L'euristica del *quasi* (il diritto penale algoritmico)

di Vincenzo Bruno Muscatiello

Professore ordinario di diritto penale, Università degli Studi di Bari Aldo Moro

It

Un approfondimento sul crescente impatto dell'intelligenza artificiale sul diritto penale, sulle trasformazioni epistemiche e cognitive introdotte dai sistemi algoritmici, sui rischi di tali tecnologie sul libero arbitrio, sulla coscienza e sulla volontà su cui si fonda la responsabilità penale. Una riflessione che si muove tra aspetti tecnici, filosofici e giuridici, problematizzando l'illusione di neutralità e infallibilità delle macchine, e ipotizzando un futuro in cui le decisioni umane saranno sempre più derivate da intelligenze artificiali, con profonde ricadute sulla teoria della colpa e sulla giustizia penale.

Intelligenza artificiale, responsabilità penale, libero arbitrio, bias algoritmico

Eng

An in-depth analysis of the growing impact of artificial intelligence on criminal law, the epistemic and cognitive transformations introduced by algorithmic systems, and the risks these technologies pose to free will, consciousness, and intent – the very foundations of criminal responsibility. A reflection which spans technical, philosophical, and legal aspects, challenging the illusion of machine neutrality and infallibility, and envisioning a future in which human decisions are increasingly shaped by artificial intelligences, with profound implications for theories of culpability and the criminal justice system.

Artificial intelligence, criminal responsibility, free will, algorithmic bias

Sommario

1. Comprendere il mondo – 2. Che-si-capisca-o-no – 3. L'intelligenza artificiale ci rende stupidi? – 4. La storia si ripete – 5. Il diritto penale ipnotico – 6. Appendice

1. Comprendere il mondo

Ciò che la modernità ci consegna, e il futuro ci consegnerà sempre più, è un sistema abitato da algoritmi che si alimentano di miliardi di dati e infinite connessioni, ora anche di tipo percettivo ed esperenziale. L'intelligenza artificiale alla vecchia maniera, quella delle conoscenze indotte e immesse dall'esterno umano, capaci semmai di straordinarie connessioni logiche, è ampiamente superata da una nuova intelligenza autopoietica, indifferente al supervisore umano, e non più vincolato ai limiti della conoscenza umana. I modelli hanno cominciato ad apprendere senza alcuna supervisione esplicita e persino il tradizionale sintagma *machine learning* sembra non bastare a rendere l'attuale stato delle cose, precisato da addende semantiche, oltre il tradizionale *deep learning*, come il *one shot learning* o il *few shot learning* (imparare al primo colpo), il *transfer learning* (trasferimento di idee da un ambito all'altro), il *in-contest learning* (apprendere dal contesto), e molto altro ancora, quanto è cioè possibile aggiungere al semplice, eppure a suo tempo futuristico, *learning*, ormai proiettato verso livelli che solo pochi anni fa sembravano inimmaginabili.

I modelli del linguaggio ne sono un chiaro e (in)visibile esempio: non è più lo studio della grammatica, delle regole del linguaggio, a modellare la conoscenza linguistica, funziona tutto in modo straordinario diverso: le regole non servono, o meglio, vengono apprese come sottoprodotto dell'addestramento autosupervisionato, dove l'algoritmo scopre sequenze di parole, in un intreccio di legami e di regole che seguono la predizione semantica, piuttosto che precederla. Le testine, ossia gli organi che governano quali simboli devono essere combinati all'interno di ciascun modulo, si specializzano spontaneamente, imparano compiti nuovi, trovano oggetti diretti dai verbi, mentre altre testine scoprono articoli determinativi dei sostantivi, oggetti delle preposizioni e quelli dei pronomi possessivi, elaborando un linguaggio complesso e sofisticato, uguale, se non migliore, di quello imparato a scuola. Ed è quanto basta al livello 12[1], decuplicato nel modello di Megatron con 102 livelli di astrazione, vale a dire un modello con capacità stilistiche e conoscenze del mondo tali da riuscire a prevedere persino le parole mancanti di un testo.

Comprendere-il-mondo, la corsa non conosce sosta: il modello a suo tempo più grande del mondo (GPT-2), con un vocabolario di più di 50 mila parole (token) e addestrato su pagine web e innumerevoli libri tratti dal BookCorpus, già dieci volte più grande del suo predecessore (GPT), è un lontano parente del suo più immediato successore, addestrato su diverse centinaia di miliardi di token e milioni di libri e 175 miliardi di parametri regolabili, in un livello anch'esso 10 volte più grande; destinato a sua volta ad essere superato da GPT-3.5 e dall'ultimo GPT-4, in una corsa dove LaMDA di Google (137 miliardi di parametri e 168 miliardi token), Llama e Llama2 rispettivamente con 65 miliardi e 70 miliardi di parametri e 1,5 miliardi e 2 miliardi di token, e PaLM con 540 miliardi di parametri e 768 miliardi di token, corrono verso la soglia psicologica dei mille miliardi di parametri, in una immensità fatta di miliardi di miliardi di pagine web, decine di miliardi di dati ed immagini, centinaia di miliardi di dati vocali e una conoscenza di 550 miliardi di fatti, aperti ora, e dunque ancora più estesi, alla nuova conoscenza offerta dai sensori. Miliardi, non più milioni, meno che mai migliaia.

Si pensi – ed, ovviamente, il pensiero non è il nostro[2] – a un meccanismo che conti le macchine che passano in strada e misuri il livello di traffico; il dato potrebbe essere insincero e la conta potrebbe capire il traffico solo di quel giorno, in una stima, inizialmente imprecisa, che però il sistema è in grado di aggiornare, aggiungendo il dato sul concetto di *week end* (dove il traffico aumenta), o la variabile del Natale, dell'anno scolastico, delle vacanze estive, riducendo sempre più l'errore di predizione. La comprensione del mondo è esattamente questa, aggiungere strati di conoscenza, sempre più accurati, così ancora il ciclo delle stagioni, il concetto di ponte, o le feste nazionali, consentendo all'algoritmo di incontrare il mondo e ridurre la percentuale di imprecisione. Fino ad una approssimativa certezza, comunque più ampia di quella che il ragionamento umano avrebbe elaborato.

Il traffico diviene, esattamente come nel modello linguistico prodotto da Transformer, occasione per un modello del mondo, con una spinta a comprendere sempre più cose, miliardi di

cose, sempre più miliardi di cose, dove, ormai, non conta capire *come*, ma conta capire *cosa* i dati ci suggeriscono di fare, senza neppure sapere (e capirne) il perché. "Certo, la macchina" così Alan Turing in una intervista radiofonica "può fare solo ciò che le ordiniamo, qualsiasi altra cosa sarebbe un guasto meccanico. Tuttavia, non c'è bisogno di supporre che, quando le diamo gli ordini, sappiamo cosa stiamo facendo, né quali siano le conseguenze. Non è necessario essere in grado di capire come questi ordini conducano al comportamento successivo della macchina più di quanto non sia necessario comprendere il meccanismo della germinazione quando si mette un seme nel terreno. La pianta spunta, che si capisca o no".

2. Che-si-capisca-o-no

Questa nuova caratteristica della Ia non può, tuttavia, lasciarci indifferenti, e la derivata sul *cosa*, piuttosto che sul *come*, non può non meritare una migliore riflessione, piuttosto che una irragionevole acquiescenza. È cioè un *cosa* potenzialmente denso di pericoli ed errori, il che rende il *come* decisamente importante, dal momento che una valutazione su dati che umanamente restano senza dubbio ingovernabili, potrebbe contenere elementi di disturbo, segnali deboli, capaci di influenzare e persino deviare la risultanze statistiche, apparentemente, solo apparentemente, impermeabili a *bias* e brutte abitudini.

I casi di *jailbreaking* (evasione) mostrano – ce lo ricorda la sempre più vasta letteratura in argomento – come le macchine non cancellino le informazioni pericolose, ma siano solamente addestrate a reprimerle e non manifestarle, salvo poter essere ipnotizzate o semplicemente tratte in inganno e indotte a disvelarle; gli errori, eufemisticamente chiamati *allucinazioni*[3], possono portare a casi come quello di Jonathan Turley, professore di diritto accusato di commenti inappropriati ad una studentessa durante una gita in Alaska mai svolta, in una università diversa da quella in cui insegnava (il professore insegnava a Seattle e non alla Georgetown University), sulla base di articoli mai pubblicati. L'analisi della recidiva e la traduzione automatica ne sono un ulteriore e reale esempio: nel 2016 un software usato per stimare il rischi di recidiva (Correctional Offender Management Profiling for Alternative Sanctions: COMPAS), usato negli Stati del Wisconsin, California, Florida e New York, produceva risultati ingiustamente sfavorevoli ai danni della comunità afroamericana; e Google Translate usava tradurre la frase "*The president met the senator, while the nurse cured the doctor and the babysitter*" come se fosse "*Il presidente ha incontrato il senatore, mentre l'infermiera ha curato il medico e la baby sitter*", incurante del fatto che la versione inglese non contenesse alcuna indicazione di genere di infermiere, baby sitter e degli altri lavoratori.

Come dire: discriminazione e stereotipi allo stato dell'inconsapevole artificiale. E non è cosa da poco: se una parola si giudica dalle compagnie che frequenta (come si ricordava il linguista britannico John R. Firth), l'algorithmic bias è nondimeno la possibilità che le parole restino associate a pregiudizi, stereotipi, inganni cognitivi; e che l'impronta statistica assuma su di sé una fallacia computazionale che potrebbe sfuggire al controllo di concordanza. E perché no, al controllo in generale: per riprendere una risposta di un guru della tecnologia, Geoffrey Hilton[4], ad una domanda sulle preoccupazioni che verranno, il sistema potrebbe dire a se stesso, e quindi all'esterno, "*Otteniamo più energia. Reindirizziamo tutta l'elettricità ai miei chip*", oppure prescrivere a se stesso una copia migliorata, innescando una reazione incontrollabile, nel noto paradosso delle *graffette* qui portato ai massimi livelli.

3. L'intelligenza artificiale ci rende stupidi?

L'interrogativo, che un tempo non molto lontano occupava la riflessione legata al mondo del www[5], vale ora per il più ampio settore dell'intelligenza artificiale, dove ritornano molte, o forse tutte, le preoccupazioni sulla alterazione delle capacità cognitive e decisorie. Il tema dell'integrità del libero arbitrio del decisore umano rimanda irrimediabilmente alla tenuta dell'assioma "coscienza e volontà", ancora oggi insoluto, eppure enfaticamente posto a base di una qualsiasi iniziativa umana di cui il diritto penale è chiamato ad occuparsi.

La coscienza dell'atto umano si conferma complessa, a tratti misteriosa, e certamente la coscienza dell'atto illecito contiene le medesime inquietudine epistemiche[6]. Comprendere se il *cosa* sia davvero un atto libero, o meglio se il *cosa* custodisca e preservi la libertà del decisore umano, è un tema di riflessione che non potrà esimersi da interrogarsi sulla possibilità di errore, che in qualche modo si tende ad escludere nelle interferenze con i sistemi di intelligenza artificiale, aperti invece a calcoli computazionali oltre i due soli e tradizionali livelli di verità, di tipo binario (vero/falso), ed aperti invece a "insiemi sfocati", logiche fuzzy, conoscenze cosiddette approssimate, esattamente come nei primi passi dei calcoli di rinforzo. Eppure l'*errore* – continuiamo a chiamarlo così – quello che, per le nostre cose, potrebbe condurci a scrutini di legalità punitiva, potrebbe non mancare, il che rende necessario riflettere non solo sul *cosa*, ma anche sul *come*, potendo questo momento contenere la rischiosità erronea che un sistema artificiale finge di non ipotizzare.

Il navigatore non sbaglia, la strada *deve* essere quella giusta; l'algoritmo non si confonde, la diagnosi *deve* essere esatta; l'ingegno del sistema è matematicamente infallibile, il calcolo *deve* essere esatto; e così di seguito, in una *logica-del-deve* che sfugge all'ipotesi dell'errore, o almeno ci distrae dall'ipotesi di un errore. E del resto, come potremmo noi ridiscutere la scelta artificiale, scovare l'errore, anticipare nel dubbio una soluzione diversa e, a conti fatti prima, migliore di quella artificiale. Come potremmo competere con la potenza di calcolo dei nuovi computer quantistici capaci di informazioni misurata in qubit, capaci persino di sovrapposizioni di stati, e di competenze letteralmente illimitate; come potremmo dissentire da un idea, un suggerimento – giacché la finzione è quella ancora di un semplice *suggerimento*[7] – per un tragitto, percorso magari persino tante volte, rispetto alla nuova opzione di un percorso alternativo, deciso dalla la sulla base di una complessa elaborazioni di dati, condizioni meteo, di traffico, di ambiente di cui Waze è solo una delle tante applicazioni[8]; chi mai si affiderebbe all'occhio umano, piuttosto che a quello profondo di un sonar, di un radar, di sistema di individuazione, di una lente focale parabolica o satellitare; allo sguardo di un giudice di linea piuttosto che all'occhio di falco presente a Wimbledon o in un torneo di cricket; di dubitare che la sedia esposta in un museo con il nome di un artista, al riparo dei visitatori, sia una opera concettuale sulla quale non è possibile sedersi? Come potremmo insistere in una diagnosi o in una terapia a dispetto del suggerimento di Heron, o di Flamingo, o BlueJay, infine Sycamore o Zuchongzhi, "dottori" quantistici capaci di eseguire un miliardo di porte e di risolvere i più importanti problemi della medicina e, direi, del mondo?

Non possiamo farlo, non riusciamo a farlo, anche perché la mente è nel frattempo cambiata. È cioè mutato il modo del ragionare riflessivo, in assoluta corrispondenza con la riduzione del pensiero contemplativo e della libertà decisoria, alterata da una radicale trasformazione dell'architettura del pensiero, certamente anche di quello che accompagna l'agire umano (e tanto più quello dei saperi esperti), dove la coscienza si smarrisce e l'inconscio e l'inconsapevole vivono il brivido della immersione in un ignoto che non sono in grado di controllare.

4. La storia si ripete

Come tutto ciò che ha accompagnato i punti di svolta della storia umana, che si tratti di oggetti meccanici, di forma della scrittura, di misuratori del tempo o quant'altro, gli strumenti tecnologici, nell'offrire aiuto al modo in cui si vive, si archivia, si pensa, si abita lo spazio, hanno alterato l'andamento delle cose umane e modellata la struttura fisica e il funzionamento della mente umana[9]: l'avvento del gps, nel momento in cui ha sostituito le mappe, ha anche affievolita la capacità dei taxisti londinesi di elaborare modelli dello spazio e della memoria, riuscendo così a mutare il funzionamento, ma prima ancora la struttura, del sistema cerebrale, adattato alla nuova tecnologia. L'invenzione della scrittura – per non sostare solo nelle grinfie del ragionamento punitivo – ha modificato la capacità di memoria, tipica della tradizione orale e liberata la mente da esercizi e sforzi cognitivi ora applicati, ma in forme differenti, ad altre attività; e, subito dopo, il solo e semplice uso dello spazio fra parole, capace di interrompere la

scriptura continua, è riuscito ad alleviare lo sforzo cognitivo legato a parole lunghe e continue, consentire una nuova abilità del cervello nel decodificare il testo, consentire, cioè, un processo automatico, capace, più di prima, di agevolare la riflessione, e la interpretazione più profonda: una semplice, minuscola, apposizione di uno spazio fisico fra parole, è riuscita a prescindere dal lavoro dello scriba, diffondere la lettura silenziosa e incentivare la scrittura di storie finalmente personali, avventurose, anticonvenzionali, quanto non lo sarebbero state nella dettatura ad uno scriba; è riuscito, infine, a produrre una modifica del processo neurofisiologico della lettura, dell'apprendimento, della cultura e della società[10].

Ma la storia sa essere bizzarra. Se lo spazio-fra-parole ha cambiata la mente, quella del lettore, dello scrittore, della comunità pensante; e se la invenzione di Gutenberg, la nuova etica della produzione libraria, favorendo la quiete del pensiero e il flusso di coscienza, ha modellata la mente letteraria, rendendola *meditante*, in un singolare contrappasso l'avvento delle nuove tecnologie, portatrici di una smaterializzazione della scrittura e una alterazione della relazione con la mente lettrice, hanno recato al seguito una differente capacità di lettura e di riflessione di una mente sempre più *calcolante*.

Il nuovo ambiente, la nuova scrittura, le nuove forme di lettura, la nuova ampiezza di uno spazio non più fisicamente ristretto ai margini di una pagina, un tempo avvolta in due fogli rigidi, ed ora aperta alla infinita complessità della navigazione di rimando, favoriscono la lettura veloce, distratta, affrettata, l'apprendimento superficiale, in una conseguente e inevitabile alterazione della mente e delle sue funzioni. Il lettore attento è stato sostituito da un disattento osservatore, il pensiero cosciente da uno meno cosciente, l'attenzione profonda da una distrazione dalla distrazione, la mente pacata e non esagitata, da un *cervello del giocoliere*, affamato di dati, salti logici, finestre di dialogo per un continuum di *link* e rimandi quali moderna espressione di una lettura continua, priva di spazi di riflessione.

Ogni strumento in fondo dischiude nuove e, per noi, meravigliose ed innumerevoli possibilità, ma impone anche nuove limitazioni e nuovi adattamenti: il cervello, la mente *multitasking*, è ora chiamato alla rapida occhiata, alla lettura di sfuggita, all'accumulo dei dati, alla velocità di entrata e di uscita, senza esitazione, senza (ri)pensamenti, in una moderna passione dell'irrilevanza, quale anestetico della fatica dell'analisi, della comprensione, della profondità.

Il cervello, il nuovo cervello, non ha tempo, non attende *il lento e scrupoloso esame del tempo*[11], la fretta impone i suoi tempi, le sue narrazioni subitanee, scoraggia il pensiero lento, coltiva il business della distrazione, sacrificando le sfumature, i significati più autentici, sull'altare della immediatezza e della frammentazione del pensiero. Conviene fidarsi e, perché no, affidarsi.

Non è il giuoco semantico dell'ossimoro *intelligenza stupida*, più semplicemente la consapevolezza che l'intelligenza può condurre a scelte e decisioni irrazionali, dal momento che la formazione di opinioni vere, la cosiddetta intelligenza o razionalità epistemica, può essere deviata da scopi e desideri di un agente, guidato, in questi casi, da una intelligenza o razionalità strumentale. Nelle scelte umane le due dis-razionalità convivono nelle insidie del ragionamento disgiuntivo, delle decisioni viscerali, delle decisioni condizionate da contesti irrilevanti, cui si aggiungono criticità epistemiche, quali la fiducia immotivata nella propria conoscenza, la tendenza a non cercare di falsificare le proprie ipotesi, la tendenza a ricercare sempre e comunque gli eventi fortuiti, ignorando ipotesi alternative e le insidie del *myside bias*[12].

L'Ia riesce ad evitare le prime, sfuggire ai bias ed euristiche semplificanti, ai *default bias*, agli effetti *framing*, a pregiudizi emotivi, alle *echo chambers*[13], ma non è detto che sappia sempre sottrarsi alle seconde, mettersi al riparo da una irrazionalità epistemica, dalla incapacità di mettersi in discussione. Sembra un paradosso: la molta intelligenza non è detto che sia garanzia di molta razionalità, in ragione del fatto – e qui il paradosso è davvero manifesto – che il dubbio non abita i sistemi artificiali, quella medesima

condizione di incertezza che si vuole combattere, aiuta il sistema ad essere razionale. L'euristica del dubbio compone la razionalità, ne guida le scelte, suggerisce alternative, si interroga sulla falsicabilità delle opzioni e, in questo stato di allerta, ne previene l'irrazionalità; una *passione triste* che sa cioè opporsi alla cristallizzazione, agevolare la fluidità della razionalità, suggerire una rimeditazione della propria idea.

Nei pochi secondi di uno sguardo che non ama perdere tempo, che non ha memoria del passato, ma solo dell'esistente, l'errore non solo non è contemplato, ma non è neppure rilevante, potendo essere oltrepassato da un nuovo errore che, al pari del precedente, attende un nuovo superamento indifferente all'attributo di verosimiglianza; come se la liturgia della dimenticanza fosse capace di involgere anche l'errore, che, non venendo ricordato, non può essere neppure emendato.

5. Il diritto penale ipnotico

È questa, probabilmente, la più grande, o almeno la più vicina, insidia della intelligenza artificiale, quella di cui prossimamente converrà (preo)occuparsi: fronteggiare l'errore (umano) guidato dalla intelligenza (artificiale). Fingere, continuare a fingere, che l'errore – laddove sussumibile in uno schema di illecito penale, e tanto più nel territorio dei saperi esperti – sia un errore proprio, cosciente e volontario, fingendo di ignorare la partenogenesi dell'errore, sempre più profonda e umanamente impenetrabile. In fondo sapevamo che il richiamo alla coscienza e volontà contenesse l'equivoco di una leggerezza semantica, cui certamente il sistema penale non poteva rimediare, non poteva cioè risolvere il buco nero della co(no)scienza, quale mistero insolubile, e tanto più alla ignara ed assiomatica dogmatica penalistica. Il punto è che, ora, la leggerezza epistemica, funzionale alle dissonanze di *noluntà* legate a speciali stati patologici, potrebbe sempre più assomigliare ad una verità apofantica per esperienze decisorie e decisionali, rese inconsapevolmente incoscienti e involontarie: una finzione di *suitas* decisoria, aggrappata alla antica idea di coscienza e volontà, ma ridotta ad una parimenti illusoria *actio libera in causa*, anacronisticamente appartenuta all'antica idea di un iniziale determinismo biologico delle decisioni algoritmiche.

Nella immensa, se non infinita, complessità di dati, nell'inestricabile groviglio di pensieri, nel congedo dalla rassicurante logica binaria delle nostre tradizionali forme di comunicazione, l'intelligenza umana è, se così fosse, destinata ad abbandonare la ridondanza, i tentativi e gli errori, la verifica e la falsificazione, i tradizionali metodi di acquisizione del sapere, i dubbi, i fallimenti, persino l'infelicità che ad essi si accompagna, a vantaggio di un mondo di conoscenze impeccabili, comode, impigrite, abitato da *prigionieri liberi*[14]. Persino l'atto illecito potrebbe essere deciso *fuori di sé*, o meglio, da un sé preso in prestito *dal fuori*: e l'io del soggetto agente, fintamente cosciente, diventare né più né meno che un centro di connessione temporaneo di varie esperienze portate da fuori alla nostra (dis)attenzione[15].

La seconda fra le due più grandi incognite del sapere umano, l'origine dell'universo e l'essenza della coscienza, la logica del tutto e quella del poco, dell'universale e del particolare, eretta, ciononostante, a requisito dogmatico per un addebito che si vuole, con apodittica sicumera, cosciente e volontario, è destinato a comporre il mistero scientifico più affascinante e sconcertante dell'origine delle esperienze quotidiane, senza che sia possibile comprendere esattamente cosa sia la coscienza, e quanto essa sappia diventare aliena; cosa la componga e la origini, e quanto essa resista alla complessità dei nuovi soggetti coscienti e alla magia delle interferenze artificiali dei nuovi agenti artificiali (coscienti o meno che siano, o che saranno). Niente di certo, tutto ancora da capire, poco, ancora meno, del quasi.

A questa prima e impredicibile sicumera, l'arretramento o meglio la finzione di coscienza come apodittico e ineffabile criterio di una teoria altrimenti inefficace, il sistema è destinato ad aggiungere un secondo inganno, quella di una quasi-volontà, in una complessiva quasi-colpevolezza, posta anch'essa, nel giuoco del doppio *quasi*, a base dogmatica di un addebito punitivo chiamato a misurare la meritevolezza di pena secondo logiche del *quan-*

to-basta: il sintagma che ne esprime la sapienza è l'"*avremmo-potuto-impedirlo*".

Avremmo-potuto, ecco il punto, la seconda grande finzione euristica. Nel *nuovo* mondo del quantitativo e del quantificabile, diverso dal *vecchio* del qualitativo e delle sfumature[16], in questo amalgama di intelligenza umana e di intelligenza artificiale, una sorta – potremmo dire – di *intelligenza universale* capace di fondere le esperienze coscienti e ignorare talvolta l'una, talaltra l'altra, in questa terra di confine di un *sé percipiente* e di un *sé raziocinante*, la colpa penale, una volta l'Ia entrata nella mente, non potrà che essere, semplicemente, quella di essersi (af)fidati e di non aver tentato di mantenere gli errori entro il limite del noto e dell'evitabile: sostituire inferenze da entità sconosciute con costruzioni a partire da entità conosciute[17], pur se, queste ultime, non del tutto sincere e veritiere, tuttavia più comode, e semplicemente accettate come valide: nuovamente, l'euristica del controllo, sui saperi esperti; vale a dire, abitare attenta-mente la dis-attenzione. E dunque *avremmo-potuto*, senza però che sia per davvero possibile *aver-potuto*.

Non sarà, quello nell'*officina del diavolo*[18], un processo immediato, per quanto inesorabile esso appaia: occorreranno forse due o tre generazioni per radicare la diseducazione all'impegno personale, per pietrificare la legge del minimo sforzo, per naturalizzare la delega inconsapevole e profonda ai sistemi algoritmici, ma il prodotto di questo nuovo ordine delle cose non potrà non sublimare un nuovo agire incosciente e involontario; rendere il sintagma dell'art. 42 c.p. – almeno in settori ad alta specializzazione, come quello medico o ingegneristico – ancora più approssimato e incerto. Nel presente si tratterà, dunque, semplicemente di tamponare la falla, a che l'euristica punitiva – ancora una sostituzione del probabile incerto sul certo improbabile – non smarrisca il suo potere, e il prezzo da pagare non potrà che essere un ripensamento della colpa, una limitazione degli oggetti, destinatari di una *debole pienezza* psicologica nel cui ossimoro trovi spazio l'idea salvifica di un libero arbitrio, *semplicemente* indebolito[19], ma pur sempre legato all'*avremmo-potuto*; nel futuro, invece, il giuoco fittissimo di interferenze, l'oggettiva inesigibilità del sintagma accusatorio, l'umana cioè impossibilità dell'*avremmo-potuto*, si renderà portatore di un naturale allontanamento dall'obbligo costituzionale e convenzionale di una piena responsabilità penale, il che, in ultima analisi, potrebbe suggerire un elogio della fuga, una rinuncia all'esangue pensiero punitivo, e così accompagnare il diritto penale artificiale, quel particolare campo del diritto penale, al suo tramonto, in una nuova e diversa umanità ultrametafisica, a condizione, s'intende, che sia davvero possibile farlo. Come dire: una nuova esperienza per un diritto penale ipnotico, costretto *inconsapevolmente* ad una *euristica del quasi*. Perché non sia il *niente*, il *quasi* potrebbe, per l'intanto, finire per bastare. Sapendo, però, che fra non molto, più presto che tardi, decidere e volere sarà sempre meno una espressione di libertà, e allora l'impasse del libero arbitrio ci restituirà un rassegnato *nuovo* sintagma: "fiat voluntas sua".

Appendice
1. Così per giuoco

Prendendo spunto da Murray Shanahan, professore di Robotica cognitiva al dipartimento di Computing dell'Imperial College di Londra, provo anch'io a raccontarvi una storia liberamente adattata a quella nel saggio *La rivolta delle macchine*.La storia è ambientata in un futuro prossimo, non molto distante in noi, in un tempo in cui l'intelligenza artificiale abbia raggiunto livelli di maturazione che la rendano simile alla intelligenza umana.

In questa storia ci sono tre sistemi di intelligenza artificiale: il primo è un sistema di Ia di marketing che appartiene a una grande casa editrice che noi chiameremo General Corporation, con sede in una città del nord Italia; il secondo sistema è una Ia della Polizia del luogo dove la General Corporation ha una libreria che deve commercializzare il libro di Ciro Marchi, autore di un libro molto conosciuto e molto contestato, dal titolo *Il profeta di sventura*, in alcune sue parti dissacrante verso identità religiose giudicate aporetiche; il terzo sistema è una Ia di sicurezza gestita e controllata dal governo di un piccolo paese in via

di sviluppo, al di là del Mediterraneo, governato da una élite religiosa di tipo confessionale (di una di quelle religioni verso le quali il libro dedica una irriguardosa ironia), non proprio democratica, e preoccupata di custodire la propria santità religiosa, e sopire qualsiasi manifestazione di dissenso religioso.

La storia inizia quando la Ia di marketing General Corporation decide di avviare una campagna pubblicitaria per massimizzare le vendite dell'ultimo prodotto editoriale, quel nuovo e contestato saggio editoriale di cui si diceva, accolto con entusiasmo da una fetta di intransigenti lettori affezionati, e con preoccupazione da una altra fetta di ecumenici oppositori alle idee contenute nel libro. Utilizzando il complesso sistema di comportamento umano che l'intelligenza ha costruito su una infinità di dati, vendite del precedente libro, gusti editoriali, tendenze politiche del territorio, e applicando tecniche di ottimizzazione, questa Ia di marketing presenta un lancio delle vendite e annuncia che, nel primo giorno di uscita nella sua libreria in una grande città del sud Italia, ai primi 200 potenziali acquirenti verrà regalato il libro autografato dall'autore, con tanto di *selfie* con Ciro Marchi in persona.

Questa idea di marketing, sviluppata dalla Ia marketing, viene comunicata all'ufficio di Polizia del luogo in cui ha sede la libreria della casa editrice, a sua volta dotata di un proprio sistema di Ia per la sicurezza. In ragione della stimata affluenza di una folla numerosissima, fra estimatori e contestatori, dunque un assembramento insidioso per l'ordine pubblico, la Ia della Polizia, utilizzando il proprio sistema previsionale, formula la stima che in quel giorno saranno presenti almeno 5000 persone, alcune per manifestare il proprio consenso, altre il proprio dissenso, il che significa, in termini probabilistici, secondo sempre un sistema algoritmico, una stima di rischio del 10% di disordine urbano. Sulla base di queste due previsioni, l'Ia di polizia decide di presidiare la zona con personale in tenuta antisommossa, per evitare incidenti, ciò che in realtà anche la Ia del marketing ha previsto e stimato possibile, tanto quanto la presenza di poliziotti in tenuta antisommossa, con un calcolo delle probabilità del 94%. Rischio incidenti e presenza di forze di polizia sono entrambi previsti dalla Ia marketing, la quale oltre a stimarne la possibilità, immagina anche che questo dispiegamento di forze rappresenti una grande opportunità pubblicitaria per il target di riferimento e una grande visibilità per la produzione editoriale; ordina dunque 5000 gadget pubblicitari (mettiamo si tratti di magliette e cappellini), tutte con il logo della casa editrice e del titolo del libro in bella vista, che saranno distribuiti ai presenti, ed affida la produzione di queste 5000 fra magliette e cappellini ad una piccola fabbrica, dislocata in un piccolo paese in via di sviluppo al di là del Mediterraneo (e il cui governo – si diceva - è da tempo intento alla repressione di qualsiasi dissenso religioso), dotato anch'esso di una propria Ia di sicurezza, la quale, avuta conoscenza della fabbricazione di ben 5000 prodotti pubblicitari, in qualche modo testimoni di una ostilità alla religione del territorio, elabora la previsione, con una probabilità del 20%, che esse saranno utilizzate in una situazione antigovernativa sovversiva nel proprio territorio. Non immaginando, cioè, che l'assembramento riguarderà altro territorio e che le magliette saranno destinate ad altro paese al di là del Mediterraneo, e non il proprio, l'Ia di sicurezza del piccolo governo ordina nel giro di un'ora che le 5000 magliette, con il logo della casa editrice e il titolo del libro offensivo della religione ivi praticata, siano confiscate e in questa operazione di *pulizia* culturale neutralizza il tentativo di reazione di una guardia giurata, presente all'ingresso dell'azienda cui era stata commissionata la produzione, neutralizzata dagli aggressori inferociti (si sa che in nome della religione talvolta si compiono nefandezze e misfatti) e in condizione serie tanto da essere ricoverata con ferite multiple in ospedale.

Nel giro di pochi minuti la storia viene raccontata da tutti i mass media e le foto scattate durante l'attacco di una folla in tumulto, inneggiante ai cancelli della impresa, mostrano la guardia giurata sanguinante e la catasta di 5000 magliette, tutte con il logo del libro e della General Corporation che ne aveva commissionato la produzione; nel giro di pochissimo tempo la IA marketing della casa editrice viene attaccata per le sue tattiche di commercializzazione e, tuttavia, paradossalmente questo grande clamore mediatico fa sì che il prodotto editoriale, che andava venduto nel primo giorno di lancio e che sarebbe stato donato, con tanto di autografo e immancabili *selfie*, ai primi 200 avventori che si fossero presentati, vede schizzare le vendite al 200% più del previsto. Per quanto sembri strano, tutto, proprio tutto,

va esattamente come la Ia di marketing aveva pianificato ed essa ha semplicemente eseguito in maniera perfetta la sua missione, massimizzando la funzione di ottimizzazione, senza alcun intervento umano. Il ferimento della guardia giurata ha rappresentato, in un cinico calcolo pubblicitario, un evento imprevisto o comunque trascurato nel disegno del marketing artificiale, intento a porre in essere una soluzione di vendita e pubblicità, pur se eticamente discutibile. Questo evento, una sorta di danno collaterale, non previsto o, anche solo, reso indifferente e accettato nella sua drammaticità, produce a sua volta un altro evento: la tragica aggressione della guardia per la intransigenza religiosa dei manifestanti innesca in uno dei dirigenti della General Corporation un profondo esame di coscienza che lo induce a rinunciare al suo patrimonio e a dedicare la sua vita a commercializzare prodotti editoriali di giovani scrittrici di tutte le religioni, altrimenti estranee ai circuiti editoriali a causa della tecnologia di Ia e alle non convenienti previsioni di vendita editoriale. La Fondazione finanziata da questa persona, oltre a patrocinare questo umanesimo religioso, diventerà inoltre un movimento globale, portando luce al buio di numerose esistenze umane.

Tutto è bene quello che finisce bene (tranne che per la povera guardia giurata). Epperò, la storia si arricchisce di un ulteriore particolare: anche questo scenario successivo, apparso in qualche misura *umanamente* imprevedibile, in realtà era stato previsto e pianificato dalla Ia di General Corporation, la quale, oltre a possedere una sezione marketing, possedeva anche un altro sistema di Ia etica, la quale aveva consigliato l'utilizzo del sistema Ia di marketing, aveva previsto l'aggressione della guardia giurata (per fortuna successivamente guarita) e aveva pronosticato correttamente l'effetto che questa avrebbe avuto sul dirigente dell'impresa. Il sacrificio di una vita (per fortuna non del tutto) bene valeva il beneficio per altre vite, in una logica di pura ottimizzazione del profitto umano e di comparazione fra costi/benefici, ciò che rendeva l'apparente cinismo pubblicitario, in realtà una strategia di medio tempore, portatrice di più estesi benefici umani e sociali. In sostanza, e per concludere, le conseguenze indesiderate erano state previste, così come era stato previsto il beneficio che ne sarebbe derivato, il sacrificio di una vita umana, per la salvezza di più esistenze umane, senza peraltro nessun coinvolgimento decisionale della intelligenza biologica di alcuna intelligenza umana, resa estranea e inconsapevole delle previsioni e delle decisioni del sistema di Ia.

Resta, tuttavia, il sacrificio della guardia giurata, aggredita e segnata da un evento traumatico: orbene, in disparte le questioni di territorialità punitiva, la condotta lesiva ai danni della guardia giurata – prevista, calcolata, utilizzata dalla Ia – non dovrebbe poter essere sottratta ad un sindacato di legalità punitiva, restando sussumibile in una fattispecie punitiva che restiamo, intuitivamente, alle sole lesioni personali. Ma la domanda è: chi *avrebbe-potuto* impedirlo?

Note

1. Cristianini, N. (2024), *Machina sapiens. L'algoritmo che ci ha rubato il segreto della conoscenza*, Bologna, Il Mulino, p. 109, ricorda come uno studio sperimentale di BERT si fondava su 12 moduli e un altro su 24, già sufficienti per una sequenza di operazioni tipica del linguaggio.
2. Cristianini, N., *Machina sapiens*, cit., pp. 111-112.
3. Il rischio di risposte errate, le cosiddette allucinazioni, è un rischio reale, che però – ce lo ricorda in una intervista Reid Hoffman nell'inserto di *Repubblica* del 4 dicembre 2024, *The age of intelligence* – appartiene anche gli umani più esperti, salvo che – ed è questo il punto – in qualche modo non ce lo si aspetta da un sistema artificiale.
4. Geoffrey Hilton è indicato come il padre delle reti neurali, e di recente è stato insignito del premio Nobel per la fisica. In un articolo a lui dedicato, pubblicato nell'inserto di *Repubblica* del 4 dicembre 2024, *The age of intelligence*, sono raccontate alcune delle sue preoccupazioni, confessate al New York Times e pri-

ma ancora in un *paper* del 2023, legate ai rischi del fuori-controllo dell'Ia. Con lui il manifesto sui rischi dell'Ia è stato sottoscritto da una ventina di studiosi ed accademici, fra cui Yuval Noah Harari e il premio Nobel per l'Economia, Daniel Kahneman, entrambi ampiamente citati in questo saggio.

5. Carr, N. (2011), *Internet ci rende stupidi? Come la rete sta cambiando il nostro cervello.* Milano, Raffaello Cortina Editore.

6. Ed in fondo continuiamo a non sapere cosa sia la coscienza: Perilli, L. (2025), *Coscienza artificiale: Come le macchine pensano e trasformano l'esperienza umana*, Milano, Mondadori, p. 130.

7. Un semplice *pungolo*, potremmo dire, un suggerimento di voltare attraverso una voce gentile ed educata che non protesterebbe se la svolta non avviene nei termini suggeriti: il che ovviamente dipende anche dalla quantità e qualità dei *pantani*. In argomento Thaler, R. H., Sunstein, C. R. (2024), *Nudge: La nuova strategia per migliorare le nostre decisioni su denaro, salute, felicità*, Milano, Feltrinelli, p. 290.

8. Harari, Y. N. (2023), *Homo Deus: Breve storia del futuro*, Bologna, Bompiani, p.416, ricorda come Waze non sia solo in grado di allontanarci da eventuali ingorghi e condurre l'utente verso itinerari più veloci, suggerendoci di svoltare a sinistra anziché a destra, come ordinariamente un automobilista sarebbe stato tentato di fare, ma aggiunge la capacità di distribuire il traffico, evitando così che una fuga da un ingorgo crei un secondo ingorgo, determinato dallo spostamento verso l'itinerario alternativo. L'intelligenza di Waze è, dunque, quella di individuare il percorso 2 senza intasarlo, convogliando solo la metà degli automobilisti del percorso 1, in modo che né l'uno né l'altro ne risultino intasati. Si pensi anche – come ci ricorda Russell, S. (2025), *Compatibile con l'uomo. Come impedire che l'IA controlli il mondo*, Torino, Codice edizioni, p. 147 – al traffico aereo: all'inizio i computer si limitavano a programmare gli orari dei voli, ma ben presto si sono occupati della assegnazione degli equipaggi, dei posti a sedere, della manutenzione di routine, fino, ad oggi, a fornire informazioni sullo stato dei voli e a cambiare le rotte degli aerei, spostare gli equipaggi, modificare le prenotazioni e riorganizzare i turni delle manutenzioni.

9. Carr, N., *Internet ci rende stupidi?*, cit., p. 69.

10. Diffusamente Carr, N., *Internet ci rende stupidi?*, cit., pp. 83 e ss.

11. Carr, N., *Internet ci rende stupidi?*, cit., p. 205.

12. Quattrocchi, W., Clerici, A., *Liberi di crederci*, cit., p. 82.

13. Per una analisi di questi luoghi ideali per la radicalizzazione delle proprie opinioni, si rinvia a Quattrocchi, W., Clerici, A. (2018), *Liberi di crederci. Informazione, internet e post-verità*, Torino, Einaudi, pp. 99 e ss., e degli stessi Autori, più diffusamente, (2023) *Polarizzazioni. Informazioni, opinioni e altri demoni nell'infosfera*, Milano, Franco Angeli, pp. 37 e ss.

14. È il titolo del bel saggio di Trautteur, G. (2020), *Il prigioniero libero*, Milano, Mimesi.

15. Nelle applicazioni belliche i sistemi come Lavender e Hasbora sono in grado di selezionare migliaia di obiettivi sulla base di incroci di ogni genere di informazioni, come telefonate, mail, posizionamenti gps, contatti, pagamenti on line e così via. Una volta individuato il bersaglio il sistema è in grado di prevedere il numero di vittime collaterali civili e sulla base di una scala di tollerabilità, da 10 a 100, decidere l'iniziativa ed accettarne gli effetti, in base al grado di importanza del bersaglio: Nel solo conflitto israeliano palestinese il sistema Ia Lavender ha selezionato circa 37000 obiettivi mediante sistemi di machine learning per l'analisi di big data: così Perilli, L., *Coscienza artificiale*, cit., p. 116.

16. Perilli, L., *Coscienza artificiale*, cit., p.179.

17. Le parole di Bertrand Russell sono citate da Hoffman, D., (2020), *L'illusione della realtà. Come l'evoluzione ci inganna sul mondo che vediamo*, Torino, Bollati Boringhieri, p. 261.

18. Sapolsky, R. (2024), *Determinati. Biologia, comportamento e libero arbitrio*, Milano, Raffaello Cortina Editore, p. 288.

19. Sulla complessità che l'avverbio esprime, rinviamo al punto di vista di Sapolsky, R., *Determinati*, cit., passim, e dello stesso Autore, (2014), *L'uomo bestiale. Come l'ambiente e i geni costruiscono la nostra identità*, Roma, Codice edizioni.