programmazione originale» (*ibidem*).

4. Le tipologie di Intelligenza Artificiale: IA «forte» vs IA «debole»

Sulla base di questa costante e rapidissima evoluzione, storicamente si è giunti a una vera e propria divisione dottrinale – all’interno di un vasto panorama di analisi composto tanto da giuristi quanto da tecnici del mondo IOT e ICT, psicologi e filosofi della mente – tra due diverse concezioni dell’Intelligenza Artificiale e dei suoi possibili sviluppi futuri: la teoria funzionalista, a favore di una nozione «forte» di IA, e la teoria strutturalista, piuttosto orientata verso la nozione «debole» della stessa (A. Giannini, 2021).

Analizzando brevemente entrambi i punti di vista, l’approccio funzionalista valorizza le «funzioni svolte dalla mente umana e dalle macchine, in grado di svolgere funzioni simbolico computazionali» (M.B. Magro, 2023): di anno in anno, infatti, informatici, filosofi, scienziati sociali, neuroscienziati, imprenditori e giuristi si sono sempre più convinti che i progressi nella programmazione e nella velocità di elaborazione dei dati e di memorizzazione degli stessi avrebbero portato, a un certo punto, a un certo momento in cui i calcolatori elettronici – qualsiasi tipologia di macchina in grado di elaborare dei simboli, cioè informazioni codificate, in maniera automatica, cioè sulla base di regole e procedure formali di calcolo, e quindi dotata di razionalità – sarebbero diventati intelligenti come gli esseri umani e hanno così deciso di adottare una formulazione generale o «forte» della stessa Intelligenza Artificiale. Questi autori – tra cui compaiono nomi quali A. Turing²⁵ e G. Hallevy, sostenitore di una tesi espansionistica, e G. Sartor, con una dottrina più moderata e costruita intorno ad un modello B.D.I., *Belief Desider Intention* – erano fortemente convinti che prima o poi l’umanità sarebbe stato in grado di ricreare agenti dotati di «tutte le funzioni cognitive di un essere umano e quindi, di fatto (...)

²⁵ M.B. MAGRO, 2023 in A. CADOPPI, S. CANESTRARI A. MANNA, M. PAPA, op.cit. Per stabilire se una macchina fosse intelligente o meno, A. Turing aveva predisposto un apposito test, il “test di Turing” o “*imitation game*” (menzionato anche nella prima sezione di questo capitolo, è riportato per la prima volta in un testo scientifico in A. Turing, 1950). L’obiettivo di questo test era accertarsi che il *computer* fosse un sistema computazionale in grado di interagire con l’ambiente circostante (requisito della reattività); che fosse in grado di prendere decisioni e di conseguenza di agire, in modo autonomo, per raggiungere un obiettivo predefinito o negoziato (requisito della proattività); e che fosse in grado di comunicare, coordinarsi, cooperare, negoziare con altri agenti e/o esseri umani (requisito della capacità di interazione sociale); cioè, che fosse un’intelligenza di tipo comportamentale e interattivo, cioè un “agente intelligente” (M.B. Magro, 2023).

indistinguibili dalla sua mente»: potendo assimilare il cervello all’*hardware* e la mente al *software*, ciò che contava maggiormente non era più la struttura biologica ma la mera “funzione” cognitiva, esistendo d’altronde già allora intelligenze in grado di agire indipendentemente dalla sede fisica in cui risiedevano, con la diretta conseguenza che «un *computer* è paragonabile ad un essere umano quanto ad intelligenza» in tutti i casi in cui «le prestazioni svolte dal *computer* non possono essere distinte da quelle svolte dagli esseri umani» (A. Turing, 1950). Su questa stessa scia, I.J. Good, nel 1960, auspicava che la tecnologia giungesse a creare un sistema in grado di evolvere e migliorarsi autonomamente, accrescendo le proprie capacità cognitive fino a diventare una “superintelligenza”, con capacità, velocità di esecuzione e obiettivi completamente al di fuori del livello umano, di modo tale che «la prima macchina superintelligente» potesse divenire «l’ultima invenzione che l’uomo avrebbe bisogno di fare, purché (...) si riveli abbastanza docile da indicarci come tenerla sotto controllo» (I.J. Good, 1960).

D’altro canto, invece, la nozione di IA circoscritta o «debole» sposa una tesi prettamente strutturalista, che «si focalizza su compiti abbastanza ristretti e specifici» e «sostiene la non autenticità del pensiero meccanico e della diversità ontologica tra intelligenza artificiale e intelligenza naturale» (M.B. Magro, 2023). Secondo questa concezione – sostenuta da autori come Searle, Salvadori, Gless e Weigend, i quali formano la c.d. scuola della dottrina “Scettica” tedesca e italiana –, infatti, l’IA non sarebbe in grado di comprendere veramente il reale significato dei comportamenti che pone in essere, limitandosi invece a simulare e a riprodurre soltanto quei processi intellettuali umani così come li ha osservati e appresi nel corso della sua programmazione; più approfonditamente, gli Scettici «affermano una nozione di intelligenza come *intenzionalità*, che a sua volta è la dimensione essenziale della coscienza, caratteristica che contraddistingue certi stati mentali, e che differenzia l’uomo dalle macchine» (M.B. Magro, 2023), per cui non è sufficiente che un programma esegua un’attività su un dato *input* per poter dire che esso sia dotato di reale intenzionalità²⁶; anche in quanto qualsiasi tentativo di imporre una responsabilità penale ai sistemi di IA è in ogni caso destinato a naufragare

²⁶ Questo aspetto è provato dal c.d. “test della stanza cinese”, in cui si osserva come il sistema simbolico-computazionale tenda a comportarsi “come se” capisse il cinese, ma non riproduca mai l’attività umana non conoscendo veramente i significati nascosti dietro alle diverse interazioni (*ibidem*).

contro gli scogli dell’articolo 27 della Costituzione, il principio della responsabilità penale personale e colpevole, e dell’articolo 85 c.p., la capacità di intendere e volere (A. Giannini, 2023)²⁷.

Traendo alcune conclusioni, il fulcro dell’intera dissertazione pare ruotare intorno a un unico perno: che cosa si intenda per intelligenza²⁸. Sebbene le discipline psicologiche concordino nel ritenere che ciò che viene comunemente inteso come “intelligenza” non possa essere ridotto a una definizione rigorosa e stringata, esse però sottolineano anche come si possa comunque riuscire a tracciare un confine per delinearne i tratti salienti, e, tra questi, un ruolo fondamentale è quello giocato dalle emozioni: «strettamente intrecciate alla razionalità computante, ma anche alle funzioni fisiologiche, alla memoria, all’esperienza, sono profondamente innestate nel corpo, inteso sia come insieme di organi sia come depositario della nostra identità e della nostra storia (...) sono tanto pervasive che ogni nostro atto si colora di esse e ogni nostra relazione con noi stessi e con l’ “altro” ne è condizionata», al punto che «sia impossibile riprodurre l’intelligenza umana (...) senza che il suo possessore abbia un corpo, (...) senza ridurre quest’ultimo ad un profilo comportamentale monodimensionale e rigido.» (M.B. Magro, 2023); il che significa che ai robot dotati di intelligenza manca, in ogni caso e per necessità di cose, quell’emotività (o, più correttamente, “intelligenza emotiva”) descrivente la «capacità di governare le pulsioni interne con relazioni esterne, che orienta verso il “senso comune”». Queste macchine, infatti, sono incapaci di collegare conoscenze

²⁷ In A. CADOPPI, S. CANESTRARI A. MANNA, M. PAPA, op.cit.

²⁸ «È possibile chiarire tale idea evocando i due significati principali del termine. Intelligenza può significare: (1) *intus legere*, cioè comprendere in profondità; ovvero: (2) *inter legere*, cioè scegliere tra. Il primo significato allude alla capacità della mente di comprendere e di rappresentare la realtà. Il secondo allude alla capacità della mente di scegliere, in modo razionale, tra ipotesi alternative. Questi due significati connotano anche le linee di fondo della ricerca che si svolge nell’ambito dell’intelligenza artificiale e privilegia due diversi e complementari paradigmi: un paradigma epistemologico, che pone al centro delle proprie ricerche i temi dei meccanismi della mente, della rappresentazione della realtà e della strutturazione della conoscenza, e un paradigma euristico, che pone al centro della propria attenzione i temi della produzione di scelte razionali, della soluzione di problemi ed acquisizione di esperienza. Alla luce di tali considerazioni e dopo aver sottolineato la dipendenza concettuale della nozione di intelligenza artificiale da quella di intelligenza, è opportuno precisare il significato dell’espressione intelligenza artificiale. Con tale espressione si intende dunque fare riferimento all’intelligenza intesa a sviluppare modelli computazionali del comportamento intelligente, e quindi a far sì che gli elaboratori possano eseguire compiti che richiederebbero intelligenza da parte dell’uomo.» (U. Pagallo, 2013; il quale a sua volta si rifà alle parole di McCarthy-Hayes e di Sartor).

specialistiche di campi diversi e di affrontare i problemi e risolverli «senza la rigidità tipica dell’approccio simbolico dell’intelligenza intesa come computazionale» (*ibidem*); motivo per cui, in fin dei conti, l’approccio più corretto pare ancora essere quello adottato dalla tesi strutturalista e, dunque, dalla scuola Scettica.

5. L’Intelligenza Artificiale “nemica”: l’IA quale strumento e/o autore del reato

«L’idea che gli agenti intelligenti possano provocare danni ingenti senza che esista un adeguato sistema di tutela provoca forte disagio. Il disagio diviene inquietudine quanto più percepiamo le intelligenze artificiali come agenti, da un lato indipendenti, ma dall’altro complementari, se non necessari, per le nostre vite.» (M.B. Magro, 2020)²⁹; e proprio da questa inquietudine crescente ha origine il problema di stabilire «se la tecnologia moderna ha creato una nuova tipologia di soggetto, l’agente non-umano, per il quale è possibile declinare il linguaggio giuridico dei diritti e delle responsabilità» (*ibidem*). In altre parole, si tratta della sempre più sentita esigenza di definire lo “statuto ontologico delle IA” o, analogamente, di tutti quegli strumenti particolarmente evoluti; un dibattito che trova piede quasi istantaneamente anche in dottrina, in quanto da un lato – secondo la tesi strutturalista – «i *robot* intelligenti sono comunque oggetti strumentali e dunque, sul piano giuridico, meritevoli di tutela solo per il valore che esprimono», mentre dall’altro – secondo la tesi funzionalista – «potrebbero essere considerati come “soggetti” e dunque meritevoli, sul piano filosofico, morale e giuridico, di una eventuale tutela per soggettività, sia pure non umana.» (M.B. Magro, 2023).

In via preliminare, allo stato attuale le regole applicabili ai *robot* sono principalmente quelle concernenti la responsabilità per i danni causati da prodotto difettoso (rientrante in via generale nell’articolo 2050 c.c. e, appositamente, negli artt. 114 e seguenti del Codice del Consumo), una forma di responsabilità civilistica “speciale” attribuita al produttore a prescindere da una sua colpa e in presenza esclusivamente di determinate circostanze a riprova dell’effettiva imputabilità degli effetti dannosi in capo a questi; l’attenzione è stata infatti rivolta – e ciò sia nella dimensione giuridica dell’UE sia oltreoceano

²⁹ M.B. MAGRO, *Decisione umana e decisione robotica: un’ipotesi di responsabilità da procreazione robotica*, in *La legislazione penale – Giustizia penale e nuove tecnologie*, 1/2020, pp. 1–22.

– soprattutto all’accertamento degli effettivi danni causati direttamente dal malfunzionamento o dal difetto di un prodotto fabbricato e venduto, piuttosto che su quelli derivanti da un suo “corretto” funzionamento. Così, in un primo momento, si è ricostruita anche la responsabilità per i danni causati dall’utilizzo delle IA in assimilazione a quelle categorie summenzionate applicabili a determinate tipologie di *res*, strumenti di controllo o comunque di dominio del rischio da parte di diversi soggetti a titolo di responsabilità oggettiva, che il nostro ordinamento già riconosceva quale «plesso di regole relative alla *product liability* che distribuiscono la responsabilità tra chi produce un bene e chi ne sperimenta un utilizzo improprio» (M. Bassini, L. Liguori, O. Pollicino, 2025)³⁰.

Il problema della responsabilità del programmatore o dell’utilizzatore di un sistema di Intelligenza Artificiale si pone, però, anche sotto il profilo penalistico: il rapporto tra diritto penale e IA, infatti, si configura in parte in termini di conflitto e in parte in termini di collaborazione, laddove «troviamo l’area di conflitto, riconducibile al diritto penale sostanziale, dove i costrutti classici faticano ad indirizzarci verso un soggetto (umano) responsabile» (A. Giannini, 2021). Ciò si può ben vedere nel fatto che, se anche l’ipotesi dolosa pare essere di facile soluzione, in quanto «l’azione “volontaria” del (...) *robot* si identifica e rappresenta una lunga *manus* della persona umana» – anche perché in ogni caso non si può mai escludere a prescindere che «l’agente intelligente, una volta sviluppato, sia dolosamente utilizzato all’insaputa degli utenti, con finalità di raccogliergli un profilo emotivo e di sfruttare a fini commerciali le risposte emotive indotte in base alle informazioni raccolte» (M.B. Magro, 2023) – non si può dire lo stesso della responsabilità a titolo di colpa dell’utilizzatore,

³⁰ O. POLLICINO, G. MUTO, *La legislazione delegata in materia di intelligenza artificiale: la costruzione di una disciplina organica al confine tra scelte governative, controllo parlamentare e vincoli europei*, in MediaLaws – Rivista di diritto dei media, 2/2025, disponibile online su: <https://www.medialaws.eu/rivista/la-legislazione-delegata-in-materia-di-intelligenza-artificiale-la-costruzione-di-una-disciplina-organica-al-confine-tra-scelte-governative-controllo-parlamentare-e-vincoli-europei>. Nello specifico, oltre all’articolo del codice civile e alle leggi speciali, la normativa europea applicabile in tal senso è riconducibile essenzialmente a tre direttive: la direttiva “Macchine” (Dir. 2006/42/CE), che «consente di considerare i robot nell’alveo dell’ampia categoria dei beni di consumo», da cui la «piena compatibilità del settore di produzione afferente alla robotica con le norme su cui si fonda la disciplina generale in tema di sicurezza e commercializzazione dei prodotti» (*ibidem*); la Direttiva 99/44/CE, «in tema di garanzia commerciale e di rimedi in caso di difetto di conformità (...) considerando il robot quale mero oggetto di scambio» (*ibidem*); e la Risoluzione del Parlamento europeo del 16 febbraio 2017, in tema di diritto civile sulla robotica, che riguarda «la necessaria regolamentazione della responsabilità civile per i possibili danni causati da un robot» (*ibidem*).

del venditore o del programmatore qualsivoglia imputato di aver causato un danno nell’utilizzo dell’agente intelligente, la quale sorge di *default* per il solo fatto di non aver saputo prevedere un agire che rimane, per le caratteristiche intrinseche della macchina, impreveduto, imprevedibile, autonomo e sottratto al controllo umano. In questa circostanza, infatti, «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», dal momento che spesso le regole decisionali sono «solo abbozzate dal programmatore e comunque elaborate solo dopo una rilevazione statistica da parte del robot.» (M.B. Magro, 2023).

Seguendo questa impostazione, si arriverebbe però all’estremo opposto, ovvero ad ammettere l’ingresso nell’ordinamento attuale di un approccio precauzionale, il quale a sua volta inibirebbe l’utilizzo delle intelligenze artificiali in molti settori della vita quotidiana, soprattutto nel momento in cui il progressivo depauperamento del contenuto della colpa soggettiva «a favore di una eccessiva normativizzazione in termini di riduzione a pura inottemperanza cautelare» sfoci in «modelli di allocazione della responsabilità di tipo oggettivo-assicurativo-indennitario» (M.B. Magro, 2023). Pertanto, per evitare che ciò accada, «sarebbe quantomeno necessario stabilire un livello di cautele differenziato a seconda dell’uso e del tipo di tecnologia intelligente immessa nel mercato: tanto maggiori sono i benefici, tanto sono seri i pericoli, maggiori saranno le misure cautelari richieste», fino ad arrivare al punto in cui gli agenti intelligenti avranno raggiunto «un livello di evoluzione da essere quasi perfetti», mentre gli eventi dannosi derivanti dal loro agire saranno «rarissimi, eccezionali, equiparabili ad un caso fortuito o forza maggiore (art. 45-46 c.p.)» (M.B. Magro, 2023); eppure, a ben vedere, anche in questo scenario non si potrebbe escludere la responsabilità penale dell’operatore “inattivo”, se non altro «per avere immesso nell’ambiente una tecnologia innovativa che aumenta il rischio per la salute umana violando il dovere di continuo e costante monitoraggio del prodotto e dei suoi possibili effetti dannosi, anche se non

identificati durante la programmazione e produzione, ma evidenziati una volta che il prodotto viene immesso in un ambiente aperto, con l’uso» (*ibidem*)³¹.

In fin dei conti, dunque, il punto focale della questione verte sulla possibilità di attribuzione o meno una forma di responsabilità personale all’umano, piuttosto che al sistema intelligente; ovvero: «a) in caso di condotta omissiva, sull’esistenza di un dovere giuridico e di un potere fattuale di impedimento dell’evento, e dunque sulla possibilità che l’umano detenga una posizione di garanzia; b) sull’esistenza di una regola cautelare, positivizzata o basata sull’esperienza, e dunque sulla configurabilità di un agente modello; c) sulla conoscibilità della regola e la prevedibilità dell’evento; d) sull’esigibilità del comportamento conforme da parte dell’umano.» (A. Giannini, 2021). Un tale rimprovero in termini di colpa, però, si scontra *ab origine* sia con la scarsa conoscenza scientifica di tali tecnologie sia con la quasi assoluta mancanza di usi e convenzioni in materia, anche dovute al fatto che «la macchina sulla quale egli dovrebbe esercitare il controllo è stata – in realtà – costruita per prevalere su di lui, e non viceversa» (*ibidem*); motivo per cui le ricerche si stanno concentrando sempre di più sull’aspetto dell’*accountability*³² – piuttosto che sui due concetti, autonomi e separati, di *responsibility* e *liability* –, più legata anche allo sviluppo di un’IA ispirata a «un’etica allineata a quelli degli umani e che quindi non produca effetti non desiderati» (M.B. Magro, 2023).

Rispetto alla responsabilità penale dell’uomo «dietro la macchina», «si sarebbe portati a concludere che i danni siano in ogni caso da ricondurre al produttore/programmatore, anche quando il meccanismo eziologico non appaia riconducibile alle azioni/decisioni *ab inizio* programmate» (C. Piergallini, 2025); tuttavia, ciò si scontra anche con non pochi problemi dal punto di vista tanto della causalità – come il «fenomeno della *black box* causale» o della «causalità sorpassante», oltre a non potersi applicare in alcun caso il criterio della causalità

³¹ Fattispecie di cui all’art. 437-bis c.p., quale omessa adozione di misure di sicurezza per immissione di strumentazioni digitali nel mercato, introdotta appositamente dal d.lgs. n. 160 del 9 settembre 2026.

³² Tale nozione si può al meglio ricondurre alla “rendicontabilità”, cioè l’essere chiamati a dar conto di certi atti e il doverne subire le conseguenze con riferimento a un determinato obbligo presente in un determinato sistema sociale (A. Giannini, 2021); «in altre parole, secondo questa accezione, *accountability* “is the process aimed at a subsequent public assessment of a person’s conduct in a given case in order to evaluate whether this conduct was require and/or justified by this person’s responsibility and, once this evalutaion is executed, to establish who can be held liable for all consequences of the conduct (...)”» (*ibidem*; la quale riprendeva I. Giesen, F.G.H. Kristen, 2014).

adeguata, dato che l’azione iniziale non è di regola adeguata a produrre l’evento lesivo – quanto della colpa, con il rischio di trovarsi di fronte a una catena di fattori causali alternativi in grado di rilasciare effetti di tipo interattivo o sinergico non riconducibili alle classiche leggi della causalità, ma soprattutto di non poter improntare il rimprovero nei termini della tipicità soggettiva, stante l’imprevedibilità degli eventi dovuta all’auto-apprendimento del sistema pensante.

Rispetto invece a un’ipotetica forma di responsabilità penale delle IA, questa idea si è sviluppata soprattutto con riferimento alle IA «forti» e pone l’attenzione non tanto sul fatto che le IA siano “libere” o meno di scegliere come comportarsi sulla base di una reale comprensione del senso ultimo delle proprie azioni, quanto piuttosto sulla possibilità di considerarle responsabili dal punto di vista, *in primis*, penalistico; d’altronde, seppur a patto di accettare una concezione normativa estremista della responsabilità in sé intesa, ciò sarebbe perfettamente plausibile, dal momento che la libertà di azione non è un fenomeno naturale di cui occorre individuare i referenti ontologici quanto piuttosto un mero attributo normativo dell’agire del responsabile autore del reato, funzionale agli scopi dell’organizzazione sociale di riferimento. Così, seguendo la scia di A. Turing e H. Putnam, G. Hallevy teorizzò proprio in tal senso il suo modello di responsabilità, assumendo cioè il riconoscimento di personalità giuridica in capo alle macchine dotate di IA, sulla base dei tre distinti modelli della *perpetration through another* – secondo cui «le entità che sfruttano Intelligenza Artificiale sono assimilabili a quegli esseri che presentano una degradazione o una menomazione mentale, come per esempio i bambini o gli incapaci naturali, e che in quanto tali difettano di un elemento soggettivo compatibile con l’attribuzione di responsabilità penale», ma che comunque «possono tuttavia costituire il *medium* tramite il quale una condotta illecita si perfeziona» (M. Bassini, L. Liguori, O. Pollicino, 2025) –, della *natural probable consequence* – applicabile nel caso in cui sia utilizzata una forma di IA senza fare ricorso alle sue capacità avanzate e nel caso di versioni ormai risalenti, basato «sull’esistenza di una stretta interrelazione tra le attività eseguite su base quotidiana dall’Intelligenza Artificiale e il ruolo del programmatore e/o utente, i quali però (...) non serbano alcuna specifica intenzione di commettere, servendosi dell’Intelligenza Artificiale, un reato» (*ibidem*) – e della *direct liability*, che scinde l’IA dal programmatore e dall’utente. Seppur l’obiettivo sotteso da

Hallevy e dagli altri strutturalisti pare proprio «quello di evitare il prodursi di zone franche rispetto al verificarsi di accadimenti che, quantomeno sotto un profilo esteriore (e dunque, avendo riguardo all’elemento oggettivo), integrano fattispecie di reato» (M. Bassini, L. Liguori, O. Pollicino, 2025), a onor del vero bisognerebbe sottolineare come al massimo, «la previsione di una diretta responsabilità a carico di un *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.»; e questo perché «il sistema di IA non agisce ma “è agito” (...) il decision making non è frutto di una scelta ma è la obbligata risultante dei meccanismi oscuri che alimentano l’algoritmo, a sua volta forgiato dall’uomo. (...) Difetta, in definitiva, il requisito, primordiale, della *suitas*.» (C. Piergallini, 2025).

A tal proposito, infatti, poiché tradizionalmente «il diritto penale, (...) pare originariamente disegnato per gli uomini» – dal momento che «solo gli uomini possono davvero delinquere, scegliendo un male per la società o il prossimo, pur di conseguire un proprio tornaconto personale (...), possono ragionevolmente essere puniti in segno di biasimo per la loro scelta colpevole (...), guardando al lato passivo di chi subisce il reato, detengono direttamente o indirettamente interessi giuridici così importanti da poter essere presidiati da una minaccia di sanzione criminale in caso di loro provocato pregiudizio» (A. Cappellini, 2019)³³ – l’opinione maggioritaria in dottrina si è schierata a difesa dello Scetticismo, affermando che macchine, *robot* e in generale qualsiasi altro soggetto artificiale, intelligente o no, non può essere direttamente responsabile per la commissione di un reato; al massimo, si sono concessi di accettare che il tema potrebbe presentare forti analogie con quello della responsabilità da reato degli enti collettivi, dal momento che renderebbe responsabile un’agente artificiale soggettivato per il suo operare e per quello del suo utilizzatore, programmatore, designer o produttore in quanto soggetto che si serve di esso.

D’altro canto, anche in questo senso, la visione di Hallevy e dei suoi predecessori «pare viziata da (...) due semplificazioni: anzitutto, l’idea che, se la responsabilità penale può essere affermata andando oltre il fronte delle persone fisiche, con riguardo cioè a imprese ed enti, la sua applicazione possa investire

³³ A. CAPPELLINI, *Machina delinquere non potest? Brevi appunti su intelligenza artificiale e responsabilità penale*, in *DisCrimen. Criminalia – Annuario di scienze criminalistiche*, 2019

anche entità diverse – il che rischia di tralasciare il fondamentale dettaglio per cui la responsabilità d’impresa è una responsabilità della *societas* e ripete comunque il suo fondamento da comportamenti di individui, pur avvinti da un legame sociale –; in secondo luogo, al cuore della teorizzazione di Hallevy si ritrova un’assimilazione tra caratteristiche umani e dei sistemi di Intelligenza Artificiale che prova apparentemente troppo e che non è suffragata da evidenze scientifiche ma solo da *fictio* giuridiche fondate sul funzionamento di questi diversi “esseri”» (M. Bassini, L. Liguori, O. Pollicino, 2025). In altre parole: poiché «gli agenti non sono costituiti da agenti umani (...) 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, (...) la punizione penale del robot intelligente (ovvero la sua distruzione o riprogrammazione) sarebbe inoltre inutile sotto il profilo dell’efficacia deterrente anche se rivolta all’ente artificiale in sé considerato (...) 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)» (M.B. Magro, 2025)³⁴.

³⁴ Così anche in A. Cappellini, 2019: «La persona giuridica, cioè, a differenza del robot, non ha realtà naturalistica indipendente dai suoi membri: e valorizzando questo dato, ci sembra sia forse possibile superare l’opposta prospettiva che accomunava i due tipi di responsabilità. (...) Il comando sotteso alla norma, che dunque solo apparentemente si appunta all’ente, è invece inevitabilmente rivolto ai soli soggetti che davvero esistono nel mondo fisico: cioè gli uomini, suoi rappresentanti. In questa logica, le persone giuridiche non sono che degli schermi, delle maschere, che vengono colpite dalla minaccia di sanzione (ovviamente economica) al fine di influenzare quei burattinai umani che la manovrano (...). Le intelligenze artificiali, invece, non sono costrutti di rilievo soltanto sociale, ma esistono davvero nel mondo fisico. L’uomo si limita a crearle e programmarle. Dopo di ciò, con il distacco dal proprio artefice, il loro centro decisionale si autonomizza. Così, esse iniziano ad agire nella realtà fisica “di propria mano”, autodeterminandosi, senza fili che le muovano, in tempo reale, come marionette: libere della libertà che è stata loro elargita dal proprio creatore (...). Una sanzione rivolta alle IA, così, si ripercuoterà essenzialmente su nessun altro all’infuori di loro medesime. Non influenzerà il comportamento di nessun umano dotato di libertà di scelta. E se anche ciò fosse possibile, in ogni caso, resterebbe sempre il fatto che l’uomo, pur minacciato di riflesso dalla pena della macchina, non può influenzare direttamente il comportamento del soggetto artificiale, spingendolo in tempo reale a non commettere delitti.»

Laddove poi si voglia continuare a utilizzare la costruzione “classica” di un reato – costruita sui due pilastri dell’elemento oggettivo e soggettivo risultanti dal combinato disposto tra gli articoli 43 e 85 del codice penale – bisogna poi confrontarsi anche con fenomeni quali: «l’*automation bias* e l’*automation complacency*», cioè l’eccessiva fiducia nella macchina con conseguente riduzione dell’attenzione del suo supervisore, rispetto al requisito della *mens rea*; il «*problem of many hands*», cioè la gestione di un unico sistema informatico da parte di molteplici agenti umani, anche in modalità spazialmente e temporalmente dislocate e distanti l’una dall’altra; il «*problem of the black box*», cioè l’opacità dei sistemi di *machine learning*, che rende generalmente impossibile o comunque estremamente complessa la ricostruzione causale di un evento dannoso; e il «*problem of the shortcuts*», per cui i sistemi di IA più sviluppati tendono a basarsi su correlazione errate nel *data set* per risparmiare energia, rispetto invece a quello dell’*actus reus*. Dacché, più che sulla fattibilità di un meccanismo d’imputazione di responsabilità penale diretta in capo ai sistemi di IA – che, sulla base di quanto analizzato finora, è reso astrattamente possibile già sulla base della schematizzazione di Hallevy, ma non è del tutto attualizzabile per la configurazione attuale dell’ordinamento giuridico, soprattutto penalistico –, è fondamentale interrogarsi sulla sua effettiva necessità: si può anche essere disposti a rinunciare alle nozioni giuridiche tradizionali e antropomorfe, in vista di nuove categorizzazioni “neutrali” applicabili *in toto* anche ad agenti “moralì” artificiali; tuttavia, al fine di poter ritenere veramente penalmente responsabili i sistemi di IA, si dovrebbe anche abbandonare del tutto il concetto di razionalità intesa così come capacità di rispondere alla legge (A. Giannini, 2023); ed è questa la ragione per cui, a conti fatti, molti autori ribadiscono la non necessità di una tale antropomorfizzazione dei sistemi di IA, perlomeno fino al momento in cui sarà ancora possibile individuare dei soggetti umani dietro all’azione artificiale e, dunque, non si sarà ancora giunti a ritenere le macchine quali soggetti autonomi, responsabili e auto-pensanti, parte “attiva” della comunità (*ibidem*)³⁵.

³⁵ «The succinct ansie to the first prong of the RQ is the following: criminal liability of AI agents is feasible to the extent that we are willing to give up our anthoropomorphic legal notions (...) My answer is that criminal liability of AI agents is not needed as long as AI systems are not treated as members of our community and, as such, as subject to the same rules as us.» (A. Giannini, 2023)

Al netto di quanto esposto e continuando a difendere la posizione di una non necessità e non utilità di una forma di responsabilità autonoma per l’IA, in conclusione pare opportuno ritornare sulla principale obiezione mossa, mettendo in risalto proprio la diversità “ontologica” tra decisione umana e decisione robotica; e ciò sull’onda di quei recenti studi di psicologia cognitiva, che hanno evidenziato come gli aspetti emozionali canalizzati nel processo decisionale «non sono da considerare un’interferenza fuorviante, originata da una “razionalità limitata”». Al contrario, infatti, «essi sono quel potenziale neurologico in grado di offrire un *trade off* che colma il *gap* tra limiti cognitivi della mente umana e la complessità della situazione 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 (...) tende a vincere o imporsi quella inconscia emotiva (...). 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» (M.B. Magro, 2020); il che, in altre parole, significa che «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.» (*Ibidem*). Dunque, «se escludiamo l’intelligenza, la coscienza, i sentimenti, il libero arbitrio, cosa rimane ancora di irriducibilmente umano? Qual è lo *specificum* dell’uomo?» (F. Basile, 2025)³⁶; «come stabilire il punto esatto in cui una IA cessa di essere *res* e diviene persona?» (A. Cappellini, 2019, che a sua volta richiama P. Asaro, 2011). I robot, infatti, «non sono fragili, mortali, moralmente bisognosi» e «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 equiparate alle straordinarie potenzialità della mente umana, che funziona sulla base di circuiti neurologici stabilizzati, ma non ne è affatto schiava» (M.B. Magro, 2020).

³⁶ F. PIZZETTI (a cura di), *Intelligenza artificiale, protezione dei dati personali e regolazione*, Torino, 2025, I edizione.

Il tutto, per arrivare ad affermare con un grado di certezza sufficientemente elevato e fondato che la scelta dell’ordinamento, comunitario o nazionale, di riconoscere una forma di responsabilità propria e autonoma ai sistemi di IA, trasformandoli così da soggetti “agiti” a soggetti “agenti”, potrebbe valere solo quale scelta di opportunità, senza poter davvero raggiungere, almeno, allo stato attuale della tecnica, quella consapevolezza e auto-coscienza profondamente legate al senso comune e alla sensibilità dell’agire di ogni essere umano, alla base delle radici più antiche del rimprovero di colpevolezza penale. Il che è esattamente ciò che mira a fare il Parlamento Europeo, nel momento in cui «contempla altresì la possibilità che nel lungo termine possa esser istituito uno *status* giuridico specifico per i robot autonomi più sofisticati» – parlando cioè di «persone elettroniche dotate di una capacità giuridica e, quindi, potenzialmente responsabili e in grado di essere chiamati a risarcire qualsiasi danno da loro causato, nonché eventualmente il riconoscimento della personalità elettronica dei robot che prendono decisione autonome o che interagiscono in modo indipendente con terzi.» (M. Bassini. L. Liguori, O. Pollicino, 2025) – o si appropria alla regolazione dell’intelligenza artificiale anche tramite veri e propri atti normativi – tra cui l’*AI Act* e il progetto *Accountability Principles for Artificial Intelligence (AP4AI)*, che riunisce Europol, Eurojust, centro di ricerca CENTRIC e centro CEPOL nella «creazione di un quadro di principi per (...) un utilizzo etico, trasparente e responsabile di sistemi di IA da parte delle forze dell’ordine» (A. Giannini, 2021) – .

6. L’Intelligenza Artificiale “amica”: il suo utilizzo nella lotta alla criminalità digitale

Se, come visto, da un lato il progresso tecnologico e il massiccio impiego dell’Intelligenza Artificiale hanno fornito strategie e tattiche più veloci, precise ed efficaci agli stessi *cyber*-criminali, rendendone più sofisticate le operazioni di attacco e riducendone i tempi e le possibilità di reazione, dall’altra, però, l’IA sta progressivamente assumendo anche un ruolo fondamentale quale alleato strategico nel contrasto alla criminalità digitale³⁷. Secondo i *report* I.B.M. del

³⁷ N. MOHAMMED NASAR, A.F. MOHAMMED, B. SRUTHI op.cit. F.A. MOHANAD, S.M. MOHAMMED, S.B. SALMAN, B. REX, A. JAMIL, M.I. MASRULLIZAM, G.H. SAFARUNDIN, *HCPA: Hybrid cyber attack prediction model for securing healthcare applications*, in PLOS-One, 2025 (e0321941), disponibile online su: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0321941> (articolo pubblicato il 12 maggio 2025); A. MAHA, K.P. SATHEESH, M. GOPINATH, V. GNANA SRI MANASA,

2023, nel contesto odierno l’IA interviene «sia sul lato degli attaccanti sia su quello delle difese (...) con una riduzione del costo delle effrazioni a livello globale del 9%» (I.B.M, 2023, riportato in AgendaDigitale, M. Dal Co, articolo pubblicato il: 2 febbraio 2026, ultimo accesso il: 30 maggio 2026) e mostra tutto il suo potenziale benefico nel momento in cui riesce ad analizzare quantità di informazioni non più governabili dall’intervento umano, a intercettare comportamenti anomali in tempo reale, a sostenere processi decisionali più rapidi e selettivi e a costruire tempestivamente efficienti algoritmi di sicurezza (International Hospital Federation, articolo pubblicato il: 3 ottobre 2023, ultimo accesso il: 30 maggio 2026).

Su questa scia, si inseriscono tutte quelle diverse strategie di sicurezza della rete basate prevalentemente su strumentazioni dotate di sistemi di Intelligenza Artificiale.

In primo luogo, il meccanismo *Managed Detection and Response* – anche sottoforma *as-a-Service*, per quelle realtà più piccole e meno dotate di risorse a disposizione –, ovvero «una struttura di comando che grazie ad un’idonea infrastruttura tecnologica, adeguati processi e un ben formato team di professionisti», tra cui anche investigatori forensi ed *ethical hacker*, si occupa di «garantire la sicurezza informatica all’interno di una organizzazione indagando, analizzando, prevenendo e rispondendo in tempo reale alle minacce informatiche» (ICT Security Magazine, articolo pubblicato il: 14 giugno 2022; ultimo accesso il: 30 maggio 2026) –; infatti, questo sistema raccoglie e analizza informazioni di sicurezza IT, interviene immediatamente al verificarsi di un *cyber*-attacco, struttura un piano di difesa preventiva, gestisce gli incidenti e le *cyber*-minacce, rileva le intrusioni nel perimetro aziendale e svolge funzioni di mitigazione, *remediation*, monitoraggio, *threat hunting* e analisi delle vulnerabilità, con l’obiettivo di assicurare la continuità dell’assistenza mediante

M. ADITYA RAM, S. KHAN M.D., *A hybrid XGBoost–SVM ensemble framework for robust cyber-attack detection in the internet of medical things (IoMT)*, in Scientific Reports, 16/2026 (6855), disponibile online su: <https://www.nature.com/articles/s41598-026-37832-0> (articolo pubblicato il 31 gennaio 2026). M.J. MARTIN, P.P. PATEL, T. EGODAGE, *ransomware attacks and Cybersecurity concerns in modern hospitals: vulnerabilities and impacts on trauma centers and patient care*, in BMJ Journals, Trauma Surgery & Acute Care Open, vol. 11, 1/2026 (e002293), disponibile online su: https://tsaco.bmj.com/content/11/Suppl_1/e002293/ (articolo pubblicato il: 30 aprile 2026). INTERNATIONAL HOSPITAL FEDERATION, *Artificial Intelligence and cyber security in healthcare (YEL2023)*, disponibile su: <https://ihf-fih.org/news-insights/artificial-intelligence-and-cybersecurity-in-healthcare/>; articolo pubblicato il: 3 ottobre 2023, ultimo accesso il: 30 maggio 2026.

un efficace sistema di *incident response* e di consentire una corretta valutazione dei rischi, nonché la creazione di una strategia difensiva realmente funzionante (*ibidem*). In secondo luogo, il modello *Zero Trust* (Z.T.) per i settori critici, ad oggi uno dei paradigmi più efficaci (CyberSecurity360, F.M.R. Livelli, articolo pubblicato il: 2 luglio 2025; ultimo accesso il: 30 maggio 2026)³⁸: esso si configura come un modello di sicurezza «*never trust, always verify*», in cui nessun utente, dispositivo o flusso informativo è considerato attendibile o affidabile a priori secondo una logica di «*default deny*» che rafforza e isola i sistemi fino alla validazione dell’identità (M.J. Martin, P.P. Patel, T. Egodage, 2026)³⁹, e si caratterizza per valorizzare la protezione dell’identità di ogni utente, dei dispositivi e delle reti su cui opera con controlli multilivello, verifiche secondarie, autenticazione forte, segmentazione e revisione costante, così da poter ridurre i vettori di attacco, contenere il movimento laterale delle minacce e rafforzare la sicurezza anche in quei contesti caratterizzati da frequenti compromissioni delle credenziali e crescente sofisticazione delle minacce – i cd. “settori ad alta esposizione – (*ibidem*)⁴⁰.

Un ulteriore passo avanti nell’utilizzo della strumentazione IA a scopo difensivo è poi anche la sua integrazione in sistemi di *Open Source Intelligence* (abbreviato in O.S.I.N.T.), capaci di sfruttare i dati pubblici a disposizione sulla Rete per individuare gli autori di un reato, risalendo da *username*, *email*, *wallet*, numeri di telefono, *password*, immagini e linguaggi di programmazione dei *malware* utilizzati fino all’identità, alle infrastrutture e alle relazioni tra gli

³⁸ CYBERSECURITY 360, F.M.R. LIVELLI, *La Strategia Zero Trust*, disponibile su: <https://www.cybersecurity360.it/outlook/frodi-digitali-e-intelligenza-artificiale-che-cosa-ci-attende-questanno/>; articolo pubblicato il: 2 luglio 2025; ultimo accesso il: 24 aprile 2026.

³⁹ «Il modello ZT può, di fatto, rappresentare un approccio efficace per proteggere questi sistemi critici da avversari sempre più sofisticati, in grado di sfruttare i rapidi avanzamenti tecnologici e le mutevoli dinamiche delle minacce (...) Tale approccio non solo garantirà la continuità operativa, ma contribuirà anche a rafforzare la resilienza collettiva della comunità globale, migliorando la sicurezza, l’affidabilità e la solidità dei servizi essenziali per il benessere fisico ed economico della nostra società» (CyberSecurity360, F.M.R. Livelli, articolo pubblicato il: 2 luglio 2025; ultimo accesso il: 30 maggio 2026).

⁴⁰ «*There are seven principles or pillars of implementation of a zero-trust model (...) This includes zero assumed trust in five areas: (1) employees or other people accessing the system, (2) devices and (3) networks that access or interact with the system, (4) data and data sharing, and (5) workloads (particularly running in the public cloud). The final two principles are visibility and analytics to continuously monitor and analyze every activity on the network and automation and orchestration to enable automatic and coordinated incident response and task delegations during any cyberattack attempt or successful infiltration of the protected network*» (M.J. Martin, P.P. Patel, T. Egodage, 2026).

attaccanti (M.S. Bouflih, A. Chaibi, 2025)⁴¹; tuttavia, nonostante i sistemi OSINT siano estremamente economici e in grado di produrre piste investigative rapide ed efficaci anche in situazioni di scarsità di dati, non bisogna dimenticare il rischio non indifferente di falsi positivi, posto che i dati pubblici potrebbero rivelarsi falsati, datati o manipolati e/o manipolabili.

Oltre a ciò, l’efficacia dell’IA si manifesta soprattutto nella sua capacità di riconoscere *pattern* ricorrenti nei *log* di sistema, nei flussi di rete e nei comportamenti degli utenti connessi, individuando così i segnali sospetti prima che si traducano in un attacco conclamato (N. Mohammed, A.F. Mohammed, S. Balammagary, 2025): infatti, i meccanismi di Intelligenza Artificiale non si limitano a eseguire controlli statici, bensì, grazie ai sistemi di *Machine Learning* (M.L.) e *Deep Learning* (D.L.) di cui sono dotati, apprendono in continuazione nuove informazioni dai dati e raffinano progressivamente i propri modelli di classificazione, riuscendo così a stare al passo con l’evoluzione continua delle tecniche di attacco della criminalità cibernetica (International Hospital Federation, articolo pubblicato il: 3 ottobre 2023, ultimo accesso il: 30 maggio 2026). Pertanto, strumenti di *Machine Learning* (M.L.) e di *Ensemble Learning* (E.L., una tecnica che implica l’utilizzo combinato di più modelli, per poi prendere una decisione in base alla media delle probabilità delle classi, aumentando così l’accuratezza dell’*output* finale), come *XG-Boost* (*Extreme Gradient Boosting*, un algoritmo di M.L. basato su molti alberi decisionali in sequenza, i cd. *boosting*, ottimizzato al fine di ottenere alti livelli di accuratezza e velocità soprattutto con grandi *dataset*) e S.V.M. (*Support Vector Machine*, un modello di M.L. funzionante per classificazione, che utilizza diversi *kernel*, dei *software* fondamentali che fungono da ponte tra *hardware* del dispositivo e programmi in esecuzione, per separare al meglio le classi in uno spazio di caratteristiche) appaiono particolarmente promettenti per la rilevazione delle intrusioni ai danni di sistemi interconnessi (E.A. Al-Qarni, 2023)⁴². Ne deriva che i modelli IA, se ben progettati, non costituiscono soltanto un mezzo di

⁴¹ M.S. BOUFLIH, A. CHAIBI, *Specificity of Ransomware Investigation – Analytical Study in Light of the European Cybercrime Programmed 2022*, in *Journal of Human Sciences*, vol. 36, 2/2025, disponibile online su: <https://doi.org/10.34174/0079-036-002-020> (articolo pubblicato il: 30 giugno 2025).

⁴² E.A. AL-QARNI, *Health data sarin attitudes towards primary and secondary use of data: a systematic review*, in *I.J.A.C.S.A. International Journal of Advanced Computer Science and Applications*, vol. 14, 5/2023, disponibile online su: <https://www.ijacsa.thesai.org/>.

automazione, ma possono anzi divenire un presidio dinamico di prevenzione e supporto, soprattutto nelle prime fasi della risposta all’incidente (*ibidem*).

A tal proposito, l’architettura basata sull’integrazione tra IA e tecnologia *blockchain* – un registro distribuito, condiviso e mutabile solo con l’aggiunta di nuove informazioni in cui le transazioni sono organizzate in blocchi collegati, usato per garantire l’integrità e la tracciabilità dei dati – si articola solitamente su cinque livelli: il primo, denominato “livello dei dati”, raccoglie informazioni, acquisendo anche campioni di *malware*, successivamente inoltrati al livello di analisi (E.A. Al-Qarni, 2023); il secondo, dedicato all’analisi del *malware*, impiega strumenti specializzati – ad esempio, *Pestudio* e *Process Explorer*, due strumenti di analisi appositamente previsti per il sistema *Windows* – per esaminare i campioni raccolti, trasferendo poi i dati al terzo livello, il “livello intelligente”, dove l’intelligenza artificiale interviene in forma di S.V.M. o R.F. – *Random Forest*, un algoritmo di *machine ensemble* che combina numerosi alberi decisionali per migliorare accuratezza e robustezza nelle classificazione e regressione, molto simile al sistema *XG-Boost* – con l’obiettivo di individuare e appianare eventuali vulnerabilità (*ibidem*); il quarto, rappresentato dall’intervento della *blockchain*, si occupa dell’archiviazione dei dati in maniera sicura, garantendo integrità e tracciabilità delle informazioni (*ibidem*); e, infine, il quinto, il “livello applicativo”, consente la fruizione dei dati da parte di tutti gli interessati. Inoltre, tali sistemi possono includere anche meccanismi di rilevamento delle intrusioni basati su *autoencoder* – un insieme di reti neurali capaci di comprimere e ricostruire i dati, usate per ridurre le dimensionalità e rilevare più facilmente le anomalie – o i cd. *stacked autoencoders* – più *autoencoder* impilati tra loro per estrarre caratteristiche via via più astratte, utilizzati soprattutto per captare meccanismi di *intrusion* cibernetica –, articolati nelle tre fasi di pre-elaborazione dei dati, in cui solitamente sono definiti i modelli di comportamento intrusivo, estrazione delle caratteristiche e identificazione dei comportamenti anomali (*ibidem*).

Un ulteriore sviluppo di rilievo è rappresentato dal passaggio da architetture centralizzate basate su sistemi M.L. a modelli distribuiti, quale il modello di *Federated Learning* (F.L.), un paradigma che permette l’addestramento degli algoritmi su dispositivi locali e nodi periferici senza trasferimento dei dati grezzi al *server* centrale (A. Ghourabi, A.A. Alkhalil,

2025)⁴³; sono infatti diversi anni che in dottrina si susseguono svariate proposte su come integrare l’IA nei meccanismi di *cyber*-sicurezza. In particolare, la ricerca si è sviluppata soprattutto su tre direttrici: la prima è lo sfruttamento di *Machine Learning* (M.L.) tradizionale ed *Ensemble Learning* (E.L.) per il rilevamento delle intrusioni; la seconda è lo sviluppo di approcci *Deep Learning* per il *training* dei sistemi di IA più sviluppati; e la terza è lo studio di soluzioni ibride o distribuite, in grado di combinare tutte le opzioni possibili nella risposta più soddisfacente (M.F. Ali, M.S. Mohmood, B.S. Shukur, R. Bacarra, J. Abedalrahim, J. Alsayaydeh, M.M. Ibrahim, S.G. Herewan, articolo pubblicato il: 12 maggio 2025; ultimo accesso il: 30 maggio 2026)⁴⁴. Nel complesso, però, l’opinione condivisa è che le soluzioni ibride superino di gran lunga i metodi di IA più tradizionali, pur rimanendo sempre presenti alcuni problemi di struttura legati alla *privacy*, ai falsi positivi e alla generalizzazione degli attacchi reali (*ibidem*): modelli leggeri, distribuiti e ottimizzati si presentano dotati di un 98,04% di *accuracy*, con un buon equilibrio tra questo aspetto, la robustezza del sistema, il costo computazionale e la sua leggerezza e ottimizzazione (*ibidem*), mentre l’adozione di soluzioni basate su M.L. e *blockchain* richiede invece una notevole quantità di dati per l’addestramento e pone diverse questioni di interoperabilità, implementazione, costi e opacità (*ibidem*).

Accanto al piano strettamente tecnico, permane però anche una necessaria dimensione giuridica ed etica, che impedisce di considerare l’IA come un modello di risposta autosufficiente: l’uso di sistemi intelligenti nella difesa dalle minacce digitali deve infatti essere governato da regole di tracciabilità, *audit* continuo, responsabilizzazione dei processi di sviluppo e controllo umano effettivo; in tal senso, l’efficienza dell’algoritmo non può mai sostituire la qualità della *governance* umana, che ne garantisce uno sviluppo affidabile in linea con il principio di *accountability*. Da cui l’esigenza di integrare l’utilizzo dell’IA con *policy* di *cyber*-sicurezza autonome, procedure di *incident response*, *standard* di conformità rigorosi e appositi meccanismi di *risk management* e *risk assessment*, di modo che la tecnologia non sostituisca mai del tutto

⁴³ G. ABDALLAH, A. ADEL, *A Federated Learning Model for Detecting Cyberattacks in Internet of Medical Things Networks*, in IEEE-Xplore, Journals & Magazines, 13/2024 (11079590), disponibile online su: <https://ieeexplore.ieee.org/abstract/document/11079590/>.

⁴⁴ N. MOHAMMED NASAR, A.F. MOHAMMED, B. SRUTHI op.cit. F.A. MOHANAD, S.M. MOHAMMED, S.B. SALMAN, B. REX, A. JAMIL, M.I. MASRULLIZAM, G.H. SAFARUNDIN, op.cit.

l’organizzazione umana alla base, bensì la rafforzi fino a renderla inattaccabile e invincibile (E.A. Al-Qarni, 2023).

Si può dunque concludere affermando che l’IA diviene davvero “amica” dell’uomo nella lotta alla criminalità digitale se e solo se viene impiegata non come strumento isolato, ma quale parte di un ecosistema più ampio, fatto di prevenzione, rilevazione e reazione; infatti, una difesa efficace non dipende solo dal grado di evoluzione degli algoritmi intelligenti, quanto piuttosto dalla sua integrazione con attività di supervisione umana, solidi presidi regolatori e già di partenza solide fondamenta *hi-tech*, capaci di rendere l’innovazione un fattore di vera e propria resilienza sistemica.

NC