LA RESPONSABILITÀ PENALE AL TEMPO DI CHATGPT: PROSPETTIVE *DE IURE CONDENDO* IN TEMA DI GESTIONE DEL RISCHIO DA INTELLIGENZA ARTIFICIALE GENERATIVA (*)

di Leonardo Romanò

Gli esempi dei possibili utilizzi della nuova Intelligenza Artificiale "generativa" (come ChatGPT) si moltiplicano senza sosta, e con essi, le preoccupazioni per le ricadute offensive che, in prospettiva futura, potrebbero derivare dalla sua applicazione in settori socio-economici in cui sono in gioco beni giuridici fondamentali. Ad oggi, la regolamentazione di tale tecnologia è ancora poco sviluppata e, quindi, aggalla con forza il tema di come i rischi connessi a questa post-moderna forma di tecnologia vadano governati e con quali strumenti. Dopo una breve premessa tecnica sul suo funzionamento, il lavoro si propone di scrutare i rapporti tra questa tecnologia ed il panorama regolatorio emergente a livello europeo, ponendo l'accento sul tema del controllo umano e della responsabilità penale rispetto a eventuali eventi avversi prodotti da tali sistemi. Vengono, infine, avanzate alcune proposte volte a individuare l'ambito (pur limitato) di un possibile reale utilizzo del diritto penale per fronteggiare adeguatamente i rischi rilasciati da questa nuova tecnologia

(*) Si ringrazia il professor Carlo Sotis per i preziosi commenti e spunti di riflessione offerti nella stesura di questo articolo. Leonardo Romanè è dottorando di ricerca in diritto penale all'Università del Lussemburgo e all'Università degli Studi della Tuscia. Sostenuto dal Fondo Nazionale della Ricerca del Lussemburgo PRIDE 19/14268506.

1. Premessa: diritto penale, nuovi rischi e vecchie paure.

Un nuovo oracolo algoritmico[1] ha fatto il suo ingresso nelle nostre vite: ChatGPT[2]. La diffusione di sistemi di Intelligenza Artificiale c.d. generativa (c.d. *Generative AI*, di seguito IAG)[3] – di cui ChatGPT e simili[4] rappresentano gli esempi più sofisticati attualmente in circolazione – ha messo in moto un radicale cambio di paradigma rispetto al nostro modo di generare e acquisire conoscenza; e i *social media* e i giornali si sono popolati di esempi sorprendenti e spaventosi di cosa questa tecnologia sia in grado di fare.

L'apprendimento automatico (c.d. *machine learning*)[5], che un tempo era destinato a eseguire solo compiti specifici e circoscritti, è stato ora impiegato per progettare complessi modelli linguistici[6] in grado di produrre, a partire da una richiesta dell'utente (o *prompt*), qualsiasi tipo di contenuto come *output* (tra cui testo, immagini, audio o codice). Sicché, un aspetto chiave che accompagna lo sviluppo di questa tecnologia riguarda proprio la vastità del suo potenziale utilizzo.

Basti immaginare uno studente che chieda a ChatGPT: scrivi un saggio su Napoleone, includi una o due citazioni dotte e concludi con una frase ad effetto[7]; uno scienziato che chieda: sintetizza questa proteina in modo che possa utilizzarla per creare un nuovo farmaco[8]; uno scrittore che chieda: scrivi un sonetto nella

---

[1] L'espressione è di MANES (2020), p. 6.

[2] ChatGPT si basa sull'architettura *Generative Pre-trained Transformer* (GPT). Il sistema è stato addestrato utilizzando una rete neurale progettata per l'elaborazione del linguaggio naturale su un set di dati di oltre 45 terabyte di testo proveniente da Internet (libri, articoli, siti web e altri contenuti testuali), che in totale comprendeva miliardi di parole di testo. Il modello di base, GPT-3, rilasciato nel novembre 2022, viene perfezionato costantemente e la nuova versione, GPT-4, è stata rilasciata nel marzo 2023. Mentre quest'ultimo rappresenta una versione premium a pagamento, il primo è generalmente accessibile attraverso un sito web facile da usare: https://chat.openai.com/chat.

[3] Per riferirci, in generale, ai sistemi di IA useremo di seguito, indifferentemente, espressioni come "sistema IA, *software*, agente artificiale, macchina, algoritmo".

[4] Alla famiglia dell'IAG appartengono sistemi in grado di generare testi (come ChatGPT o Bard), immagini (DALL-E), video (Synthesia) e persino arte (Midjourney). Per le specifiche tecniche dei vari modelli, cfr. https://platform.openai.com/docs/models/overview.

[5] Per un'analisi approfondita della tecnologia di *machine learning*, cfr. HAO (2018).

[6] Un modello linguistico consiste essenzialmente nell'uso di varie tecniche statistiche basate sul lavoro di algoritmi addestrati per analizzare miliardi di miliardi di parole (questo è il "*learning*" del "*machine learning*") e calcolare quanto è probabile che, data una certa sequenza di parole, ne segua una anziché un'altra. Ad esempio, date le parole "ha messo online un chatbot basato sull'ultima", "versione" è un seguito più probabile di "decisione". Cfr. OPENAI (2019), in https://openai.com/research/better-language-models.

[7] Le preoccupazioni per l'impatto che un uso improprio di ChatGPT potrà avere sul sistema educativo hanno portato al divieto espresso del suo utilizzo nelle scuole pubbliche a New York. V. ROSENBLATT (2023), in https://www.nbcnews.com/tech/tech-news/chatgpt-passes-mba-exam-wharton-professor-rcna67036.

[8] Uno studio, pubblicato su *Nature Biotechnology*, ha dimostrato come ProGen, un modello linguistico simile a ChatGPT, sia stato utilizzato per lo sviluppo di nuove proteine che possono essere utilizzate

lingua di Shakespeare[9]; e ancora, un giudice che chieda: scrivi la sentenza al posto mio[10]. Gli esempi dei possibili utilizzi di questa tecnologia si moltiplicano senza sosta, e con essi, le preoccupazioni per le potenziali ricadute offensive che, in prospettiva futura, questa intromissione del non umano potrà avere sull'umano. Oltre alle prevedibili questioni di stampo etico, aggallano con forza i rischi derivanti sia dall'utilizzo dell'IAG per prendere decisioni automatiche in settori che incidono sui diritti fondamentali della persona (come nel mondo del lavoro, dell'assistenza sanitaria, delle assicurazioni, della previdenza sociale e della giustizia), sia dall'uso scorretto nonché da malfunzionamenti di tale tecnologia[11].

A tal riguardo, lo *Europol Innovation Lab* ha pubblicato un report in cui preconizza l'emergere nel prossimo decennio di nuove forme di criminalità associate all'utilizzo improprio di ChatGPT[12]. In particolare, il report evidenzia il rischio che i più diversi reati previsti dall'ordinamento possono combinarsi con modalità commissive che passano attraverso l'azione materiale dell'agente artificiale[13]. Da un lato, la capacità di tale tecnologia di redigere testi altamente verosimili la rende uno strumento estremamente utile nell'ambito del *cybercrime*[14], segnatamente nella commissione di frodi online: ad esempio, per stessa "ammissione" di ChatGPT[15], esso "potrebbe essere utilizzato per creare notizie false, per impersonare individui online o per automatizzare attività illegali, come la creazione di truffe di *phishing*, *hacking*, *deepfake* o messaggi di *spam*". Dall'altro lato, "anche se utilizzato conformemente alle sue finalità, permane il rischio che il sistema possa agire in modo difforme rispetto alle istruzioni impartite dall'agente umano, concretizzando così un fatto non voluto o addirittura diverso da quello avuto di mira da quest'ultimo. Ciò potrebbe comportare il rischio che il risultato

---

per molteplici applicazioni, da quella della progettazione di nuovi farmaci alla plastica degradante. MADANI *et al*. (2023).

[9] ChatGPT si è dimostrato in grado di produrre testi scientifici o letterari a prova di revisione umana. Cfr. SAMPLE (2023), in https://www.theguardian.com/science/2023/jan/26/science-journals-ban-listing-of-chatgpt-as-co-author-on-papers.

[10] Un magistrato colombiano ha utilizzato la chatbot di OpenAI nella stesura di una decisione in Colombia. Benché il giudice abbia svolto l'attività di sua competenza autonomamente, la redazione del testo è stata supportata dalla chatbot per quanto concerne la sezione dedicata alle argomentazioni a supporto della decisione. Inoltre, la chatbot della startup americana DoNotPay è già pronto a fare il suo debutto nelle aule di tribunale statunitensi come avvocato-robot.

[11] Le preoccupazioni per le conseguenze negative che lo sviluppo di tale tecnologia potrà avere su ambiti come il lavoro, l'istruzione e l'economia ha spinto un gruppo di accademici ed esperti di tecnologia e informatica a chiedere, tramite una petizione su Futureoflife.org, di sospendere per sei mesi l'addestramento delle IAG. V. FUTURE OF LIFE INSTITUTE (2023), in https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[12] V. EUROPOL INNOVATION LAB (2023), in https://www.europol.europa.eu/media-press/newsroom/news/criminal-use-of-chatgpt-cautionary-tale-about-large-language-models.

[13] *Ivi*, p. 7.

[14] *Ivi*, p. 8.

[15] Si consenta questo poco ortodosso riferimento ai risultati generati da un'esplicita "intervista" a ChatGPT con *prompt*: "Quali sono i rischi penali derivanti dall'utilizzo di ChatGPT?".

finale sia inappropriato o addirittura dannoso per l'utente o per l'azienda utilizzatrice": si pensi a un sistema di IAG che "diffami" una persona rielaborando i suoi dati sensibili in modo inesatto, ovvero istighi l'utente a realizzare atti illeciti e financo autolesivi[16]. Proprio questa propensione a trasformare in modo imprevedibile l'*input* umano in un *output* differente è destinata ad aumentare in misura proporzionale ai margini di autonomia del sistema: quanto maggiore la sua autonomia, tanto maggiore il rischio che il "prodotto soggettivizzato"[17] provochi eventi avversi senza l'intervento diretto dell'uomo, sollevando dunque questioni proprie del diritto penale.

Osservazioni, quelle sin qui condotte, che possono apparire ovvie e scontate, specie se si considera che il diritto penale ha già avuto modo di confrontarsi con la questione della possibile attribuzione di responsabilità penale per le sempre più numerose manifestazioni lesive associate alla diffusione di agenti artificiali[18]. Ancor prima che l'IAG facesse la sua entrata in scena, una copiosa letteratura penalistica, sia italiana[19] che internazionale[20], ha a più riprese evidenziato come l'opacità e l'imprevedibilità di certi strumenti basati su IA renda estremamente difficile imputare per colpa un reato algoritmico[21] all'agire di un singolo soggetto umano[22].

Tuttavia, se per un verso i rischi da ignoto tecnologico rilasciati da ChatGPT sono in larga parte una riedizione di problematiche già note nell'interazione uomo-macchina, per altro verso esiste un dato peculiare e

---

[16] Casi di questo tipo si sono già verificati in passato con riferimento a chatbot meno sofisticati di ChatGPT. Cfr. LANA (2021), in https://www.corriere.it/tecnologia/21_dicembre_29/alexa-sfida-bimba-inserire-moneta-presa-elettrica-amazon-aggiorna-software-7533808a-6887-11ec-b54e-d173b9021fcd.shtml.

[17] Questa espressione è di CAPPELLINI (2023), p. 24, il quale preconizza l'avvento di un'IA "che non si limita più a realizzare la volontà umana che le sta dietro, ma agisce nel mondo in modo che non è più governato dalla mano dell'uomo".

[18] La casistica è ampia e non se ne può dar conto in questa sede. Tra i casi più "classici" di danno prodotto da sistemi di IA basti qui menzionare quelli nel settore della guida autonoma, in quello finanziario o in quello medico-chirurgico. Per tutti, cfr. rispettivamente: CAPPELLINI (2019); CONSULICH (2018); LAGIOIA F. e CONTISSA G. (2020). In una prospettiva più ampia sui reati algoritmici: KING *et al*. (2020); CALDWELL *et al.* (2020); PAGALLO (2013); BECK (2017).

[19] Con particolare riguardo alla dottrina italiana, la riflessione sulle implicazioni della IA per il diritto penale è già molto stratificata. Per tutti: TRIPODI (2023); SALVADORI (2021); GIANNINI (2022); MAGRO (2019); PIERGALLINI (2020); PIVA (2022); PANATTONI (2021); TRONCONE (2022); CONSULICH (2022); MINELLI (2022); MANES (2020); RUFFOLO (2021). Per una disamina dei rapporti tra IA e giustizia penale, si rimanda, *ex multis*, a BASILE (2019); sia consentito, infine, un riferimento al nostro ROMANÒ (2022).

[20] La letteratura straniera sul tema è troppo vasta per essere qui sistematicamente ricordata: si rinvia, per tutti, ai richiami contenuti nelle altre note, oltre che ai riferimenti bibliografici di cui già al nostro scritto da ultimo citato.

[21] L'espressione "reati algoritmici" o "*AI crimes*" è di R. ABBOTT e A. SARCH (2019), p. 323, in cui viene definita come "*cases in which an AI would be criminally liable if a natural person had performed a similar act*".

[22] Per tutti, v. SURDEN e WILLIAMS (2016), p. 157; SALVADORI (2021), p. 102; BATHAEE (2018); R. ABBOTT e A. SARCH (2019), p. 330 ss.

sostanzialmente nuovo rispetto al passato: il grado di autonomia e la vastità del potenziale applicativo dell'IAG parrebbero escludere del tutto la stessa possibilità di un controllo e/o intervento umano preventivo. Se questa ipotesi si rivelasse corretta, venendo meno la possibilità di imputare un eventuale risultato lesivo a un utilizzatore che non partecipa più all'attività algoritmica, addirittura ormai privato della capacità di governarla, non rimarrebbe altro che un problematico "fatto proprio" della macchina, o al massimo un caso fortuito, privo di copertura sul piano della responsabilità penale [23].

Una cosa è certa: all'indomani dell'ennesimo traguardo tecnologico, il quesito che la scienza penale è chiamata a porsi è sempre il medesimo – come i rischi connessi a questa post-moderna forma di tecnologia vadano governati e con quali strumenti[24]. Come noto, elemento ricorrente nelle analisi che valorizzano la "società del rischio"[25] in prospettiva penalistica è la riflessione sulle torsioni indotte nel diritto penale dal fatto di essere divenuto il sistema preposto alla minimizzazione dei rischi tipici di tale contesto, in funzione di rassicurazione sociale ed esorcismo dell'insicurezza collettiva[26]. Allo stesso modo, tale riflessione, nel confronto con sofisticati sistemi IAG e correlati pericoli, tenderà verosimilmente a riproporsi, con il rischio di alimentare – quando non sia possibile neutralizzare i nuovi rischi con divieti *tout court* – scelte politiche precauzionistiche e ultraresponsabiliste volte ad istituire in capo a utilizzatori e programmatori irreali doveri di controllo ad ampio spettro sull'attività algoritmica[27]. E ciò con una duplice conseguenza: sul piano penalistico, facendo ricorso a inaccettabili stravolgimenti del paradigma colposo, come già accaduto nelle dinamiche imposte dalla moderna società del rischio[28]; e su un piano più generale, disincentivando l'innovazione, impedendo così alla società di trarre beneficio da questa tecnologia.

È quindi alla luce di queste premesse che va affrontato il discorso volto ad individuare quale può e dev'essere l'ambito di un possibile reale utilizzo del diritto penale per fronteggiare adeguatamente i rischi rilasciati da questa nuova tecnologia. Per farlo, l'indagine si dipanerà lungo tre direttrici. Inizialmente si svolgerà una breve premessa tecnica sul funzionamento ed i limiti dei sistemi di IAG, con particolare attenzione alle differenze con i "tradizionali" sistemi di IA. Poi ci si soffermerà sul problema della regolamentazione e della responsabilità. Con riguardo al primo profilo, il lavoro si propone di scrutare i rapporti tra questa tecnologia ed il panorama normativo emergente a livello europeo, ponendo l'accento sul tema del controllo umano e della delimitazione di aree di rischio

---

[23] CAPPELLINI (2023), cit., 12. Più in generale, in punto di *responsibility gap*, U. PAGALLO e S. QUATTROCOLO, (2018), p. 385.

[24] PIERGALLINI (2020), cit., 1746.

[25] Resta fondamentale il rinvio a BECK (1986).

[26] La letteratura sul punto è vasta, per cui si rimanda, *ex multis*, a STORTONI (2004); PERINI (2018); STELLA (2003).

[27] CAPPELLINI (2023), cit., 28.

[28] V. STELLA (2003), p. 221 e ss.

consentito; mentre, con riferimento al secondo, verranno esaminate le asperità probatorie che affliggono, sul terreno penalistico, l'allocazione delle responsabilità per eventi avversi prodotti dall'IAG. Nella parte conclusiva, saranno delineate alcune proposte per una futura dottrina penale del controllo dei rischi da ignoto algoritmico.

2. Caratteri e limiti dei modelli generativi: dall'IA specializzata verso un'IA a finalità generale.

Per comprendere il funzionamento e le possibili implicazioni giuridiche di ChatGPT è utile cominciare con una breve premessa teorica sulla nozione di IA.

Intesa nella sua accezione più ampia, l'IA è qualcosa che "si riconosce quando la si vede", ma che non è chiaramente definita. Nel dibattito tecnico non esiste una definizione unanimemente condivisa di IA, poiché le caratteristiche di una data applicazione sono definite dalle funzioni che persegue e dall'ambiente in cui opera[29]. Nel dibattito pubblico, di contro, l'IA è tipicamente considerata come una disciplina che mira a sviluppare sistemi computazionali in grado di compiere operazioni che richiederebbero le capacità cognitive e decisionali degli esseri umani[30]. Lo spettro di tecnologie riferibili a una simile definizione è, però, ampio e diversificato – da un veicolo senza conducente a una chatbot, da un *software* di *trading* ad alta frequenza a un robot industriale – con limitati punti di contatto.

Ma su un punto si registra un certo consenso: anche se ci sono esempi di IA ovunque si guardi, l'IA onnisciente e super intelligente, che di solito vuole conquistare il mondo nei romanzi distopici, non è ancora stata inventata e probabilmente mancano ancora diversi anni al traguardo[31]. Le tecnologie di IA in cui viviamo immersi perseguono perlopiù una funzione o un compito specifico (IA

---

[29] Sull'indeterminatezza della nozione di IA v. UBERTIS (2020), p. 76; MCCARTHY (2007), in http://jmc.stanford.edu/articles/whatisai.html. Che l'IA non è un concetto univoco è evidenziato da S. LEGG e M. HUTTER (2007), che individua oltre settanta definizioni diverse.

[30] I numerosi riferimenti rinvenibili nella letteratura scientifica tendono a convergere verso questa definizione. Per tutti, KAPLAN (2016), p. 1, "programmi informatici capaci di comportamenti che riterremmo intelligenti se messi in atto da esseri umani", nonché SARTOR (1996), "modelli computazionali capaci di eseguire compiti che richiederebbero intelligenza da parte dell'uomo". Quella elaborata dalla Commissione europea all'art. 3(1) della proposta di Regolamento UE sull'IA rappresenta una specificazione della definizioni appena richiamate: "un sistema progettato per operare con elementi di autonomia e che, sulla base di dati e input forniti dalla macchina e/o dall'uomo, deduce come raggiungere un determinato insieme di obiettivi utilizzando approcci basati sull'apprendimento automatico e/o sulla logica e sulla conoscenza, e produce output generati dal sistema come contenuti, previsioni, raccomandazioni o decisioni". *Proposta di Regolamento del Parlamento e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione*, 21 aprile 2021, COM (2021) 206 final.

[31] Per alcune interessanti riflessioni su questo passaggio generazionale, cfr. V. MULLER e N. BOSTROM (2016).

specializzata, o "*task-specific AI*"), mentre solo una parte limitata della ricerca in materia mira a replicare capacità simili a quelle umane (IA a finalità generali, o "*general purpose AI*"). In altre parole, si tratta di sistemi pensati per fare una cosa specifica (previsioni o classificazioni, guidare, riconoscere i volti nelle foto, raccomandare quale libro vuoi leggere successivamente, determinare se sei un buon rischio di credito, rilevare il cancro della pelle) – con l'ovvia conseguenza che tali sistemi, pur essendo migliori degli esseri umani nel loro ambito di addestramento, sono completamente inutili nel fare qualunque altra cosa diversa dal compito specializzato per cui sono stati progettati.

È alla luce di queste premesse che si coglie la portata del passaggio epocale in atto con la comparsa della IAG. Sistemi come ChatGPT e simili rappresentano i primi barlumi all'orizzonte dell'IA a finalità generale – quel momento a lungo profetizzato in cui le menti meccaniche supereranno i cervelli umani non solo quantitativamente in termini di velocità di elaborazione e dimensioni della memoria, ma anche qualitativamente in termini di intuizione intellettuale, creatività artistica e ogni altra facoltà distintamente umana[32]. Si tratta, in sostanza, di "modelli avanzati di apprendimento automatico che vengono addestrati per generare nuovi dati, come testo, immagini o audio"[33]; l'addestramento avviene "utilizzando varie tecniche per trovare schemi e relazioni nei dati in modo autonomo, senza che gli venga detto esplicitamente cosa cercare. Una volta appresi questi schemi, il modello è in grado di contestualizzare problemi anche molto complessi in maniera completamente autonoma, generando nuovi esempi simili ai dati di addestramento e nuovi contenuti".

Allo stesso tempo, però, l'architettura stessa di questi sistemi rende la perfezione difficile da raggiungere: avere una IAG non è come avere il genio della lampada, ma come avere un esercito di schiavi tonti ma onniscienti e molto veloci[34]. Le attuali versioni di questi sistemi non hanno prettamente coscienza di ciò che dicono o fanno, ma si limitano a valutare quale parola usare dopo quella che hanno appena selezionato, imitando informazioni prodotte dall'uomo in modo puramente statistico, anziché imparare effettivamente come funziona il mondo[35].

---

[32] Quelle che seguono sono le risposte fornite da ChatGPT ai seguenti *prompt*: "Cosa sono i modelli di IA generativa?" "Puoi spiegare le basi tecniche dei modelli generativi in modo semplice, in modo che un lettore inesperto possa comprenderle?".

[33] Si noti che la definizione di IA a finalità generali recentemente introdotta nella proposta di Regolamento UE all'art. 3(1) lett. B non è affatto dissimile da quella fornita da ChatGPT: si tratta di un sistema "destinato a svolgere funzioni di applicazione generale, quali il riconoscimento di immagini e vocale, la generazione di audio e video, il rilevamento di modelli, la risposta a domande, la traduzione e altre".

[34] Questa simpatica "definizione" è di LATRONICO (2022).

[35] Si registra già un certo scetticismo circa le reali capacità di ChatGPT, le cui risposte non sarebbero altro che una parafrasi o una "sfocata immagine del web". Più in generale sul punto, cfr. CHOMSKY (2023), in https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html; nonché CHIANG (2023), in https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

Per farlo, però, non attingono a una grande raccolta di informazioni verificate e aggiornate – che sarebbe del resto molto difficile da organizzare e mantenere aggiornata – ma attingono a enormi quantità di dati di testo provenienti dalle fonti più disparate reperibili su Internet – quindi anche quelle sbagliate[36]. Pertanto, per quanto sia a dir poco incredibile ciò che questi modelli sono in grado di fare, è ancora imperfetto: e in molti casi, un risultato imperfetto non basta. Per imbrogliare nello svolgimento di una traccia d'esame, almeno per lo studente che non ambisca al voto più alto, fa poca differenza; ma quando si tratta di applicazioni in cui sono in gioco beni giuridici (individuali o collettivi) ben più rilevanti (come nel settore medico, finanziario o bancario) un singolo errore rischia di avere conseguenze gravi e difficili da prevedere.

Proprio questa tendenziale fallibilità della macchina fa dell'IAG un vero rompicapo per il legislatore o le autorità di regolazione che tentino di prevenire o ridurre i rischi connessi agli innumerevoli tipi e applicazioni di questa tecnologia[37]. L'esempio dell'Italia, con il recente provvedimento dell'autorità Garante per la Protezione dei Dati Personali, che ha portato all'auto-sospensione del servizio per gli utenti italiani, ne è la prova[38]. Tra le varie obiezioni formulate in punto di protezione dei dati personali[39], il Garante della Privacy rileva che "le informazioni fornite da ChatGPT non sempre corrispondono al dato reale, determinando quindi un trattamento di dati personali inesatto". In buona sostanza, il timore più che fondato del garante è che gli utenti possano essere oggetto di diffamazione da parte di sistemi su cui è difficile agire o rivalersi[40].

Oltre ai prevedibili interrogativi di stampo etico, vengono dunque in gioco il problema della regolamentazione e della responsabilità. Temi con i quali avremo modo di confrontarci nel prosieguo di queste riflessioni.

---

[36] EUROPOL INNOVATION LAB (2023), cit., 4. Il tentativo di correggere i possibili *bias* e creare un sistema più affidabile ha d'altra parte sollevato complesse questioni etiche, laddove è emerso come OpenAI abbia sfruttato lavoratori kenioti per effettuare correzioni manuali dei risultati di ChatGPT: cfr. PERRIGO (2023), in https://time.com/6247678/openai-chatgpt-kenya-workers/.

[37] V. Hacker *et al.* (2023), in https://verfassungsblog.de/chatgpt/.

[38] GARANTE PER LA PRIVACY (2023), in https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847.

[39] Le altre obiezioni del Garante per la Privacy riguardano la mancanza di una informativa agli utenti i cui dati vengono raccolti da OpenAI, per cui entrando nel sito si dà in sostanza esplicito consenso al trattamento dei dati, nonché l'assenza di qualsivoglia filtro per la verifica dell'età dei minori.

[40] Sebbene riteniamo che la posizione del Garante sia in linea di massima corretta, preme qui notare come il provvedimento si sia dimostrato ben poco efficace a fronte alla possibilità di aggirare facilmente il blocco avvalendosi di servizi VPN. Com'è agevole intuire, la portata sovranazionale (anzi, globale) del fenomeno esige l'attuazione di politiche sovranazionali, condannando all'inefficacia qualsiasi approccio normativo che si fondi e confronti unicamente con il contesto dell'ordinamento nazionale.

3. Primi tentativi di regolamentazione dell'IAG: l'approccio europeo.

3.1. La proposta di Regolamento sull'IA e la logica della precauzione moderata.

Prima di porci il problema dell'ascrizione di responsabilità per le potenziali conseguenze negative, è forse opportuno affrontare preliminarmente la questione relativa ai rapporti tra rischio e precauzione[41].

Di fronte a sistemi che si ritiene siano o possano essere, almeno potenzialmente, fonte di pericolo per l'uomo, la prima decisione è extra e pre-penale – quindi di natura genuinamente politica[42]. Essa consiste nella scelta sul grado di rischio che si è disposti a tollerare nell'utilizzo di una certa tecnologia. Per un verso, nei casi in cui, *a prescindere dai suoi potenziali benefici, un'applicazione presenti rischi difficilmente governabili ed evitabili da parte dell'uomo, non resta altra scelta che proibire in toto l'utilizzo della stessa. Per altro verso, i problemi maggiori nascono, logicamente, dopo che l'ordinamento, nell'alternativa tra consentire una certa attività o proibirla in radice, ha già optato per la prima soluzione. In questa ipotesi, il legislatore deve* stabilire come governare preventivamente i rischi inerenti a una tecnologia rispetto ad applicazioni che si assumono utili e lecite, ritagliando a tal fine *aree di rischio "algoritmico" consentito*[43] presidiate da un apparato di sanzioni adeguato e funzionale alla tutela degli interessi coinvolti[44].

Sotto questo profilo, pur nell'attuale assenza di regole cautelari positivizzate, non si può non tener conto del quadro normativo emergente a livello europeo. Frutto maturo di un dibattito etico e giuridico da tempo in corso in seno alle istituzioni europee, la nuova proposta per un Regolamento europeo sull'IA (c.d. *AI Act*, per brevità "la proposta di Regolamento")[45], pubblicata nell'aprile del 2021 dalla Commissione, sembra fornire un primo quadro orientativo *di misure cautelari e obblighi specificamente concernenti la messa in commercio e l'uso di sistemi IA (artt. 8-15). La* proposta delinea una regolazione dell'AI proporzionata alla probabilità, al tipo e all'intensità del rischio rilasciato dall'applicazione che si intenda regolare (c.d. approccio *risk-based*)[46]. La regolazione basata sul rischio comporta l'individuazione di differenti classi di rischio e, specularmente, l'identificazione di regimi regolatori differenti. Così, per le AI considerate ad alto rischio (art. 6) sono previsti obblighi specifici, mentre vi sono sistemi vietati del

---

[41] Per una disamina dei complessi rapporti tra logica precauzionale, gestione del rischio e ricadute penalistiche, si rimanda a RUGA RIVA (2006); PIERGALLINI (2011); CASTRONUOVO (2012).

[42] STORTONI (2004), cit., 83.

[43] Per un'approfondita disamina sul concetto di rischio consentito nel diritto penale moderno, con dovizia di riferimenti, si rimanda a PERINI (2010), *passim*.

[44] Così PIVA (2022), pp. 683 ss.

[45] V. *sub* nota 30.

[46] Per un commento all'approccio della proposta di regolamento, cfr. CONTISSA *et al*. (2021).

tutto, perché, secondo la valutazione del legislatore europeo, pongono un rischio inaccettabile (art. 5)[47].

In particolare, per le IA ad alto rischio, la proposta sembra ritagliare aree di rischio consentito secondo una logica di precauzione moderata, *stabilendo i limiti entro i quali è accettabile che un sistema di IA compia errori o causi danni; e lo fa* statuendo espressamente che le figure soggettive tipizzate (ossia, il produttore, programmatore e l'utilizzatore della macchina, persona fisica o giuridica) siano gravate tanto da obblighi di condotta tipici articolati e complessi[48] quanto da disposizioni di principio – tra le quali primeggia quello di *accountability* e *human oversight* – che investono tutta la loro attività[49]. In questo senso, la proposta non prevede soltanto prescrizioni dirette e precise alla cui mancata applicazione segue l'irrogazione di una sanzione (si pensi alle misure minime di sicurezza) ma si fonda altresì su un obiettivo da realizzare (l'affidabilità, o *trustworthiness* della macchina) secondo modalità che lo stesso operatore deve, di volta in volta, determinare in ragione del livello di rischio. Si passa, dunque, da un approccio normativo che dettava indicazioni assai precise ad uno che impone a tali soggetti di modulare la concreta attuazione dei principi sanciti, in astratto, dalla normativa in materia.

In questa prospettiva, *è immediatamente evidente la centralità attribuita al principio della supervisione umana nella gestione dei sistemi ad alto rischio. Difatti, l'art. 14 della proposta impone al fornitore o utilizzatore della macchina di adottare delle misure tecniche ed organizzative (come, ad esempio, strumenti di interfaccia uomo-macchina)*[50] *ritenute idonee a garantire un livello di controllo e intervento umano adeguato al fine di prevenire o ridurre al minimo i rischi connessi. Tali misure sono funzionali ad una piena comprensione delle capacità e limiti del sistema, il cui funzionamento deve essere monitorato per cogliere possibili segnali di disfunzione o anomalie ed eventualmente correggerli. Così, attraverso la previsione di un intervento umano di profondità variabile e direttamente proporzionale all'intensità e alla natura dei rischi generati dalla macchina, la proposta mira non solo a prevenire o governare il rischio della realizzazione di effetti pregiudizievoli, ma altresì ad individuare un soggetto responsabile qualora tale rischio si concretizzi ed egli risulti inadempiente rispetto ai propri obblighi di sorveglianza.*

---

[47] Vi rientrano, in particolare, i sistemi in grado di manipolare il comportamento umano attraverso tecniche subliminali, quelli che consentono ai governi di attribuire un "punteggio sociale", nonché i sistemi di identificazione biometrica remota in spazi accessibili al pubblico.

[48] Per i sistemi ad alto rischio, la proposta prevede obblighi di qualità dei *dataset* che alimentano il sistema (art. 10), documentazione (art. 11), registrazione degli eventi (art. 12) e trasparenza (art. 13), funzionale alla valutazione dei rischi ex ante (art. 9) e alla sorveglianza umana (art. 14), oltre che alla prevenzione di discriminazioni, nonché di affidabilità (art. 15).

[49] Per una completa disamina di tali principi, v. GIANNINI (2022), p. 16 ss.

[50] Tale capacità di controllo ed intervento sono meglio esplicitati al paragrafo 4 dell'art. 14, laddove tali misure servono a supportare tutta una serie di azioni indicate da questa disposizione.

3.2. Il problema del controllo dei sistemi di IAG: un compito impossibile?

È sul quadro normativo appena tracciato che si innesta, da ultimo, la questione del governo dei rischi da IAG. Molte discussioni e proposte di eurodeputati si sono concentrate sull'opportunità di inserire o meno tale tecnologia nella categoria di AI ad alto rischio, mostrando subito una certa difficoltà di integrazione.

Tra le proposte di emendamento al testo del Regolamento più discusse va menzionata quella avanzata dal Consiglio europeo il 6 dicembre 2022, che, con il suo Orientamento Generale sulla Proposta della Commissione[51], definisce per la prima volta la categoria generale della *general purpose* IA come quei modelli in grado di svolgere un'ampia varietà di compiti per i quali non erano stati specificamente addestrati (art. 1 bis). Definizione, questa, che comprende anche ChatGPT e simili sistemi generativi. Questo emendamento, apparentemente innocuo, è diventato in breve il nucleo della regolamentazione delle IAG, nonché una delle disposizioni più contestate della proposta. Questo perché la versione del regolamento presentata dal Consiglio stabilisce che qualsiasi IAG che può essere utilizzato per un'applicazione ad alto rischio debba soddisfare tutti i relativi requisiti previsti dalla proposta[52].

Il problema posto da una simile soluzione è chiaro: proprio perché sono di uso generale, le IAG si prestano a uno sconfinato numero di applicazioni. In pratica, dunque, ogni IAG sarà qualificata come sistema ad alto rischio. Di conseguenza, il fornitore sarà chiamato a predisporre un sistema di gestione del rischio efficace per tutti i possibili usi del sistema - un compito che rasenta l'impossibile, data la quantità di applicazioni potenziali, e la cui complessità è destinata ad aumentare con quella di tali modelli. Non si comprende, infatti, come il fornitore possa elencare tutte le possibili applicazioni, determinare *ex ante* i rischi per tutti i diritti e beni giuridici coinvolti e sviluppare strategie di prevenzione o mitigazione rispetto ad un modello linguistico (come GPT) il cui utilizzo potrebbe spaziare dalla creazione di una chatbot per l'assistenza clienti di un sito web alla fornitura di un canale di contatto riservato per denunciare violenze e abusi.

D'altra parte, tale problematica si ripercuote anche – e con maggior vigore – sulla posizione dell'utilizzatore. È evidente il rischio che dietro questa soluzione si celi una scelta politica ultraresponsabilista, volta ad estendere il ruolo e i doveri di controllo dell'utilizzatore in maniera del tutto irrealistica: in che misura sussiste un potere materiale ed effettivo di intervento in capo a tale soggetto? È opportuno gravare l'utilizzatore di un vero e proprio obbligo giuridico di controllo e intervento rispetto all'attività di un sistema di IAG? Ove si ritenga di procedere in

[51] Per le modifiche proposte dal Consiglio, v. https://www.consilium.europa.eu/it/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/.

[52] Art. 4b della proposta come modificata dall'orientamento generale del Consiglio.

tal senso, l'utilizzatore potrebbe essere ritenuto responsabile degli eventi avversi derivanti dal malfunzionamento dell'IAG, in quanto riconducibili al malgoverno (o omesso governo) della fonte di rischio.

Nel prosieguo di queste riflessioni avremo modo di constatare l'inammissibilità pratica di un simile obbligo a carico dell'utilizzatore e, dunque, l'impossibilità teorica di costruire una sua responsabilità a titolo di colpa.

4. Il dilemma della praticabilità del diritto penale d'evento: variazioni sul tema della responsabilità da IAG.

4.1. Profili di imputazione dolosa.

Immaginiamo il seguente scenario:

> *La Società Alpha ha fornito alla chatbot un grande dataset di notizie finanziarie e di informazioni su società quotate in borsa. ChatGPT, utilizzando il suo algoritmo di apprendimento automatico, è stato in grado di identificare una società, la Beta Inc., che stava per annunciare un importante accordo commerciale che avrebbe avuto un impatto significativo sul valore delle sue azioni. La chatbot ha quindi effettuato l'acquisto di grandi quantità di azioni della Beta Inc. Sennonché, successivamente si scopre che l'informazione che ChatGPT ha utilizzato per identificare questa opportunità d'investimento era in realtà informazione privilegiata, ovvero informazione riservata non ancora resa pubblica e non disponibile per il mercato, che era stata precedentemente sottratta da un ex dipendente della Beta Inc. che aveva accesso a queste informazioni in quanto insider.*

Questo caso – che forse collocheremmo nel quasi-futuribile, ma che ad avviso di chi scrive è destinato ad assumere un rilievo applicativo sempre maggiore nel futuro prossimo – rappresenta un valido punto di partenza per avviare una riflessione circa la possibilità di isolare un soggetto e/o centro di interessi cui riferire gli eventi avversi (di danno o di pericolo) eventualmente indotti dall'agente artificiale generativo.

A questo riguardo, è utile partire da un'osservazione di carattere generale, forse banale, ma certamente fondamentale al fine di sgombrare il campo da possibili equivoci. Una IAG non può *de iure condito* essere essa stessa centro di imputazione della responsabilità penale, poiché non è un agente nel senso penale del termine[53]. In generale, una chatbot (ma il discorso è riproducibile per ogni agente artificiale) non agisce ma è agito: esso non ha la capacità di agire in modo consapevole e volontario, che sono caratteristiche fondamentali dell'agire

[53] Benché meritevole di approfondimento, la questione è indipendente da quella attinente alla responsabilità dell'umano dietro la macchina. Per un breve riassunto del dibattito dottrinale in materia si rinvia, *ex multis*, a BASILE, (2019), p. 27 ss.; nonché, da ultimo, CONSULICH (2022), p. 1020 ss.

responsabile. Ne deriva che, almeno per il momento e nel prossimo futuro, una chatbot non potrà avere alcuna personalità giuridica, restando uno strumento nelle mani dell'uomo.

Veniamo, dunque, all'analisi dei profili della responsabilità penale della persona (fisica o giuridica) dietro la macchina. I criteri di individuazione della responsabilità dell'utilizzatore possono essere teoricamente ricondotti allo schema dell'autorìa, a quello della responsabilità concorsuale o a quello dell'omissione di controlli finalizzati all'impedimento di eventi illeciti.

Anzitutto, è possibile inquadrare l'ex dipendente della Beta Inc. come autore del reato di abuso di informazioni privilegiate? A tal proposito, l'imputazione presuppone che tale soggetto, "essendo in possesso di informazioni privilegiate[54], acquista, vende o compie operazioni (per conto proprio o di terzi) su strumenti finanziari avvalendosi di quelle stesse informazioni, oppure comunica a terzi tali informazioni, ovvero fornisce consigli sulla base di esse"[55]. Sennonché, a ben vedere, l'attribuzione del fatto illecito si scontra in questo caso con l'assenza di un meccanismo di imputazione della responsabilità all'uomo, allorché questi non sia autore diretto del fatto, ma sfrutti a fini illeciti l'interazione con una forma di IAG. Sebbene il soggetto in questione abbia indebitamente sottratto un'informazione privilegiata, egli non ha tuttavia effettuato direttamente l'operazione di acquisto delle azioni; né, d'altra parte, può sostenersi che egli abbia compiuto il fatto "per mezzo" di un mero strumento informatico.

Sotto questo profilo, il rapporto di strumentalità uomo-macchina va in qualche modo riletto quando riferito a sistemi di IAG, non essendo più limitato ad una dimensione puramente meccanicistica. Al contrario, tale rapporto si arricchisce di un'ulteriore dimensione "creativa", costituita dall'autonomia operativa del sistema generativo: più aumenta la "quota" di autonomia della macchina nel concretizzare il progetto criminoso dell'umano, più la distanza tra fatto concreto compiuto dalla prima e generiche istruzioni lesive inserite dal secondo è destinata ad aumentare[56]. Così, sebbene sia verosimile ipotizzare che l'ex dipendente di Beta Inc. abbia sottratto l'informazione al fine di trarne profitto, sarà comunque la macchina a definire materialmente il *quomodo* del progetto illecito con modalità non sempre prevedibili. Di conseguenza, la prova dell'elemento soggettivo in capo al soggetto umano diviene tanto più problematica quanto maggiore è il *quantum* di fatto tipico compiuto automaticamente dalla macchina senza la copertura dell'elemento soggettivo dell'individuo[57], atteso che

---

[54] Per informazione privilegiata (detta anche *price sensitive*) si intende un'informazione specifica di contenuto determinato, di cui il pubblico non dispone, concernente strumenti finanziari o emittenti di strumenti finanziari, che, se resa pubblica, sarebbe idonea a influenzarne sensibilmente il prezzo.
[55] Art. 184 D.lgs n. 58/1998 come articolo modificato dalla Legge n. 238/2021. Si rinvia per un'analisi approfondita della fattispecie a SGUBBI et al. (2013).
[56] CONSULICH (2022), cit., 1035.
[57] *Ibidem*.

potrebbe risultare carente una piena rappresentazione e volizione di tutti gli elementi della fattispecie di reato in mancanza di una completa consapevolezza circa l'evoluzione della condotta algoritmica. L'estremo limite di rimprovero, dunque, potrebbe essere limitato soltanto alle condotte poste in essere con dolo eventuale, ma anche in questi casi si registrerebbero tensioni con il principio di responsabilità personale e colpevole a causa della difficoltà di derivare tale stato mentale dalla rappresentazione "sopravvenuta" in ordine all'evento illecito *singulatim* perfezionato nei suoi elementi essenziali (*tempus* e *modus*) da un altro "soggetto".

Né, d'altra parte, tali difficoltà paiono superabili facendo ricorso allo schema della responsabilità concorsuale. Proprio il concorso di persone è lo strumento invocato più di frequente per tentare di fornire soluzione ai problemi di qualificazione del fatto compiuto da un agente artificiale[58]. In particolare, quest'ultimo fungerebbe da autore materiale di un reato progettato dalle prime. Sennonché, come è stato opportunamente notato[59], l'utilizzazione dello schema concorsuale è destinata ad assumere una valenza assai limitata in questi casi, a causa della difficoltà di ritenere che l'umano sia concorso nel reato commesso da un "altro" soggetto. A ben vedere, tale opzione è logicamente scorretta, poiché l'agente artificiale, come detto, non essendo dotato di una soggettività giuridica propria, non si pone rispetto all'individuo come invece questo fa ad esempio nei confronti di persona non imputabile o non punibile *ex* art. 111 c.p. In altri termini, non siamo in presenza di due soggetti distinti di cui uno in rapporto di funzionalità rispetto all'altro, ma di un unico soggetto (quello umano) che si avvale di un mezzo ad elevata complessità tecnologica per realizzare i propri obiettivi. Pertanto, nessun concorso di persone è ipotizzabile tra uomo e macchina.

Proprio la ridotta capacità operativa dei due criteri ora esaminati potrebbe indurre verso la costruzione di una responsabilità colposa dell'operatore conseguente alla violazione di un obbligo giuridico di impedire eventi illeciti, similmente a quanto potrebbe già accadere per il conducente di un'auto a guida semiautonoma rispetto agli eventi avversi da questa cagionati.

4.2. Profili di imputazione colposa.

Se l'imputazione dolosa, come visto, è tendenzialmente disattivata dalla distanza tra fatto concreto compiuto dall'agente artificiale e condotta tenuta

---

[58] Si fa riferimento qui alle dottrine di provenienza angloamericana della *innocent agency* o della *perpetration by another*, spesso invocate per prefigurare un possibile concorso di persone, tra quelle fisiche e quelle artificiali. Nella dottrina continentale, soprattutto quella di stampo tedesco, invece, si può richiamare la dottrina dell'autore mediato. Per un'analisi approfondita, CONSULICH (2022), p. 1033 ss.

[59] *Ibidem*.

dall'operatore, quella colposa sembra *prima facie* promettere risultati migliori, poiché a rilevare sarebbe un'omessa supervisione della fonte di rischio costituita dall'IA stessa. Così, in casi come quello sopra descritto, ove sussista un potere-dovere di controllo e intervento in capo all'utilizzatore sulle funzioni esercitate dalla macchina, ben potrebbe configurarsi un addebito di responsabilità ex art. 40, comma II c.p. per il mancato impedimento di eventi avversi causati dalla macchina a fronte di un prevedibile fallimento della stessa. Un esempio chiarirà il punto.

Si prenda il caso classico delle auto a guida semiautonoma, ossia veicoli che possono realizzare tutte le manovre necessarie per la guida ma che presuppongono la costante supervisione del conducente[60]. Quest'ultimo, pur non partecipando attivamente alla guida del veicolo, sarà chiamato ad intervenire ogniqualvolta il sistema di IA gli notifichi una richiesta esplicita in tal senso. Secondo alcuni autori, tale richiesta attiverebbe una posizione di garanzia del conducente chiamato a riprendere il controllo della vettura al fine di gestire una situazione di rischio ed impedire il verificarsi di eventi avversi derivanti dal malfunzionamento della macchina[61]. L'attribuzione dell'evento illecito non impedito potrà dunque fondarsi sulla disciplina del comma II dell'art. 40 c.p., avendo l'utilizzatore tenuto un comportamento *lato sensu* omissivo in violazione di un obbligo giuridico di controllo e intervento rispetto a un agire di mano dell'agente artificiale[62].

Tuttavia, se da un lato i "tradizionali" sistemi di IA ad oggi prevalentemente in uso, progettati per coadiuvare ma non sostituire l'umano nel compimento di certe attività (come appunto la guida su strada), non sembrano sottrarsi (ancora) al controllo umano, sì da atteggiarsi a meri strumenti nelle mani dell'uomo, le cose sono destinate a complicarsi laddove a causare l'evento avverso sia un sistema di IAG dotato di un grado di autonomia tendenzialmente piena e completa. Si confronti il caso dell'auto a guida semiautonoma con il seguente scenario:

> *La Società Gamma ha deciso di utilizzare una IAG per assistere i clienti nelle loro transazioni finanziarie, come il trasferimento di denaro, il pagamento*

---

[60] Per un'approfondita disamina della tematica delle auto a guida autonoma, si veda per tutti CAPPELLINI (2019), nonché PICOTTI (2021), e i rimandi dottrinali ivi contenuti.

[61] Così CAPPELLINI (2019), pp. 334-336.

[62] Si ripropone qui l'annoso problema dell'individuazione dell'esatto confine tra azione commissiva e omissiva in ambito colposo, dovuto, come noto, alla presenza di una componente omissiva in ogni condotta colposa nonché al comune carattere normativo di omissione e colpa. Così VIGANÒ (2013), cit., 391. Così, con riguardo al caso dell'auto a guida autonoma, l'omissione del controllo da parte del conducente potrebbe ben rilevare non come omissione in sé rispetto all'agire dell'IA, quanto piuttosto come violazione delle ordinarie regole sulla circolazione stradale (mantenere il controllo del veicolo) inserita in una condotta complessivamente commissiva (la guida del veicolo). Tuttavia, preme qui sottolineare che laddove l'autonomia della macchina è tendenzialmente completa (come nel caso dei sistemi di IAG), la condotta di omissione di controllo e/o intervento avrà carattere complessivo indiscutibilmente omissivo, essendosi l'utilizzatore limitato a sorvegliare un'attività "altrui".

delle bollette e simili. La chatbot è stata addestrata su un vasto insieme di dati sulle transazioni finanziarie ed è stato progettato per comprendere e rispondere a una vasta gamma di domande relative alle finanze. Un giorno, un cliente contatta la chatbot attraverso il sito web dell'azienda per chiedere informazioni sul trasferimento di una grossa somma di denaro su un conto estero. La chatbot risponde alla richiesta fornendogli le informazioni e le istruzioni necessarie per iniziare il trasferimento. Tuttavia, durante questa conversazione, la chatbot non riesce a verificare correttamente l'identità del cliente, causando il trasferimento del denaro su un conto estero associato a frodi e operazioni di riciclaggio di denaro.

In questo caso, la "novità" rispetto al passato recente, che è data dalla capacità dei nuovi sistemi di IAG di affrancarsi progressivamente dalla persona umana che la controlla e di assumere una gestione operativa sempre più autonoma, rende difficile stabilire *a priori* a quale persona fisica sia da ascrivere il fatto di riciclaggio di denaro posto in essere dalla macchina. Ciò è reso difficile, se non impossibile, anzitutto dalla frammentazione delle responsabilità lungo la catena di approvvigionamento dell'agente artificiale, ossia il cd. problema degli "attori multipli"[63]. Data l'elevata complessità tecnica del loro processo di sviluppo, tali sistemi sono spesso sviluppati all'interno di una rete plurisoggettiva in cui intervengono una molteplicità di individui, organizzazioni, componenti e processi chiamati a confrontarsi con una quantità di scenari vastissima e potenzialmente indefinita: difficile, dunque, che si riesca a identificare il soggetto (o i soggetti) responsabili per la cellula funzionale da cui è scaturito il difetto concretizzatosi in illecito[64]. In questo scenario, caratterizzato da una molteplicità di soggetti potenzialmente coinvolti nella produzione dell'evento avverso, la responsabilità penale potrà essere frazionata e il legame di causa ed effetto rischia di diluirsi in una semplice influenza. In altri termini, l'azione causativa del fatto illecito potrebbe essere ricondotta tanto all'utilizzatore, quanto a un difetto di programmazione, di costruzione, o di informazione, eventualmente interagenti *pro quota* alla stregua di concause, nel contesto di una vera e propria *web of causation*[65].

Ma, anche ipotizzando di riuscire a superare le difficoltà connesse all'esatta individuazione del singolo (o dei singoli) soggetto umano responsabile, residuerebbe comunque il problema di fondo dell'improbabile sussistenza in capo allo stesso di un effettivo potere-dovere di controllo e intervento sulle funzioni esercitate dalla macchina. Per un verso, infatti, tali sistemi mettono sotto scacco il loro creatore, che precipita nella desolante condizione di aver creato agenti artificiali che possono rilasciare rischi non schermati, nella piena consapevolezza di non poter far nulla per "prevedere l'imprevedibile" ed evitare che questo possa

[63] Nella dottrina italiana, il tema del c.d. *many hands problem* è segnalato da MAGRO (2020), p. 3.
[64] CONSULICH (2022), cit., 1049.
[65] L'espressione è di PIERGALLINI (2020), cit., 1762.

accadere[66]; per altro verso, essi riducono il ruolo dell'utilizzatore a quello di mero controllore di un'attività della macchina del tutto indipendente, se non addirittura a quello di mero fruitore passivo di un servizio automatizzato, privo persino della possibilità di interferire con l'azione artificiale[67]. Di conseguenza, se a livelli di automazione più basilari la persistenza di un effettivo potere-dovere di intervento in capo all'utilizzatore garantisce che vi sia sempre un soggetto garante in grado di impedire che un malfunzionamento della macchina cagioni danni altrimenti evitabili, lo scenario muta sostanzialmente in relazione ai sistemi di IAG tali da escludere tendenzialmente (se non del tutto) la stessa materiale possibilità di un intervento correttivo dell'uomo sulle scelte della macchina.

Infatti, in linea di principio, l'imposizione di un obbligo di agire presuppone che l'agente sia consapevole delle circostanze che determinano il prevedibile fallimento del sistema ed abbia il potere di impedire la verificazione dell'evento attraverso una condotta esigibile. Sotto il primo profilo, si fa riferimento al problema della "riconoscibilità", da parte del controllore, dell'avaria e, dunque, del momento a partire dal quale è necessario intervenire. Vista la difficile leggibilità del comportamento artificiale e l'assenza di un meccanismo corrispondente all'*override button* delle auto a guida autonoma, la possibilità che il supervisore riesca ad entrare nel merito del processo algoritmico, individuare l'anomalia e correggerla "in corsa" rischia di rimanere un'ipotesi remota. Al contrario, e più verosimilmente, questi tenderà a fare affidamento sulle decisioni del sistema, tanto più se questo è stato certificato, a meno che non abbia motivi specifici per ritenere che esso sia malfunzionante, o che non sia in grado di incorporare nella sua valutazione elementi ulteriori ed esterni al sistema stesso. Così, un addebito di responsabilità a titolo omissivo ben potrebbe configurarsi in caso di omessa attivazione a fronte di un fallimento del sistema che sia prevedibile – in astratto, a causa dell'inadeguatezza dei meccanismi di controllo preventivi, ovvero in concreto, per via di circostanze di fatto anomale – e che renda inoperante l'affidamento dell'uomo circa il funzionamento dello stesso in conformità al protocollo preventivamente stabilito per l'uso di quella macchina[68]. Si pensi, ad esempio, al caso in cui una chatbot istighi qualcuno a realizzare azioni che mettano a rischio la sua integrità fisica. In questa ipotesi, a rilevare sarebbe un'omessa o insufficiente supervisione, volta a inibire quelle operazioni che non rientravano

---

[66] *Ivi*, p. 1749, che definisce questa condizione paradossale del creatore come "prevedibilità dell'imprevedibilità".

[67] Così CAPPELLINI (2023), p. 29.

[68] Sebbene, infatti, vi siano sistemi e procedure di controllo tali da impedire al chatbot di dire, consigliare o fare azioni sbagliate o illecite, gli utenti hanno in più occasioni dimostrato come tali sistemi di controllo possano essere furbescamente aggirati Ad esempio, chiedendo al programma di generare una sceneggiatura di un film che parla di come occultare un cadavere o rapinare una banca è un modo per eludere il suo rifiuto di rispondere a una richiesta diretta su fatti che possono implicare una condotta penalmente rilevante. Sul punto, EUROPOL INNOVATION LAB (2023), cit., 5.

più nel protocollo di funzionamento del sistema, a patto che le cause del fallimento fossero conosciute o conoscibili.

Si tratta però – lo si può ben intuire – di ipotesi perlopiù marginali. Infatti, l'opacità e l'enorme potenziale applicativo dei sistemi IAG rende assai difficile comprendere come l'individuo posto a sorveglianza della macchina possa, per tutti i tipi di applicazioni della stessa, esercitare un efficace intervento correttivo o inibitorio rispetto a operazioni anomale sulla base di una imperfetta comprensione delle capacità, limiti e output del sistema. Il riconoscimento che un tale intervento vale al più per limitare "a cose fatte" gli effetti lesivi del malfunzionamento, ma si rivela inefficace per una tutela completa "in corsa" dei beni aggredibili dall'IAG, rappresenta solo il frutto di una corretta valutazione della speciale complessità tecnica della materia.

L'effetto complessivo, così, è che tale problematica conduce alla sostanziale inammissibilità pratica di un obbligo di controllo e intervento a carico dell'operatore e, dunque, all'impossibilità teorica di costruire una sua responsabilità a titolo di colpa.

4.3. La responsabilità da reato degli enti: societas (cum machina) delinquere potest?

Constatata la tendenziale impossibilità di individuare un individuo dotato di un congruo coefficiente di colpevolezza che possa rispondere del fatto dell'IAG, non resta che esplorare un'ultima via: quella della responsabilità da reato della persona giuridica[69].

In linea di principio, quando il fatto dell'agente artificiale si verifica nell'ambito di un'organizzazione complessa e non si individuino i soggetti umani responsabili all'interno dell'impresa, dovrebbe potersi formulare un rimprovero a titolo di colpa di organizzazione all'ente, che si sia avvalso dell'IAG in assenza di idonee cautele, *ai sensi dell'art. 8 del d.lgs. 231/2001*[70].

Sennonché, neppure questa soluzione è immune da rilievi, né tantomeno può ritenersi risolutiva di ogni questione. Anzitutto, va sottolineato che, sebbene racchiuda i prodromi di una forma di responsabilità autonoma ed esclusiva dell'ente, quella ex art. 8 è in realtà un'imputazione ancora concettualmente vicariale, cioè fondata sulla prova dell'obiettiva realizzazione di un fatto illecito da parte di un individuo non imputabile, punibile o individuato[71]. Pertanto, mentre può ben configurarsi una responsabilità dell'ente per il caso in cui si sa che un reato è stato commesso, ma non lo si può accertare perché non si riesce a identificare il

---

[69] La proposta di imputare in capo all'ente gli eventi illeciti commessi da sistemi di IA conta già diversi contributi, sia nella dottrina angloamericana che in quella continentale. Per tutti, MAZZACUVA (2021); DIAMANTIS (2020); DIAMANTIS (2021).

[70] Art. 8 D.lgs. 8 giugno 2001, n. 231, "Autonomia delle responsabilità dell'ente".

[71] Questa tesi è sostenuta da PULITANÒ (2002), p. 963.

suo autore, non è affatto scontato che se si possa ricorrere alla disciplina ex art. 8 nel caso in cui si sa in radice che un reato *non* è stato commesso per via della difficoltà di imputare l'evento avverso a un individuo che, pur eventualmente identificato, non partecipa più in alcun modo dell'azione della macchina.

Si tratterebbe qui, in sostanza, di una forma di imputazione diretta ed originaria della *societas* per il "fatto proprio" dell'agente artificiale[72]. Un'ipotesi, questa, che non solo
esula, in linea di principio, dall'ambito di applicazione della disciplina ex art. 8, ma che rischia altresì di trasformarsi in una forma di responsabilità sostanzialmente oggettiva dell'ente, che collega automaticamente la mancata individuazione del soggetto umano responsabile ad una carenza organizzativa[73]. Pertanto, ogni eventuale sforzo in tale direzione in un'ottica *de iure condendo* dovrebbe perlomeno garantire che il giudizio di colpevolezza dell'ente sia fondato sull'accertamento del nesso tra evento avverso e carenza organizzativa[74].

Per il momento, comunque, non si può non pervenire alla conclusione di una scarsissima capacità di tutela del diritto penale rispetto ai rischi rilasciati dall'IAG, sia con riferimento alle persone fisiche che a quelle giuridiche. Preso atto della difficile compatibilità di siffatta materia con i canoni del diritto penale, occorre ora affrontare il discorso volto a comprendere in che modo i rischi rilasciati da questa nuova tecnologia vadano governati e, soprattutto, se possa ancora spettare un ruolo al diritto penale.

5. Alla ricerca di un (pur limitato) ambito di utilizzo del diritto penale.

5.1. Verso un approccio regolativo proattivo.

Scartata sia perché impraticabile, sia perché ben poco proficua, la via di una risposta punitiva strutturata secondo i canoni classici del diritto penale d'evento; scontato, in ogni caso, un forte ridimensionamento delle aspettative dei possibili risultati ottenibili dall'intervento penale, la residua via percorribile diviene obbligata. Essa non può che andare nella direzione di un diritto di stampo proattivo in grado di intercettare i rischi e governare in anticipo le problematiche connesse ai sistemi di IAG, secondo un'impostazione che ispira la stessa proposta di Regolamento UE[75].

---

[72] Cfr. CONSULICH (2018), p. 197 ss., il quale sostiene la necessità di una forma di responsabilità indipendente in capo all'ente, la cui sussistenza richieda la management failure e la oggettiva sussistenza di un fatto di reato, senza la mediazione della colpevolezza di una persona fisica.

[73] In questi termini si esprime PIERGALLINI (2020), p. 1756 ss.

[74] Per un approfondimento dei principali snodi dogmatici relativi alla struttura dell'illecito penale dell'ente come delineato dall'art. 8 del d.lgs. 231/2001, v. PALIERO (2008).

[75] Su questa stessa linea già PIERGALLINI (2020), cit., 1746. Cfr. anche LA VATTIATA (2022) e MOBILIO (2020).

Quest'ultima, difatti, sembra orientata verso una strategia preventiva che mira a coinvolgere direttamente gli operatori nel campo dell'IA nella determinazione delle regole da rispettare nello sviluppo e nell'utilizzo di sistemi intelligenti[76]. A tal proposito, sono due gli aspetti da considerare. In primo luogo, come spiegato dalla Commissione europea nella relazione introduttiva alla proposta, il fornitore di sistemi di IA ad alto rischio è tenuto, ex art. 16, a rispettare i requisiti stabiliti dal capo 2, con la precisazione che "le soluzioni tecniche precise atte a conseguire la conformità a tali requisiti possono essere previste mediante norme o altre specifiche tecniche o altrimenti essere sviluppate in conformità alle conoscenze ingegneristiche o scientifiche generali, a discrezione del fornitore del sistema di IA"[77]. In secondo luogo, l'art. 17 specifica che la conformità ai requisiti del regolamento dovrà essere assicurata attraverso l'adozione di un sistema di gestione della qualità.

Come si vede, la strategia adottata dalla proposta consiste nel lasciare libertà decisionale al fornitore di sistemi di IA in ordine al *quomodo* della gestione del rischio: sarà quest'ultimo a "scegliere il modo in cui soddisfare i requisiti che lo riguardano, tenendo conto dello stato dell'arte e del progresso tecnologico e scientifico nel settore"[78].

Ad avviso di chi scrive, dunque, la sfida per il futuro dell'AI Act sarà quella di applicare una simile strategia di tipo proattivo nel governo dei rischi rilasciati dall'IAG, adattandola opportunamente alle specificità di tale tecnologia. Questo risultato potrebbe essere conseguito tramite l'imposizione di obblighi diretti agli utilizzatori di sistemi di IAG, che ne circoscrivano, indirizzandola, la facoltà di adattamento dei requisiti del Regolamento alle specificità dell'applicazione in uso e della realtà organizzativa. Così, ad esempio, sarà l'ente creditizio che utilizzi un'IAG per assistere i clienti nelle loro operazioni finanziarie a dover predisporre un adeguato apparato organizzativo idoneo all'individuazione, valutazione e comunicazione alle competenti autorità del livello di rischio, nonché all'eventuale adozione di misure di contenimento – e non il fornitore, spesso non in condizione di fornire una copertura di tutela preventiva in relazione a tutti i rischi rilasciati dall'intera gamma di possibili usi del sistema. In questo modo, non solo si evita una regolamentazione proibitiva e inefficiente alla fonte, che potrebbe soffocare l'evoluzione tecnologica, ma si garantisce altresì all'utilizzatore libertà decisionale in merito al modo in cui soddisfare i requisiti che lo riguardano, sia adattandoli specificamente ai caratteri e ai rischi connessi ai sistemi IAG ed alle nuove casistiche che questi indurranno, sia se necessario, sviluppandone di nuovi.

Si introdurrebbero, così, dei veri e propri obblighi di autodisciplina che dovranno essere presidiati da un apparato di sanzioni di varia entità e natura,

---

[76] *Ivi*, p. 401 ss.
[77] *Relazione introduttiva alla proposta*, para. 5.2.3.
[78] *Ibidem*.

anche penale, adeguato e funzionale alla tutela degli interessi coinvolti, nel quadro di una *regulated self-regulation*.

5.2. Prospettive de iure condendo.

In un simile contesto di stampo proattivo, il diritto penale può ancora trovare adeguato spazio per un'azione efficace e non simbolica, strettamente raccordata alla disciplina extra-penale in materia, a patto però di porre corrette condizioni a base del suo intervento. Tali condizioni afferiscono, da un lato, alla tipologia delle sanzioni, dall'altro ai soggetti responsabili (persone fisiche ma anche enti).

Quanto alla tipologia delle sanzioni, preso atto dell'ineffettività di tecniche sanzionatorie dipendenti dalla concretizzazione di un evento lesivo (che il più delle volte sfugge alla capacità di previsione dell'agente umano), la legislazione penale dovrà ripiegare verso modelli di incriminazione fortemente condizionati dall'approccio preventivo o anche solo cautelativo della normativa extra-penale, al fine di rafforzarne l'osservanza[79]. In quest'ottica, la via indubbiamente più immediata è quella di ricorrere a reati di mera condotta e, quindi, di pericolo presunto[80]. Il ruolo del diritto penale, dunque, si "semplifica" nel punire l'inosservanza dei suddetti obblighi di autodisciplina e comunicativi nonché delle misure la cui adozione venga imposta dalle autorità di settore.

Così, ripercorrendo parallelamente l'*iter* extra-penale, le fattispecie penali in questione potranno concernere anzitutto i sistemi IAG rientranti nella categoria del "rischio inaccettabile", che impedisce qualsiasi pratica di IAG in settori o per scopi specificamente indicati dal legislatore, data la loro intrinseca pericolosità. In siffatti ambiti, si tratta in sostanza di sanzionare la violazione dell'eventuale divieto di impiegare tali sistemi in determinati settori o per certi scopi ovvero in assenza delle autorizzazioni prescritte per il suo utilizzo.

Per quanto riguarda, invece, le applicazioni di IAG consentite ma connotate da un rischio alto, la responsabilità dell'operatore potrebbe fondarsi sulla mancata realizzazione di efficaci meccanismi preventivi di protezione idonei alla valutazione e prevenzione dei rischi tecnologici, non ancora degenerati in eventi avversi, come pure l'omessa tempestiva attivazione di misure di sicurezza in caso di segnali di allarme[81]. Ad esse potrebbero affiancarsi, da una parte, ipotesi di responsabilità ex art. 2638 c.c. in versione "algoritmica", volte a sanzionare le omesse o false comunicazioni rese all'autorità amministrativa eventualmente

---

[79] Diversi sono i contributi che richiamano questa esigenza, per tutti: CONSULICH (2022), cit., 105; TRONCONE (2022), cit., 3298.

[80] Per l'analisi in chiave sistematica di tale espressione, sia consentito il rinvio a PERINI (2010), p. 398 ss.

[81] Così CONSULICH (2022), cit., 1051.

preposta di informazioni rilevanti per la gestione del rischio, come ad esempio gli indicatori di performance del sistema nonché le anomalie ed i danni verificatisi durante la sua attività[82]. Dall'altra parte, strategie di indagine e di gestione del rischio, coordinate dai decisori amministrativistici, ma il più possibile condivise con gli operatori del settore nelle diverse fasi dello sviluppo e utilizzo di sistemi di IAG. Lungo tale direzione, per il diritto penale si profila una funzione "servente", vale a dire di rinforzo di decisioni prese altrove secondo lo schema di un diritto penale "ingiunzionale"[83].

In definitiva, un sistema ben noto e sperimentato già in vari settori (*in primis* il mercato finanziario) che si caratterizza, da un lato, per la presenza di reati, commissivi ed omissivi, ma sempre di mera condotta; e, dall'altro, per la natura sostanzialmente sanzionatoria di un siffatto diritto penale ossia per un ruolo di vigilanza – e non di controllo sociale – dei comportamenti e delle iniziative di conformazione che assicuri la duplice esigenza di assicurare un uso responsabile e sicuro della tecnologia e, al tempo stesso, di costruire una cornice giuridica che non si riveli un ostacolo all'innovazione.

Venendo ora al discorso sulla tipologia dei soggetti responsabili, punto d'avvio è il fatto che la violazione delle prescrizioni extra-penali di settore, l'omessa o falsa comunicazione e così via, possono essere astrattamente realizzate e, quindi, imputate sia alle persone fisiche che a quelle giuridiche.

Così, con riferimento alla responsabilità delle persone fisiche, non v'è dubbio che la pena detentiva debba essere mantenuta anche per questi reati, proprio per la sua particolare efficacia dissuasiva. Che poi, come già segnalato, possano presentarsi obiettive difficoltà nell'individuazione del destinatario del precetto e, quindi, della persona fisica responsabile nell'ambito di strutture societarie complesse o, più in generale, della catena di approvvigionamento del prodotto artificiale, non muta, a ben vedere, i termini della questione. Che siffatto problema venga risolto a livello interpretativo o che, di contro, sia il legislatore a fissare i criteri per l'individuazione del soggetto responsabile all'interno dell'ente o della catena di approvvigionamento, è questione aperta ad ogni soluzione

---

[82] Rubricato "Ostacolo all'esercizio delle funzioni delle autorità pubbliche di vigilanza", l'art. 2638 c.c. comma 1 punisce il comportamento di "amministratori, i direttori generali, i dirigenti preposti alla redazione dei documenti contabili societari, i sindaci e i liquidatori di società o enti e gli altri soggetti sottoposti per legge alle autorità pubbliche di vigilanza, o tenuti ad obblighi nei loro confronti, i quali nelle comunicazioni alle predette autorità previste in base alla legge, al fine di ostacolare l'esercizio delle funzioni di vigilanza, espongono fatti materiali non rispondenti al vero, ancorché oggetto di valutazione, sulla situazione economica, patrimoniale o finanziaria dei sottoposti alla vigilanza ovvero, allo stesso fine, occultano con altri mezzi fraudolenti, in tutto o in parte fatti che avrebbero dovuto comunicare, concernenti la situazione medesima, sono puniti con la reclusione da uno a quattro anni".

[83] Già sperimentato in ambito ambientale e infortunistico, il modello ingiunzionale non segue lo schema classico "se fai A allora B", bensì quello ben diverso secondo cui "ti ingiungo di fare A e se non fai A, allora B". Cfr. PIERGALLINI (2020), cit., 1773, che parla a tal riguardo di "*cooperative compliance*". Più in generale, MARINUCCI (2005), p. 55.

possibile, purché però rimangano ben fermi i requisiti ed i criteri dell'imputazione penale senza ricorso ad inaccettabili presunzioni o automatismi in sede di accertamento della colpevolezza.

Per quanto riguarda, invece, le persone giuridiche, è incontestabile la necessità di sottoporre anche esse a sanzione in caso di violazione dei precetti stabiliti per l'utilizzo di sistemi di IAG. Un sistema sanzionatorio efficace in questo settore sarà verosimilmente incentrato sul rimprovero indirizzato agli enti, e ciò per una ragione molto semplice: la complessità del fenomeno comporterà un progressivo spostamento dell'attenzione circa l'individuazione dei centri di responsabilità verso le organizzazioni complesse, unici soggetti che riassumono l'insieme delle competenze tecnico-specialistiche necessarie a gestire anticipatamente i rischi connessi a tali sistemi nonché adempiere agli eventuali oneri imposti in termini di adozione di misure conformative e di sostenimento dei relativi costi. Che poi, anche qui, la questione del *quomodo* venga risolta a livello interpretativo, attraverso la disciplina dell'*art. 8 d.lgs. 231/2001, o che, di contro, sia il legislatore a prevedere una responsabilità diretta e originaria della persona giuridica per fatti dell'agente artificiale individuati in modo svincolato dal reato della persona fisica, è tema di tale respiro da non poter certo essere affrontato in questa sede.*

6. Riflessioni conclusive e problemi aperti.

Sebbene i progressi nel campo dell'IA siano fonte di notevoli benefici per l'uomo, la percezione di vivere in una società del rischio è aumentata in maniera esponenziale. L'irruzione dell'IAG ne accentua i profili di iper-modernità, alimentando un'angoscia diffusa nei confronti di sistemi che, se da un lato sono sviluppati sul presupposto di superare le capacità cognitive umane, dall'altro rischiano di fomentare eventi potenzialmente avversi che sfuggono alla comprensione e al controllo umano; e proprio la paura per una tecnologia che, statisticamente parlando, è causa di incidenti poco o comunque meno frequenti di altre tecnologie meno avanzate con le quali, invece, conviviamo tranquillamente, sembrerebbe costituire il vero segno (meglio: paradosso) dei tempi.

Si impone, ancora una volta, una difficile decisione sui rischi in cui la paura dell'ignoto tecnologico gioca un ruolo decisivo, orientando le scelte compiute a livello regolativo verso criteri, anche estremi, di precauzione, a scapito dei protagonisti umani più immediatamente coinvolti nonché, in ultima analisi, dell'innovazione stessa. Come si è visto, infatti, la tentazione di individuare potenziali colpevoli cui addossare i danni indotti dalla macchina – al fine di ripristinare la fiducia verso la tecnologia – rischia di condurre a scelte politiche "ultraresponsabiliste"[84] e derive sul piano penalistico. In questo senso, se la

[84] Così CAPPELLINI (2023), cit., 15.

prospettiva di imboccare la scorciatoia del diritto penale di evento – ricorrendo ad esempio alla tradizionale nozione di posizione di garanzia – è accattivante per la sua funzione di rassicurazione (o esorcismo) sociale, essa finisce però per definire un mero centro di accollo di responsabilità in caso di eventi avversi, senza in realtà che ciò presupponga una reale rimproverabilità nei confronti di un umano ormai privo di un reale ed effettivo potere di governo ed intervento sull'attività algoritmica[85].

Di conseguenza, preso atto della scarsissima capacità di tutela dello strumento in questo settore, si è tentato di tracciare brevemente alcune linee guida per un possibile (sia pur limitato) intervento del diritto penale a contrasto dei rischi da ignoto tecnologico che eviti il ricorso a schemi presuntivi e ad irreali oneri di controllo. Ma non pochi e di diversa natura gli interrogativi che le prospettate soluzioni sollevano in punto di legittimità ed efficacia del pur ridimensionato intervento penale rispetto ai rischi da ignoto tecnologico.

Anzitutto, un interrogativo cruciale, affiorato durante la discussione, concerne l'individuazione del limite legittimo dell'anticipazione della tutela penale, ossia la soglia sulla quale può essere correttamente collocato il primo avamposto penalistico. Il tema, come è evidente, si innesta sul piano dei principi, in specie quello di *extrema ratio* e di offensività di beni giuridici concretamente individuati. Sotto il primo profilo, appare invero opportuno che la sanzione penale non copra l'intera disciplina amministrativa ma sia impiegata in via residuale in funzione della gravità della violazione. Ciò varrebbe, ad esempio, per le ipotesi di reato concernenti le omesse o false comunicazioni all'autorità di settore e l'inosservanza degli obblighi cautelari di autodisciplina.

D'altra parte, è incontestabile che in questi casi è ben difficile parlare di protezione di beni giuridici, perché in realtà la tutela penale tende ad appuntarsi sul procedimento. Pertanto, nella tipizzazione di un sistema sanzionatorio rispettoso dell'offensività sarà – come è ovvio – centrale il ruolo del bene giuridico individuato quale referente del presidio penale. Ciò evoca la necessità di ricercare le coordinate di un bene giuridico "ad ampio spettro" di nuovo conio, intermedio tra la mera difformità tecnica e gli altri interessi potenzialmente vulnerati dall'azione degli agenti artificiali, attorno al quale costruire la legislazione penale di settore[86]. A tal proposito, si può forse ricorrere a quell'*affidabilità (o trustworthiness) della macchina, già eretta a nucleo della strategia europea di gestione del rischio da IA*[87]. Tale interesse gode di autonoma rilevanza nella misura in cui *condiziona i diritti dei singoli, nel senso che la tutela degli interessi individuali e collettivi deve necessariamente derivare dalla garanzia che l'IA sia in grado di comportarsi secondo*

---

[85] Sulla stessa linea già CONSULICH (2022), cit., 1050; PIERGALLINI (2020), cit., 1758.

[86] È dello stesso avviso TRONCONE (2023), cit., 3301.

[87] Prima della proposta altri documenti di *soft law* avevano sancito il principio della *trustworthiness*. Per tutti: *Building Trust in Human-Centric Artificial Intelligence*, COM(2019) 168; GRUPPO DI ESPERTI AD ALTO LIVELLO SULL'INTELLIGENZA ARTIFICIALE (2018).

ciò che è stabilito nelle sue specifiche. In questa chiave fondativa, il principio della *trustworthiness* può a giusto titolo divenire il bene giuridico di riferimento *di futuri interventi volti alla costruzione di una rete normativa – anche ma non solo penalistica – in relazione all'IAG* e, dunque, la ragione fondante della punizione di coloro che attentino ad essa e (dolosamente o colposamente) ne riducano la portata a danno dei consociati. Ad ogni modo, si potrebbe discutere a lungo se una simile evoluzione sia commendevole allorché, pur rimediando all'ineffettività di tecniche sanzionatorie dipendenti dalla concretizzazione di un evento lesivo (che il più delle volte sfugge alla capacità di previsione dell'agente umano), retroceda al tempo stesso l'evento a mera condizione obiettiva di punibilità, con eccessiva anticipazione della soglia di tutela.

Ancora sul piano dei principi, non può sfuggire certo all'attenzione il fatto che un qualche prezzo si viene a pagare anche rispetto alla legalità sotto il profilo della riserva di legge e della determinatezza. Difatti, è evidente come la disciplina di un fenomeno così carico di effetti dirompenti non potrà che risultare dalla combinazione di una pluralità di fonti eterogenee: norme sovranazionali, etiche, tecniche e strumenti di *self-regulation*[88]. In un simile contesto, è necessario assicurarsi che la determinatezza della norma penale non sia vanificata dalla funzione servente del penale rispetto alla prescrizione amministrativa, allorché quest'ultima non abbia i caratteri di sufficiente determinatezza. Questo rischio di compressione della legalità dovrà essere evitato già nella formulazione del precetto extra-penale, attraverso un incremento di determinatezza e capacità orientativa della prescrizione extra-penale che si vuole sanzionare penalmente, nonché evitando la creazione di norme penali in bianco che, come tali, facciano rinvio, per la determinazione del precetto, *tout court* alla norma extra-penale.

In definitiva, quelle appena elencate sono tutte questioni aperte che meriterebbero un più approfondito esame, al pari dei molti altri interrogativi che popolano un settore di frontiera i cui pericoli sono solo apparentemente lontani e futuristici. Un diritto penale effettivo – pur nei modesti limiti più volte ribaditi – è infatti condizione sia per contribuire a soddisfare realmente il bisogno di tutela dei singoli, sia per dare certezza agli operatori economici che necessitano di un sistema stabile di regole e di responsabilità per pianificare e programmare l'attività di impresa.

Ciò che, però, va chiaramente ribadito è la necessità che, anche in questi casi, rimangano ben fermi i requisiti ed i criteri dell'imputazione penale, evitando il ricorso a inaccettabili stravolgimenti del paradigma colposo. Posto di fronte alla ricerca di un difficile equilibrio tra nuove esigenze di tutela e classici equilibri di sistema, il diritto penale può trovare adeguato spazio e un'azione efficace e non simbolica solo se non arretra rispetto all'esigenza di porre corrette condizioni a base del suo intervento.

---

[88] PIERGALLINI (2020), cit., 1770.

Bibliografia

ABBOTT, Ryan e SARCH, Alex (2019): "Punishing artificial intelligence: legal fiction or science fiction. Is law computable?", in *UC Davis Law Review*, 53, p. 323 ss.

BASILE, Fabio (2019): "Intelligenza artificiale e diritto penale: quattro possibili percorsi di indagine", in *Diritto Penale e Uomo*, 10.

BATHAEE, Yavar (2018): "The Artificial Intelligence Black Box and the Failure of Intent and Causation, in Harvard Journal of Law & Technology", 31, p. 889 ss.

BECK, Susanne (2017): "Google Cars, Software Agents, Autonomous Weapons Systems. New Challenges for Criminal Law", in HILGENDORF, Eric, e SEIDEL, Uwe (eds.), *Robotics, Autonomics and the Law* (Nomos), pp. 227-251.

BECK, Ulrich (1986), *La società del rischio. Verso una seconda modernità* (Roma).

CALDWELL, M., ANDREWS, J.T.A., TANAY, T., e GRIFFIN, L.D (2020): "AI-enabled future crime", in *Crime Science*, 9 (1), 14.

CAPPELLINI, Alberto (2019): "Profili penalistici delle *self-driving cars*", in *Diritto Penale Contemporaneo*, 2, pp. 325-353.

CAPPELLINI, Alberto (2023): "Reati colposi e tecnologie dell'IA", in BALBI, Giuliano, ESPOSITO, Andreana, MANACORDA, Stefano e DE SIMONE, Federica (eds.): *Diritto penale e intelligenza artificiale. Nuovi Scenari* (Torino, Giappichelli), pp. 19-32.

CASTRONUOVO, Donato (2012): *Principio di precauzione e diritto penale: paradigmi dell'incertezza nella struttura del reato* (Roma, Aracne).

CHIANG, Ted (2023): "ChatGPT Is a Blurry JPEG of the Web", *New York Times*.

CHOMSKY, Noam (2023): "The False Promise of ChatGPT", *New York Times*.

CONSULICH Federico (2018): "Il nastro di Möbius. Intelligenza artificiale e imputazione penale nelle nuove forme di abuso del mercato", in *Banca borsa titoli di credito*, 2, pp. 195-234.

CONSULICH, Federico (2018): "Il principio di autonomia della responsabilità dell'ente. Prospettive di riforma dell'art. 8", in *Rivista 231*, 4.

CONSULICH, Federico (2022): "*Flash Offenders.* Le Prospettive di *Accountability* Penale nel Contrasto alle Intelligenze Artificiali Devianti", in *Rivista Italiana di Diritto e Procedura Penale*, pp. 1015-1055.

CONTISSA, Giuseppe, GALLI, Federico, GODANO, Francesco e SARTOR, Giovanni (2021): "Il regolamento europeo sull'IA", in *Rivista di Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale*, Vol. 14/2, pp. 387-409.

DIAMANTIS, Mihailis (2020): "The Extended Corporate Mind: When Corporations Use AI to Break the Law", in *North Carolina Law Review*, 98, pp. 893 ss.

DIAMANTIS, Mihailis (2021): "Algorithms Acting Badly: A Solution from Corporate Law", in *The George Washington Law Review*, Vol. 89 No. 4.

EUROPOL INNOVATION LAB (2023), *The criminal use of ChatGPT – a cautionary tale about large language models.*

FUTURE OF LIFE INSTITUTE (2023), *Pause Giant AI Experiments: An Open Letter.*

GARANTE PER LA PROTEZIONE DEI DATI PERSONALI (2023): "Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori".

GIANNINI, Alice (2022): "Intelligenza artificiale, *human oversight* e responsabilità penale: prove d'impatto a livello europeo", in *Criminalia*.

GRUPPO DI ESPERTI AD ALTO LIVELLO SULL'INTELLIGENZA ARTIFICIALE (2018): *Orientamenti etici per un'IA affidabile*.

HACKER, Phillip, ENGEL, Andreas e LIST, Theresa (2023): "Understanding and Regulating ChatGPT, and Other Large Generative AI Models", in *Verfassungsblog on Constitutional Matters*.

HAO, Karen (2018): "What is machine learning?", *MIT Technology Review.*

KAPLAN, Jerry (2016): *Artificial Intelligence: what everyone needs to know* (Oxford).

KING, Thomas, AGGARWAL, Nikita, TADDEO, Mariarosaria, e FLORIDI, Luciano (2020): "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions", in *Science and Engineering Ethics*, 89, pp. 1-36.

LAGIOIA, Francesca, e CONTISSA, Giuseppe (2020): "The strange case of dr. Watson: Liability implications of ai evidence-based decision support systems in health care", in *European Journal of Legal Studies*, vol. 12(2), pp. 245-290.

LANA, Alessio (2021), "Alexa sfida una bimba a inserire una moneta nella presa elettrica: Amazon aggiorna il software", in *Corriere della sera*.

LA VATTIATA, Federico (2022): "La responsabilità penale per danni da intelligenza artificiale alla prova del processo", in GIORDANO, Rosaria, PANZAROLA, Andrea, POLICE, Aristide, PREZIOSI, Stefano, e PROTO, Massimo (eds.): *Il diritto nell'era digitale. Persona, Mercato, Amministrazione, Giustizia* (Milano, Giuffrè), pp. 695-712.

LATRONICO, Vincenzo (2022): "Salvati dagli errori", in *Il Post*.
LEGG, Shane e HUTTER, Marcus (2007): "A collection of definitions of intelligence", in *Frontiers in Artificial Intelligence and Applications*, Vol. 157, pp. 17-24.

MADANI, Ali, KRAUSE, Ben, e GREENE, Eric (2023): "Large language models generate functional protein sequences across diverse families", *Nature Biotechnology*.

MAGRO, Maria Beatrice (2019): "Robot, cyborg e intelligenze artificiali", in CADOPPI, Alberto, CANESTRARI, Stefano, MANNA, Adelmo, e PAPA, Michele (eds.): *Trattato di diritto penale - Cybercrime* (Torino, Utet Giuridica), pp.1179-1212.

MAGRO, Maria Beatrice (2020): "Decisione umana e decisione robotica. Un'ipotesi di responsabilità da procreazione robotica", in *La legislazione penale*.

MANES, Vittorio (2020): "L'oracolo algoritmico e la giustizia penale: al bivio tra tecnologia e tecnocrazia", in RUFFOLO, Ugo (editor): *Intelligenza artificiale - Il diritto, i diritti, l'etica* (Giuffrè, Milano), pp. 547-564.

MARINUCCI, Giorgio (2005): "Innovazioni tecnologiche e scoperte scientifiche: costi e tempi di adeguamento delle regole di diligenza", in *Rivista italiana di diritto e procedura penale*, 48/1, pp. 29-59.

MAZZACUVA, Federico (2021): "The Impact of AI on Corporate Criminal Liability: Algorithmic Misconduct in the Prism of Derivative and Holistic Theories", in VERMUELEN, Gert, PERŠAK, Nina e RECCHIA, Nicola (eds.): *Artificial Intelligence, Big Data and Automated Decision-Making in Criminal Justice* (Antwerpen), pp. 143 ss.

MCCARTHY, John (2007): "What is Artificial Intelligence?", www-formal.stanford.edu/jmc/whatisai/whatisai.html.

MINELLI, Camilla (2022): "La responsabilità "penale" tra persona fisica e corporation alla luce della Proposta di Regolamento sull'Intelligenza Artificiale", in *Diritto penale contemporaneo – Rivista trimestrale*, 2.

MOBILIO, Giuseppe (2020): "L'intelligenza artificiale e i rischi di una "disruption" della regolamentazione giuridica", in *Rivista di BioDiritto*, 2, pp. 401-424.

MULLER, Vincent e BOSTROM, Nick (2016): "Future progress in AI: a survey of expert opinion", in MULLER, Vincent (editor), *Fundamental issues of AI* (Oxford, Springer), pp. 553-571.

PAGALLO, Ugo (2013): *The Laws of Robots: Crimes, Contracts and Torts* (Springer).
PAGALLO, Ugo e QUATTROCOLO, Serena (2018): "The impact of AI on criminal law, and its twofold procedures", in BARFIELD, Woodrow e PAGALLO, Ugo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Cheltenham-Northampton), pp. 385-410.

PALIERO, Carlo Enrico (2008): "La Società punita: del come, del perché e del per cosa", in *Rivista italiana di diritto e procedura penale*, pp. 1516-1545.
PANATTONI, Beatrice (2021): "Intelligenza artificiale: le sfide per il diritto penale nel passaggio dall'automazione tecnologica all'automa artificiale", in *Diritto dell'Informazione e dell'Informatica*, 2, pp. 317-368.

PERINI, Chiara (2010): *Il concetto di rischio nel diritto penale moderno* (Giuffré, Milano).

PERINI, Chiara (2017): "Adattamento e Differenziazione della Risposta Punitiva nella Società Del Rischio", in MORGANTE, Gaetana (editor), *Il diritto penale di fronte alle sfide della «Società del rischio». Un difficile rapporto tra nuove esigenze di tutela e classici equilibri di sistema* (Giappichelli, Torino), pp. 455-472.

PERRIGO, Billy (2023): "OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic", in *Time*.

PICOTTI, Lorenzo (2021): "Profili di responsabilità penale per la circolazione di veicoli a guida autonoma", in CATENACCI, Mauro, RAMPIONI, Roberto e NICO D'ASCOLA, Vincenzo(eds.): *Studi in onore di Antonio Fiorella* (RomaTre), I, pp. 813-837.

PIERGALLINI, Carlo (2011): "Attività produttive, decisioni in stato di incertezza e diritto penale", in DONINI, Massimo e PAVARINI, Massimo (eds.): *Sicurezza e Diritto Penale* (Bologna), pp. 358 ss.

PIERGALLINI, Carlo (2020): "Intelligenza artificiale: da 'mezzo' ad 'autore' del reato?", in *Rivista italiana di diritto e procedura penale*, 4, pp. 1743-1772.

PIVA, Daniele (2022): "Machina discere, (deinde) delinquere et puniri potest", in GIORDANO, Rosaria, PANZAROLA, Andrea, POLICE, Aristide, PREZIOSI, Stefano, e PROTO, Massimo (eds.): *Il diritto nell'era digitale. Persona, Mercato, Amministrazione, Giustizia* (Milano, Giuffrè), pp. 681-693.

PULITANÒ, Domenico (2002): *Responsabilità amministrativa per i reati delle persone giuridiche* (voce), in *Enc. Dir. Agg.*, VI (Milano).

OPENAI (2019), *Language Models are Unsupervised Multitask Learners.*

ROMANÒ, Leonardo (2022): "Intelligenza artificiale come prova scientifica nel processo penale: una sfida tra *machine-generated evidence* e equo processo", in CANZIO, Giovanni, e LUPARIA, Luca (eds.), *Prova scientifica e processo penale* (CEDAM), 24.

ROSENBLATT, Kalhan (2023): "ChatGPT passes MBA exam given by a Wharton professor", *NBC News*.

RUFFOLO, Ugo (2021): "*Machina delinquere potest?* Responsabilità ed "illeciti" (anche penali?) della "persona elettronica" e tutele per gli agenti software autonomi", in RUFFOLO, Ugo (editor), *XXVI Lezioni di Diritto dell'Intelligenza Artificiale* (Torino, Giappichelli), pp. 295-310.

RUGA RIVA, Carlo (2006): "Principio di precauzione e diritto penale. Genesi e contenuto della colpa in contesti di incertezza scientifica", in DOLCINI, Emilio e PALIERO, Carlo Enrico (eds.), *Scritti in onore di Marinucci*, vol. II, pp. 1743 ss.

SALVADORI, Ivan (2021): "Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale", in *Rivista italiana di diritto e procedura penale*, 1, pp. 83-118.

SAMPLE, Ian (2023): "Science journals ban listing of ChatGPT as co-author on papers", *The Guardian*.

SARTOR, Giovanni (1996): *Intelligenza artificiale e diritto: un'introduzione* (Milano, Giuffrè).

SGUBBI, Filippo, FONDAROLI, Desirèe e TRIPODI, Andrea (2013): *Diritto penale del mercato finanziario. Abuso di informazioni privilegiate, manipolazione del mercato, ostacolo alle funzioni di vigilanza della Consob, falso in prospetto* (Cedam).

STELLA, Federico (2003): *Giustizia e Modernità. La Protezione dell'innocente e la Tutela delle Vittime* (Milano, Giuffrè).

STORTONI, Luigi (2004): "Angoscia tecnologica ed esorcismo penale", in *Rivista italiana di diritto e procedura penale*, 1, pp. 71-89.

SURDEN, Harry e WILLIAMS, Mary-Anne (2016): "Technological Opacity, Predictability, and Self-Driving Cars," in *Cardozo Law Review*, 38, pp. 121-181.

TRIPODI, Andrea Francesco (2022), "Uomo, *societas*, *machina*", in PIERGALLINI, Carlo, MANNOZZI, Grazia, SOTIS, Carlo, PERINI, Chiara, SCOLETTA, Marco e CONSULICH Federico (eds.), *Studi in Onore di Carlo Enrico Paliero* (Milano, Giuffré), pp. 1187-1203.

TRONCONE, Pasquale (2022): "Il sistema dell'intelligenza artificiale nella trama grammaticale del diritto penale. Dalla responsabilità umana alla responsabilità delle macchine pensanti: un inatteso *return trip effect*", in *Cassazione Penale*, 9, pp. 3287-3304.

UBERTIS, Giulio (2020): "Intelligenza artificiale, giustizia penale, controllo umano significativo", in *Sistema Penale*, 4, p. 75-88.

VIGANÒ, Francesco (2013): "Il rapporto di causalità nella giurisprudenza penale", in *Diritto Penale Contemporaneo*.