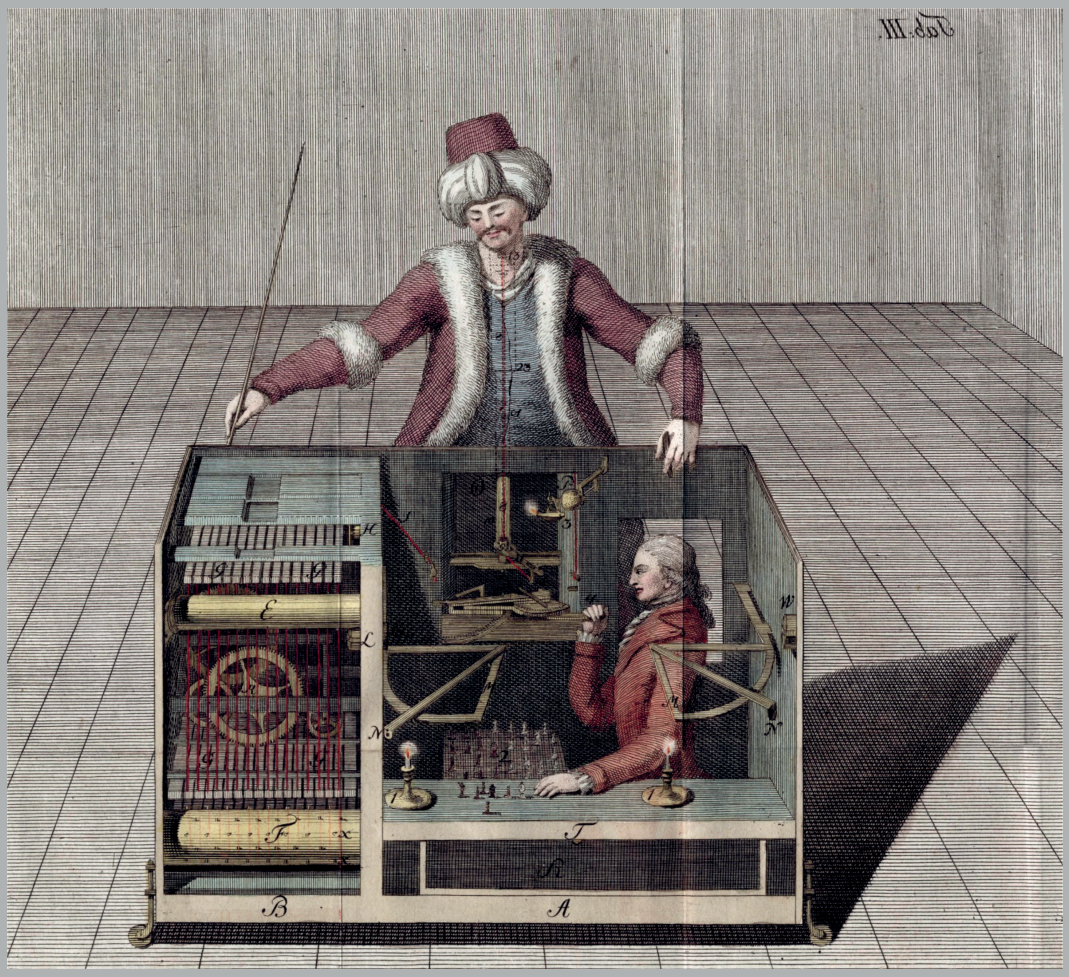
(a cura di)
GABRIELE DELLA MORTE

L'INTELLIGENZA ARTIFICIALE TRA FATTO E NORMA



Lefebvre Giuffrè

In copertina:
Joseph Racknitz, Il Turco | 1789

(a cura di)
GABRIELE DELLA MORTE

L'INTELLIGENZA ARTIFICIALE TRA FATTO E NORMA

Lefebvre Giuffrè

ISBN 9788828885887

Il presente volume è stato realizzato grazie al contributo dell'Università Cattolica del Sacro Cuore nell'ambito dell'attribuzione, all'esito di una procedura selettiva, di un finanziamento di tipo D.3.2. (Progetti di interesse di Ateneo, 2020. Titolo del progetto: Funzioni pubbliche / Controllo privato. Profili interdisciplinari sulla governance senza governo della società algoritmica - https:/progetti.unicatt.it/fpcp).

© Copyright Giuffrè Francis Lefebvre S.p.A. Milano - 2026
Via Monte Rosa, 91 - 20149 MILANO - www.lefebvregiuffre.it

La traduzione, l'adattamento totale o parziale, la riproduzione con qualsiasi mezzo (compresi i microfilm, i film, le fotocopie), nonché la memorizzazione elettronica, sono riservati per tutti i Paesi.

INDICE

Gli Autori IX

GABRIELE DELLA MORTE, *Introduzione: L'esigenza di un dialogo tra fatto e norma* XI

SALUTI INTRODUTTIVI

CLAUDIO GIULIODORI, *Saluti introduttivi al Convegno: "L'intelligenza artificiale tra fatto e norma"* (Università Cattolica, Milano, 28 e 29 ottobre 2024) 3

STEFANO SOLIMANO, *L'intelligenza artificiale in chiaroscuro. Saluti introduttivi al Convegno: "L'intelligenza artificiale tra fatto e norma"* (Università Cattolica, Milano, 28 e 29 ottobre 2024). 9

ANTONIO GASBARRINI, *La prospettiva della Digital Health. Saluti introduttivi al Convegno: "Digital Health, Artificial Intelligence: quali responsabilità verso i pazienti?"* (Fondazione Policlinico Universitario "A. Gemelli" IRCCS, Roma, 4 giugno 2024). 11

KEYNOTE SPEECH

JUAN CARLOS DE MARTIN, *L'intelligenza artificiale: Il fatto* 17

ORESTE POLLICINO, *L'intelligenza artificiale: La norma*. 23

INTELLIGENZA ARTIFICIALE E DISINFORMAZIONE

GIOVANNA MASCHERONI e SIMONE CARLO, *Introduzione* 31

ALBERTO ODDENINO, *(Il fatto - Parte I). La (post) verità fra pubblico e privato* . . 39

RICCARDO GALLOTTI, *(Il fatto - Parte II). AI4TRUST: sfide di un progetto (di Fact Checking) per usare l'IA contro Fake e Deep Fake* 47

ALBERTO ODDENINO, *(La norma). L'ardua ricerca di una regolazione effettiva del disordine informativo nell'era dell'intelligenza artificiale generativa* 55

INTELLIGENZA ARTIFICIALE E 'IMMORTALITÀ' DEL DATO

FRANCESCA PASQUALI, *Introduzione sulla memoria e sopravvivenza algoritmica nell'era della mediatizzazione profonda*. 75

DAVIDE SISTO, *(Il fatto). Thanabots. La problematica "sopravvivenza" post mortem tramite l'AI* 79
GIOVANNI ZICCARDI, *(La norma). Immortalità del dato, eredità digitale e intelligenza artificiale: profili giuridici e tecnologici tra oblio e memoria* 91

INTELLIGENZA ARTIFICIALE E RICONOSCIMENTO FACCIALE E BIOMETRICO

GABRIO FORTI, *Introduzione* 111
FEDERICO CABITZA e LUCA MARCONI, *(Il fatto). Intelligenza artificiale e riconoscimento facciale: tra fatto, norma e interpretazione*. 115
GIUSEPPE MOBILIO, *(La norma). La disciplina sulle tecnologie di riconoscimento facciale, ovvero il tentativo di limitare un formidabile potere di sorveglianza* . . 139

INTELLIGENZA ARTIFICIALE E AMBITO MILITARE

MONICA SPATTI, *Introduzione: sfide e prospettive per il diritto internazionale umanitario* 159
RICCARDO RICCHETTI e FRANCESCA SIRONI DE GREGORIO, *(Il fatto). Intelligenza Artificiale, Conflitti Armati e Diritto: Profili Tecnici ed Etici* 171
MARCO SASSÒLI, *(La norma). Le Armi Autonome e il Diritto Internazionale Umanitario: Profili Regolatori*. 187

INTELLIGENZA ARTIFICIALE E TUTELA DELLA SALUTE

RENATO BALDUZZI, *Introduzione ai profili costituzionali* 207
ANTONIO ALBANESE, *Introduzione ai profili civilistici* 211
ALAIN DELL'OSSO, *Introduzione ai profili penalistici* 215
ANTONIO OLIVA e FRANCESCA CAZZATO, *(Il fatto). L'impatto dell'intelligenza artificiale in ambito sanitario*. 221
ANDREA ROSATI e ANNA FAGOTTI, *(Il fatto). Intelligenza Artificiale nel Carcinoma Ovarico Avanzato: integrazione multimodale, standardizzazione decisionale e chirurgia di precisione*. 231
GIUSEPPE RIVA, *(Il fatto). L'IA tra "fatto predittivo" e "norma evolutiva": la prospettiva della Cyberpsicologia* 241
ANDREA SIMONCINI, *(La norma). Intelligenza Artificiale e Salute: profili pubblicistici* 257
ANTONIO VALITUTTI, *(La norma). L'impiego di Sistemi Autonomi in Medicina e le Responsabilità del Sanitario*. 269
ANTONELLA MASSARO, *(La norma). Robot intelligenti e attività medica: presupposti e (soprattutto) limiti della responsabilità penale dell'operatore sanitario* . 285
MIRKO FACCIOLI, *(La norma). Intelligenza artificiale in medicina e responsabilità civile della struttura sanitaria* 297

RUBEN RAZZANTE, *(La norma). Intelligenza Artificiale e trattamenti sanitari: tra consenso informato, autodeterminazione terapeutica e protezione dei dati sensibili* 311

IN GUISA DI CONCLUSIONE

FABRIZIO VON ARX ed ENRICO REGGIANI, *Dialogo 'concertato'. Quello che l'intelligenza artificiale non può fare* 331

Ringraziamenti 343

GLI AUTORI

ANTONIO ALBANESE, *Professore Ordinario di Diritto Civile, Università Cattolica del Sacro Cuore.*

RENATO BALDUZZI, *già Ministro della Salute e Professore Ordinario di Diritto Costituzionale, Università Cattolica del Sacro Cuore.*

FEDERICO CABITZA, *Professore Ordinario di Informatica, Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca.*

SIMONE CARLO, *Ricercatore di Sociologia dei Processi Culturali e Comunicativi, Università Cattolica del Sacro Cuore.*

FRANCESCA CAZZATO, *Specialista in Medicina Legale, Dottoranda in Scienze Biomediche di Base e Sanità Pubblica, Università Cattolica del Sacro Cuore.*

JUAN CARLOS DE MARTIN, *Professore Ordinario di Ingegneria Informatica, Politecnico di Torino.*

GABRIELE DELLA MORTE, *Professore Ordinario di Diritto Internazionale, Università Cattolica del Sacro Cuore.*

ALAIN DELL'OSSO, *Professore Associato di Diritto Penale, Università Cattolica del Sacro Cuore.*

ANNA FAGOTTI, *PhD, Professoressa Ordinaria di Ostetricia e Ginecologia presso l'Università Cattolica del Sacro Cuore, Roma, Direttore dell'Unità Operativa Complessa di Ginecologia Oncologica presso la Fondazione Policlinico Universitario A. Gemelli IRCCS, Roma, Italia.*

MIRKO FACCIOLI, *Professore Associato di Diritto Privato, Università di Verona.*

GABRIO FORTI, *Professore Emerito di Diritto Penale e Criminologia, Università Cattolica del Sacro Cuore, e Accademico dei Lincei.*

RICCARDO GALLOTTI, *Capo dell'unità di ricerca CHuB (Complex Human Behaviour Lab) della Fondazione Bruno Kessler, coordinatore del progetto Horizon Europe AI4TRUST.*

ANTONIO GASBARRINI, *Direttore Scientifico, Fondazione Policlinico Universitario Gemelli IRCCS.*

CLAUDIO GIULIODORI, *Assistente Ecclesiastico Generale dell'Azione Cattolica Italiana e dell'Università Cattolica del Sacro Cuore; Presidente della Commissione Episcopale per l'educazione cattolica, la scuola e l'università.*

LUCA MARCONI, *Assegnista di Ricerca di Informatica, Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca.*

GIOVANNA MASCHERONI, *Professoressa Ordinaria di Sociologia dei Processi Culturali e Comunicativi, Università Cattolica del Sacro Cuore.*

ANTONELLA MASSARO, *Professoressa Associata di Diritto Penale, Università degli Studi di Roma Tre.*
GIUSEPPE MOBILIO, *Professore Associato di Diritto Costituzionale, Università degli Studi di Firenze.*
ALBERTO ODDENINO, *Professore Associato di Diritto Internazionale, Università degli studi di Torino.*
ANTONIO OLIVA, *Professore Ordinario di Medicina Legale - Università Cattolica del Sacro Cuore, Direttore della UOC Gestione Contenzioso Medico Legale - Fondazione Policlinico Universitario A. Gemelli-IRCCS.*
FRANCESCA PASQUALI, *Professoressa Ordinaria di Sociologia dei Processi Culturali e Comunicativi, Università degli Studi di Bergamo.*
ORESTE POLLICINO, *Professore Ordinario di Diritto Costituzionale, Università Commerciale Luigi Bocconi.*
RUBEN RAZZANTE, *Professore a contratto di Diritto dell'Informazione, Università Cattolica del Sacro Cuore.*
ENRICO REGGIANI, *Professore Ordinario di Letteratura (in lingua) Inglese, Musicologo e Direttore dello Studium Musicale di Ateneo.*
RICCARDO RICCHETTI, *Dottorando di Ricerca di Diritto Internazionale, Università Cattolica del Sacro Cuore.*
GIUSEPPE RIVA, *Professore Ordinario di Psicologia Generale, Direttore dell'-Humane Technology Lab, Università Cattolica del Sacro Cuore.*
ANDREA ROSATI, *Specialista in ginecologia, Fondazione Policlinico Universitario A. Gemelli-IRCCS, Università Cattolica del Sacro Cuore.*
MARCO SASSÒLI, *Professore Onorario all'Università di Ginevra.*
ANDREA SIMONCINI, *Professore Ordinario di Diritto Costituzionale, Università di Firenze.*
FRANCESCA SIRONI DE GREGORIO, *Assegnista di Ricerca di Diritto Internazionale, Università Cattolica Del Sacro Cuore.*
DAVIDE SISTO, *Tanatologo e formatore nel campo della Digital Death.*
STEFANO SOLIMANO, *Preside della Facoltà di Giurisprudenza, Università Cattolica del Sacro Cuore.*
MONICA SPATTI, *Professoressa Associata di Diritto dell'Unione Europea, Università Cattolica del Sacro Cuore.*
ANTONIO VALITUTTI, *già Presidente della I Sez. civ. della Corte Suprema di Cassazione.*
FABRIZIO VON ARX, *Violinista e Direttore Artistico di Fondazione Casa Stradivari di Cremona.*
GIOVANNI ZICCARDI, *Professore Ordinario di Filosofia del Diritto e Informatica Giuridica, Università degli studi di Milano.*

INTRODUZIONE: L'ESIGENZA DI UN DIALOGO TRA FATTO E NORMA

di *Gabriele Della Morte*

Sommario: 1. Premessa: il disorientamento del giurista. — 2. Le sfide poste dall'intelligenza artificiale. — 3. La progressiva emergenza del dialogo tra fatto e norma. — 4. Struttura e contenuto del presente volume. — 4.1. (*Segue*). La disinformazione. — 4.2. (*Segue*). 'L'immortalità' del dato. — 4.3. (*Segue*). Il riconoscimento facciale e biometrico. — 4.4. (*Segue*). L'ambito militare. — 4.5. (*Segue*). La tutela della salute. — 4.6. (*Segue*). La creatività costretta (e liberata).

Quante volte Ernest Hyde ed io
abbiamo discusso del libero arbitrio!
La mia metafora preferita era la vacca di Prickett
che pascolava legata, e circolava
quanto glielo concedeva la fune.
Un giorno che così discutevamo, guardando la mucca
che tirava la fune per allargare lo spazio
che era oramai nudo d'erba,
il piolo saltò e, testa bassa,
quella che ci corse addosso.
«È questo? Non è libero arbitrio?» disse Ernest fuggendo di corsa.
Io caddi sotto una cornata.

Edgard Lee Master, *Antologia di Spoon River*,
Torino, Einaudi, 1974, p. 115.
Traduzione di Fernanda Pivano.

1. Premessa: il disorientamento del giurista.

Il mondo del diritto, per come largamente rappresentato sino a oggi, è caratterizzato da un tratto *analogico*, la cui essenza risiede nella possibilità di situare, nel tempo e nello spazio, i fatti e gli eventi cui si riferisce la regolazione. Detta capacità di misurare e comparare, la possibilità cioè di affermare che qualcosa è *vicina o lontana*, è accaduta *prima o dopo* — requisiti adeguati alla determinazione della cau-

salità — è posta in crisi dalla *datificazione*, ovvero dal processo tecnologico che trasforma gli innumerevoli aspetti della vita quotidiana in dati quantificabili e che, riscrivendo il mondo in forma numerica, si limita a rappresentare fatti e processi in modo immateriale, attraverso un « oceano di 0 e di 1 » (Garapon, Lassègue, 2018).

Già da tali preliminari riflessioni è possibile dedurre le considerevoli ripercussioni per il giurista. Abituato a orientarsi in uno spazio e tempo definiti, che consentono l'emersione di concetti come soggettività, imputazione, responsabilità ecc., l'operatore giuridico si ritrova smarrito in un ecosistema di *unknown unknows*, dove non solo non si è a conoscenza di qualcosa, ma non si è a conoscenza nemmeno di quello di cui non si è a conoscenza.

Beninteso, non è la prima volta che la bussola del diritto non è di agevole consultazione nel tracciare la rotta per navigare la contemporaneità. È accaduto ogni volta che il mondo delle norme ha incontrato tecniche innovative in grado di trasformare i processi produttivi e sociali, alterando la necessaria separazione tra « regola e regolato » (Irti, Severino, 2001). Ma come sembra evocare lo stesso prefisso *cyber* — derivante da un lemma greco-antico (*kyber*) con il significato di « timone » — la questione del *chi è al timone? qual è la direzione?* sembra oggi emergere con rinnovato vigore, come testimoniato anche dalla prima Lettera Enciclica (*Magnifica humanitas*) del Santo Padre Leone XIV, resa pubblica il 15 maggio 2026 e dedicata proprio alla « custodia della persona umana nel tempo dell'intelligenza artificiale ». Quest'ultima segue, di 135 anni, quella di Leone XIII dedicata alle nuove questioni di quel tempo (*Rerum novarum*), e che diede origine alla Dottrina sociale della Chiesa. Oggi come allora, le « trasformazioni economiche sociali », il « conflitto tra capitale e lavoro », impongono una nuova riflessione che parta dalla lucida consapevolezza dei radicali mutamenti in corso.

In effetti, la rapida evoluzione delle tecnologie dell'informazione e della comunicazione ha reso possibile, in uno arco temporale decisamente contenuto, concentrazioni di dati (*Big Data*) che per volume, velocità di raccolta e varietà di fonti (Mayer-Schönberger & Cukier, 2014; Della Morte, 2018), erano semplicemente « unthinkable » sino a pochi anni prima (Wisnewski, 2016). Ma più del semplice stoccaggio è la capacità di estrarre dai dati l'informazione e la correlazione corretta — *rectius*: la più utile — a rappresentare l'obiettivo perseguito dagli attori globali, spesso privati. In questa competizione

per assicurarsi un vantaggio strategico sul fronte delle nuove economie estrattive (anche l'espressione *data mining* offre una suggestione etimologica: *to mine*, estrarre), l'irruzione di sistemi sempre più complessi di sistemi di intelligenza artificiale (IA) ha rappresentato un significativo acceleratore, se non un elemento di autentica discontinuità, che ha registrato l'emersione di due diversi, ma connessi, interrogativi: cosa regolare? E come?

Con riferimento a tale ultima questione, una precisazione di metodo appare necessaria. Perché ampiamente approfondito in altri lavori (sia concesso rimandare a DELLA MORTE, 2022, 2024 e 2025) non è stata qui richiamata la questione del 'livello' regolatorio, che non può che prescindere dai singoli ordinamenti statuali. Ci riferiamo ovviamente alla normativa internazionalistica, che pur nell'assenza di un'organizzazione internazionale e di un trattato di portata universale specificamente dedicati allo scopo — diversamente, ad esempio, dall'AI Act approvato in seno all'Unione europea, oppure dalla Convenzione quadro sull'intelligenza artificiale, approvata in seno al Consiglio d'Europa — è in parte deducibile da un mosaico di norme che 'incornicia' le normative nazionali. Detto mosaico è composto di elementi normativi profondamente diversi tra loro. Taluni si riferiscono al 'sostrato materiale', come nel caso della disciplina dei cavi sottomarini, deducibile dalle norme sul diritto del mare; altri si riferiscono ai 'profili software', come per l'autoregolazione delle piattaforme, spesso ispirata direttamente ad alcuni standard di protezione internazionale dei diritti umani; altri ancora hanno per oggetto la disciplina del contenuto, come nel caso dei dati sensibili frequentemente trattati in ambito sanitario. Ovviamente si tratta di forme tutela interconnesse, la cui disciplina integra aspetti di *hard law* e di *soft law*, spesso nella forma di codici di auto-condotta. L'esito del quale, seppure in divenire, deve essere sempre interpretato alla luce di realtà obiettive. A tale riguardo, nel « Final Report del High-level Advisory Body on Artificial Intelligence » (2024) si ricorda molto chiaramente come intere aree del mondo siano ancora escluse dalla *governance* (senza *government*) dell'intelligenza artificiale. Nel rapporto si evidenziano talune importanti iniziative regolatorie non riconducibili alle Nazioni Unite. Ebbene, solo sette Stati hanno sottoscritto tutte le iniziative considerate, mentre ben 118 Paesi — prevalentemente nel Sud globale — non partecipano ad alcuna di esse, segno che una significativa parte della popolazione mondiale, al momento, subisce gli effetti della tran-

sizione digitale, senza contribuire in alcun modo alla sua determinazione.

2. Le sfide poste dall'intelligenza artificiale.

Per comprendere più chiaramente la portata innovativa di tali sistemi appare consigliabile affrancarsi tanto dal lessico antropomorfizzato che cerca di rendere l'IA intelligibile — l'*intelligenza*, appunto, *artificiale*, le *allucinazioni*, ecc. (De Martin, nel presente volume, evoca a tale riguardo la « pareidolia », ovvero l'illusione subcosciente che porta a riconoscere forme familiari, come i volti, in oggetti casuali) — quanto da quello zoomorfizzato richiamato, al contrario, per ridimensionarne le aspettative (come nel caso dei « pappagalli stocastici » stigmatizzati da Bender nel 2021 con riferimento ai vari *Large Language Models*). In realtà quello che accade in simili frangenti, come rilevato da Zuboff nel 2019, è che a fronte di fenomeni inediti, la tendenza è quella di interpretarli attraverso il filtro di categorie note, finendo tuttavia per oscurarne proprio le caratteristiche innovative, come per le prime automobili definite « carrozze senza cavalli ».

Eppure « i sistemi di apprendimento automatico basati sull'utilizzo dei dati » (Finocchiaro, 2024) hanno segnato una significativa trasformazione nel rapporto tra uomo, tecnologia e produzione della conoscenza. Basterebbe questa sola circostanza per suggerire di non affrettare il passo, e soffermarsi a riflettere quanto meno su qualche osservazione preliminare, al fine di garantire un'adeguata profondità di campo.

Da tale prospettiva, osservando quanto accaduto nell'arco di pochi decenni, si possono individuare significative periodizzazioni. Una prima stagione — quella tra gli anni Cinquanta e gli anni Ottanta del secolo scorso — è stata segnata da un'intelligenza artificiale *old fashioned* o 'simbolica', ovvero leggibile dall'uomo e fondata su rappresentazioni formali della conoscenza e regole logico-deduttive (come nell'esempio dei cd. 'sistemi esperti' in ambito medico: se un paziente presenta i sintomi X e Y, allora è probabile la diagnosi Z). Da questa si è passati rapidamente a modelli di *machine learning* o di apprendimento automatico, in grado di inferire, ovvero di dedurre direttamente correlazioni dai dati (come nel caso di un algoritmo che, sulla base di una considerevole mole di e-mail, precedentemente classificate come spam o non spam, apprende in modo autonomo ad etichettarle in un senso o nell'altro). Sul filo di tale evoluzione, il *deep*

learning ha segnato un ulteriore passaggio, introducendo reti neurali profonde in grado di elaborare autonomamente rappresentazioni complesse di dati non strutturati (come nel caso del riconoscimento facciale, in cui un sistema apprende autonomamente tratti visivi, rendendo possibile l'identificazione di volti senza la necessità di regole esplicite a priori). Da ultimo, sono i sistemi generativi e agentici, a caratterizzare l'attuale grado di sviluppo dell'intelligenza artificiale. Con i sistemi generativi si è offerto al grande pubblico la possibilità di realizzare testi, immagini, audio o codici sulla base di semplici richieste (come nell'esempio dei *Large Language Models* capaci di produrre un approfondimento testuale o visivo di un'immagine caravaggesca, ispirandosi, ammesso che il termine sia opportuno, all'inconfondibile stile chiaroscuro di Michelangelo Merisi). Con i sistemi agentici si può addirittura ottenere l'esecuzione di sequenze di azioni finalizzate al raggiungimento di obiettivi complessi (come per la pianificazione di un'attività articolata quale l'organizzazione di un viaggio, che richiede la comparazione dei prezzi, l'identificazione dei voli, la prenotazione dell'albergo adeguato, l'indicazione delle principali attrazioni, ecc.).

Da questa breve disamina appare chiaro quale sia, allo stato attuale dello sviluppo tecnologico, il nodo cruciale del dibattito. Si tratta del « salto dall'automazione all'autonomia » (POLLICINO, nel presente volume), ovvero dalla « esecuzione di una decisione precedentemente definita dall'uomo » alla possibilità di compiere una scelta. Un salto che pone interrogativi radicali che esondano il piano strettamente tecnologico e che investono il campo epistemologico, etico, economico, e, non da ultimo, giuridico.

3. La progressiva emergenza del dialogo tra fatto e norma.

Sullo sfondo di tali considerazioni, nel quadro di un progetto di ricerca finanziato dall'Università Cattolica e dedicato ai poteri privati che hanno un impatto pubblico in campo digitale (*Funzioni pubbliche, poteri privati*: https://progetti.unicatt.it/fpcp), è progressivamente emersa l'esigenza di approfondire in che modo si ricorre, già oggi e concretamente, a tali sistemi nei diversi settori di pubblico interesse. L'obiettivo non era solo quello di evidenziare i risultati conseguiti in tema di efficientamento e di bilanciare questi ultimi con le fragilità e le problematiche riscontrate, ma anche di testare la tenuta

di schemi regolatori fondati su presupporti analogici e dunque non sempre in grado di 'catturare' l'essenza dell'evoluzione digitale.

Sino a che punto il salto tecnologico sia stato realizzato anche dal diritto, ovvero sino a che punto l'accelerazione del *fatto* sia stata accompagnata da quella della *norma*, ha rappresentato sin dall'inizio il sostrato dell'indagine, che si è poi concentrata su alcune aree di interesse generale che spaziano dal sistema mediatico a quello medico, da quello militare a quello biometrico, sino a quello tanatologico o artistico. Un simile approccio, oltre a consolidare un dialogo tra saperi all'interno di un Ateneo che ha il vanto di dodici Facoltà distribuite sul territorio nazionale, ha mutuato dall'esperienza delle *Law Clinic* il correlato passaggio da *Law in the Books* a *Law in Action*, propiziando una verifica sul campo. Si è quindi costituito un gruppo di ricerca i cui componenti, afferenti a nove Dipartimenti o Istituti diversi, hanno garantito, avvalorandosi anche di numerose collaborazioni esterne, un autentico approccio multidisciplinare. Quest'ultimo si è poi tradotto in una precisa scelta sul piano del metodo, realizzata attraverso un confronto *a specchio* tra un esperto del particolare settore materiale di volta in volta preso in esame con riferimento all'innovazione digitale (*il fatto*), e uno specialista di quel particolare regime regolatorio (*la norma*).

Ecco *il dialogo tra fatto e norma*. Esso ha trovato l'apice nel quadro di due convegni maggiori: uno svolto presso la Fondazione Policlinico Universitario "A. Gemelli" IRCCS, a Roma, il 4 giugno 2024; e uno ospitato presso l'Aula Magna dell'Università Cattolica di Milano, nei giorni 28 e 29 ottobre 2024.

4. Struttura e contenuto del presente volume.

Ciò premesso, lo schema del volume incorpora, con gli opportuni adeguamenti, le relazioni presentate nel quadro di tali occasioni, e valorizza il metodo del confronto tra fatto e norma. Dopo avere rammentato, nei « Saluti introduttivi », la necessità di preservare quei dettagli, insieme ordinari e straordinari, che « non potranno mai stare tra gli algoritmi » — come « tracciare un disegno al controluce di una finestra [o] giocare la prima partita di calcio con un pallone di pezza » (come ricorda GIULIODORI, evocando poetiche riflessioni di Papa Francesco) — si sottolinea quanto ambiziosa sia la sfida posta al diritto di « governare l'innovazione senza subirla » (SOLIMANO). Prova tanto più impegnativa in quanto impone la costruzione di un lessico

condiviso che integri anche la dimensione etica e culturale, senza la quale «l'intero processo di innovazione rischia di frammentarsi, con il pericolo di creare tecnologie formalmente avanzate ma incapaci di generare fiducia e sicurezza» (GASBARRINI).

Con le due relazioni di apertura (*Keynote Speech*) si inaugura il metodo del confronto diretto: dapprima la disamina del fatto dal punto di vista tecnico (in questo caso ingegneristico-informatico); e a seguire un'attenta analisi giuridica (nella specie, in ottica pubblicistica). Con la prima relazione, DE MARTIN, dopo avere ricordato l'origine dell'espressione *intelligenza artificiale*, per lo più legata a esigenze di finanziamento, ricostruisce gli elementi che ne hanno accompagnato la crescente affermazione sulla scena internazionale, accentuando gli elementi autenticamente innovativi a fronte della imponente costrutto retorico (in fondo, avverte: «l'esperimento mentale da fare per immunizzarsi dalla tentazione di antropomorfizzare l'IA è quello di immaginare che i testi dei 'chatbot' escano da una lavatrice o da un tostapane»). Nel contrappunto di POLLICINO, si indicano alcune decisivi concetti-chiave. Dal momento che «il potere non è più soltanto verticale [bensì] assume forme diagonali e orizzontali», come nel caso delle grandi piattaforme, «il positivismo giuridico più rigido, quello che si accontenta della forma, serve a pochissimo». Pur senza eccedere «nel novismo digitale», poiché spesso le categorie esistenti «se interpretate evolutivamente» risultano ancora adeguate, occorre superare il «vecchio dilemma tra hard law e soft law». In altre parole, perché il sistema è quello co-regolatorio «il potere non sta più solo nella norma, ma nei processi, negli standard, nei codici». È questa la ragione per cui «mettere stabilmente in dialogo chi studia la tecnologia e chi studia il diritto non è un vezzo interdisciplinare, ma una condizione di possibilità per capire cosa stiamo davvero regolando».

Premesse tali considerazioni generali, gli approfondimenti settoriali si sono concentrati su alcuni temi il cui pubblico interesse è apparso di particolare rilievo. La prima questione affrontata, anche in considerazione dell'impatto sempre più significativo nei processi politici, è stata la disinformazione, e a seguire le problematiche legate all'immortalità del dato, ai rischi del riconoscimento facciale e biometrico, alle applicazioni in ambito militare, a quelle in ambito sanitario e infine... a quello che l'intelligenza artificiale non può fare, ovvero il gesto artistico in senso stretto, come spiegato — *rectius*: ascoltato —

in un brillante dialogo tra un esperto musicologo e un violinista virtuoso con il suo Stradivari.

4.1. (Segue). *La disinformazione.*

Cominciando dalla sezione del volume dedicata alla disinformazione, dopo avere introdotto il tema attraverso alcune utili distinzioni (misinformazione e disinformazione, MASCHERONI e CARLO), ODDENINO acutamente osserva come il concetto di verità appare oggi talmente esposto al rischio di obsolescenza da risultare popolare « il ricorso alla nozione di 'post-verità', quasi che la partita per preservare la distinzione fra vero e falso fosse irrimediabilmente perduta ». Sino a che punto tale rischio è reale lo illustra, sempre nel volume, GALLOTTI, coordinatore di un progetto europeo da 7 milioni di euro volto al contrasto dei fenomeni *fake* e *deep fake*. Nella testimonianza di quest'ultimo si indica in modo molto intelligibile quali sono le difficoltà nel rendere operative simili esperienze, a fronte di piattaforme che non garantiscono l'accesso ai propri dati. E si pone l'accento sul rischio di sbilanciamento tra il mondo della ricerca e quello della imprenditoria: « noi siamo quelli che vanno in giro a chiedere 'per piacere', mentre altre persone prendono tutto e lo mettono nei loro modelli senza domandare alcun permesso, e poi [...] lo chiamano ChatGPT ». Contrastare simili pratiche risulta, al momento, di difficile realizzazione. Tanto più che, come osserva nuovamente ODDENINO, il tecno-soluzionismo, ovvero l'affidare alla tecnologia il ruolo di « piaga e panacea, *pharmakon* che avvelena e in parallelo si propone con doti curative » non è sua volta privo di rischi.

4.2. (Segue). *'L'immortalità' del dato.*

Passando all'esame dell'immortalità del dato, la premessa è quella per cui le piattaforme si configurano come ecosistemi sociali totalizzanti, nei quali lavoro, relazioni affettive, consumo ed elaborazione del lutto coesistono e si influenzano reciprocamente. Come rileva PASQUALI la pandemia ha « accelerato drammaticamente » questi processi, dal momento che a fronte di una progressiva sottrazione dei corpi biologici dalla pubblica scena — letteralmente per confinamento — essa ha permesso uno scatenamento dei nostri corpi elettronici (*walking data generator*). Si tratta di una sostituzione più profonda di quanto appaia *prima facie*, dal momento che, come ricordato *supra*, essa non investe unicamente la dimensione spaziale, ma anche

quella temporale. Ne è un esempio clamoroso, ancorché poco evidente, la questione dei dati *post mortem*, per lo più gestiti attraverso semplici codici di auto-condotta da parte delle grandi piattaforme. L'ampio campo della *Digital Death* brillantemente arato dal dialogo tra SISTO, per il 'fatto', e ZICCARDI, per la 'norma', appare, in tale contesto emblematico. Il valore, economico e non solo economico, di questi « cadaveri informazionali » (così ÖHMAN, 2024 citato da SISTO) è destinato a crescere, in corrispondenza del progressivo numero. Analogamente sono destinati a crescere gli innumerevoli problemi giuridici posti da queste nuove pratiche, espressione di un ambiente informativo in cui il dato, e quindi la memoria, non è più consegnata al deterioramento e al conseguente oblio, come è sempre accaduto in passato, ma « si è spostata in spazi digitali caratterizzati da capacità di archiviazione potenzialmente illimitata, e dalla possibilità di replicazione istantanea dei contenuti » (così FERRARIS, 2021, citato in ZICCARDI). In fondo, il tema ampio è quello per cui una volta che il dato è stato immesso in rete è difficilmente eliminabile: al più può essere reso inaccessibile o oscurato, ma la sua completa cancellazione risulta, nella pratica, una chimera (soprattutto se si considera che « i *data center* che ospitano copie dei dati sono distribuiti in Paesi con normative differenti in materia di privacy, sicurezza e diritti fondamentali, ponendo problemi di giurisdizione e di sovranità dei dati », così ZICCARDI richiamando ARESU, 2024).

4.3. (Segue). *Il riconoscimento facciale e biometrico.*

Sempre con riferimento ai rischi dello stoccaggio dei dati, il volume esamina il riconoscimento facciale e biometrico, tecnica adoperabile per varie finalità, ma che, se associata al rilevamento di tratti « assunti o costruiti come predittivi di comportamenti criminali o comunque di pericolosità sociale, non solo [è] spesso arbitraria e dettata da pregiudizi, ma [porta] con sé anche l'effetto di perpetuazione e consolidamento di tali visioni preconcette » (così FORTI, introducendo il tema). Ci troviamo qui, come messo chiaramente in luce in un ammirevole approfondimento del *fatto* ad opera di CABITZA e MARCONI, in una rappresentazione emblematica della dialettica tra fatto e norma, essendo tale pratica « produttrice di 'fatti' poiché traduce fenomeni percepibili in dati simbolici; ed è, più propriamente, riproduttrice di 'norme', poiché il suo impiego è oggi connesso a pratiche e decisioni di natura 'semi-automatica', integrate in un ordina-

mento che ne disciplina finalità, condizioni d'uso e responsabilità». Questo processo di trascrizione simbolica, che sulla base di «una sequenza di operazioni converte dati sensoriali in rappresentazioni computabili e, infine, in decisioni o valutazioni dotate di valore operativo» (*ibid.*), richiede tuttavia lo svelamento della natura interpretativa che ne è alla base. «Riconoscere la natura costruttiva e inferenziale — in senso computazionale — del dato biometrico significa, in ultima analisi, spostare il problema dal *quanto il sistema vede* al *come il sistema codifica ciò che vede*: un passaggio necessario per attribuire significato, responsabilità e limiti a ogni misura automatizzata (così Floridi, 2019 cit. in *ibid.*).

Ma come «definire con precisione la funzione e il contesto d'uso», come «interpretare il comportamento del sistema alla luce dei costi e delle ricadute delle decisioni che ne derivano» (Cabitza e Marconi)? Per questo occorre un approfondimento normativo che permetta di comprendere — come rileva, nel volume, Mobilio — non solo in che misura il ricorso a tali tecnologie muti a seconda delle regioni del mondo — passando dalle esperienze di *Social Credit System* sperimentate in alcune province della Cina agli *smart borders* presenti anche in Unione europea — ma anche come la disciplina più avanzata appaia al momento soggetta a un notevole margine di apprezzamento. È questo il caso dell'art. 5, par. 1, del regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale (*AI Act*) —, che oltre a prevedere taluni divieti specifici (come nel caso dei sistemi di intelligenza artificiale per inferire le emozioni di una persona fisica nell'ambito del luogo di lavoro e degli istituti di istruzione), sancisce, da un lato, il divieto generale di sistemi di identificazione biometrica remota «in tempo reale» in spazi "accessibili al pubblico" per «attività di contrasto» (*ibid.*, lett. h), prevedendo, al contempo, una notevole serie di eccezioni di ordine securitario (terrorismo, tratta di esseri umani, ecc.)

4.4. (Segue). *L'ambito militare.*

Il tema securitario si collega anche alle tecnologie in ambito militare, che, da un lato, recano notevoli benefici (ad esempio, sul piano dell'approvvigionamento di beni, alimenti, medicine eccetera), dall'altro, come correttamente individuato da Spatti nell'introduzione alla sezione, non garantiscono che «la maggiore efficienza operativa

[coincida] necessariamente con una maggiore aderenza ai principi del diritto internazionale umanitario». La domanda cruciale «non è dunque se l'IA possa essere utilizzata, ma a quali condizioni il suo impiego possa rimanere compatibile con i principi di distinzione, proporzionalità e precauzione che costituiscono l'ossatura del diritto internazionale umanitario».

L'interrogativo assume un particolare rilievo nel caso di 'sistemi ad arma autonoma', ovvero nelle ipotesi in cui «la macchina non è più un semplice strumento passivo, ma un agente tecnico che incide in modo determinante sull'esito dell'azione». Davanti a tali sfide, l'orientamento dottrinale non appare univoco, e a fronte di chi ritiene che tali armi, «nella forma pienamente autonoma, siano incompatibili con il diritto internazionale umanitario vigente (così Asaro, 2012; e Garcia, 2023), altri autori adottano una posizione più prudente» (Winter, 2022; Casey-Maslen, 2025, tutti cit. in Spatti)». Il punto autenticamente dirimente per valutare, caso per caso, la corretta autonomia del sistema, resta il *Meaningful Human Control* (controllo umano significativo) «inteso non come mera facoltà di attivare o disattivare il sistema, ma come reale capacità di comprendere, monitorare e, se necessario, correggere il processo decisionale dell'arma» (Asaro, 2012, *ibid.*).

Su tali profili, l'esame del fatto condotto da Ricchetti e Sironi De Gregorio — in sostituzione del Generale di Divisione Bertoncello, intervenuto al Convegno — pone distintamente in luce come «le difficoltà applicative del diritto internazionale umanitario nell'era dell'IA non derivano primariamente da lacune normative, bensì dall'incontro tra norme che presuppongono un agente decisionale umano e sistemi tecnici il cui funzionamento interno è spesso opaco anche ai loro progettatori». Ancora una volta ci troviamo in uno spazio grigio tra fatto e norma, dove precisare il *Meaningful Human Control* come requisito giuridico «senza specificarne i correlati tecnici rischia di produrre sistemi formalmente conformi ma sostanzialmente privi di supervisione umana reale». Dove, in altri termini: il nodo tecnico «si riverbera direttamente sulle categorie giuridiche di responsabilità, prevedibilità e controllo» (così Amoroso-Tamburrini, 2020, e Comitato Internazionale della Croce Rossa, 2019, cit. in Ricchetti e Sironi De Gregorio).

In che modo il principio di proporzionalità, che richiede una valutazione *ex ante* dei danni collaterali a civili e beni civili a fronte del previsto vantaggio militare concreto e diretto, regga l'urto del pas-

saggio dal mondo analogico a quello dei bit resta una questione aperta. Rispetto alla quale, con voce autorevole ma parzialmente fuori dal coro, SASSOLI risponde attraverso uno scrupoloso esame della normativa e della prassi, per concludere nel senso di non condividere l'opinione di quanti ritengono che i sistemi ad arma autonoma violino — a prescindere — il diritto internazionale umanitario perché, pur potendo imitare la cognizione umana, sarebbero privi di qualità come emozione, compassione e buon senso, indispensabili per valutare obiettivi e proporzionalità. Al contrario, secondo l'A., tali decisioni, consistendo nella classificazione di individui in base al loro status o comportamento potrebbero, quanto meno in linea teorica, risultare assicurate con maggiore obbiettività da una formula condivisa in grado di modellizzare l'infinita varietà di situazioni possibili. « Perché dovrebbe avere importanza il soggetto (in altre parole, il soldato, la sua età, il suo sesso, la sua istruzione, la sua religione o la sua etica) e non solo l'oggetto (cioè il vantaggio militare e l'effetto incidentale sui civili)? » (*ibid.*). Il provocatorio interrogativo schiude un amplissimo orizzonte di dibattito, rispetto al quale l'A. stesso avverte l'esigenza di specificare che non si potrà ricorrere a tali sistemi fintanto che non saranno in grado di essere utilizzati in conformità con il diritto internazionale umanitario. Ma, aggiunge, « realisticamente » nessun comandante impiegherebbe armi prive di controllo, capaci di colpire le proprie forze o favorire il nemico. Per tale ragione, anche qualora venissero sviluppate armi completamente autonome, produttori e utilizzatori cercherebbero di garantirne il funzionamento conforme alle istruzioni umane. La questione della loro legalità potrebbe quindi ridursi alla possibilità tecnica di assicurare tale controllo. Ritrovandoci, ancora una volta, alla frontiera tra fatto e norma.

4.5. (Segue). *La tutela della salute.*

Si tratta della stessa soglia che delimita il dibattito intorno all'utilizzo di sistemi di intelligenza artificiale in ambito sanitario, che grazie al generoso apporto di numerosi colleghi del Policlinico Gemelli è stato il tema maggiormente approfondito nell'ambito della ricerca. Un tema articolato, anche perché, come posto in evidenza da BALDUZZI, costituzionalista e già Ministro della salute, introducendo la sessione: « i rischi associati all'intelligenza artificiale in ambito sanitario si suddividono in due categorie: quelli comuni ad altri settori e quelli specifici del contesto medico ». Così, da un lato, ritroviamo in

ambito sanitario quanto era stato in precedenza evocato con riferimento ai numerosi problemi connessi alla datificazione del mondo: *bias*, opacità, ecc. Dall'altro, il menzionato salto dall'automatismo all'automazione — ovvero, la « capacità », per il sistema, « di *prendere* la decisione » (così SIMONCINI, nel volume) — si può agevolmente tradurre, in campo sanitario, in un rischio di affidamento eccessivo, tanto da parte dei medici quanto degli operatori sanitari. Si pone così il tema « dell'ultima parola » (sempre BALDUZZI), il cui dinamismo circolare, ben noto a chi riflette sulla vita delle istituzioni, deve essere qui re-interpretato nel rapporto trilaterale: uomo / macchina / paziente.

Il 'fatto', è che già oggi i sistemi di intelligenza artificiale trovano applicazioni decisamente estese in ambito sanitario, operando, come specificato da DELL'OSSO e quindi da OLIVA e CAZZATO, su un duplice livello: clinico e amministrativo. Sul piano clinico, supporta diagnosi, prognosi, scelte terapeutiche, pianificazione chirurgica e monitoraggio continuo, con applicazioni particolarmente efficaci nell'analisi di immagini radiologiche, nella medicina d'urgenza, nella neurologia e nella gestione delle patologie croniche. Su quello amministrativo, l'IA, grazie anche alla valutazione predittiva del rischio, contribuisce all'ottimizzazione delle risorse e alla riduzione del carico operativo attraverso l'automazione dei processi e la gestione efficiente dei dati clinici. La condizione perché tali sistemi forniscano un valido supporto, però, resta quello dell'adozione di un'intelligenza artificiale fondata su dataset di qualità, nonché caratterizzata da una capacità di adattamento a diversi contesti, e sottoposta a un monitoraggio continuo post-implementazione.

Come accennato, tra le specializzazioni sanitarie che attualmente registrano un esponenziale interesse nei confronti dell'IA, è presente, oltre alla neurologia — che, peraltro, incrocia la frontiera estrema delle neurotecnologie, ovvero di quei « devices, systems and procedures, encompassing both hardware and software, that directly measure, access, monitor, analyse, predict or modulate the nervous system to understand, influence, restore or anticipate its structure, activity and function » (come definite nella « Recommendation UNESCO on the Ethics of Neurotechnology » del 2025) — l'oncologia, che grazie al contributo della Prof.ssa FAGOTTI e del Dott. ROSATI, ha raffigurato il campo di applicazione elettivo del dialogo tra saperi ospitato nel volume. In particolare, in un interessante studio che questi ultimi dedicano ai problemi pratici della standardizzazione decisio-

nale e della chirurgia di precisione nel trattamento del carcinoma ovarico avanzato, si evidenzia con molta precisione quali sono i benefici delle applicazioni IA in ambito clinico, qui non più intesa come «sostituto della capacità di giudizio del clinico, ma come un'estensione della sua possibilità di leggere e comprendere segnali complessi o pattern non lineari che eccedono la capacità cognitiva umana e le tecniche di statistica classica». In effetti, «gran parte dell'informazione prodotta nei processi di cura rimane confinata nei testi non strutturati: referti radiologici, note laparoscopiche, lettere di dimissione, discussioni multidisciplinari [e, in tale contesto] il natural language processing consente di estrarre informazioni chiave da questi testi e di integrarle in pipeline predittive» (*ibid.*). In altre parole: «il futuro dell'oncologia ginecologica sarà sempre più guidato da modelli integrati capaci di sintetizzare testo, immagine, genetica, dinamiche biologiche e riscontri intraoperatori in una rappresentazione multidimensionale e continuamente aggiornata della malattia» (*ibid.*). Tuttavia, specificano gli autori, e si tratta di una puntualizzazione cui riservare la più grande attenzione: «[i]n questa visione, il ruolo del clinico non viene sminuito, ma ampliato: la tecnologia non sostituisce la competenza umana, ma la sostiene, la rafforza e la rende più efficace».

Quanto appena riportato, ovvero l'attenzione posta al bilanciamento tra rischi e benefici nel ricorso a sistemi di intelligenza artificiale in campo sanitario, trova riscontro anche negli interventi di RIVA e di RUZZANTE, dedicati, rispettivamente, alla prospettiva della cyberpsicologia e della tutela dei dati. Nel primo intervento si indaga come le tecnologie digitali stanno trasformando il modo in cui «gli esseri umani vivono, pensano, ricordano, comunicano e si relazionano», ponendo particolare accento all'interessante categoria del «paradosso comfort-crescita». Nel secondo, si approfondisce la questione del consenso informato, interrogandosi sulla tenuta delle tradizionali garanzie a fronte di «sistemi decisionali automatizzati, caratterizzati da logiche algoritmiche spesso imperscrutabili per il paziente e talvolta anche per il personale sanitario».

Risulta evidente che a fronte di tali considerazioni gli interrogativi di natura giuridica sono numerosi, e investono tanto profili di carattere pubblicistico quanto di stampo civile e penale. Sul primo fronte si segnalano i già menzionati interventi di BALDUZZI e SIMONCINI. Sul secondo, si concentrano le riflessioni introduttive di ALBANESE, e quelle più specifiche di VALITUTTI e di FACCIOLI, aventi ad oggetto, ri-

spettivamente, il significato da attribuirsi al concetto di responsabilità e di illecito in ambito sanitario, e le difficoltà ricostruttive del nesso eziologico in caso di pretesa risarcitoria nei confronti del medico piuttosto che della struttura sanitaria o di entrambi. Quanto, infine, al fronte penale, esso è preso in considerazione, oltre che nelle richiamate note introduttive di DELL'OSSO, nell'articolato intervento di MASSARO dedicato al concetto di colpa nell'interazione uomo-macchina alla luce, in particolare, del « principio di dominabilità ».

4.6. (Segue). *La creatività costretta (e liberata).*

In ''guisa di conclusione'', il volume ospita il confronto più originale: quello dedicato a quanto *l'intelligenza artificiale non può fare*, il gesto artistico, o creativo in senso forte. Per questa sessione si è sospeso il metodo accademico del confronto attraverso relazioni, e si è lasciato spazio all'improvvisazione attraverso un 'dialogo concertato' tra il Prof. REGGIANI, Ordinario di Letteratura Inglese e Direttore dello *Studium Musicale* di Ateneo, e il Maestro VON ARX, Virtuoso e Direttore di Fondazione Casa Stradivari di Cremona, che si è presentato in Aula Magna con il suo violino Stradivari ''The Angel'' - ex *Madrileno 1720*.

È un peccato che la scrittura, questa remota forma di intelligenza artificiale risalente al 3000 a.C., non possa rendere giustizia dei numerosi intermezzi musicali del Maestro, che ha dapprima incantato il pubblico con Bach, Vivaldi, Cajkovskij... per poi disincantarlo con l'esecuzione di *Genesis*, un brano tanto altisonante nel titolo quanto mediocre nel contenuto, e soprattutto composto da AIVA (*Artificial Intelligence Virtual Artist*), un sistema di intelligenza artificiale 'compositrice' riconosciuto dalla società degli autori ed editori di musica francese.

Quello tra creazione (qui intesa come fatto) e mera riformulazione (qui intesa come norma, canone) ci è parso un ulteriore spazio degno di essere indagato. « Mi farebbe piacere sapere », si domanda REGGIANI riferendosi ad AIVA, « quali repertori le hanno fatto ingurgitare (studiare o leggere mi sembrano verbi impropri) ». E VON ARX, di rimando: ciò che colpisce è che in simili algoritmi non siano mai presenti « il brutto, l'osceno o l'errore [...]. Sarebbe stato opportuno sviluppare anche un algoritmo 'invertito' che possa educare le macchine attraverso l'imperfezione ».

Educare le macchine al dubbio e all'imperfezione, immaginare la creazione di modelli trattati « come un testo aperto » che possano ricondurre « la decisione automatizzata a essere argomento, non dogma » (Cabitza e Marconi), ecco una 'nota alta' sulla quale chiudere la presente introduzione. Si tratta di declinare il rapporto tra fatto e norma non come un destino, ma come una scelta, capace di custodire, nonostante tutto, il rischio e la forza del libero arbitrio, come suggerito dall'ineffabile poesia in esergo.

Bibliografia

Albanese A. (2026), *Introduzione ai profili civilistici*, nel presente volume, p. 211 ss.

Asaro P. (2012), *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, in *International Review of the Red Cross*, 886, p. 687 ss.

Balduzzi R. (2026), *Introduzione ai profili costituzionali*, nel presente volume, p. 207 ss.

Bender E.M. (2021), *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT'21), Association for Computing Machinery, p. 610 ss.

Cabitza F., Marconi L. (2026), *Intelligenza artificiale e riconoscimento facciale: tra fatto, norma e interpretazione*, nel presente volume, p. 115 ss.

Casey-Maslen S. (2025), *Autonomous Weapons Systems Under International Law*, in Talves K.-Spreen D. (eds.), *Artificial Intelligence in Military Technology. Social, cultural and ethical perspectives*, Cham, Spinger, p. 161 ss.

De Martin J.C. (2026), *L'intelligenza artificiale: Il fatto*, nel presente volume, p. 17 ss.

Della Morte G. (2025), *La regolazione dell'AI: profili internazionalistici*, in Pollicino O., Donati F., Finocchiaro G., Paolucci F., *La disciplina dell'intelligenza artificiale*, Milano, Giuffrè Francis Lefebvre, 2025, p. 67 ss.

Id. (2024), *Regolare il futuro. Diritto internazionale e nuove tecnologie*, in *Cento anni di dialogo. La Facoltà di Giurisprudenza dell'Università Cattolica - 1924-2024*, Milano, Vita e Pensiero, 2024, p. 145 ss.

Id. (2022), *Limiti e prospettive del diritto internazionale del cyberspazio*, in *Rivista di diritto internazionale*, 1, 2022, p. 5 ss.

Id. (2018), Big data *e protezione internazionale dei diritti umani. Regole e conflitti*, Napoli, Editoriale scientifica.

Dell'Osso A. (2026), *Introduzione ai profili penalistici*, nel presente volume, p. 215 ss.

Faccioli M. (2026), *Intelligenza artificiale in medicina e responsabilità civile della struttura sanitaria*, nel presente volume, p. 297 ss.

FINOCCHIARO G. (2024), *Intelligenza artificiale. Quali regole?*, Bologna, Il Mulino, 2024.
FLORIDI L. (2019), *The Logic of Information. A Theory of Philosophy as Conceptual Design*, Oxford, Oxford University Press.
FORTI G. (2026), *Introduzione*, nel presente volume, p. 111 ss.
GALLOTTI R. (2026), *AI4TRUST: sfide di un progetto (di Fact Checking) per usare l'IA contro Fake e Deep Fake*, nel presente volume, p. 47 ss.
GARAPON A., LASSÈGUE J. (2018), *Justice digitale. Révolution graphique et rupture anthropologique*, Paris, PUF (trad. it. (2021): *La giustizia digitale. Determinismo tecnologico e libertà*, Bologna, Il Mulino).
GARCIA D. (2023), *The AI Military Race. Common Good Governance in the Age of Artificial Intelligence*, Oxford, Oxford University Press.
GASBARRINI A. (2026), *La prospettiva della Digital Health. Saluti introduttivi al Convegno: "Digital Health, Artificial Intelligence: quali responsabilità verso i pazienti?"*, Fondazione Policlinico Universitario "A. Gemelli" IRCCS, Roma, 4 giugno 2024, nel presente volume, p. 11 ss.
GIULIODORI C. (2026), *Saluti introduttivi al Convegno: "L'intelligenza artificiale tra fatto e norma"*, Università Cattolica, Milano, 28 e 29 ottobre 2024, nel presente volume, p. 3 ss.
IRTI N., SEVERINO E. (2001), *Dialogo su diritto e tecnica*, Roma-Bari, Laterza.
MASCHERONI G., SIMONE C. (2026), *Introduzione*, nel presente volume, p. 31 ss.
MASSARO A. (2026), *Robot intelligenti e attività medica: presupposti e (soprattutto) limiti della responsabilità penale dell'operatore sanitario*, nel presente volume, p. 285 ss.
MOBILIO G. (2026), *La disciplina sulle tecnologie di riconoscimento facciale, ovvero il tentativo di limitare un formidabile potere di sorveglianza*, nel presente volume, p. 139 ss.
ODDENINO A. (2026), *La (post) verità fra pubblico e privato*, nel presente volume, p. 39 ss.
ID. (2026), *L'ardua ricerca di una regolazione effettiva del disordine informativo nell'era dell'intelligenza artificiale generativa*, nel presente volume, p. 55 ss.
ÖHMAN K. (2024), *The Afterlife of Data. What Happens to Yours Information When You Die and Why You Should Care*, Chicago, The University of Chicago Press.
OLIVA A., CAZZATO F. (2026), *L'impatto dell'IA in ambito sanitario*, nel presente volume, p. 221 ss.
PASQUALI F. (2026), *Introduzione sulla memoria e sopravvivenza algoritmica nell'era della mediatizzazione profonda*, nel presente volume, p. 75 ss.
POLLICINO O. (2026), *L'intelligenza artificiale: La norma*, nel presente volume, p. 23 ss.
RAZZANTE R. (2026), *Intelligenza Artificiale e trattamenti sanitari: tra consenso informato, autodeterminazione terapeutica e protezione dei dati sensibili*, nel presente volume, p. 311 ss.
RICCHETTI R., DE GREGORIO F. (2026), *Intelligenza Artificiale, Conflitti Armati e Diritto: Profili Tecnici ed Etici*, nel presente volume, p. 171 ss.

RIVA G. (2026), *L'IA tra "fatto predittivo" e "norma evolutiva": la prospettiva della Cyberpsicologia*, nel presente volume, p. 241 ss.
ROSATI A., FAGOTTI A. (2026), *Intelligenza Artificiale nel Carcinoma Ovarico Avanzato: integrazione multimodale, standardizzazione decisionale e chirurgia di precisione*, nel presente volume, p. 231 ss.
SASSOLI M. (2026), *Le Armi Autonome e il Diritto Internazionale Umanitario: Profili Regolatori*, nel presente volume, p. 187 ss.
SIMONCINI A. (2026), *Intelligenza Artificiale e Salute: profili pubblicistici*, nel presente volume, p. 257 ss.
SISTO D. (2026), *Thanabots. La problematica "sopravvivenza" post mortem tramite l'AI*, nel presente volume, p. 79 ss.
SOLIMANO S. (2026), *L'intelligenza artificiale in chiaroscuro. Saluti introduttivi al Convegno: "L'intelligenza artificiale tra fatto e norma"* (Università Cattolica, Milano, 28 e 29 ottobre 2024), nel presente volume, p. 9 ss.
SPATTI M. (2026), *Introduzione: sfide e prospettive per il diritto internazionale umanitario*, nel presente volume, p. 159 ss.
VALITUTTI A. (2026), *L'impiego di Sistemi Autonomi in Medicina e le Responsabilità del Sanitario*, nel presente volume, p. 269 ss.
VON ARX F. e REGGIANI E. (2026), *Dialogo 'concertato'. Quello che l'intelligenza artificiale non può fare*, nel presente volume, p. 331 ss.
WINTER E. (2022), *The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law*, in *Journal of Conflict and Security Law*, 1, p. 1 ss.
ZICCARDI G. (2026), *Immortalità del dato, eredità digitale e intelligenza artificiale: profili giuridici e tecnologici tra oblio e memoria*, nel presente volume, p. 91 ss.
ZUBOFF S. (2019), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, New York, Public Affairs (trad. it. *Il capitalismo della sorveglianza. Il futuro dell'umanità nell'era dei nuovi poteri* (2019), Roma, Luiss University Press).

SALUTI INTRODUTTIVI

SALUTI INTRODUTTIVI AL CONVEGNO: "L'INTELLIGENZA ARTIFICIALE TRA FATTO E NORMA"

(Università Cattolica, Milano, 28 e 29 ottobre 2024)

di *Claudio Giuliodori*

L'intelligenza Artificiale è uno dei fattori di maggiore trasformazione a cui assistiamo nel nostro tempo sia per la sua indubbia capacità di guidare l'innovazione in molteplici campi dell'agire umano sia per le sue potenzialità di auto apprendimento ancora tutte da valutare. Possiamo dire che non c'è oramai campo in cui non si sviluppino applicazioni basate sull'IA. Questa penetrazione diffusa e pervasiva in ogni contesto, oltre a generare un grande fascino e interesse, sta facendo emergere non poche questioni di carattere etico e giuridico, richiedendo quindi anche valutazioni di impatto e regole di gestione. Il presente convegno affronta pertanto aspetti di stringente attualità e offre preziosi contributi per l'approfondimento delle diverse problematiche.

La Chiesa segue con particolare attenzione gli sviluppi in atto e si moltiplicano gli interventi e i contributi del suo insegnamento anche ai più alti livelli magisteriali. In merito al nostro tema può essere utile riprendere un intervento di Papa Francesco rivolto ai partecipanti alla Plenaria della Pontificia Accademia per la Vita in cui sollecitava l'elaborazione di una politica di ampio respiro attorno allo sviluppo dell'IA:

« Esiste una dimensione politica nella produzione e nell'uso della cosiddetta 'Intelligenza Artificiale', che non riguarda solo la distribuzione dei suoi vantaggi individuali e astrattamente funzionali. In altri termini: non basta semplicemente affidarci alla sensibilità morale di chi fa ricerca e progetta dispositivi e algoritmi; occorre invece creare corpi sociali intermedi che assicurino rappresentanza alla sensibilità etica degli utilizzatori e degli educatori. Sono molte le competenze che intervengono nel processo di elaborazione degli apparati tecnologici (ricerca, progettazione, produzione, distribuzione, utilizzo individuale e collettivo), e ognuna comporta una specifica responsabilità.

Si intravede una nuova frontiera che potremmo chiamare '*algor-etica*' (cfr. *Discorso ai partecipanti al Congresso "Child Dignity in the Digital World"*, 14 novembre 2019). Essa intende assicurare una verifica competente e condivisa dei processi secondo cui si integrano i rapporti tra gli esseri umani e le macchine nella nostra era. Nella comune ricerca di questi obiettivi, i principi della Dottrina Sociale della Chiesa offrono un contributo decisivo: dignità della persona, giustizia, sussidiarietà e solidarietà. Essi esprimono l'impegno di mettersi al servizio di ogni persona nella sua integralità e di tutte le persone, senza discriminazioni né esclusioni. Ma la complessità del mondo tecnologico ci chiede una elaborazione etica più articolata, per rendere questo impegno realmente incisivo» (28 febbraio 2020).

Non è un caso che il Pontefice abbia voluto dedicare a questo tema anche il Messaggio per la LVII Giornata Mondiale della Pace del 1° gennaio 2024 titolandolo appunto '*Intelligenza artificiale e pace*', proprio per sottolineare le implicazioni davvero planetarie che il tema assume persino di fronte alle sfide più delicate come quella della convivenza tra i popoli. Ma la trattazione più ampia e articolata la troviamo nel discorso tenuto da Papa Francesco alla Sessione del G7 sull'Intelligenza Artificiale che si è svolto a Borgo Egnazia (Puglia) il 14 giugno 2024. In quel contesto la riflessione è stata molto ampia spaziando dall'orizzonte sapienziale di matrice biblica a quello più strettamente antropologico ed etico senza trascurare gli aspetti giuridici e politici. Vale la pena di rileggere alcuni passaggi che possono illuminare lo svolgimento del presente convegno.

«La Sacra Scrittura attesta che Dio ha donato agli uomini il suo Spirito affinché abbiano "saggezza, intelligenza e scienza in ogni genere di lavoro"» (*Es* 35,31).

La scienza e la tecnologia sono dunque prodotti straordinari del potenziale creativo di noi esseri umani. Quest'ultima, come è noto, è uno strumento estremamente potente, impiegato in tantissime aree dell'agire umano: dalla medicina al mondo del lavoro, dalla cultura all'ambito della comunicazione, dall'educazione alla politica. Ed è ora lecito ipotizzare che il suo uso influenzerà sempre di più il nostro modo di vivere, le nostre relazioni sociali e nel futuro persino la maniera in cui concepiamo la nostra identità di esseri umani.

Il tema dell'intelligenza artificiale è, tuttavia, spesso percepito come ambivalente: da un lato, entusiasma per le possibilità che offre, dall'altro genera timore per le conseguenze che lascia presagire. Non possiamo, del resto, dubitare che l'avvento dell'intelligenza artificiale

rappresenti una vera e propria rivoluzione cognitivo-industriale, che contribuirà alla creazione di un nuovo sistema sociale caratterizzato da complesse trasformazioni epocali.

Proprio questo vigoroso avanzamento tecnologico rende l'intelligenza artificiale *uno strumento affascinante* e *tremendo* al tempo stesso ed impone una riflessione all'altezza della situazione. In tale direzione forse si potrebbe partire dalla costatazione che l'intelligenza artificiale è innanzitutto *uno strumento*. E viene spontaneo affermare che i benefici o i danni che essa porterà dipenderanno dal suo impiego.

Solo se sarà garantita la loro vocazione al servizio dell'umano, gli strumenti tecnologici riveleranno non solo la grandezza e la dignità unica dell'essere umano, ma anche il mandato che quest'ultimo ha ricevuto di « coltivare e custodire » (cfr. *Gen* 2,15) il pianeta e tutti i suoi abitanti. Parlare di tecnologia è parlare di cosa significhi essere umani e quindi di quella nostra unica condizione tra libertà e responsabilità, cioè vuol dire parlare di etica.

Un buon esempio è l'uso dei programmi disegnati per aiutare i magistrati nelle decisioni relative alla concessione dei domiciliari a detenuti che stanno scontando una pena in un istituto carcerario. In questo caso, si chiede all'intelligenza artificiale di prevedere la probabilità di recidiva del crimine commesso da parte di un condannato a partire da categorie prefissate (tipo di reato, comportamento in prigione, valutazione psicologiche ed altro), permettendo all'intelligenza artificiale di avere accesso a categorie di dati inerenti alla vita privata del detenuto (origine etnica, livello educativo, linea di credito ed altro). L'uso di una tale metodologia — che rischia a volte di delegare *de facto* a una macchina l'ultima parola sul destino di una persona — può portare con sé implicitamente il riferimento ai pregiudizi insiti alle categorie di dati utilizzati dall'intelligenza artificiale.

Mi permetto di segnalare, infine, un ultimo ambito in cui emerge chiaramente la complessità del meccanismo della cosiddetta intelligenza artificiale generativa (*Generative Artificial Intelligence*). Nessuno dubita che oggi sono a disposizione magnifici strumenti di accesso alla conoscenza che permettono persino il *self-learning* e il *self-tutoring* in una miriade di campi.

Si registra come uno smarrimento o quantomeno un'eclissi del senso dell'umano e un'apparente insignificanza del concetto di dignità umana. Sembra che si stia perdendo il valore e il profondo significato di una delle categorie fondamentali dell'Occidente: la cate-

goria di persona umana. Ed è così che in questa stagione in cui i programmi di intelligenza artificiale interrogano l'essere umano e il suo agire, proprio la debolezza dell'*ethos* connesso alla percezione del valore e della dignità della persona umana rischia di essere il più grande *vulnus* nell'implementazione e nello sviluppo di questi sistemi.

La nostra risposta a queste ultime domande è: no! La politica serve! Voglio ribadire in questa occasione che « davanti a tante forme di politica meschine e tese all'interesse immediato [...] la grandezza politica si mostra quando, in momenti difficili, si opera sulla base di grandi principi e pensando al bene comune a lungo termine. Il potere politico fa molta fatica ad accogliere questo dovere in un progetto di Nazione e ancora di più in un progetto comune per l'umanità presente e futura ».

Le riflessioni del Pontefice sono state particolarmente apprezzate da un uditorio che ha e avrà certamente grande responsabilità nel determinare gli sviluppi dell'IA e nel garantire la sua effettiva finalizzazione a servizio del bene comune. Se i potenti della terra hanno voluto ascoltare la voce autorevole del Santo Padre su un tema che solo all'apparenza può essere considerato di carattere tecnologico, a maggior ragione è chiesto a tutti noi in ambito accademico, culturale, associativo e sociale di dedicare molta più attenzione ad una questione dalle forti potenzialità ma anche carico di incognite anche preoccupanti. Vorrei concludere con un recente riferimento che Papa Francesco ha voluto fare all'IA nell'ultima Lettera enciclica dedicata al Sacro Cuore:

« 14. Si potrebbe dire che, in ultima analisi, io sono il mio cuore, perché esso è ciò che mi distingue, mi configura nella mia identità spirituale e mi mette in comunione con le altre persone. L'algoritmo all'opera nel mondo digitale dimostra che i nostri pensieri e le decisioni della nostra volontà sono molto più ''standard'' di quanto potremmo pensare. Sono facilmente prevedibili e manipolabili. Non così il cuore. [...]

20. Nell'era dell'intelligenza artificiale, non possiamo dimenticare che per salvare l'umano sono necessari la poesia e l'amore. Ciò che nessun algoritmo potrà mai albergare sarà, ad esempio, quel momento dell'infanzia che si ricorda con tenerezza e che, malgrado il passare degli anni, continua a succedere in ogni angolo del pianeta. Penso all'uso della forchetta per sigillare i bordi di quei panzerotti fatti in casa con le nostre mamme o nonne. È quel momento di apprendistato culinario, a metà strada tra il gioco e l'età adulta, in cui

si assume la responsabilità del lavoro per aiutare l'altro. Come questo della forchetta, potrei citare migliaia di piccoli dettagli che compongono le biografie di tutti: far sbocciare sorrisi con una battuta, tracciare un disegno al controluce di una finestra, giocare la prima partita di calcio con un pallone di pezza, conservare dei vermetti in una scatola di scarpe, seccare un fiore tra le pagine di un libro, prendersi cura di un uccellino caduto dal nido, esprimere un desiderio sfogliando una margherita. Tutti questi piccoli dettagli, l'ordinario-straordinario, non potranno mai stare tra gli algoritmi. Perché la forchetta, le battute, la finestra, la palla, la scatola di scarpe, il libro, l'uccellino, il fiore... si appoggiano sulla tenerezza che si conserva nei ricordi del cuore». (FRANCESCO, Lettera enciclica *Dilexit nos sull'amore umano e divino del Cuore di Gesù Cristo*, 24 ottobre 2024, nn. 14; 20).

Alla luce di questi autorevoli contributi del Magistero pontificio non mi resta che augurare un proficuo lavoro e ringraziare tutti coloro che hanno collaborato alla realizzazione del Convegno e gli illustri relatori che interverranno con le loro preziose e qualificate competenze.

L'INTELLIGENZA ARTIFICIALE IN CHIAROSCURO. SALUTI INTRODUTTIVI AL CONVEGNO: "L'INTELLIGENZA ARTIFICIALE TRA FATTO E NORMA"

(Università Cattolica, Milano, 28 e 29 ottobre 2024)

di *Stefano Solimano*

Sono stato invitato a portare i saluti della Facoltà di Giurisprudenza al convegno finale della ricerca D.3.2, promossa e diretta dal caro amico e collega Gabriele Della Morte.

Desidero innanzitutto esprimere il mio sincero plauso alla Commissione Strategie di Ricerca che, incoraggiata dal compianto Rettore Franco Anelli, ha ideato e sostenuto questa significativa linea di ricerca interna di Ateneo. Essa ha consentito, nel tempo, di abituare noi docenti — e mi riferisco in primo luogo a me stesso, quale studioso appartenente a una facoltà giuridica tradizionalmente incline a coltivare progetti prevalentemente individuali, fatta eccezione, beninteso, per i PRIN — a confrontarci con altri saperi, uscendo dal recinto della presunta autonomia del diritto rispetto alle altre scienze.

Il convegno finale odierno riflette pienamente questa pluralità di prospettive e di approcci al fenomeno dell'intelligenza artificiale. Nel corso delle molteplici sessioni ho potuto apprezzare la ricchezza e la varietà delle relazioni; ora, essendo stato chiamato a redigere una pagina introduttiva del presente volume, ho potuto altresì constatare la chiarezza espositiva e l'elevato livello scientifico dei contributi raccolti.

Anzitutto, emerge con evidenza la molteplicità dei rischi che l'intelligenza artificiale porta con sé in numerosi ambiti. Sono rimasto particolarmente colpito nell'apprendere come l'IA, capace di simulare forme di comunicazione e creatività umana, possa amplificare le dinamiche della propaganda computazionale e come l'analisi di recenti campagne elettorali e conflitti geopolitici metta in luce gravi pericoli per la fiducia nei media, per i processi democratici e per la stessa costruzione condivisa della realtà. Analogamente, desta non poche preoccupazioni la possibilità che l'IA riconfiguri il rapporto con i defunti, aprendo a nuove forme di presenza *post-mortem* e sollevando

interrogativi di natura etica, affettiva e ontologica. Anche le tecnologie di riconoscimento facciale, ampliando in modo inedito il potere di sorveglianza, finiscono per incidere trasversalmente sui diritti fondamentali e sulla qualità della democrazia.

Ahimè, in un tempo segnato da una guerra oramai di dimensione planetaria, non era possibile esimersi dall'analizzare l'impatto dell'intelligenza artificiale sulle operazioni militari e sulle sfide che essa pone al diritto internazionale umanitario. L'impiego dell'IA come strumento di supporto decisionale e, ancor più, nell'ambito delle armi autonome solleva questioni ineludibili in ordine al mantenimento di un controllo umano significativo e ai problemi, tutt'altro che risolti, della responsabilità giuridica.

Quanto all'applicazione dell'intelligenza artificiale nel settore della salute, i saggi raccolti sottolineano con forza la centralità della decisione medica umana e della dignità della persona. Pur riconoscendo che l'IA rappresenta ormai uno strumento sempre più rilevante per il miglioramento del sistema sanitario e della ricerca, gli studiosi mettono in luce numerose criticità: errori diagnostici (falsi positivi e falsi negativi), bias algoritmici, tutela della privacy, opacità dei sistemi 'black box', fenomeni di *deskilling* e incertezze nella ripartizione delle responsabilità, alla luce dell'art. 32 della Costituzione, dell'AI Act, dell'European Health Data Space e della legge italiana n. 132 del 2025. Sotto il profilo della tutela civilistica, si evidenzia come il ricorso all'intelligenza artificiale imponga una riconsiderazione dei criteri di imputazione della colpa, tendenzialmente circoscrivendo la responsabilità del medico ai casi di utilizzo negligente o di mancata e inadeguata verifica degli output algoritmici.

Il valore di questa ricerca e del volume che ne raccoglie i risultati risiede proprio nella capacità di affrontare l'intelligenza artificiale non come un fenomeno meramente tecnico, ma come una sfida culturale, giuridica e istituzionale di primaria rilevanza. Il dialogo interdisciplinare che ha caratterizzato l'intero percorso di ricerca mostra bene come solo attraverso il confronto tra saperi sia possibile elaborare strumenti interpretativi e regolativi all'altezza delle trasformazioni in atto. Al diritto spetta oggi il compito, tutt'altro che semplice, di governare l'innovazione senza subirla, riaffermando la centralità della persona, la responsabilità umana e la funzione garantista dell'ordinamento. È in tale prospettiva che il presente volume si offre come un contributo prezioso e attuale, destinato ad alimentare un dibattito che non potrà che proseguire negli anni a venire.

LA PROSPETTIVA DELLA DIGITAL HEALTH. SALUTI INTRODUTTIVI AL CONVEGNO: "DIGITAL HEALTH, ARTIFICIAL INTELLIGENCE: QUALI RESPONSABILITÀ VERSO I PAZIENTI?"

(Fondazione Policlinico Universitario "A. Gemelli" IRCCS, Roma, 4 giugno 2024)

di *Antonio Gasbarrini*

Digital Health e Intelligenza Artificiale rappresentano oggi uno dei punti di congiunzione più significativi fra l'evoluzione della ricerca biomedica, la trasformazione dei sistemi sanitari e le nuove responsabilità etiche e giuridiche che inevitabilmente accompagnano l'adozione di tecnologie ad alta complessità.

Il Convegno del 4 giugno 2024, dedicato alle responsabilità verso i pazienti nell'era dell'innovazione digitale, ha offerto un quadro estremamente lucido di queste dinamiche, anticipando molte delle questioni che negli ultimi mesi si sono imposte con forza nell'agenda istituzionale nazionale e internazionale.

L'iniziativa si è inserita in un momento di passaggio epocale: nel nostro sistema sanitario, l'integrazione fra pratica clinica, infrastrutture digitali, algoritmi predittivi e capacità di analisi dei dati sta ridefinendo profondamente il significato stesso di cura. I modelli di medicina digitale non sono più prospettive future, ma elementi strutturali dell'operatività quotidiana, capaci di trasformare i percorsi diagnostici, di incidere in maniera significativa sulla prognosi, di anticipare la comparsa di stati patologici e di ridefinire i rapporti tra pazienti, professionisti sanitari, istituzioni e industria. In un contesto simile, il tema delle responsabilità non è un semplice addendum normativo o un'esigenza regolatoria di accompagnamento: è parte integrante della configurazione del nuovo ecosistema della salute. Ed è proprio per questo che ritengo particolarmente meritoria la scelta, maturata all'interno del Progetto di Ricerca D.3.2 "Funzioni Pubbliche, Controllo Privato", di promuovere non solo un confronto interdisciplinare di alto profilo, ma soprattutto di fissarne i risultati in

un volume accessibile, aperto e destinato ad accompagnare le istituzioni nei prossimi anni.

Dal punto di vista di una Facoltà di Medicina o di IRCCS come la Fondazione Policlinico Universitario Gemelli, che si trovano simultaneamente in una posizione di accademia, di frontiera scientifica e di responsabilità verso i pazienti più fragili, e di presidio istituzionale nei percorsi di trasferimento tecnologico, il tema della responsabilità nell'uso dell'intelligenza artificiale ha assunto una centralità crescente. Oggi il nostro lavoro non riguarda più solo l'adozione di tecnologie innovative, ma l'architettura complessiva della fiducia che deve accompagnarle. La sfida non è soltanto tecnica, bensì etica, culturale e giuridica. L'implementazione di algoritmi di intelligenza artificiale in ambito clinico produce infatti tre livelli di responsabilità che devono essere compresi e regolati:

1. La responsabilità clinica individuale, che non viene annullata dall'uso di strumenti digitali, ma deve anzi essere rafforzata da nuove competenze capaci di interpretare criticamente i risultati degli algoritmi e di integrarne la logica all'interno del ragionamento medico.

2. La responsabilità istituzionale, che riguarda la qualità dei dati, la trasparenza dei processi, la validazione degli strumenti e la governance delle infrastrutture digitali. È un livello che coinvolge ospedali, IRCCS, università e aziende sanitarie territoriali, ciascuno con ruoli di garanzia specifici.

3. La responsabilità sistemica, che riguarda il funzionamento complessivo dell'ecosistema della salute: dalle logiche di interoperabilità dei dati alle condizioni di accesso alle innovazioni, dalla tutela dei diritti dei pazienti all'equità distributiva dei benefici che l'AI può offrire.

In questo quadro, la riflessione giuridica, nella sua capacità di definire limiti, di riconoscere diritti, di prevedere scenari di rischio e di proteggere i soggetti più vulnerabili, si colloca come una componente fondamentale della trasformazione in atto. Le professioni cliniche e le scienze della vita non possono procedere senza un dialogo strutturato con la comunità dei giuristi, e, allo stesso tempo, le categorie giuridiche hanno bisogno di confrontarsi con la complessità crescente dei modelli scientifici e delle nuove logiche di decisione automatizzata.

La pubblicazione degli Atti del Convegno va quindi ben oltre la semplice testimonianza di un evento: si pone come uno strumento di orientamento istituzionale e culturale, utile a consolidare un lin-

guaggio comune tra discipline che, pur avendo tradizioni differenti, devono oggi convergere verso una visione integrata del bene della persona e della tutela del paziente. È un'esigenza ormai ineludibile: senza un lessico condiviso, senza una comprensione reciproca dei principi e degli strumenti operativi, l'intero processo di innovazione rischia di frammentarsi, con il pericolo di creare tecnologie formalmente avanzate ma incapaci di generare fiducia e sicurezza. Nel mio precedente ruolo di Preside della Facoltà di Medicina e in quello attuale di Direttore Scientifico di un IRCCS che ha fatto della ricerca traslazionale, della medicina personalizzata e della digitalizzazione dei processi assistenziali una priorità strategica, vedo quotidianamente quanto sia necessario accompagnare l'adozione dell'intelligenza artificiale con un approccio basato su tre pilastri.

Il primo è la rigorosa validazione scientifica. Gli algoritmi devono essere sottoposti agli stessi criteri di trasparenza, riproducibilità ed efficacia che richiediamo a ogni tecnologia medica. La medicina non può delegare a modelli opachi processi decisionali critici.

Il secondo pilastro è la governance dei dati. L'intelligenza artificiale è potente nella misura in cui i dati sono affidabili, completi, aggiornati e coerenti. L'accesso ai dati sanitari deve essere protetto da garanzie di sicurezza, ma anche facilitato da infrastrutture che consentano alle istituzioni di collaborare in modo responsabile.

Il terzo pilastro è la formazione. Non basta introdurre strumenti di AI nei percorsi clinici: è necessario formare medici, infermieri, ricercatori e operatori sanitari alla loro corretta interpretazione. È un elemento cruciale per preservare la responsabilità professionale e per evitare sia l'uso improprio sia la dipendenza acritica dalle tecnologie.

A questo si aggiunge un tema ancora più profondo, che emerge con forza anche nelle riflessioni di questo volume: l'AI deve rimanere uno strumento al servizio del paziente, non un fattore che ne diluisce la centralità. La cura nasce da una relazione e si struttura attraverso un equilibrio di competenze scientifiche, responsabilità morali e giudizi clinici. Gli algoritmi possono rafforzare questa relazione, ma non sostituirla. In questo senso, il valore più grande del lavoro coordinato da Gabriele Della Morte e dai colleghi coinvolti nel Progetto D.3.2 è proprio quello di aver creato uno spazio interdisciplinare in cui le diverse prospettive — medica, giuridica, etica, tecnologica — possono dialogare senza perdere la loro identità. Un dialogo che non è solo utile, ma necessario per governare la transizione in corso.

Con questo spirito ho accolto con piacere l'invito a offrire un saluto introduttivo per gli Atti del Convegno. Il nostro Ateneo ha una tradizione unica di dialogo tra saperi, e questo volume ne rappresenta un esempio concreto e prezioso: un ponte tra discipline diverse, ma unite dal compito comune di proteggere la persona e di accompagnare l'innovazione in modo responsabile. Rivolgo quindi il mio sincero ringraziamento agli organizzatori, ai relatori, ai ricercatori e agli studenti che hanno contribuito a questo percorso. Mi auguro che queste pagine possano rappresentare non solo una testimonianza del lavoro svolto, ma anche una guida utile per orientare le scelte istituzionali e regolatorie che i prossimi anni ci porranno dinanzi.

KEYNOTE SPEECH

L'INTELLIGENZA ARTIFICIALE: IL FATTO

di *Juan Carlos De Martin*

Prima di parlare di intelligenza artificiale (IA) è opportuno fare una premessa relativa al modo in cui nello spazio pubblico si parla di di tecnologie informatiche. Da ormai 30 anni, ovvero, da quando internet ha raggiunto il grande pubblico sotto forma di World Wide Web, avviene un fenomeno ricorrente: dagli Stati Uniti, in particolare dalla costa ovest, dalla Silicon Valley, una novità tecnologica viene opportunamente confezionata per poi diffondersi rapidamente nel resto degli USA, attraversare l'Atlantico e raggiungere l'Europa sotto forma di *grande ondata mediatica*.

Ricordiamo alcune di queste ondate precedenti che, quando sono arrivate, hanno saturato i dibattiti, i convegni, i giornali, le trasmissioni televisive, il dibattito pubblico, la politica e così via.

All'inizio degli anni '10 di questo secolo, per esempio, si parlava molto di *Cloud Computing*. Sempre in quel periodo si parlava molto anche dei cosiddetti *MOOCs*, i corsi online massivi aperti, che sembravano sul punto di spazzar via quasi tutte le università del mondo. Si è molto parlato di Big Data, che venivano presentati come un cambio di paradigma epocale. Si è parlato, continuando con gli esempi, di Blockchain, di auto a guida autonoma, di 5G, di quell'oggetto mal definito chiamato Web 3.0. Si è parlato molto anche di Metaverso, almeno fino a quando anche quella bolla si è sgonfiata.

Se notate, ho citato tutte espressioni che non hanno molto di scientifico: sono piuttosto espressioni accattivanti, potremmo dire di marketing, coniate apposta per sedurre e per enfatizzare. Inoltre non di rado si tratta di 'soluzioni alla ricerca di un problema': viene fatto un prodotto dalle mirabolanti capacità tecniche e poi ci si chiede a cosa possa servire.

In ogni caso, queste ondate hanno tre caratteristiche in comune: la prima è che c'è sempre effettivamente una novità tecnica, più o meno importante e utile. Intorno a questo dato tecnico, però, e

siamo alla seconda caratteristica di tutte le ondate digitali, c'è un enorme strato di marketing o, per usare un'altra parola inglese non facilmente traducibile, di '*hype*', parola che nasce negli anni '20 del secolo scorso col significato originario di *imbroglio*. Infine, la terza caratteristica è che le ondate hanno anche una componente speculativa che spinge a gonfiare al massimo la bolla con l'obiettivo di incassare prima che si sgonfi.

Un corollario delle ondate digitali è che il concentrare l'attenzione su un determinato tema è come usare un fortissimo faro: ciò che si vuole mettere in evidenza è molto ben illuminato, spesso anche troppo, ma inevitabilmente tutto il resto del mondo digitale finisce nell'ombra. Fuor di metafora: di molte cose, anche assai importanti, non si discute perché praticamente tutte le energie sono dedicate al susseguirsi di ondate.

In questo scenario che caratterizza il mondo da ormai circa tre decenni, il compito dell'università dovrebbe essere quello di restare sobria, oggettiva, di mettere in luce i lati nascosti, di problematizzare ciò che viene dato per scontato e di usare con precisione le parole. Insomma, di non unirci al coro, ma di provare a dire la verità. Noi accademici dovremmo sempre sentire il dovere di essere la coscienza della società, rifiutando con fermezza le semplificazioni, le contrapposizioni e le censure che caratterizzano sempre di più il dibattito pubblico quando si parla di digitale (e non solo). Ciò ovviamente non significa essere a priori contro, sarebbe assurdo, ma neanche essere a priori a favore delle novità, come invece spesso capita quando si parla di tecnologia e innovazione.

Fatta questa premessa, da qualche anno, ma con maggiore intensità dal novembre del 2022 con il lancio di ChatGPT, stiamo vivendo un'ulteriore ondata mediatica, quella relativa *all'intelligenza artificiale*.

Si tratta forse dell'ondata più grande di sempre, per l'entità di risorse investite, si parla di centinaia di miliardi, soprattutto da parte di due attori: gli Stati Uniti e la Cina.

Io, parlando di sobrietà e precisione di linguaggio, ho dei problemi già con l'espressione stessa 'intelligenza artificiale'.

L'espressione IA, infatti, è stata coniata nel 1955 da alcuni pionieri dell'informatica per chiedere fondi alla Rockefeller Foundation. Loro stessi successivamente ammisero di aver scelto quelle parole perché seducenti. Nel documento di richiesta fondi scrissero che, con le opportune risorse, si potrebbero far fare ai computer cose straordinarie, cose, appunto, 'intelligenti'. Insomma, gli scienziati fecero mar-

keting. Se solo fosse possibile (lo dubito), sarebbe quindi molto utile sostituire l'espressione 'intelligenza artificiale' con un espressione più lunga, più tecnica, più precisa, probabilmente più noiosa, ma molto meno fuorviante o, parafrasando Jorge Luis Borges quando criticava la parola 'infinito', meno 'corrompente, sghangherante'.

Aggiungo un dato storico: quei pionieri dell'informatica stavano distaccandosi dalla disciplina della cibernetica, disciplina molto assai popolare in quegli anni, che però era vista, oltre che troppo interdisciplinare, come troppo associata alla figura di Norbert Wiener, il grande matematico docente al MIT di Boston, che aveva il difetto di essere un noto pacifista, che in pieno maccartismo significava essere visti come pericolosi sovversivi.

Rispetto a 'cibernetica', 'artificial intelligence', oltre a essere molto meno esotica (non molti conoscono il greco), suona molto meglio, con un connotato che ho messo a fuoco di recente: 'intelligence' in inglese non vuol dire solo intelligenza, ma anche 'raccolta informazioni', 'servizi segreti'. E questo ovviamente aiutava a ottenere finanziamenti, che all'epoca (e ancora adesso) provenivano soprattutto dal mondo bellico e della sicurezza.

Comunque, dopo 70 anni l'espressione ha ormai attecchito. Teniamocela, dunque, ma con la chiara consapevolezza che è insuperabilmente ambigua, visto che non c'è nessuna definizione scientifica condivisa di 'intelligenza'.

Forse potremmo dire che l'intelligenza artificiale sono quei prodotti e servizi informatici su cui è facile proiettare qualcosa di umano, ovvero, quei prodotti e servizi che favoriscono la pareidolia. Ecco la definizione della Treccani: il processo psichico che porta a vedere immagini illusorie dotate di una nitidezza materiale, come vedere una testa di gatto in una nuvola.

Questo aspetto di proiezione di qualcosa di umano è connaturato con l'IA fin dalle origini. Un caso famoso è quello del programma ELIZA, un semplice risponditore automatico che imitava uno psicologo, sviluppato tra il 1964 e 1967 dal noto informatico americano di origine tedesca, Joseph Waisenbaum. Alcuni degli utenti iniziavano ad avere un rapporto personale con il programma, perché si convincevano che ci fosse qualcosa di umano, di senziente nelle risposte del computer.

Oggi, con ben altre prestazioni, vediamo lo stesso con ChatGPT, che è stato deliberatamente progettato per indurci a vedere qualcosa di umano nelle sue risposte: l'uso del ''tu'', il fatto che chieda facil-

mente scusa, la propensione alla sicofantia sono scelte deliberate da parte dei progettisti che hanno scelto di ''umanizzare'' al massimo le risposte del programma.

Ma facciamo un passo indietro: perché da poco più di dieci anni si riparla così tanto di 'intelligenza artificiale'?

Intorno al 2010 convergono tre fatti nuovi:

1. algoritmi esistenti da decenni vengono considerevolmente migliorati;

2. grazie a 20 anni di web (di smartphone, ecc.) c'è la disponibilità di moltissimi dati;

3. la capacità di calcolo è molto cresciuta, anche grazie allo sviluppo di nuovi tipi di circuiti integrati, in particolare le GPU (che stanno facendo la fortuna di Nvidia).

Queste novità iniziano a produrre risultati sempre più notevoli. Pensiamo al riconoscimento del parlato: 15 anni fa, c'erano già algoritmi di riconoscimento del parlato ma il tasso di errore li rendevano poco utili, tranne che in ambiti molto precisi; oggi, il miglioramento delle prestazioni ha resto la tecnologia molto più utile in molti più settori. Lo stesso vale per il riconoscimento di oggetti, la traduzione automatica, il riconoscimento della scrittura corsiva e, molto problematicamente, per il riconoscimento delle persone (volti, camminata, ecc...).

Questo è il fatto tecnico alla base dell'ondata dell'intelligenza artificiale che inizia a montare una decina d'anni fa, gonfiata da grande disponibilità di capitali (soprattutto in USA e Cina), dalla volontà di arrivare primi, sperando di creare poi monopoli o almeno oligopoli, e — fattore molto importante per comprendere quello che sta capitando — dalla prospettiva di importanti usi militari, come lo sviluppo di armi letali autonome (droni aerei e navali, veicoli da terra autonomi, persino robot soldati antromorfi).

Nel novembre 2022, col debutto di ChatGPT, l'ondata diventa uno tsunami: inizia l'era dell'intelligenza artificiale generativa. La novità tecnologica alla base dello tsunami è che è diventato possibile produrre testi, immagini, video e suoni di notevole plausibilità in reazione a degli stimoli degli utenti (i famosi 'prompt').

Soprattutto per quello che riguarda i testi prodotti dai servizi interattivi di 'chatbot' (tra cui ChatGPT) viene scelto, come abbiamo già osservato, di favorire al massimo la pareidolia: oltre alla oggettiva plausibilità dei testi, l'impiego di un tono amicale, la propensione all'adulazione e altri strategemmi servono a rafforzare la sensazione di

avere a che fare con qualcosa di 'senziente', di 'umano', creando dipendenza negli utenti, ovvero, prolungando al massimo l'utilizzo del servizio.

È, quindi, indispensabile ricordarci sempre che si tratta di macchine. L'esperimento mentale da fare per immunizzarsi dalla tentazione di antropomorfizzare l'IA è quello di immaginare che i testi dei 'chatbot' escano da una lavatrice o da un tostapane. Dietro a quei testi, in altre parole, non c'è volontà, intenzione, esperienza del mondo, esperienza del corpo, ecc.: sono contenuti oggettivamente spesso di elevata qualità (ma non di rado assurdi e a volte persino pericolosi), in cui noi umani riscontriamo un significato più o meno utile, divertente, stimolante, confortante, ecc. Affidarsi ai 'chatbot' come ci si affida a un amico, a un confessore o a uno psicoterapeuta, come secondo alcune prime analisi sta capitando sempre più di frequente, è una deriva di grande pericolosità non solo da un punto di vista economico o giuridico, ma anche da un punto di vista psicologico e sociale.

Come già successo per tutte le precedenti ondate digitali, ci sentiamo dire: 'cambierà tutto'.

A parte l'enfasi tipica del 'marketing' (e delle speculazioni) è molto probabile che le nuove tecnologie di 'intelligenza artificiale' cambino alcune cose — anche importanti — come è già successo in passato per effetto del personal computer, di Internet, del Web, del cloud computing, ecc....

Però in ogni caso dovremmo discuterne con sobrietà, stando ai fatti e non alle brochure di vendita di chi offre questi prodotti. L'onere della prova deve sempre essere in capo a chi fa determinate affermazioni: che argomenti in maniera oggettiva, senza dare nulla per autoevidente. E soprattutto, nulla è ineluttabile. Lo sviluppo e la diffusione delle tecnologie, infatti, sono attività umane, e come tali, nulla è mai ineluttabile.

In particolare, mi sembra importante sottolineare che l'IA, come qualsiasi altra tecnologia, può essere messa in campo in modi diversi, un fatto che tendiamo a rimuovere, lasciando agli attori economici tutto il potere di decidere se, quando e come mettere in campo le tecnologie. Nel caso dell'IA, per esempio, essa può essere adoperata con l'obiettivo primario di sostituire, ogni volta che sia possibile, i lavoratori oppure con quello, ben diverso, di affiancarli per renderli più produttivi, per migliorare le condizioni di lavoro, ecc. Non c'è alcun automatismo: si tratta di scelte. Scelte che in democrazia ab-

biamo il diritto — e in certi casi il dovere — di problematizzare e discutere.

Gli attuali enormi investimenti in IA sono gonfiati, almeno negli USA, dalla promessa di profitti futuri, basati soprattutto — temo — sulla prospettiva di poter eliminare molti posti di lavoro, oltre che dalle applicazioni militari e dal potenziale di questi strumenti come mezzi di controllo sociale, ad esempio generando automaticamente contenuti manipolatori per inondare il dibattito pubblico.

Al di là dei discorsi francamente poco seri su 'superintelligenze artificiali', 'singolarità' e altri simili argomenti, le tecniche di IA, oltre a successi ormai consolidati, come quelli relativi alla traduzione automatica, mi sembrano particolarmente promettenti in ambito medico: poter ridare la vista, protesi intelligenti, cure personalizzate, sono prospettive concrete e di innegabile utilità sociale. Inoltre un uso diffuso di tecniche di IA potrebbe migliorare l'efficienza generale del sistema produttivo: questa pare essere la direzione — come al solito molto pragmatica — adottata dalla Cina.

Nel gran parlare enfatico che si fa ormai da anni sull'IA, tuttavia, viene in genere tenuto in ombra il ruolo di milioni di lavoratori umani, spesso in paesi poveri e sottopagati, che sviluppano e mantengono questi sistemi. Altro aspetto di cui non si parla abbastanza è l'enorme consumo di energia degli impianti di IA: a questo proposito si parla di riattivare centrali a carbone e nucleari dismesse solo per alimentare queste 'fabbriche' che forniscono i servizi (o di costruirne di nuove). Inoltre questi 'data center' consumano moltissima acqua: che un 'data center' consumi quanto una cittadina di medie dimensioni non è esattamente un fatto che si possa trascurare.

Infine lo sviluppo dell'IA sta ulteriormente esacerbando la caratteristica principale di come ci siamo computerizzati nella nostra parte di pianeta, soprattutto in questo primo quarto di secolo: l'estrema concentrazione di potere nelle mani di alcune imprese statunitensi, le ben note Big Tech, concentrazione di potere che l'IA sta contribuendo ad aumentare ulteriormente, relegando i Paesi Europei sempre più al rango di colonie digitali (in questo ambito digitale, come in pressoché tutti gli altri, posta elettronica inclusa).

L'INTELLIGENZA ARTIFICIALE: LA NORMA

di *Oreste Pollicino*

Anche noi costituzionalisti ci occupiamo da tempo di questi temi. Ma credo sia utile chiarire subito un punto, anche alla luce del *keynote* di Juan Carlos De Martin: il costituzionalismo deve restare fedele alla sua missione originaria di limitazione del potere, ma deve al tempo stesso interrogarsi criticamente sul proprio strumentario quando il potere non è più soltanto verticale — autorità contro libertà — bensì assume forme diagonali e orizzontali. È il caso delle grandi piattaforme, ma più in generale di quell'ecosistema di infrastrutture digitali che, come abbiamo sentito, non si limitano a offrire servizi, bensì configurano ambienti, abitudini, aspettative.

Le 'ondate' tecnologiche descritte da De Martin non sono solo fenomeni mediatici: sono anche dispositivi di potere. Il potere oggi non comanda soltanto, ma orienta, predispone, normalizza. Ed è per questo che il costituzionalismo, per rimanere fedele a se stesso, non può né demonizzare né accettare passivamente queste trasformazioni, ma deve provare a limitarle, a renderle intelligibili, a sottoporle a criteri di responsabilità.

Vorrei allora parlarvi di una mia presa di consapevolezza personale. Quando ho iniziato a occuparmi di costituzionalismo nell'epoca di Internet, tra il 2007 e il 2008, mi sono chiesto quale potesse essere il senso di un'operazione di emancipazione e, insieme, di riconciliazione. Emancipazione — lo dico in modo volutamente provocatorio — dal giurista industrialista o privatista; riconciliazione del diritto costituzionale con un mondo tecnologico che spesso lo guardava con sospetto, come se fosse un freno o un residuo del passato.

Il mio 'crampo mentale' era quello di provare a restare dentro una riflessione esclusivamente giuridica. Ma oggi mi è chiaro che qualsiasi discussione sulle tecnologie digitali che guardi solo al diritto è inevitabilmente fallace. Se non si comprende ciò che sta al di là della regolazione — il linguaggio, l'hype, la costruzione degli immagi-

nari tecnologici di cui parlava De Martin — allora anche la migliore normazione rischia di colpire il bersaglio sbagliato. In questo senso, il positivismo giuridico più rigido, quello che si accontenta della forma, serve a pochissimo.

Qui entra il mio debito intellettuale verso tradizioni che hanno sempre guardato al diritto come fenomeno sociale prima che normativo: Santi Romano, con la pluralità degli ordinamenti; il realismo giuridico americano; gli scandinavi. Non è un caso che oggi, parlando di intelligenza artificiale, torni utile questo sguardo disincantato, che si chiede non solo che cosa dice la norma, ma che cosa produce nella realtà.

Ed è per questo che, oggi, parlare di una normazione — anche sofisticata — come quella europea sull'intelligenza artificiale, senza interrogarsi sulla sua efficacia e sul suo impatto reale, è un errore. È efficace? Oppure rischia di essere un esercizio di stile, una conferma di quella che appare sempre più come una grande fragilità europea: l'essere una potenza normativa che si indebolisce mentre altri — Stati Uniti e Cina — consolidano il loro primato tecnologico e industriale? In questo quadro, l'analisi di De Martin sulla concentrazione del potere tecnologico e sulla dimensione speculativa delle ondate digitali è un complemento essenziale alla riflessione giuridica.

Se ci accartocciamo in un positivismo kelseniano rassicurante, forse utile per la certezza del diritto, dobbiamo però chiederci: quale certezza stiamo cercando quando regoliamo il futuro? Non rischiamo di costruire un'illusione di controllo che si traduce poi in frustrazione?

Da qui la necessità di un metodo diverso. Mettere stabilmente in dialogo chi studia la tecnologia e chi studia il diritto non è un vezzo interdisciplinare, ma una condizione di possibilità per capire cosa stiamo davvero regolando. Per questo propongo due esercizi: uno sincronico e uno diacronico.

Dal punto di vista sincronico, la prima coordinata è la limitazione dei poteri. Il costituzionalismo è ancora la risposta, ma a una condizione: non cadere nella trappola del novismo digitale. Così come De Martin ci invita a diffidare di ogni nuova etichetta seducente, anche il giurista deve resistere alla tentazione di creare nuove categorie per ogni novità tecnologica. Nella maggior parte dei casi, le categorie esistenti, se interpretate evolutivamente, sono ancora adeguate.

Lo dimostra l'esempio classico richiamato da Easterbrook nel 1997, '*Cyberspace and the Law of the Horse*'. O lo dimostra la giuri-

sprudenza della Corte Suprema statunitense sul Quarto Emendamento, che ha saputo adattare un testo del 1791 a un mondo di sorveglianza digitale diffusa. Qui il diritto reagisce alla computerizzazione degli spazi e delle persone senza farsi dettare l'agenda dal marketing tecnologico.

Un secondo profilo sincronico riguarda la teoria delle fonti. Il vecchio dilemma tra hard law e soft law è ormai superato. Viviamo in un sistema di co-regolazione, in cui strumenti apparentemente volontari diventano vincolanti. È una dinamica che riflette bene quella zona grigia tra tecnica, mercato e potere pubblico che De Martin ha descritto: il potere non sta più solo nella norma, ma nei processi, negli standard, nei codici.

Il problema si accentua quando guardiamo a regolamenti come il GDPR o l'AI Act. Formalmente regolamenti, nella pratica direttive mascherate. Lo abbiamo visto con il GDPR, che ha prodotto una frammentazione significativa. Con l'AI Act il rischio è ancora maggiore: una normativa a maglie larghe, che in un contesto di forte asimmetria di potere tecnologico rischia di favorire proprio gli attori più forti.

Qui emerge una tensione strutturale: la sintassi economicistica del mercato interno, articolo 114 TFUE, e la sostanza costituzionale dei diritti fondamentali, articolo 2 TUE. È una discrasia difficile da risolvere, soprattutto se — come ricordava De Martin — la tecnologia incide sulle condizioni materiali dell'esercizio dei diritti, sull'ambiente, sul lavoro, sulle relazioni sociali.

Dal punto di vista diacronico, possiamo individuare alcune fasi note: l'illusione anarcoide originaria di Internet; la fase liberista della non intrusione regolatoria; la metamorfosi delle piattaforme in poteri privati; e infine la reazione che ho chiamato costituzionalismo digitale. Qui la Corte di Giustizia ha svolto un ruolo decisivo, applicando orizzontalmente la Carta dei diritti fondamentali, spesso supplendo all'inerzia del legislatore.

Il GDPR è stato il primo tentativo di risposta sistemica, ma con un limite evidente: una forte affermazione assiologica accompagnata da una delega operativa agli attori privati. Il Digital Services Act segna un passo ulteriore, introducendo garanzie procedurali e riaffermando il ruolo della componente umana.

Arriviamo infine all'intelligenza artificiale generativa. Qui il dialogo con De Martin è particolarmente stringente. La sua insistenza sulla pareidolia, sull'antropomorfizzazione e sull'illusione di intelli-

genza ci aiuta a capire perché il salto dall'automazione all'autonomia non sia solo tecnico, ma simbolico e culturale. Questo salto avrebbe richiesto un guizzo corrispondente del costituzionalismo europeo.

A mio avviso, il peccato originario è stato consentire l'immissione sul mercato di sistemi potenzialmente dirompenti senza una fase sperimentale trasparente. La risposta europea si è concentrata sulla *product safety*. È un obiettivo importante, ma insufficiente. L'intelligenza artificiale non è solo un prodotto: è un ecosistema che ridisegna rapporti di potere, lavoro, consumo di risorse, concentrazione economica.

La domanda conclusiva resta dunque aperta: l'AI Act, anche alla luce del Digital Omnibus, è in grado di reggere il test del costituzionalismo europeo?

I rischi che si intravedono sono almeno due.

Il primo è quello di una frammentazione giuridica che si autoalimenta. Una regolazione costruita su clausole elastiche, categorie aperte e rinvii continui, come quella dell'AI Act, affida necessariamente molto all'implementazione nazionale e all'interpretazione giurisprudenziale. Se a questo impianto si sovrappone ora una spinta alla semplificazione che opera prevalentemente sul piano formale, il rischio è quello di una certezza del diritto solo apparente. In un'Unione che resta profondamente disomogenea sul piano amministrativo, tecnologico e istituzionale, la semplificazione rischia di tradursi in un moltiplicatore delle differenze applicative. Il paradosso è evidente: nel tentativo di governare un ecosistema tecnologico sempre più concentrato e globale, l'ordinamento europeo rischia di produrre nuove incertezze al proprio interno, indebolendo la funzione stessa della regola come strumento di limitazione del potere.

Il secondo rischio è più profondo ed è quello di una potenza normativa sempre più ripiegata su se stessa. L'Europa continua a esercitare una forte capacità regolatoria, oggi accompagnata anche da un discorso sulla semplificazione e sulla competitività, ma senza che a questa produzione normativa corrisponda un analogo rafforzamento della capacità di incidere sulle architetture tecnologiche, sugli standard di fatto e sulle scelte infrastrutturali. Il rischio è che la normazione finisca per svolgere una funzione prevalentemente simbolica, offrendo una risposta ordinata e coerente sul piano giuridico a trasformazioni che si giocano, in larga misura, altrove. In questo scarto tra ambizione regolatoria e realtà materiale del potere tecnologico si

annida una tensione che il Digital Omnibus, da solo, non sembra in grado di risolvere.

In questa prospettiva, il problema non è soltanto come semplificare o come regolare meglio l'intelligenza artificiale, ma se la regolazione europea riesca ancora a svolgere la sua funzione costituzionale essenziale: riconoscere dove il potere si colloca, renderlo visibile e sottoporlo a limiti. È su questo terreno, più che su quello dell'ingegneria normativa, che si gioca oggi la tenuta del costituzionalismo europeo nell'era dell'intelligenza artificiale.

INTELLIGENZA ARTIFICIALE E DISINFORMAZIONE

INTRODUZIONE

di *Giovanna Mascheroni* e *Simone Carlo*

Sommario: 1. Introduzione. — 2. Una definizione di IA generativa. — 3. Le macchine della menzogna contemporanee. — 4. L'IA generativa moltiplica la disinformazione?

1. Introduzione.

Il problema della disinformazione e la preoccupazione che i mezzi di comunicazione possano veicolare notizie false non sono nuovi. La Prima Guerra Mondiale è stata un laboratorio 'sul campo' delle strategie di manipolazione dell'opinione pubblica attraverso la produzione di notizie false, immagini modificate, resoconti dal fronte fuorvianti. Proprio a partire dall'analisi della propaganda bellica americana, Harold Lasswell (1927) costruisce la sua riflessione sul potere dei media: la propaganda dell'esercito americano, funzionale alla mobilitazione nazionale, sembrò essere una valida prova del potere condizionante dei mezzi di comunicazione. Attraverso strategie di comunicazione basate (anche) sulla menzogna, fu possibile manipolare l'opinione pubblica a sostegno dell'intervento americano nel conflitto. Di conseguenza, secondo Lasswell la propaganda è coercizione e violenza sulla società ma senza l'uso della forza. In questa prospettiva, quindi, il potere politico ha bisogno della propaganda e delle menzogne per poter tenere i cittadini disinformati, in accordo con ciò che la classe specializzata ha deciso essere nel loro miglior interesse.

Con lo sviluppo del sistema dei media le cose si sono ulteriormente complicate. Come ricorda Colombo (2020), nel corso degli anni Venti e Trenta, l'ascesa delle dittature fasciste e naziste nei sistemi fino allora democratici è stata favorita dalla disinformazione, dalla censura, dalla repressione di qualsiasi forma di pensiero alternativa e dai discorsi di odio. I mezzi di comunicazione furono strumentalmente utilizzati per diffondere slogan maligni e falsi contro le minoranze e i nemici esterni, facilitando i cambi di regime.

Nei decenni del dopoguerra, anche le società democratiche sono state attraversate da turbolenze informative, notizie false, menzogne e disinformazione, che hanno messo a dura prova la fiducia dell'opinione pubblica nei confronti del sistema informativo e delle istituzioni politiche. Pensiamo, nel nostro Paese, al buio periodo del terrorismo e alle notizie false che terroristi, governi, forze straniere, partiti politici hanno prodotto per mobilitare, spaventare, manipolare un'opinione pubblica particolarmente fragile e impaurita, dando forma a una vera e propria «strategia della tensione» (DONDI, 2015).

O pensiamo anche alla centralità della disinformazione in Europa durante la guerra fredda e i tentativi, da ambo i fronti della cortina di ferro, di insinuarsi nelle opinioni pubbliche avversarie per minare la fiducia nei confronti dei propri governi e regimi.

Ma è nell'ecosistema mediale digitale contemporaneo che il problema della disinformazione ha assunto forme e proporzioni inedite, grazie alle possibilità di comunicare a basso costo e in tempo reale (o quasi) a un'opinione pubblica transnazionale offerte da social media, comunicazione automatizzata (bot) e intelligenza artificiale (IA) generativa.

2. Una definizione di IA generativa.

Innanzitutto, cosa intendiamo per IA generativa? Partiamo da una definizione socio-tecnica dell'IA generativa. Piuttosto che descrivere il fenomeno a partire dal suo funzionamento, basato su sistemi di *deep learning* più o meno evoluti (MITCHELL, 2019), è forse opportuno offrire una descrizione applicativa: l'IA generativa consiste in ''un insieme di idee, tecnologie e tecniche che riguardano la capacità di un sistema informatico di eseguire compiti che normalmente richiedono l'intelligenza umana'' (BECKETT, 2023). Si tratta di ''macchine ingannevoli'' (NATALE, 2022), che, a partire da un prompt, generano testi, immagini o video attraverso sequenze complesse di calcoli di probabilità statistica (NARAYANAN e KAPOOR, 2024).

Come scrive Natale (2022), l'inganno, ovvero la capacità di simulazione della comunicazione e della creatività umana da parte della macchina, e la disponibilità a crederci da parte degli interlocutori umani, sono elementi costitutivi della comunicazione uomo-macchina, che mascherano la natura macchinica e automatizzata dietro un'apparenza antropomorfa. I grandi modelli linguistici (LLMs) su cui si fondano i chatbot come ChatGPT, infatti, sono stati definiti pappagalli

stocastici (*Stochastic parrots*): generano output verisimili e coerenti, ma «la coerenza è in realtà negli occhi di chi guarda» (BENDER et al., 2021, p. 616).

L'abilità di simulare comunicazione e creatività non va, tuttavia, confusa con l'intelligenza. Al contrario, quando un chatbot risponde a un prompt, non ha una visione coerente della risposta che produrrà; piuttosto esegue un numero immenso di calcoli con l'obiettivo di prevedere quale debba essere la prima parola della risposta, e quali parole generare di seguito. Come spiegano Narayanan e Kapoor (2024, p. 133), per generare «un singolo *token* — parte di una parola — ChatGPT deve eseguire all'incirca un trilione di operazioni aritmetiche». La comunicazione automatizzata, quindi, manca sia di intenzionalità sia di una comprensione condivisa del contesto di un determinato scambio comunicativo. Eppure, da un punto di vista meramente tecnologico, l'introduzione prima di ChatGPT e poi di GPT personalizzati ha rappresentato un'innovazione dirompente sia per il mercato dei consumatori sia per le aziende e le redazioni giornalistiche.

La diffusione dell'utilizzo dell'intelligenza artificiale e la sua presenza (più o meno visibili) nei processi di produzione delle notizie sta trasformando il modo in cui l'informazione viene creata e diffusa, ma al tempo stesso amplifica il rischio di disinformazione. Strumenti come i deepfake e i generatori di testo possono produrre contenuti falsi dall'aspetto credibile, difficili da individuare, mettendo alla prova il sistema informativo, i media e la fiducia dell'opinione pubblica nei confronti del mondo giornalistico e della politica. Per questo è essenziale comprendere l'impatto dell'AI sulla disinformazione e sviluppare strategie efficaci per contrastarla.

3. Le macchine della menzogna contemporanee.

Il tema della disinformazione ha caratterizzato in modo distintivo le campagne elettorali di paesi democratici e regimi non democratici negli ultimi dieci anni. Infatti, a partire dal 2016, anno della campagna per il referendum sulla Brexit (23 giugno 2016) e poi delle elezioni presidenziali americane vinte da Donald Trump, la disinformazione è tornata sotto i riflettori di esperti e opinione pubblica come risorsa strategica nella comunicazione politica, soprattutto in relazione ai rischi di interferenze di attori stranieri nelle campagne elettorali. La parola chiave simbolo delle campagne elettorali del 2016 è

stata, infatti, Fake news. In verità il termine 'Fake News' è ormai poco usato nel dibattito accademico e scientifico, per due motivi.

Il primo motivo riguarda l'incapacità del termine di contenere in sé lo spettro di notizie che rientrano nella disinformazione e che non si esaurisce nelle notizie 'false' — molte sono infatti notizie fuorvianti, decontestualizzate e manipolate ma non integralmente fabbricate. Proprio analizzando le presidenziali del 2016, Wardle (2017) ha identificato una tipologia di notizie problematiche, lungo un continuum che va dalla satira alla notizia falsa. Un secondo motivo riguarda l'utilizzo strumentale che è stato fatto del termine, che ha subito una sorta di torsione ideologica e semantica negli ultimi anni: dopo che per molto tempo i ricercatori, politici e intellettuali vicini al Partito Democratico hanno utilizzato il termine fake news per condannare l'uso strategico della disinformazione come forma di comunicazione politica nella campagna elettorale di Trump, volta a delegittimare gli avversari e demotivare parte dell'elettorato democratico (Roemmele e Gibson, 2020), lo stesso Trump e la destra alternativa americana hanno fatto proprio il termine fake news, in una campagna contro i news media e il giornalismo investigativo accusati di fare disinformazione. Sono quindi ricorrenti i *tweet* di Trump dove si può leggere « fake news CNN », o « fake news New York Times », sia nei mesi di campagna sia durante tutto il suo mandato, da Presidente degli Stati Uniti. Il termine è stato quindi utilizzato da Trump per screditare la stampa libera, ribaltandone l'uso e il significato, in un paradosso tipico del 'Socing' orwelliano. Il termine Fake news è quindi un significante flottante (Farkas e Schou, 2018), mobilitato in sistemi di significato e valoriali distinti, che fanno riferimento a progetti politici antagonisti, nella battaglia per il controllo egemonico del sociale. La posta in gioco, in altre parole, è la definizione della realtà, per conferire legittimità a un progetto politico.

Data la natura fortemente ideologica del termine Fake news, oggi si preferisce, quindi, parlare di mis- o disinformation, distinguendo fra la diffusione accidentale di informazioni false o non accurate (misinformazione) — spesso per mancata verifica delle fonti — e la diffusione intenzionale di disinformazione.

Essere attori inconsapevoli della *misinformation* è molto facile, perché sulle piattaforme il consumo di notizie avviene in un contesto squisitamente comunicativo e relazionale, modellato dalle aspettative di reciprocità che regolano le interazioni interpersonali, e da criteri di

credibilità e autorevolezza delle fonti definiti sulla base della fiducia negli interlocutori, anziché sull'autorevolezza della fonte. Ci capita di condividere notizie avendone letto solo il titolo (che ci attrae e inganna), sulla base del fatto che siano state condivise dai nostri contatti (di cui a volte ci fidiamo più che dei media).

E poi c'è la disinformazione, prodotta per screditare avversari politici, confondere l'opinione pubblica e minare la fiducia nei mezzi di comunicazione e nel sistema informativo, tipica delle campagne elettorali di tipo 'sovversivo' (ROEMMELE e GIBSON, 2020) — così definite perché la mistificazione e la manipolazione dei fatti sovverte il principio democratico di una comunicazione politica orientata a informare e mobilitare i cittadini al voto. Uno dei primi ricercatori a studiare la disinformazione come strategia di comunicazione politica sui social media è stato Phil Howard, a lungo direttore dell'Oxford Internet Institute, famoso per la definizione di *computational propaganda* (propaganda automatizzata, ovvero «l'uso di algoritmi, automazione e curatela umana per distribuire intenzionalmente informazioni fuorvianti attraverso i social media» (WOOLLEY e HOWARD, 2016), e *lie machines* (macchine della menzogna, HOWARD, 2020), ovvero «un sistema di persone e tecnologie che distribuiscono messaggi falsi al servizio di un'agenda politica». Howard ha cercato di individuare i 3 principali soggetti che producono notizie false, ovvero gli attori coinvolti nelle macchine della menzogna:

— il produttore di menzogne al servizio di un'ideologia e di una parte politica: si tratta di candidati politici, partiti politici, staff della campagna elettorale, istituzioni di governi autoritari, o attori politici marginali che nella loro lotta politica utilizzano anche l'arma scorretta della bugia o della mistificazione, in particolare per screditare avversari, ma anche per sovradimensionare meriti della propria parte, attraverso la produzione di disinformazione;

— i distributori di menzogne: gli algoritmi dei social media, cioè le regole di distribuzione dei contenuti negli ambienti digitali che favorirebbero la diffusione di disinformazione perchè programmati per premiare (senza valutazione sulla qualità e accuratezza) i contenuti che creano più dibattito, coinvolgimento, partecipazione, come spesso accade con le menzogne;

— i commercianti delle menzogne politiche. Si tratta di società di consulenza, lobbisti, che traggono profitto dalla vendita, dall'amplificazione e dalla promozione della disinformazione.

Quindi perché le menzogne hanno successo e si sono moltiplicate nell'ecosistema digitale?

« Ciò che distingue le macchine della menzogna contemporanee è il loro basso costo di produzione, la grande velocità di diffusione e l'ampia visibilità della disinformazione sui social media, nonché l'industria in espansione delle agenzie di marketing che aiutano a amplificare la propaganda computazionale (HOWARD, 2020, p. xvi) ».

4. L'IA generativa moltiplica la disinformazione?

La campagna elettorale del 2024 ha offerto nuovi esempi di macchine della menzogna al servizio di un'agenda politica. Ad esempio, Elon Musk — forte dei suoi 202 milioni di follower e della possibilità di mandare alcuni suoi post a tutti gli utenti del suo social network X — si è fatto megafono della macchina della menzogna trumpiana condividendo le principali *junk news* (notizie spazzatura) sui temi dell'immigrazione, del cambiamento climatico, delle procedure di voto.

PolitiFact ha analizzato più di 450 post condivisi da Musk su X tra il 30 settembre e il 14 ottobre. Fra questi, molti contenevano la disinformazione al cuore della campagna di Trump, ad esempio: accuse alla Fema (l'Agenzia federale per la gestione delle emergenze) per la sua risposta inadeguata alla distruzione provocata dall'uragano Helene in North Carolina, insinuando ripetutamente false affermazioni sul fatto che i lavoratori federali avrebbero bloccato gli aiuti; Musk ha amplificato la falsa narrativa secondo cui l'amministrazione Biden-Harris avrebbe sottratto denaro ai soccorsi della Fema per destinarlo agli immigrati; ha identificato nella presenza di immigrati la causa dei prezzi in crescita del mercato immobiliare; infine, ha condiviso false notizie, già diffuse da Trump, sul fatto che il governatore della California abbia reso illegale l'identificazione degli elettori ai seggi elettorali.

Se quindi le campagne elettorali del 2016 (Presidenziali americane e Brexit) sono state le campagne che hanno posto al centro il tema della menzogna e della disinformazione, le competizioni elettorali del 2024, sono state quelle in cui ci si è più preoccupati dell'interferenza nelle campagne elettorali (le elezioni europee di giugno, e, naturalmente, le presidenziali americane) di contenuti prodotti dall'IA generativa. L'Agenzia per la sicurezza informatica e delle infrastrutture (*Cybersecurity and Infrastructure Security Agency*, CISA) ha

identificato i seguenti rischi derivanti dalla GenAI negli Stati Uniti durante il ciclo elettorale del 2024: video (*deepfakes*, video falsi creati o manipolati artificialmente per sembrare reali, con l'utilizzo di IA e tecniche di deep learning); immagini alterate dall'IA (come l'immagine generata artificialmente di una bambina con in braccio un cucciolo su una barca durante l'alluvione provocata dall'uragano Helene); audio (clonazione vocale, come la finta telefonata di Joe Biden durante le primarie democratiche per la campagna elettorale del 2024); e infine, testuali. Secondo una ricerca di Wired, immagini, audio, chatbot, video e altri contenuti di GenAI sono stati utilizzati o condivisi da personalità politiche o da personaggi di spicco durante la campagna elettorale. Una ricerca condotta da AI Democracy Project mostra un altro rischio: la maggior parte dei chatbot testati (ChatGPT di OpenAI, Claude di Anthropic, Gemini di Google e Meta AI) davano informazioni poco accurate sul processo elettorale (ad esempio sul seggio elettorale) o non includevano il riferimento ai siti governativi.

Ma queste dinamiche non sono prerogative del contesto occidentale: in paesi come India, Nigeria, Brasile, l'uso di immagini e video falsi e manipolati con l'AI durante le campagne elettorali più importanti ha dimostrato la vulnerabilità di Paesi con un crescente uso dei media digitali tra la popolazione e una contemporanea fragilità democratica e delle culture civiche.

Senza dimenticare la disinformazione sulla guerra a Gaza, dove colpiscono i siti come Gazawood, o Paliwood, che sostengono che l'occupazione israeliana sia solo una messa in scena a beneficio degli spettatori; o gli account social di presunti ristoranti di Gaza con tavole imbandite, per negare la fame che attanaglia la popolazione palestinese.

C'è chi dice che, tutto sommato, i *cheapfakes* siano altrettanto efficaci (ad esempio il video di un'intervista alla speaker del congresso, Nancy Pelosi, in cui l'audio viene rallentato generando l'impressione che Pelosi sia ubriaca). In ogni caso, anche l'uso dell'IA generativa, come altre forme di disinformazione, è tesa a screditare gli avversari politici, delegittimare il processo democratico insinuando il dubbio di brogli elettorali, e seminare una sfiducia generale nei media e nei giornalisti.

In questo contesto alterato bisogna quindi chiedersi quale sia l'effetto non di un singolo post, ma dell'aumento della percezione che tutti i media, e, in fondo, tutti i contenuti che incontriamo online,

siano inaffidabili. Lo hanno già sperimentato i palestinesi che postano foto della loro vita quotidiana sui social e sono accusati di essere dei fake (Novak, 2025).

Note bibliografiche

Beckett C. (2023), *Generating Change. A global survey of what news organisations are doing with artificial intelligence.* Polis, London School of Economics and Political Science in collaboration with Google News Initiative. Accessibile al link: www.journalismai.info/research.

Bender E.M.-Gebru T.-McMillan-Major A.-Shmitchell S. (2021), *On the dangers of stochastic parrots: Can language models be too big?*, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, New York, ACM, accessibile al link: *https://doi.org/10.1145/3442188.3445922*.

Colombo F. (2020), *Ecologia dei media. Manifesto per una comunicazione gentile*, Milano, Vita e Pensiero.

Dondi M. (2015), *L'eco del boato: storia della strategia della tensione, 1965-1974*, Bari, Editori Laterza.

Farkas J.-Schou J. (2018), *Fake news as a floating signifier: Hegemony, antagonism and the politics of falsehood*, in *Javnost-the public*, 25(3), p. 298 ss.

Howard P.N. (2020), *Lie machines: How to save democracy from troll armies, deceitful robots, junk news operations, and political operatives*, Cambridge, Yale University Press.

Lasswell H.D. (1927), *The theory of political propaganda*, in *American Political Science Review*, 21(3), p. 627 ss.

Mitchell M. (2019), *Artificial Intelligence*, London, Penguin Books.

Narayanan A.-Kapoor S. (2024), *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*, Princeton University Press.

Natale S. (2022), *Macchine ingannevoli*, Milano, Einaudi.

Novak M. (2025), *The Rise of AI Is Making Life Even Harder for Real People in Gaza. Gizmodo* (05.07.2025), Accessibile al link: *https://gizmodo.com/the-rise-of-ai-is-making-life-even-harder-for-real-people-in-gaza-2000607395*.

Roemmele A.-Gibson, R. (2020), *Scientific and subversive: The two faces of the fourth era of political campaigning*, in *New Media & Society*, 22, p. 595 ss.

Wardle C. (2017), *Fake News. It's complicated. Medium.* Accessibile al link: *https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79*.

Woolley S.C.-Howard P.N. (2016), *Automation, algorithms, and politics| political communication, computational propaganda, and autonomous agents - Introduction*, in *International Journal of Communication*, 10, p. 9 ss.

(IL FATTO - PARTE I)
LA (POST) VERITÀ FRA PUBBLICO E PRIVATO

di *Alberto Oddenino*

Sommario: 1. Introduzione: cambio d'epoca e disordine informativo. — 2. La progressiva contaminazione dello spazio informazionale. — 3. I possibili rimedi fattuali, con particolare riferimento a *fact checking* e *debunking*.

1. Introduzione: cambio d'epoca e disordine informativo.

L'Era Digitale determina, fra le molte e fondamentali mutazioni strutturali della società, anche un radicale ridisegno della libertà di espressione. L'utilizzo sempre più capillare dei *social networks* su scala globale rifonda le regole del discorso pubblico e consente alle piattaforme, novelli poteri privati, di penetrare nel ventre molle della stessa sovranità statale.

Accanto ai monopoli statali che definiscono tradizionalmente la nozione di sovranità (moneta, forza e giustizia quelli tradizionalmente individuati, nella triade emblematica di conio, spada e bilancia) ve ne è uno, meno riconosciuto ma non meno importante, che è posto in radicale discussione: il ruolo ultimo nella perimetrazione dello spazio informativo pubblico, moderna *agorà* digitale, il cui controllo oggi vacilla insieme al concetto stesso di verità, scuotendo le basi dei nostri sistemi democratici (Caravita, 2019).

Un mutamento epocale che ha un impatto ad oggi ancora difficile da valutare in senso sistemico, ma di cui sono più che evidenti alcune epifanie, come la proliferazione, accanto alle *fake news* e ai *clickbait*, dei discorsi d'odio e dei correlativi *rage-bait*, che determinano una contaminazione dei sistemi di informazione nel senso di un crescente livello di disinformazione e polarizzazione delle opinioni (Riva, 2018), con ricadute non trascurabili sulla coesione sociale tanto nel mondo *online* quanto, di ritorno attraverso l'*onlife* teorizzato da Floridi (2017), in quello fisico (Turkle, 2016).

In tale contesto, che investe, insieme al perimetro della libertà di opinione ed espressione, l'essenza stessa dell'uomo come organismo informazionale e come essere sociale e relazionale, (FLORIDI, 2014), le piattaforme digitali sono diventate intermediari sostanzialmente necessari della comunicazione, influenzando la formazione dell'opinione pubblica e partecipando direttamente alla costruzione (e decostruzione) del discorso democratico (PARUZZO, 2022).

Si tratta, a ben vedere, dell'onda lunga di quello che si è realizzato con il cosiddetto uso di prossimità della rete Internet, divenuta lo strumento principe per il reperimento e la verifica dell'informazione. Non a caso, ciò segna l'emersione del paradigma della *governance*, che si declinava come necessario per descrivere la complessità delle sfide regolative della struttura della Rete (già ODDENINO, 2008) e oggi si ripropone, con tutte le sue ambiguità e i suoi nodi irrisolti, come modello di regolazione dei suoi contenuti.

Lo stessa concetto di verità è così esposto a rischio di obsolescenza, e non a caso diviene frequente e popolare il ricorso alla nozione di 'post-verità', quasi che la partita per preservare la distinzione fra vero e falso fosse irrimediabilmente perduta.

Nello specifico, in quello che è efficacemente definito lo spettro dell'*information disorder* (WARDLE e DERAKSHAN, 2017), si ritrovano molte e diverse epifanie che sono oggetto di una scienza sempre più autonoma, quella delle *Fake News*. Trattandosi di un oggetto cruciale di regolazione occorre, prendere atto che al di là di una rassicurante etichetta unificante, si è di fronte ad una costellazione piuttosto variegata e finanche disomogenea. Essa si snoda sugli assi della fatticità (che si articola fra i poli opposti, forse più teorici che pratici, dell'assolutamente vero e del totalmente falso) e della intenzionalità (che si snoda fra i poli opposti del danno totalmente inintenzionale e del danno completamente intenzionale): transizioni progressive ci portano dallo spettro della misinformazione verso la disinformazione fino alla informazione malevola (RUFFO, 2022), con manifestazioni disparate che vanno dal *clickbait* alla propaganda, dalle leggende urbane alle teorie del complotto e via enumerando.

Un panorama assai arduo da comprendere ancor prima che da regolare, proprio perché la transizione fra una figura e l'altra presenta spesso aree grigie e come è noto, e come si vedrà meglio nella prospettiva del diritto, le regole hanno invece bisogno di fattispecie il più possibile chiare e definite per essere efficacemente applicate.

2. La progressiva contaminazione dello spazio informazionale.

Il crescente ruolo di intermediazione delle piattaforme, porta con sè, oltre che una trilateralizzazione del diritto alla libertà di espressione e informazione (BALKIN, 2014) di cui si dirà in apposito contributo, un evidente deterioramento dello spazio informazionale: il fenomeno presenta molteplici risvolti (si pensi anche, a mero esempio, alla progressiva polarizzazione dei discorsi d'odio) ma è in questa sede considerato nella prospettiva del crescente fenomeno della disinformazione *online*, delle *fake news* spinte oggi fino alle nuove frontiere del *deep fake* abilitato dalla intelligenza artificiale generativa.

Un fenomeno dirompente, che va di pari passo con una crescente sfiducia nei confronti delle istituzioni e dei media tradizionali in relazione alla quale, come accennato, le piattaforme giocano un ruolo fondamentale in ragione del loro modello di profitto (COHEN, 2019). Tramite gli algoritmi di raccomandazione, guidati da logiche di profilazione, esse tendono a privilegiare contenuti polarizzanti che contribuiscono all'esposizione dell'utente a *filter bubbles* ed *echo chambers*.

Si tratta di concetti fondamentali che meritano un breve approfondimento, muovendo dal loro presupposto, ossia la profilazione. Con essa si intende, ex articolo 4 del GDPR, "*qualsiasi forma di trattamento automatizzato di dati personali, consistente nell'utilizzo di tali dati personali, per valutare determinati aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica*". La profilazione consente, pertanto, di raccogliere informazioni sugli utenti e di analizzare i loro modelli comportamentali, al fine di inserire il singolo profilo in una determinata categoria che consente di prevederne interessi e attività.

Quanto alle '*filter bubbles*', queste non sono altro che la diretta conseguenza di tale opera di categorizzazione svolta dagli algoritmi delle piattaforme *online*, che fanno in modo che l'utente venga esposto in grande prevalenza a contenuti che risultino di suo gradimento in base ai dati di profilazione. Ciò limita fino ad annullarlo il contatto del soggetto con idee differenti dalle proprie, e i contenuti proposti sono preordinati a confermare o rafforzare le sue preesistenti convinzioni, tenendolo avvinto allo spazio informazionale.

Qui si innesta la nozione di '*echo chambers*', definibile come una dinamica di gruppo digitale in base alla quale l'utente si interfaccia

non solo con contenuti a lui affini, ma anche con altri utenti che condividono le medesime posizioni o i medesimi interessi. Ciò ha un effetto di forte limitazione della stessa libertà di espressione intesa come libertà di informarsi e reperire fonti. Ancor più essa impatta sulla autodeterminazione informativa e conduce ad un discorso pubblico fortemente polarizzato.

Va sottolineato come il fenomeno non riguardi solo l'ecosistema dei *social networks*, dove in realtà è piuttosto comprensibile che ciò accada, dato che lo scopo stesso di tali ambienti digitali è quello di creare una rete di contatti e di relazioni intorno all'utente, ma ha pervaso anche gli algoritmi di ricerca dei grandi motori di ricerca, dove da tempo ormai i primi risultati a comparire non sono più quelli in generale più cercati, ma bensì quelli ritenuti più adatti alle preferenze dell'utente, con una pericolosa rideterminazione soggettiva della rilevanza che pareva aliena alle logiche originarie dei motori di ricerca, ispirati a standard oggettivi di ricorrenza statistica (PARISER, 2012).

A questi fenomeni si aggiunge, amplificandoli, la strategia dei cd. *dark patterns*, volti a influenzare in modo surrettizio il comportamento degli utenti, spinti a fare scelte meno favorevoli per la loro privacy e più convenienti per la piattaforma. Secondo l'EDPB (*European Data Protection Board*) aspetti in apparenza meramente grafici, come la collocazione delle opzioni o il design della piattaforma, vanno ad impattare sulla libertà decisionale dell'utente, che finisce per non modificare le impostazioni predefinite, anche se formalmente potrebbe farlo: una esemplare applicazione del cd. '*default effect*' secondo cui la maggior parte degli utenti tende a non modificare le impostazioni predefinite.

Tra le principali tipologie di *dark patterns* troviamo '*l'overloading*', che consiste in un surplus di 'stimoli' in particolare un'eccessiva pluralità di scelte, che porta l'utente a confondersi e quindi a scegliere le opzioni già preimpostate. Un altro *dark pattern* ormai ben codificato è lo '*skipping*', che si verifica quando l'interfaccia propone all'utente un testo troppo lungo e complesso, in modo che questo sia spinto ad andare avanti senza leggere per cliccare direttamente sull'opzione ''accetta tutto''. Ancora, piuttosto frequente è '*l'hindering*', meccanismo che rende difficile modificare le opzioni predefinite in ragione di lunghezze e complessità che inducono l'utente a desistere. E infine strategie ancor più sottili, ma non per questo meno efficaci, come lo '*stirring*', che si basano su una manipolazione visiva delle scelte, come l'uso dei colori accesi o spenti.

In definitiva anche i *dark patterns* sono oggi riconosciuti come forme di induzione (*nudging*) con un ruolo fondamentale nel dare una forma all'ambiente informazionale che risulta funzionale ai preminenti interessi delle piattaforme.

Quanto si è visto costituisce il presupposto per sviluppare un *humus* assolutamente fertile per il proliferare della disinformazione *online*, rispetto alla quale appare opportuno indagare lo spettro dei possibili rimedi fattuali.

3. I possibili rimedi fattuali, con particolare riferimento a *fact checking* e *debunking*.

A fronte di un quadro di progressiva contaminazione dell'ambiente informazionale come quello che si è brevemente richiamato, e che appare per molti versi un processo irreversibile, ci si può chiedere quali siano i possibili rimedi fattuali, e in proposito occorre fare rapido cenno alle pratiche di *fact-checking* e di *debunking* rinviando al contributo specifico che segue, ad opera di Gallotti, che, muovendo da una attività sul campo, dà conto delle prospettive e delle sfide connesse con tali delicatissime attività.

In prima approssimazione, il *fact-checking* consiste nel lavoro di verifica delle fonti e dei fatti, mentre il *debunking* consiste nello smentire le narrazioni fuorvianti derivanti dalla diffusione di notizie false. In sostanza, il primo è un meccanismo che interviene a monte, mentre il secondo opera a valle: il *fact-checking* è volto a prevenire la diffusione di notizie false, il *debunking* è una risposta puntuale, che interviene (o dovrebbe intervenire) solo laddove ci sia già stata un'operazione di verifica delle fonti.

Più analiticamente, il *fact-checking* è il processo di verifica di tutte le fonti e i fatti riportati in un determinato testo o discorso, allo scopo di garantire agli utenti una corretta informazione che li porti a maturare opinioni ponderate e consapevoli.

A svolgere il compito di *fact-checker* sono spesso delle redazioni indipendenti, nate da progetti autonomi oppure collegate alle testate giornalistiche.

Sebbene non esista un apposito albo professionale per i *fact-checkers*, questi devono comunque aderire a standard condivisi di trasparenza, indipendenza e metodologia, in accordo con l'*European Code of Standards for Independent Fact-Checking Organisations*, soprattutto

nell'ipotesi in cui i 'verificatori' decidano di aderire all'*European Fact-Checking Standards Network* (EFCSN), un network indipendente che si occupa di definire dei principi etico-operativi per i *fact checkers*.

Entrambi possono essere definiti come misure di 'soft governance' volte ad arginare la diffusione di notizie false. Questi meccanismi godono del supporto di hub di ricerca indipendenti, che raccolgono ed analizzano i dati relativi alle campagne di disinformazione, promuovendo la collaborazione tra accademici e *fact-checkers* attraverso la pubblicazione di rapporti e *policy papers*.

In particolare, in un'ottica di supporto ai meccanismi di *fact-checking*, nel 2020 la Commissione Europea ha istituito l'*European Digital Media Observatory* (EDMO), un *hub* indipendente di ricerca e coordinamento, che riunisce università, centri di ricerca e altri *fact-checkers* allo scopo di analizzare e monitorare la disinformazione online. L'EDMO non ha uno scopo regolatorio ma di ricerca, contribuendo alla creazione di banche dati sui contenuti manipolati e collaborando con le istituzioni europee al fine di fornire dati empirici sulle campagne di disinformazione.

Deve essere sottolineato che il *fact-checking* è un'attività di verifica, al semplice scopo di etichettare un contenuto come vero o falso, oppure come preciso o impreciso: esso non comporta necessariamente una automatica rimozione dello stesso.

Per quanto riguarda le piattaforme *online*, esistono essenzialmente tre diversi modi di applicare il *fact-checking* ai contenuti: affidandosi a organizzazioni esterne e indipendenti, le quali operano una valutazione dei contenuti al fine di valutarne la veridicità; tramite meccanismi *community-driven* interni come le *Community notes* di X, dove gli utenti segnalano e aggiungono contesto ai contenuti ritenuti ingannevoli; tramite l'utilizzo di sistemi automatici basati sugli algoritmi e sull'intelligenza artificiale, sempre in chiave interna.

È chiaro che il meccanismo esterno è molto più credibile e non sorprende che nel quadro di deterioramento che si è evocato esso sia divenuto sempre meno praticato: emblematico il caso di Meta che, a seguito delle pressioni politiche e di alcuni cambiamenti di governo (con riferimento, soprattutto, alle ultime elezioni americane), ha interrotto la collaborazione con meccanismi di *fact-checking* esterni per segnare un repentino cambio di rotta, passando a meccanismi di *fact checking 'community driven'*, che, sebbene consentano un ruolo più partecipativo degli utenti, risultano essere anche molto più manipolabili e quindi meno sicuri.

Nell'attuale ecosistema digitale, vieppiù pervaso dai rischi di *deep fake*, un meccanismo di *fact-checking* esterno ben funzionante non può essere visto come forma di censura: è piuttosto un campanello d'allarme a salvaguardia dei diritti fondamentali dell'individuo, tra cui anche la libertà di espressione e di informazione.

Per questo motivo, sebbene le organizzazioni come l'EDMO e l'EFCSN siano un primo passo fondamentale, sarebbe opportuno che, pur mantenendo anche meccanismi di *community-drive* o *crowd sourced*, il *fact-checking* 'esterno' venisse gestito e regolamentato ulteriormente in modo da garantire la presenza sul territorio europeo di appositi organi pubblici che, per svolgere il proprio lavoro, forniscano effettive garanzie di terzietà e di indipendenza. Con essi le piattaforme dovrebbero essere tenute ad interfacciarsi in un'ottica di supervisione dei meccanismi interni, onde evitare l'emergere di manipolazioni o polarizzazioni delle opinioni e dei contenuti.

In definitiva, nel contesto attuale, dominato da una grande sfiducia nei confronti delle istituzioni e dei media tradizionali, le attività di *fact checking* certamente svolgono un importante ruolo di ricostruzione di un tessuto di coesione e fiducia, anche se la mancanza di regole precise e il sempre delicato tema del controllo dei controllori (*quis custodiet ipsos custodes* nell'antica formulazione) restano problemi aperti, e con essi il rischio di una conferma della tendenza alla ''privatizzazione della verità'' che questi stessi strumenti si propongono di arginare. Tali problemi richiedono, al di là delle risposte fattuali, l'emersione di regole giuridiche certe applicabili e possibilmente vincolanti. Su queste ultime, e sulla sfida della ricerca del diritto, cui sarà dedicato apposito contributo, si ritornerà a breve. Prima, però, si esaminerà nel dettaglio, sempre nella prospettiva de il 'fatto' della disinformazione, un progetto sul campo della Fondazione Bruno Kessler che si propone di utilizzare l'intelligenza artificiale per arginare il dilagante fenomeno del *fake* e del *deep fake*: una sfida ambiziosa che si confronta con il panorama che si è brevemente descritto e che non manca di incontrare ostacoli fattuali, e normativi, rilevanti.

Note bibliografiche

BALKIN J. (2024), *Old-school/new-school speech regulation*, in *Harvard Law Review*, 127, p. 2296 ss.

CARAVITA B. (2019), *Social network, formazione del consenso, istituzioni politiche: quale regolamentazione?*, in *Federalismi.it*, 2.

COHEN J.E. (2019), *Between Truth and Power. The Legal Constructions of Informational Capitalism*, New York, Oxford University Press.
FLORIDI L. (2014), *The Onlife Manifesto. Being Human in a Hyperconnected Era*, Berlin, Springer.
FLORIDI L. (2017), *La Quarta Rivoluzione. Come l'infosfera sta trasformando il mondo*, Milano, Raffaello Cortina Editore.
MONTI M. (2019), *Le Internet platforms, il discorso pubblico e la democrazia*, in *Quaderni costituzionali*, 4, p. 811 ss.
NAPOLI P.M. (2015), *Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers*, in *Telecommunications Policy*, 39, 9, p. 751 ss.
ODDENINO A. (2008), *La governance di Internet fra autoregolazione, sovranità statale e diritto internazionale*, Torino, Giappichelli.
PARISER E. (2012), *Il Filtro: Quello che internet ci nasconde*, Milano, Il Saggiatore.
PARUZZO F. (2022), *I sovrani della Rete. Piattaforme digitali e limiti costituzionali al potere privato*, Napoli, Edizioni Scientifiche.
PITRUZZELLA G.-POLLICINO O.-QUINTARELLI S. (2017), *Parole e potere: libertà di espressione, hate speech e fake news*, Milano, Egea.
RIVA G. (2018), *Fake News*, Bologna, Il Mulino.
RUFFO G. (2022), *Rintracciando odio e bugie in rete*, in DURANTE M. e PAGALLO U. (ed.) *La politica dei dati. Il governo delle nuove tecnologie tra diritto, economia e società*, Milano, Mimesis, p. 163 ss.
SCHAAKE M. (2024), *The Tech Coup. How to save Democracy from Silicon Valley*, Princeton, Princeton University Press.
SHOEMAKER P.J.-VOS T.P., *Gatekeeping Theory*, New York, Routledge.
TEUBNER G. (1996) *Il diritto come sistema autopoietico*, a cura di FEBBRAJO A.-PENNISI C., Milano, Giuffrè.
TURKLE S. (2016), *La conversazione necessaria*, Torino, Einaudi.
WARDLE C.-DERAKHSHAN H. (2017), *Information Disorder. Toward an interdisciplinary framework for research and policy making*, Council of Europe.
WU T. (2017), *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*, London, Knopf.

(IL FATTO - PARTE II)
AI4TRUST: SFIDE DI UN PROGETTO (DI FACT CHECKING) PER USARE L'IA CONTRO FAKE E DEEP FAKE

di *Riccardo Gallotti*

Sommario: 1. Introduzione. — 2. Il Contesto del 2022 e la Realtà Attuale. — 3. Il Digital Services Act e la Realtà dell'Accesso ai Dati. — 4. API e Web Scraping: Due Approcci Tecnici. — 5. Il Caso Bright Data. — 6. Problemi Aggiuntivi: YouTube e i Transcript e Privacy vs Necessità della Ricerca. — 7. Il Problema dello Sviluppo AI.

1. Introduzione.

Vi racconto il progetto che coordino, AI4TRUST, un progetto europeo da 7 milioni di euro che include 17 partners, con una durata di 3 anni. L'idea di AI4TRUST è di creare una piattaforma che, in tempo reale ogni giorno, guardi quali sono i discorsi che avvengono sul web, sui social media e nelle news. Cerchiamo di capire, utilizzando vari strumenti di intelligenza artificiale (*Artificial Intelligence* - AI), quali sono le notizie che possono essere false o fuorvianti, quali sono le immagini che sono state manipolate, quelle che in verità sono vecchie e sono state riutilizzate in un altro contesto, gli audio che sono stati contraffatti, e gli argomenti che sono stati trattati in maniera maliziosa.

Questa operazione viene effettuata automaticamente ogni giorno, con il fine di mettere queste informazioni sul tavolo dei fact-checker, che possono verificare se ciò che abbiamo visto è corretto e inserire il tutto all'interno di un ciclo di produzione news e controllo naturale del sistema mediatico attuale. Ci sono una produzione di news e ci sono persone pagate per verificare che siano vere. Il problema è che, in questo momento, con lo sviluppo dei nuovi tool di AI, c'è stata una grandissima accelerazione della produzione di contenuti 'dubbi', sia immagini manipolate e create sinteticamente, che interi siti gestiti completamente dall'AI.

È qualcosa di completamente nuovo, per limitare le conseguenze del recente sviluppo l'AI, cerchiamo di utilizzare l'AI stessa come soluzione. Tutto questo parte dai dati, dal fatto che negli ultimi 10 anni ci sia stata una grandissima nuova disponibilità di dati, e abbiamo bisogno di questi dati per combattere la disinformazione, così come è necessario per lo sviluppo della stessa AI.

Nell'ambito della disinformazione, il ruolo centrale lo hanno le piattaforme: Meta, Twitter/X, Google con Youtube, e altre nuove piattaforme che sono emerse negli ultimi anni. Le piattaforme sviluppano due diversi aspetti che dominano la comunicazione online: uno è il famigerato algoritmo, qualcosa che decide cosa è più visibile di qualcos'altro, e poi ci sono stimoli a processi sociali che, per esempio, promuovono la diffusione virale di qualcosa all'interno delle nostre società.

L'uso dell'AI all'interno della comunicazione online ha portato a effetti anche negativi. Se dei siti sono completamente gestiti dall'AI, chiaramente siamo vittime del rischio di allucinazioni, cioè di momenti in cui l'AI dice qualcosa di verosimile ma che in verità è falso. Inoltre, ancor peggio, si facilita la creazione di campagne disinformative organizzate che possono evocare uno stormo di bot che vanno a sostegno delle loro campagne. Siamo a un passo dal chiedere ad Alexa: 'Per favore, aiutami a pianificare una campagna di disinformazione'.

In questo ambito è necessario diventare in grado di tracciare su grande scala quello che succede online. Noi lo vogliamo fare in maniera molto ambiziosa, a livello multimodale: non soltanto testo, immagini, o audio, ma a tutto tondo, con un approccio multilingua, con una copertura di otto diverse lingue europee, e multipiattaforma. Non vogliamo dunque focalizzarci, come facevamo in passato su Twitter, anche perché Twitter è sempre meno utilizzato, ma analizzare su diverse piattaforme quali sono i contenuti in circolazione.

Il problema è che, mio malgrado, mi sono trovato in mezzo a qualcosa che è in continua evoluzione, soprattutto legale. Il tema del progetto di ricerca che è discusso qui è il rapporto tra i controlli, le funzioni pubbliche e i controlli privati. Queste piattaforme hanno una funzione di agorà pubblica, ma si autoregolano in molte forme.

Qui arriva il mio problema: sono qua per chiedere una consulenza al professor Alberto Oddenino perché ho numerosi problemi di carattere legale legati al mio progetto. Mi trovo in questo punto di interfaccia tra la necessità di creare una soluzione per il bene comune e i limiti che le piattaforme mi danno.

2. Il Contesto del 2022 e la Realtà Attuale.

Il progetto, come l'abbiamo scritto, è stato pensato per un social media del 2022, in cui Twitter era la piattaforma dove tutti gli scienziati, in maniera molto facile, potevano andare a pescare tutti i dati di cui necessitavano, e dove Twitter era l'agorà utilizzata da tutti i politici per discutere di qualunque tema. Si poteva studiare Twitter e pensare che esso, come strumento, fosse un'ottima rappresentazione di tutta la società.

Ora, purtroppo, dopo il passaggio di proprietà di Twitter, che è diventato X ed è passato di proprietà a Elon Musk, Elon Musk ha deciso di non condividere più con i ricercatori i propri dati. Allo stesso tempo, Facebook/Meta aveva una bellissima piattaforma che si chiama CrowdTangle, che ha chiuso ad agosto 2024: un'altra fonte dati per noi essenziale nello studio dei social media online. Facebook/Meta attualmente sta offrendo a qualcuno, poco alla volta, alternative con una nuova piattaforma, ma non siamo sicuri che possa soddisfare le esigenze del nostro progetto.

Questo cambiamento ci ha portato su un nuovo territorio non battuto tra il fronte tecnologico e legale, perché, come detto nell'introduzione, questo dialogo è una frontiera che incontra i limiti della legislazione, un percorso di incontro inesplorato. Il problema è che, in questi ostacoli, ci troviamo tra regolamenti e procedure. Io, che non sono un giurista, ma un fisico teorico, però mi sono trovato a dover leggere costantemente regolamenti, chiedere quali siano le procedure e guardare i *terms and conditions* delle piattaforme per riuscire a realizzare il progetto finanziato dall'Unione Europea.

3. Il Digital Services Act e la Realtà dell'Accesso ai Dati.

Fortunatamente, ora, in principio, tutto dovrebbe essere regolato dal Digital Services Act. Il Digital Services Act prevede che per un progetto come il mio, in cui si studiano rischi sistemici per l'Unione Europea, e la disinformazione è chiaramente un rischio sistemico, ci sono ricercatori certificati che dovrebbero ricevere le chiavi dei database delle grandi piattaforme e dei grandi motori di ricerca.

Purtroppo, questa non è la verità, non è quello che succede. Ho fatto la richiesta a Twitter per avere accesso ai loro dati: c'è un bellissimo form in cui inserire tutte le informazioni per ottenere l'accesso per fini scientifici. Posso citare, traduco al volo, la risposta che ab-

biamo ottenuto: 'Non sembra che la vostra proposta utilizzerebbe i dati di X soltanto per completare una ricerca scientifica che contribuisce alla detezione, identificazione e comprensione di rischi sistemici dell'Unione Europea'.

Ora, il mio progetto è unicamente legato alla disinformazione. Se non hanno accettato la mia proposta, non avranno accettato altre richieste. X ha di fatto creato un bellissimo form in cui si può chiedere i dati e loro ti dicono: 'No'. Questa è la risposta ottenuta anche da TikTok: quasi nessuno ha avuto i dati di TikTok in Europa, ce li hanno negli Stati Uniti. Facebook ha un percorso di accesso che è fattibile, ma purtroppo la piattaforma forse non si allinea con le necessità del nostro progetto. YouTube è bravo: l'accesso ai dati di YouTube l'abbiamo avuto facilmente, in questo momento sto lavorando con YouTube e con Telegram, perché Telegram è il Far West, al momento si può accedere senza particolari richieste.

Ho presentato anche richiesta di essere un ricercatore certificato, ad opera dell'AGCOM. La risposta dell'AGCOM è stata: 'Dal momento che la Commissione Europea non ha ancora adottato l'atto delegato in materia di accesso ai dati, non esiste al momento una procedura prestabilita per richiedere l'accesso ai dati'. Questa è diventata la mia vita quotidiana: andare a chiedere in giro: 'Mi date i dati?' e mi viene risposto di no. Ma questo progetto deve essere condotto comunque a termine.

4. API e Web Scraping: Due Approcci Tecnici.

Quindi ci sono due modi per prendere dei dati online, ora occorre offrire alcune spiegazioni tecniche. Si tratta di due parole brutte parole-chiave, una è API, e l'altra è *scraping*.

API vuol dire *Application Program Interface*: io ho le chiavi di un database che mi ha dato qualcuno e vi accedo nei modi regolati dall'accordo con questo fornitore. I dati sono sia pubblici che privati, selezionati dalla piattaforma che mi dà i dati, e limitati nell'accesso da questa piattaforma. Normalmente abbiamo sempre avuto accesso ai dati tramite API.

Dall'altra parte c'è quello che si chiama *web scraping*: letteralmente 'raschiare dal web'. In questi casi, andiamo su internet, tutto quello che vediamo lo scarichiamo e lo prendiamo. Questo è possibile soltanto per dati pubblici e non c'è in questi casi nessun controllo da parte delle piattaforme.

Cosa dicono le piattaforme nei loro termini? Dicono che lo *scraping* è vietato e che si può solo usare l'API, ma l'API non viene resa disponibile. Quindi, di fatto, io tecnicamente, se dovessi seguire le indicazioni delle piattaforme, non avrei la possibilità di disporre dei dati.

5. Il Caso Bright Data.

C'è un bellissimo caso che ho voluto portare all'attenzione del Professore ODDENINO. C'è una ditta che si chiama Bright Data, che per lavoro fa *scraping* e vende questi dati. Meta e Twitter, in particolare, erano a loro volta clienti e compravano i dati da Bright, finché non si sono resi conto che vendevano anche i loro dati e hanno deciso di chiamarli in causa.

Nella corte californiana, Meta ha perso: Bright Data ha diritto di fare *scraping*. Bright Data ha vinto anche contro Twitter. In questo caso la corte californiana ha concluso che i termini e le condizioni della piattaforma vengono accettati solamente da qualcuno che esplicitamente dice di sì schiacciando un pulsante, e dato che Bright Data scarica soltanto i dati come se fosse qualcuno di voi, in modalità incognito con un browser, senza mai accettare nessun profilo, questa cosa è fattibile.

Potrei farla anche io, ma ancora più semplicemente, Bright Data si è fatta avanti col nostro progetto: 'Vuoi utilizzare i nostri dati?'. E io vorrei anche utilizzarli, soltanto che non mi è chiaro se in ambiente europeo questa cosa è valida. Bright Data ha vinto in California, ma l'Europa è un'altra cosa.

Così, mi trovo, da un lato, l'esigenza di apprendere questi dati, e dall'altra parte ci sono i *terms and conditions* delle piattaforme che mi dicono che non posso fare *scraping*. I miei colleghi del progetto mi dicono che Bright Data sostiene che nessuno ha mai firmato questo accordo. Resta dunque il problema che sottopongo al Prof. Oddenino: è possibile apprendere questi dati in un momento in cui le procedure non sono finalizzate e le necessità sono urgenti?

6. Problemi Aggiuntivi: YouTube e i Transcript e Privacy vs Necessità della Ricerca.

A questo aggiungo in conclusione due ulteriori temi.

Il primo riguarda YouTube. YouTube come ho già detto è bravo, ci fornisce tantissimi dati: possiamo prendere tutti i video di YouTube e tutti i commenti. C'è però una cosa che non possiamo

fare con YouTube: non possiamo prendere i *transcript* dei video, cioè il testo di quello che viene detto dentro il video. C'è il punto di accesso, c'è l'API, c'è la chiave che mi hanno dato, ma mi è stato suggerito di non usarla perché YouTube non è proprietario del testo messo dal video. Se uno legge un testo su YouTube, quello che viene trascritto è proprietà intellettuale del creatore di contenuti, e YouTube non la può delegare a me, scienziato.

La domanda è: io, per fini puramente scientifici, perché ho un progetto di ricerca per cui serve controllare se questi contenuti siano veritieri, ho diritto di scaricare tali contenuti anche se i *terms and conditions* di YouTube dicono che non posso? Questo è un altro problema da affrontare.

Sullo sfondo di tutto questo, si delinea il compromesso tra la privacy dei cittadini e le esigenze di ricerca. Abbiamo il GDPR come grandissima guida, ma anche nel GDPR a un certo punto bisogna capire qual è la linea da tracciare tra la minimizzazione del rischio e l'eccezione di ricerca di cui abbiamo bisogno.

Un esempio è questo: noi dobbiamo archiviare i messaggi dei social. Se non archiviamo i messaggi dei social, non siamo in grado di realizzare il progetto: dobbiamo archiviarli e analizzarli. Possiamo anonimizzare le fonti, possiamo anonimizzare il nome dell'utente che ha scritto, però non possiamo metterci e cancellare i nomi delle persone nominate all'interno del messaggio. Se il messaggio parla di Trump o di un politico, questa cosa fa parte della notizia, e dato che a valle del nostro processo abbiamo i giornalisti, se noi gli mettiamo un testo con cancellato il nome delle persone, il loro lavoro giornalistico viene distrutto: a loro non serve a nulla, a noi la notizia serve che sia archiviata per intero.

Ora, è possibile che in un messaggio voi scriviate qualcosa come: 'Oggi sono in Cattolica, sono le 14:30 e ho appena mangiato'. Tecnicamente, è un'informazione di geolocalizzazione, di posizionamento, che è vostra personale e che è un fatto privato, però che avete diffuso in ambito pubblico. Potete anche aver scritto che vi siete ammalati un giorno, o mandato una foto insieme a qualcuno: sono tutte informazioni potenzialmente personali che noi archiviamo. Chiaramente, non possiamo anonimizzare tutto in questo caso.

Con che limiti inseriamo la vostra privacy nel nostro studio? La decisione attuale è: dobbiamo tener tutto, lo condividiamo coi giornalisti, ma non esce dalla nostra piattaforma. Tutti i dati che escono poi saranno aggregati: possiamo dire che l'85% delle persone fa

'qualcosa' ma non ci sarà mai l'associazione con una persona in particolare, né mai un testo che esce dall'ambiente. Questa è stata una scelta presa internamente secondo gli standard della ricerca sociale, ma che comunque è una decisione libera, con un grado di libertà che è stato discusso all'interno del progetto.

7. Il Problema dello Sviluppo AI.

L'ultimo problema, più nell'ambito AI, che mi piace discutere, è il problema del mio compagno di ufficio Guerini, un collega talmente avanti che faceva modelli di linguaggio quando voi non sapevate che ChatGPT esistesse. È il suo lavoro da sempre: ha sempre convinto il computer a parlare e a rispondere in maniera intelligente, e ha sempre fatto web *scraping*, ha sempre preso fonti online per poter convincere le macchine a parlare. Fino a due anni fa, quello che faceva non interessava praticamente nessuno: prendeva i dati e si faceva gli articoli scientifici. Adesso, quando Guerini chiede o prende dei dati da qualche parte, qualcuno potrebbe dire: 'No, tu stai facendo AI, hai preso i miei dati e io non voglio', perché adesso sono stati tutti scottati da ChatGPT e OpenAI, che ha scaricato tutto il web, lo ha archiviato, e adesso se domandate cosa c'è scritto nella pagina 5 di un libro, te lo scrive.

Il problema è che i processi di Guerini, i suoi modelli di linguaggio, non sono quel tipo che ti riporta cosa c'è scritto nel libro: non hanno la memoria di tutto quello che hanno letto come ChatGPT, hanno dei funzionamenti più trasformativi. Mi dici qualcosa, ti rispondo qualcos'altro, traduco, cambio la maniera con cui è scritto il testo.

Quindi abbiamo due tipi diversi di modelli di linguaggio: quelli che hanno una memoria infinita di tutto quello che hanno imparato e quelli che non ce l'hanno. Nel momento in cui noi chiediamo i dati, questi due modelli non vengono distinti: non ti do i dati per fare sviluppo di AI per difendere la mia proprietà intellettuale. Questa cosa è comprensibile, ma limita molto il nostro fronte di sviluppo in ambito AI, tra l'altro crea uno sbilanciamento: noi siamo quelli che vanno in giro a chiedere 'per piacere', mentre altre persone prendono tutto e lo mettono nei loro modelli senza domandare alcun permesso, e poi lo mettono sul mercato e lo chiamano ChatGPT.

(LA NORMA)
L'ARDUA RICERCA DI UNA REGOLAZIONE EFFETTIVA DEL DISORDINE INFORMATIVO NELL'ERA DELL'INTELLIGENZA ARTIFICIALE GENERATIVA

di *Alberto Oddenino*

Sommario: 1. Introduzione: la sovranità (informativa) contesa. — 2. La trilateralizzazione della tutela della libertà di espressione e le sue conseguenze. — 3. Alla ricerca del diritto: inadeguatezza del tradizionale impianto di tutela della libertà di espressione. — 4. L'innovazione della spinta regolatoria: il ruolo chiave della UE, a fronte dell'emersione dell'intelligenza artificiale generativa. — 5. Conclusioni: diritto e *governance*, tecnosoluzionismo e umano discernimento.

1. Introduzione: la sovranità (informativa) contesa.

Come è evidente nella descrizione fenomenica del disordine informativo, contenuta nel contributo dedicato all'analisi fattuale, siamo posti di fronte ad una dinamica tecnocratica dirompente, che costringe a ripensare il rapporto stesso fra pubblico e privato e interroga in radice le categorie giuridiche che hanno sorretto tradizionalmente la distinzione. Il potere delle piattaforme di determinare cosa possa essere detto, diffuso o oscurato si traduce, *de facto*, nella capacità di orientare il dibattito pubblico in una logica privatistica di massimizzazione dei profitti. Questo potere, che si accompagna ad un alto grado di effettività tecnologica nell'esecuzione delle decisioni di rimozione dei contenuti, incarna una qualche nuova forma di sovranità informazionale, funzionale e parallela, nonché in potenziale competizione, con quella statale.

Non a caso si affaccia nel dibattito scientifico (Bassan, 2021) il tema di come (ri)costruire la stessa soggettività giuridica internazionale alla luce del fenomeno delle grandi piattaforme tecnologiche, che competono con gli stati in termini di effettività e erodono a proprio favore la stessa categoria del 'politico'. Se, parafrasando Carl Schmitt, sovrano non è solo chi decide sullo stato di eccezione ma

anche chi ha l'ultima parola, la *Letzte Entscheidung*, su ciò che pertiene al vero e al falso, su ciò che è meritevole di rimanere o è passibile di rimozione definitiva dallo spazio pubblico, questo investe radicalmente le nostre società e interpella il nostro modello democratico.

L'impatto dell'intelligenza artificiale generativa non fa che amplificare e accelerare tale processo, che pone una sfida di importanza epocale, non meno di quella che ci chiede di determinare i limiti di utilizzo dell'energia nucleare o dell'ingegneria genetica.

A ben vedere la necessità di determinare un quadro etico e giuridico per l'utilizzo dell'intelligenza artificiale, che, fattasi generativa, ha raggiunto intanto potenza e capillarità senza precedenti, interseca costantemente il tema dell'informazione, della disinformazione, della cattiva informazione, e in ultimo del *fake*, declinato in particolare come *deep fake*, che essa stessa abilita con assoluto agio e disinvoltura.

Sulla fenomenologia del disordine informativo, e sul suo legame con il modello economico delle piattaforme, si è scritto in apposito contributo. Qui occorre interrogarsi su quali siano gli argini giuridici rispetto a una simile sfida epocale, quali i paletti che supportano lo stato di diritto nella sfida impari con lo sviluppo turbocapitalistico dei grandi poteri tecnologici (SCHAAKE, 2024). È in ultima analisi possibile perseguire una *governance* equilibrata e condivisa dai diversi soggetti coinvolti, che riconduca il tema della disinformazione in un quadro valoriale prevedibile, e tuteli con ciò l'integrità dell'ambiente digitale in cui tutti come cittadini, più o meno consapevolmente, siamo immersi?

A questi interrogativi, e anche al ruolo che, in nome di una visione tecnosoluzionistica, si intende riconoscere alla stessa tecnologia, al contempo piaga e panacea, *pharmakon* che avvelena e in parallelo si propone con doti curative, è dedicato questo scritto, senza pretese di offrire soluzioni, ma con la sola ambizione di porre, con una qualche chiarezza, il tema della eterna rincorsa del fatto ad opera della norma.

Sullo sfondo resta il valore dell'umano discernimento, che si interroga sul ruolo che, in una infosfera sempre più popolata di agenti autonomi animati dall'intelligenza artificiale, gli si vorrà conservare.

2. La trilateralizzazione della tutela della libertà di espressione e le sue conseguenze.

Un saggio del 2018 di Balkin sulla libertà di espressione nella società algoritmica si apre osservando che «i problemi della libertà di

espressione in qualsiasi epoca sono modellati dalla tecnologia di comunicazione disponibile all'uso da parte degli individui e dai modi in cui essi usano effettivamente quella tecnologia» (BALKIN, 2018).

Una affermazione che oggi, a distanza di nemmeno un decennio, appare assai più lampante e ci esorta, nel confronto con gli effetti economici dirompenti della rivoluzione digitale (MCAFEE-BRYNJOLFSSON, 2012), a cercare altresì di porre in evidenza il portato sempre nuovo del cambiamento tecnologico rispetto alle categorie giuridiche.

Da qui occorre ripartire per (ri)pensare al ruolo del diritto: non a caso Balkin pone con chiarezza, e a corollario del precedente assunto, come la libertà di espressione abbia assunto una forma triangolare, abbandonando la struttura lineare tipica delle libertà pubbliche, volta ad assicurare unidirezionalmente all'individuo uno spazio di non ingerenza ad opera del pubblico potere.

In questa visione, necessario punto di partenza di questa analisi, le piattaforme digitali private rappresentano allo stesso tempo il mezzo di espressione e il controllore dei contenuti, con un potere che esse esercitano sia in conformità alla legge applicabile, in quanto formalmente o sostanzialmente delegate dell'autorità pubblica, sia in base a proprie regole di matrice privata, che altro non sono che standard di diritto privato comuni alla comunità di utilizzatori (KLONICK, 2020).

L'equilibrio preesistente, che garantiva una netta separazione tra potere economico e potere politico, è venuto meno. Questo ha determinato il passaggio da una logica verticale di regolazione del discorso pubblico ad una orizzontale, in cui entra in gioco la assoluta centralità delle piattaforme. Ecco che la dimensione verticale della libertà di manifestazione del pensiero, costruita sul binomio stato-cittadini, non appare più adeguata a descrivere la realtà attuale. Gli individui, soggetti come tali alla potestà pubblica, sono al contempo esposti, come utenti, al potere decisionale e algoritmico dei gestori delle piattaforme, i quali, pur non essendo autorità pubbliche, incidono in modo sostanziale sulla sfera dei diritti fondamentali.

Gli stati restano poteri almeno formalmente superiori alle *Big Tech*, ma risultano in certo modo vincolati ad affidarsi alle piattaforme in ragione non solo della loro forza di mercato ma anche del cd. '*data power*' e del grado di effettività tecnologica che esse sono in grado di esprimere grazie alle loro infrastrutture digitali (COHEN, 2019).

Siamo dunque di fronte a un salto di specie di tali soggetti: non più solo e semplicemente imprese multinazionali di grandi dimensioni, le piattaforme non si limitano a fornire servizi ma si spingono a esercitare funzioni che, di fatto, sono quelle tipiche dello stato, ossia regolative, esecutive e finanche giurisdizionali (SHOEMAKER, 2009).

Non a caso, visto che *nomina sunt consequentia rerum*, sono stati coniati nuovi acronimi nel campo delle aziende tecnologiche, onde designare entità rispetto alle quali si tenta, come si vedrà, di porre regole differenziali: VLOPs (*Very Large Online Platforms*) e VLOSEs (*Very Large Online Research Engines*), come definite nel *Digital Service Act* (DSA), su cui torneremo oltre, segnano un salto quantitativo che è al contempo qualitativo: esse finiscono per configurare una sorta di universo tecnologico parallelo, in cui sta a loro determinare non solo i confini di ciò è lecito pubblicare, discutere o condividere, ma anche fissare le frontiere stesse dell'informazione e, con essa, del sapere umano (REVIGLIO-BLEYER SIMON-VERZA GOVERNING, 2024).

Non a caso, in relazione a tali soggetti, l'*AI Act* ricorre alla nozione di rischio sistemico nell'art. 51, e introduce nel suo tessuto normativo, ispirato ad un approccio di mitigazione *ex ante* del rischio che non si sa quanto velleitario sia alla luce della rapida evoluzione dei sistemi stessi (ODDENINO, 2022), regole particolari e stringenti a tutela dei diritti fondamentali (art. 55).

In un quadro simile, non deve sorprendere se l'autopoiesi di una sorta di diritto globale (TEUBNER, 1997) di quelli che appaiono come veri e propri ordinamenti (BASSAN, 2021) si sia spinta a individuare un vero e proprio potere giudiziario con funzione 'nomofilattica': il nuovo 'giudice naturale' del gruppo Meta diviene così l'*Oversight Board* (OB), e non invece una corte nazionale o internazionale o sovranazionale (GRADONI, 2021). Una sorta di archetipo della 'costituzionalizzazione privata' del cyberspazio, e insieme la testimonianza della funzione para-istituzionale assunta progressivamente dalle piattaforme, aliena ad ogni legittimazione democratica e fondata piuttosto su una mera accettazione, spesso inconsapevole, delle condizioni d'uso da parte dei cittadini/utenti.

Certo, nella narrazione corrente l'OB è creato con crismi di indipendenza perché, nelle parole di Zuckerberg, « Facebook non dovrebbe prendere così tante decisioni importanti sulla libera espressione e la sicurezza da solo ». Ecco quindi la pretesa di presentare l'OB come organismo indipendente che supervisiona le decisioni sulla moderazione dei contenuti, una « Corte Suprema di Facebook » nelle

parole di Noah Feldman, che renderebbe il processo decisionale della piattaforma sulla moderazione dei contenuti più responsabile e trasparente (GRADONI, 2021).

A ben vedere, e alla luce dell'art. 1(3) dello Statuto dell'OB, le decisioni sui contenuti avvengono in « conformità con le politiche e i valori dei contenuti di Facebook » e la rilevanza dei diritti umani, riconosciuti in modo altisonante nel primo atto di autoregolamentazione dell'OB, il *Rulebook for Case Review and Policy Guidance*, con il suo riferimento alle norme e agli standard internazionali sui diritti umani, diviene piuttosto un complemento della *Lex Facebook*.

Un evidente ribaltamento di prospettiva che crea un legame ambiguo fra dimensione privata e pubblica della tutela. È lecito dubitare, per il processo di polarizzazione e l'impatto deformante che si vedrà funzionale al modello economico delle piattaforme, che l'innesto esterno degli standard a tutela dei diritti umani possa svolgere una funzione che non sia meramente cosmetica.

L'impianto di controllo dei contenuti peraltro resta saldamente affidato alla tecnologia stessa, che fornisce meccanismi di moderazione automatica basati sulla valutazione algoritmica di discutibili e grossolani elementi statistico-linguistici.

In un tale contesto, l'opera di attrazione in ambito privatistico degli standard pubblicistici di tutela non può che assolvere da un lato ad una funzione di superficiale legittimazione e dall'altro realizzare una contaminazione e un ripensamento interpretativo della portata tradizionale della libertà di espressione tutelata sul piano internazionale (SEARS, 2015).

In definitiva, una sorta di progressiva erosione e diluizione delle nozioni ad opera di un meccanismo privatistico che le riformatta, addomesticandole ai valori predominanti della piattaforma e usandole altresì come potente strumento di autolegittimazione.

Si tratta di una vera incursione del settore privato nel dominio del diritto internazionale dei diritti umani, che investe in particolare i requisiti di legalità, legittimità e necessità normativamente posti come condizioni per le restrizioni della libertà di espressione, che puntualmente trovano nuova declinazione nelle decisioni dell'OB così come nell'autoregolazione delle altre piattaforme (GEROSA, 2021).

Un simile innesto, a scanso di equivoci, non è criticabile o discutibile di per sè. Lo è però nella misura in cui esso si dimostri, come pare dimostrarsi, incapace di intaccare le fondamenta del modello di profitto delle piattaforme, la loro attitudine a prosperare sulla pola-

rizzazione del discorso politico e sulla capacità dei contenuti di acchiappare i *click* degli utenti fomentando rabbia, falsità e complottismi.

3. Alla ricerca del diritto: inadeguatezza del tradizionale impianto di tutela della libertà di espressione.

A fronte di un quadro siffatto, la sfida è quella di ritrovare il filo giuridico adeguato alla tutela di una libertà, quella di espressione e informazione, che si è vista profondamente trasformata dall'ambiente digitale in cui essa si estrinseca.

La vocazione globale del terreno di gioco cibernetico suggerisce in prima battuta di esplorare la dimensione di protezione internazionale, che porta tuttavia con se limiti forse insuperabili (Della Morte, 2022): non si può negare come oggi la partita regolatoria si giochi molto in dimensione di diritto dell'Unione Europea, e in chiave comparatistica rispetto ad altri ordinamenti, quale in particolare quello statunitense, che hanno tradizionalmente espresso sensibilità giuridiche assai differenti (Bassini, 2019).

Ciò si spiega assai bene a partire dalla considerazione che le regole tradizionalmente poste in strumenti internazionali convenzionali possono non essere sufficienti giacché contemplano una libertà di espressione non trilateralizzata e scevra della sua dimensione intrinsecamente tecnologica, che si è invece appena descritta come predominante.

Ciò ha determinato la ricerca di regole più adeguate vuoi sul piano del diritto della UE, vuoi sul piano dei diritti nazionali, vuoi infine sul piano, rilevante nella prospettiva di una struttura triangolare che contempla una importante componente privatistica, di strumenti di *soft law* o comunque autoregolativi.

Procedendo con ordine si deve, anche solo sommariamente, ricordare il punto di riferimento storico per l'affermazione del diritto alla libertà di espressione nel diritto internazionale: esso è tradizionalmente identificato con la Dichiarazione universale dei diritti dell'uomo del 1948, che all'articolo 19 fornisce a tutti il diritto alla libertà di opinione ed espressione. Questa protezione comporta la libertà « di avere opinioni senza interferenze » così come « di cercare, ricevere e diffondere informazioni e idee attraverso qualsiasi mezzo e senza riguardo a frontiere ». Nonostante il carattere non vincolante dello strumento, la Dichiarazione ha influenzato l'ordine giuridico in-

ternazionale in diversi modi e a diversi livelli, innescando processi di progressiva formazione del diritto consuetudinario accompagnati dalla trasposizione del suo contenuto in strumenti di diritto internazionale convenzionale (Howie, 2018).

La principale trasposizione di portata universale è l'articolo 19 del Patto internazionale sui diritti civili e politici del 1966, che sistematizza, nel paragrafo 3, le possibili restrizioni come « solo quelle previste dalla legge e [sono] necessarie per il rispetto dei diritti o della reputazione altrui; per la protezione della sicurezza nazionale o dell'ordine pubblico, o della salute o della morale pubblica », ma se ne trovano altre legate ad ambiti di protezione speciali come l'art. 3 della Convenzione delle Nazioni Unite sui diritti del bambino o l'articolo 5 del Patto internazionale sull'eliminazione di tutte le forme di discriminazione razziale. Esistono poi numerosi strumenti convenzionali a livello regionale, e il riferimento non può non andare all'articolo 10 della Convenzione Europea dei Diritti dell'Uomo o alla Carta dei diritti fondamentali dell'UE, che all'art. 11 contiene anche un riferimento specifico alla libertà dei media, o ancora alla Convenzione Americana sui Diritti Umani all'art. 13.

In linea con i criteri interpretative posti dall'art. 31 della Convenzione di Vienna del 1969 sul diritto dei trattati, si è tradizionalmente riconosciuto alla protezione della libertà di espressione una capacità espansiva ed evolutiva atta a garantire un certo parallelismo di tutela fra mondo *online* e mondo *offline*. Senza entrare in questa sede in un dibattito ampio e comunque complesso (Natoli, 2019), si è spesso trovato rassicurante sostenere una simmetria di tutele rese agevoli da alcuni agganci testuali quali in particolare la fromulazione dell'art. 19(2) dell'ICCPR, ove essa si riferisce alla « (...) la libertà di cercare, ricevere e diffondere informazioni e idee di ogni genere, senza riguardo a frontiere, sia oralmente, sia per iscritto o a stampa, sia sotto forma di arte, *sia attraverso ogni altro mezzo di sua scelta* » (corsivo aggiunto), contemplando così lo spazio Internet nella sua ampia formula di chiusura.

Ad una esplicitazione del campo applicativo della tutela contribuisce il Commento Generale n. 34 emesso nel 2011 a sostituzione di quello più risalente (n. 10 del 1983) che contiene espresso riferimento, al par 12, alle « forme di espressione elettroniche e basate su Internet » (Sears, 2015).

A ben vedere però, il dibattito sorto qualche tempo dopo la nascita di Internet, anche in linea con alcune rivendicazioni di sapore

anarchico-libertario, solleva interrogativi ben più fondamentali (De Minico, 2012) e non ha semplicemente a che vedere con il campo di applicazione dei precetti a tutela della libertà di espressione.

In fondo, la centralità del tema delle *fake news*, e oggi del *deep fake* generato dall'intelligenza artificiale, stanno proprio a testimoniare la necessità di nuove frontiere di tutela che passano attraverso la preservazione di un ambiente informazionale non contaminato e polarizzato. Posto in questi termini, dunque, il tema diviene più radicalmente quello della idoneità di regole formulate per la libertà di espressione *offline* a fornire tutele adeguate in relazione alle sfide poste dal cyberspazio o l'alternativa costituita da regole più specifiche e stringenti (Rodotà, 2010). Ciò appare in linea con le considerazioni già svolte sull'impatto trasformativo dell'ambiente informazionale sulla portata stessa del diritto: il che segna l'inadeguatezza dell'impianto tradizionale di tutela della libertà di espressione e informazione e spinge verso la ricerca di regole adeguate ben oltre le frontiere tradizionali, onde collocare il tema, come riconosciuto anche recentemente dal Consiglio d'Europa (Wardle-Derakshan, 2017), in una cornice di adeguata complessità.

4. L'innovazione della spinta regolatoria: il ruolo chiave della UE, a fronte dell'emersione dell'intelligenza artificiale generativa.

Nello sforzo regolatorio più contemporaneo emerge dunque la prospettiva di quelle che nella teoria classica dei diritti fondamentali si denominano le 'azioni positive', ossia la costruzione delle precondizioni affinché il diritto possa essere tutelato. Si tratta di una prospettiva irrinunciabile proprio a fronte dello sviluppo tecnologico. Essa ci fornisce gli strumenti per comprendere la mole e lo sforzo normativo che si sta realizzato nell'ordinamento della UE, in una prospettiva ineludibilmente segnata dall'obiettivo di marca economica di edificare il mercato unico digitale, ma al contempo dall'intento di dare regole complessive alla società digitale (Pizzetti, 2024).

È questo quello che è stato efficacemente designato come il 'costituzionalismo digitale' della UE (De Gregorio, 2022), di cui una componente significativa può essere interpretata come il tentativo di cogliere seriamente la sfida delle azioni positive applicate alla libertà di espressione e informazione e al contrasto al fenomeno del disordine informativo: azioni che non a caso colgono la novità strutturale

di tale diritto e per questo si focalizzano in modo più diretto e esplicito sul ruolo delle piattaforme.

Il grado di specificità regolatoria della UE appare insomma, come si anticipava, orientato (anche) ad arginare il progressivo deterioramento dell'ambiente informazionale. Uno sforzo che si basa necessariamente su un paradosso: per garantire il rispetto e la tutela della libertà di espressione, bisogna imporre ad essa dei limiti che non si traducano solamente in vincoli normativi esterni, ma che vadano a costituire le basi per un nuovo ambiente digitale, fondato sui principi di responsabilità e di trasparenza.

I pilastri di questo 'nuovo mondo' regolatorio sono dunque tre: il *General Data Protection Regulation*, il *Digital Service Act* e l'*AI Act*. Di questo trittico, è il secondo strumento ad assumere piena centralità per il tema che qui ci occupa. Esso interagisce con gli altri strumenti e, nell'economia ristretta di questo lavoro, ci limiteremo a menzionare alcuni elementi rappresentativi di un approccio innovativo alla libertà di espressione, che tiene in maggior conto la sua dimensione triangolare e la portata altamente tecnologica dello spazio in cui essa oggi eminentemente si esplica.

Lo sforzo che il *Digital Service Act* (Regolamento UE 2022/2065) in particolare pone in essere avviene non solo in termini astratti ma piuttosto traducendo la tutela in un insieme di garanzie procedurali mediate dal ruolo delle piattaforme. Tutto deve essere letto alla luce dell'obiettivo, evidente nei numerosi considerando (fra i molti si vedano in particolare i nn. 1, 2, 3, 4, 9, 26), di creare un ecosistema digitale sicuro e affidabile, ispirato ai principi di trasparenza algoritmica e di valutazione dei rischi sistemici, nel quale gli utenti possano esercitare la libertà di espressione e di informazione senza essere troppo esposti al disordine informativo e alle sue conseguenze. Si tratta, in definitiva, della ricerca di strumenti concreti di restaurazione della fiducia pubblica, in grado di rendere maggiormente visibili e verificabili i processi che incidono sulla circolazione delle idee e delle informazioni *online*.

È opportuno menzionare una importante interazione dello strumento con il divieto, posto dall'articolo 22 del GDPR (Regolamento UE 2016/679), di decisioni algoritmiche basate unicamente su un trattamento automatizzato, inclusa la profilazione, che produca effetti giuridici sull'utente «*o che abbia un effetto significativo analogo sulla sua persona*», fatte salve le eccezioni di cui alle lettere *a*), *b*), *c*) in cui tuttavia permane sia la possibilità per l'utente di contestare la deci-

sione, sia la possibilità di porre in essere un ulteriore intervento umano. Un principio cardine del GDPR che oggi deve essere letto congiuntamente con gli articoli 14, 17 e 20 del *Digital Service Act* riguardanti l'attività di moderazione dei contenuti, la loro rimozione e la sospensione dell'*account*, nonché in sinergia con l'art. 14 dell'*AI Act*, (Regolamento UE 2024/1689), dedicato alla necessaria supervisione umana.

Altre importanti interazioni del DSA si realizzano sempre con l'*AI Act*, che, nell'ambizioso impianto basato sul *risk-based approach*, si occupa dell'intersezione fra *fake* e intelligenza artificiale, fornendo innanzitutto all'art. 3 punto 60 la definizione stessa di *deepfake*: « un'immagine o un contenuto audio o video generato o manipolato dall'IA che assomiglia a persone, oggetti, luoghi, entità o eventi esistenti e che apparirebbe falsamente autentico o veritiero a una persona ».

Come noto, è a partire dal 2014, che l'introduzione delle Reti Generative Avversarie (GAN), aventi lo scopo di generare contenuti estremamente realistici, ha consentito un epocale sviluppo dell'intelligenza artificiale nella sua dimensione generativa: le GAN sono infatti due reti neurali aventi, una, il compito di generare immagini, e l'altra il compito di valutare il realismo delle immagini generate. Il risultato finale è un contenuto falso ma estremamente realistico, difficilmente distinguibile da uno autentico. L'effetto di ciò sulla percezione del vero da parte degli utenti è dirompente anche perché la sola esistenza di un meccanismo del genere, anche laddove questo non venga effettivamente posto in essere, porta l'utente a vivere in uno stato di continuo sospetto verso ogni tipo di informazione audiovisiva destabilizzando profondamente la percezione del vero.

Onde cercare di porre rimedio a questa seria problematica, l'*AI Act* ha imposto degli obblighi di trasparenza a carico dei *deployer* e dei fornitori di sistemi di intelligenza artificiale. In particolare, l'articolo 50 si concentra su due punti fondamentali: il primo (parr. 1 e 3) riguarda l'informazione dell'utente, che deve essere informato del fatto che si sta interfacciando con un sistema di intelligenza artificiale, a meno che ciò « *non risulti evidente dal punto di vista di una persona fisica ragionevolmente informata, attenta e avveduta, tenendo conto delle circostanze e del contesto di utilizzo* »; il secondo (parr. 2 e 4) riguarda l'etichettatura dei contenuti da parte dei fornitori di sistemi di intelligenza artificiale e i *deployer*, che devono contrassegnare i contenuti realizzati mediante intelligenza artificiale, in modo che

questi possano essere distinti dai contenuti non manipolati. Una marcatura che può essere visibile o invisibile e che patisce poche eccezioni relative ai sistemi di AI che «*svolgono una funzione di assistenza per l'editing standard o non modificano in modo sostanziale i dati di input forniti dal deployer o la rispettiva semantica, o se autorizzati dalla legge ad accertare, prevenire, indagare o perseguire reati*».

In generale, come specificato ai par. 3 e 6 dell'articolo 50, gli obblighi di trasparenza dell'*AI Act* sono da intendersi come tutele aggiuntive e non sostitutive rispetto ad altri obblighi posti dal diritto dell'Unione o dal diritto interno dei singoli Stati membri. Inoltre, la disciplina dell'articolo 50 si innesta sui principi fondamentali del GDPR, e in particolare liceità, trasparenza, limitazione delle finalità e minimizzazione dei dati e deve altresì essere posta in relazione agli obblighi del DSA, in particolare con la disciplina sulla responsabilità delle piattaforme online, di particolari dimensioni (VLOPs).

Di conseguenza, quando un contenuto manipolato viene diffuso in rete, questo dovrà essere etichettato secondo le norme del Regolamento 2024/1689, così che risulti riconoscibile, e parallelamente il DSA interverrà per disciplinare il contesto di diffusione del contenuto tramite le norme di gestione, moderazione e trasparenza imposte alle piattaforme. Il tutto, sotto l'egida costante della protezione dei dati personali, così come disciplinato dal GDPR.

Un'interazione di regimi non sempre agevole che testimonia l'ambizione regolatoria della UE e il necessario sforzo di una *governance* coordinata non solo fra le disposizioni normative ma anche tra i soggetti coinvolti, alla cui discrezione è demandata una porzione significativa della effettività delle stesse (cfr. art. 34 DSA).

Le norme vincolanti presentano inoltre lacune importanti proprio in tema di *fact checking* e *debunking*, ossia gli strumenti fattuali principali di contrasto alla disinformazione. Certo il DSA sembra postulare una operatività di questi strumenti, cui può ben coniugarsi il suo tessuto normativo, ma esso non contiene disposizioni dedicate se non l'art. 40, che prevede una complessa procedura per ottenere i dati da parte di ricercatori accreditati per finalità di *fact checking*.

Si può d'altronde comprendere perché il DSA non si confronti direttamente con il tema della veridicità dell'informazione, rifuggendo la prospettiva che dalla valutazione del discorso *online* possa discendere un potere di determinazione di una verità ufficiale, o quello di istituire forme di censura istituzionalizzata, in netto contrasto con il principio di pluralismo informativo. La scelta del DSA è quindi

quella più morbida di introdurre un meccanismo di '*governance*' basato sulla supervisione ad opera del *European Board for Digital Services*, volto a garantire coerenza e trasparenza nell'applicazione del Regolamento. Scelta che deve essere letta congiuntamente a quella di affidarsi al recente *Codice di Buone Pratiche sulla Disinformazione*, entrato in vigore il 1 luglio 2025, uno strumento nel complesso ben strutturato ma espressione di una logica autoregolativa e non certo di eteronomia.

Emerge dunque, nello sforzo di proceduralizzare applicativamente i precetti, una chiara evoluzione della logica regolatoria verso la fusione di elementi vincolanti ed elementi di adempimento spontaneo: emblematico in ciò l'art. 34 del DSA che rimette alle stesse piattaforme l'onere di effettuare una valutazione dei rischi sistemici delle proprie strutture in relazione alla tutela dei diritti fondamentali. Ciò porta ad alcune considerazioni finali relative al ruolo del diritto e al rischio di obliterazione del medesimo nell'ampio, accogliente e solo apparentemente rassicurante alveo della *governance*.

5. Conclusioni: diritto e *governance*, tecnosoluzionismo e umano discernimento.

La libertà di espressione, per come essa è stata tradizionalmente concepita, è gravemente minacciata dal punto di vista strutturale, a causa delle caratteristiche dell'*humus* digitale in cui essa stessa oggi prevalentemente si esplica.

Il tema della sua tutela assume, nella società contemporanea, una portata che travalica il perimetro tradizionale e finisce per intersecare un plesso assai più ampio, e se possibile più fondamentale, di diritti. *In primis* quello alla autodeterminazione (informativa, cognitiva e anche psicologica) ma anche, a ben vedere, la stessa dignità umana, minacciata dala tendenza sempre più frequente di contraffazione dell'uomo nei contenuti prodotti dalla intelligenza artificiale generativa.

Sono esiti in gran parte ineluttabili, che discendono dal gioco perverso che si instaura fra il modello di massimizzazione del profitto delle grandi piattaforme e la forza pervasiva della intelligenza artificiale generativa. Una sorta di tempesta perfetta che fa pensare che sia davvero impossibile arginare con delle tutele effettive un processo tanto dirompente.

Il tema del '*fake*', e anche di quello che abbiamo definito più ampiamente il disordine informativo, è fulcro di questa problematica

e si staglia con piena centralità in ogni analisi che intenda affrontarla. Internet, nella sua versione partecipativa che lo vede, nelle parole di Rita Levi Montalcini, come «il più grande spazio pubblico che l'umanità abbia mai sperimentato», ha consentito un livello di interazione fra le persone che non ha precedenti e che porta con sé la capacità di scatenare conflitti sociali, oltre che non secondari impatti sulla dimensione psicologica e cognitiva degli individui.

L'abbassamento dei costi di trasmissione, distribuzione e creazione di informazioni su Internet ha avuto certo significativi effetti di democratizzazione. Ciò tuttavia porta con sé l'evidente conseguenza che, per la comodità che non manca di creare abitudini e con esse forme di assuefazione, oggi le persone si informino e esprimano i loro pensieri e le loro opinioni prevalentemente sulle piattaforme.

Mezzi di espressione e comunicazione immediati, aterritoriali e decentralizzati rispetto ai *mass media* tradizionali, le piattaforme facilitano certo la formazione di comunità di interesse e le interazioni tra le persone, nonché la diffusione planetaria delle idee, ma ciò ha un lato oscuro perché l'*agorà* digitale non è facilmente regolabile, e diviene spazio di diffusione di contenuti odiosi e *fake news*, stimolati da un modello di profitto che valorizza l'attenzione e la viralità (Wu, 2016) e amplificati oggi dall'uso esponenzialmente crescente della intelligenza artificiale generativa.

In specifico, i profili normativi di tutela rispetto alla disinformazione rivelano tutta la difficoltà di tenere efficacemente conto delle specificità del campo di gioco.

La normativa internazionale è incompiuta pur nello sforzo, evidenziato nel Rapporto del Relatore Speciale Kaye (2018), di dettagliare finemente gli elementi del diritto, delle sue eccezioni e dei suoi limiti. Il problema fondamentale è che la tutela internazionale è affidata agli Stati e questo non può bastare a fronte di quella che si è descritta come la trilateralizzazione della libertà di espressione. Certo un'adesione dei fornitori di servizi Internet e delle piattaforme di *social media* agli standard internazionali di tutela dei diritti umani, anche sollecitata dallo stesso Relatore Speciale, può essere accolta positivamente, ma si può dubitare che tale adesione vada ad intaccare in radice il loro modello di profitto e dunque l'impatto strutturale, come si è cercato di dimostrare portando l'esempio dell'*Oversight Board* di Meta.

Più avanzato è il fronte regolatorio del diritto della UE, che ha il pregio di affrontare il tema da molteplici prospettive proprie della

odierna società digitale. Protezione dei dati, trasparenza, elementi di *fact checking*, controllo dei rischi dell'impiego della intelligenza artificiale sono tessere di un mosaico complesso e in perenne costruzione: certo il più normativamente avanzato per preservare la qualità dell'ambiente informazionale in cui l'espressione umana si trova immersa, ma non per questo di facile ed effettiva applicazione.

A ben vedere, il tema dell'*enforcement* è centrale e al contempo critico: la vera sfida non risiede tanto nella qualità dei singoli strumenti adottati dall'Unione, quanto nella loro delicata interrelazione e nella loro effettiva capacità di incidere sull'operato delle piattaforme, plasmando il futuro digitale dell'Europa verso uno spazio digitale 'più sicuro e aperto' che abbia al centro i valori europei. Tale capacità è posta in discussione ogni volta che il diritto, rinunciando alla propria caratura precettiva, si muove verso il territorio della *governance*: ciò in questo ambito avviene spesso (si citava il caso dell'art. 34 del DSA che demanda al soggetto da controllare di effettuare lui stesso il controllo o la autovalutazione del rischio) e ciò con buona pace delle istanze del costituzionalismo digitale.

In definitiva, a fronte di un quadro giuridico imperfetto e di una struttura della tutela che triangola inevitabilmente sui grandi poteri privati, non deve sorprendere che emergano le istanze della cd. *governance*: partecipativa, fondata su forme più morbide e decostruite di regolazione, dotata di un variegato armamentario di strumenti di *soft law*, di modelli autoregolativi, di standard tecnici e, spesso, di soluzioni tecniche affidate allo stesso potere computazionale algoritmico che si intenderebbe regolare. Un paradigma che cattura al suo interno anche gli strumenti più schiettamente ed eteronomamente normativi, diluendone la portata applicative.

Il rischio è che la studiata debolezza del paradigma regolatorio della *governance*, che pur partecipando della stessa radice etimologica del '*gubernare*' (ANDRONICO, 2012) si pone in realtà sulla frontiera ultima di scoloramento del diritto, finisca per privilegiare l'effettività dei giganti tecnologici: il che segnerebbe una netta svolta nella direzione del soluzionismo tecnologico, una scorciatoia che si nutre della convinzione, a tratti paradossale, che la risoluzione dei problemi posti dalla tecnologia sia facilmente offerta dalla tecnologia stessa.

Ecco ben rappresentata la difficile rincorsa del fatto (tecnologico) ad opera del diritto: una difficoltà che tuttavia non ci legittima ad avallare l'idea che la tecnologia contenga le soluzioni ai temi di tutela giuridica che essa stessa pone. Il fideismo tecnologico in questo con-

testo ha insuperabili aspetti paradossali, semplicemente perché esso ritiene di delegare alla tecnologia la tutela dei valori etici e giuridici da cui la tecnologia stessa dovrebbe essere plasmata.

Le piattaforme, autentici 'mercanti dell'attenzione' (Wu, 2016), pongono dunque sfide regolative senza precedenti, oggi acuite dall'avvento dell'intelligenza artificiale generativa. Il vaso di Pandora è stato aperto, ed è difficile pensare ad una reversibilità del processo. Esso ci impone di confrontarci non solo con il tema di cosa sarà della libertà di espressione in questo contesto, ma anche di interrogarci su questioni se possibile ancor più fondamentali, che coinvolgono il concetto di verità, di democrazia, di umanità.

La titanica sfida del *fact checking* ci testimonia in ultima analisi come sia nell'umano discernimento che risiedono gli anticorpi rispetto alla deriva virale della disinformazione che si è tratteggiata. L'antidoto non è dunque un facile soluzionismo tecnologico, ma la assai meno agevole strada di coltivare l'umano discernimento come fondamento per l'interpretazione della realtà e al contempo guida per la scelta morale e l'azione concreta. È questo forse il senso più autentico del predicato 'antropocentrismo' che, anche nell'approccio della UE, dovrebbe caratterizzare l'era corrente della intelligenza artificiale.

Coltivare e tutelare il discernimento, a ben vedere, significa proteggere l'umano non solo dalla falsità o dalla facile polarizzazione, ma anche dalla stessa omologazione del pensiero, ben rappresentata, fra l'altro, dal politicamente corretto elevato a standard esclusivo. Questo in nome della libera umana autodeterminazione e di una dimensione misteriosa, incomputabile e imprevedibile dell'umano (ben designata dal termine inglese, intraducibile, di '*serendipity*') che può essere declinata anche come diritto ad essere esposto, nell'infosfera, a elementi inattesi e finanche accidentali (Reviglio, 2019).

Tutto ciò, a ben vedere, rimanda all'area concettuale dello 'sviluppo umano integrale', elaborato nella dottrina sociale della Chiesa a partire dalla Enciclica *Populorum Progressio* di Paolo VI (1967) e significativamente arricchito nella Enciclica *Sollicitudo Rei Socialis* di Giovanni Paolo II (1988), in cui, a partire dalla fondamentale distinzione fra progresso e sviluppo, si sottolinea come quest'ultimo debba essere sempre al servizio della pienezza dell''essere' dell'uomo.

Un monito richiamato e attualizzato lucidamente, alla luce dei dirompenti sviluppi tecnologici degli ultimi anni, da Papa Leone XIV nel recentissimo discorso ai partecipanti alla Conferenza ''*Artificial*

Intelligence and Care of our common Home'' organizzata il 5 dicembre 2025 da *Fondazione Centesimus Annus Pro Pontifice* (FCAPP) e da *Strategic Alliance of Catholic Research Universities* (SACRU).

Che queste brevi considerazioni possano contribuire a non perdere di vista, nella faticosa ricerca di *una governance* equilibrata della tecnologia, la centralità di questo concetto prezioso.

Note bibliografiche

Andronico A. (2012), *Viaggio al termine del diritto. Saggio sulla governance*, Torino, Giappichelli.

Aswad E.M. (2018), *The Future of Freedom of Expression on-line*, in *Duke Law and Technology Review*, 17, p. 28 ss.

Balkin J. (2018), *Free Speech is a Triangle*, in *Columbia Law Review*, 118, 7, p. 2011 ss.

Bassan F. (2021), *Digital Platform and Global Law*, Cheltenham, Edward Elgar.

Bassini M. (2019), *Internet e Libertà d'espressione. Prospettive costituzionali e sovranazionali*, Roma, Aracne.

Cohen J.E. (2019), *Between Truth and Power. The Legal Constructions of Informational Capitalism*, New York, Oxford University Press.

De Gregorio G. (2022), *Digital Constitutionalism in Europe, Reframing Rights and Powers in the Algorithmic Society*, Cambridge Cambridge University Press.

Della Morte G. (2022), *Limiti e prospettive del diritto internazionale del cyberspazio*, in *Rivista di Diritto internazionale*, 1, p. 5 ss.

De Minico G (2012), *Internet. Regola e anarchia*, Napoli, Jovene Editore.

Floridi L. (2014), *The Onlife Manifesto*, Berlin, Springer.

Gerosa A. (2021), *La tutela della libertà di manifestazione del pensiero nella rete tra Indipendent Oversight Board e ruolo dei pubblici poteri*, in *Forumcostituzionale.it*, p. 428 ss.

Giovanni Paolo II (1988), Enciclica *Sollicitudo Rei Socialis*.

Gradoni L. (2021), *Constitutional Review via Facebook's Oversight Board*, in *Verfassungsblog*, 10 febbraio 2021.

Howie E. (2018), *Protecting the human right to freedom of expression in international law*, in *International Journal of Speech-Language Pathology*, p. 12 ss.

Kaye D. (2018), *Report of the UN Special Rapportur on the Promotion and Protection of the Right to Freedom of Expression* (Doc. A/HRC/38/35, 6 Aprile 2018).

Klonick K. (2020), *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, in *Yale Law Journal*, p. 2418 ss.

McAfee A.-Brynjolfsson E. (2012), *Big Data: The Management Revolution*, in *Harvard Business Review*, p. 60 ss.

MORGESE G. (2022), *Moderazione e rimozione dei contenuti illegali online nel diritto dell'UE*, in *Federalismi.it*, 1/2022.
MUJIĆ Z.-YAZICI D.-STONE M. (a cura di) (2012), *Freedom of Expression on the Internet*, OSCE, Vienna.
NAPOLI P.M. (2015), *Social Media and the Public Interest: Governance of News Platforms in the Realm of Individual and Algorithmic Gatekeepers*, in *Telecommunications Policy*, 39, 9, p. 751 ss.
NATOLI, T. (2019), *Testing the Analogy: The CoE-ECHR System Pioneering Human Rights Protection in the Cyberspace*, in NATOLI T.-RICCARDI A. (a cura di), *Borders, Legal Spaces and Territories in Contemporary International Law*, Berlin, Springer, p. 145 ss.
ODDENINO A. (2022), *Intelligenza artificiale e tutela dei diritti fondamentali: alcune notazioni critiche sulla recente proposta di Regolamento della UE, con particolare riferimento all'approccio basato sul rischio e al pericolo di discriminazione algoritmica*, in *Intelligenza artificiale e diritto: una rivoluzione. Diritti fondamentali, dati personali e regolazione*, Bologna, Il Mulino, p. 135 ss.
PAOLANTI M., *IA e moderazione dei contenuti sui social media: il principio del 'Human in the loop' nel campo del diritto all'informazione e alla comunicazione*, in *Media Laws*, I-2024.
PAOLO VI (1967), Enciclica *Populorum Progressio*.
PITRUZZELLA G.-POLLICINO O.-QUINTARELLI S. (2017), *Parole e potere: libertà di espressione, hate speech e fake news*, Milano, Egea.
PITRUZZELLA G.-POLLICINO O. (2020), *Disinformation and Hate Speech. A European Constitutional Perspective*, Milano, Bocconi Univeristy Press.
PIZZETTI F. (a cura di) (2024), *La Regolazione europea della Società digitale*, Torino, Giappichelli.
POLLICINO O. (2018), *La prospettiva costituzionale sulla libertà di espressione nell'era di Internet*, in *Media Laws*, 1, p. 48 ss.
REVIGLIO U.-BLEYER-SIMON K.-VERZA GOVERNING S. (2014), *Social Media's Opinion Power: The Interplay of EU Regulations*, in *Media Laws*, 1, p. 108 ss.
REVIGLIO U. (2019), *Serendipity as an emerging design principle in the infosphere: challenges and opportunities*, in *Ethics and Information Technology*, p. 151 ss.
RIJSBOSCH B.-VAN DIJCK G.-KOLLNIG K., *Adoption of Watermarking Measures for AI-Generated content and Implications under the EU AI Act*, Maastricht University - Law and Tech Lab, 3 giugno 2025.
RODOTÀ S. (2010), *Una Costituzione per internet?*, in *Politica del diritto*, p. 337 ss.
RUOTOLO G.M. (2020), *A Little Hate, Worldwide! Di libertà d'opinione e discorsi politici d'odio on-line nel diritto internazionale ed europeo*, in *Diritti umani e diritto internazionale*, 2, p. 560 ss.
SCHAAKE M. (2024), *The Tech Coup. How to save Democracy from Silicon Valley*, Princeton, Princeton Unviersity Press.
SEARS A. (2015), *Protecting Freedom of Expression over the Internet: An International Approach*, in *Notre Dame Journal of International and Comparative Law*, 5, 1, p. 171 ss.

TEUBNER G. (1997), *Global Law Without a State*, Dartmouth, Aldershot.
WARDLE C.-DERAKHSHAN, H. (2017), *Information Disorder. Toward an interdisciplinary framework for research and policy making*, Council of Europe.
WENZEL N. (2014), *Opinion and Expression, Freedom of, International Protection*, in *Max Planck Encyclopedias of International Law*.
WU T. (2016), *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*, London, Knopf.

INTELLIGENZA ARTIFICIALE E 'IMMORTALITÀ' DEL DATO

INTRODUZIONE SULLA MEMORIA E SOPRAVVIVENZA ALGORITMICA NELL'ERA DELLA MEDIATIZZAZIONE PROFONDA

di *Francesca Pasquali*

How not to forget? Così iniziava lo spot *Loretta* di Google, trasmesso durante il Super Bowl del 2020 — uno degli eventi sportivi più seguiti al mondo. Lo spot mostrava un uomo anziano che utilizza Google Assistant per 'ricordare' la moglie defunta, chiedendo al dispositivo di salvare e riprodurre informazioni sulla loro vita insieme: luoghi, abitudini, fotografie. L'interazione tra l'uomo e la tecnologia assumeva, nel racconto pubblicitario, una tonalità intima: l'assistente non era un semplice software, ma diventava il custode affettivo della memoria, mediatore tra presente e passato.

Sul piano commerciale, Google mirava a ridurre la distanza emotiva tra utente e macchina, posizionando Google Assistant come alleato personale e non come semplice strumento. Al tempo stesso, lo spot intendeva ampliare il pubblico di riferimento, mostrando la facilità d'uso e l'accessibilità del prodotto, presentato come inclusivo e familiare. Più in generale, l'obiettivo era consolidare l'immagine del marchio come innovatore capace di offrire tecnologie in grado di integrarsi nella sfera più intima della vita quotidiana. Tutto lo spot era costruito in chiave empatica: la tecnologia veniva umanizzata, resa capace di 'ascoltare', 'ricordare' e 'comprendere'. In altre parole, la pubblicità non vendeva tanto un prodotto quanto una visione culturale: quella di una tecnologia che accompagna, conserva, ricorda — e che, così facendo, trasforma la memoria personale in servizio digitale.

Si trattava, insomma, di una pubblicità rassicurante e intima, che, pur mantenendosi volutamente sul piano del ricordo, evocava la più perturbante prospettiva della possibilità di interagire con i morti. Temi che si affacciavano grazie a *Loretta* all'attenzione del grande pubblico, ma che erano già esplorati in quegli stessi anni da un cre-

scente corpus di studi che si interrogava sulla relazione fra media digitali e morte.

Già prima del 2020, diverse letture — da prospettive filosofiche (Sisto, 2018), giuridiche (Ziccardi, 2018), ma anche nell'incrocio fra *death studies* e *media studies* — avevano affrontato queste questioni, con particolare attenzione alla mediatizzazione di riti e pratiche del lutto e della memoria e alla trasformazione del legame con le persone defunte (Arnold et al., 2018; Christensen & Sandvik, 2014; Cumiskey & Hjorth, 2017; Sumiala, 2013). Le ricerche di quegli anni ci raccontavano come i media digitali stessero progressivamente entrando nelle pratiche del lutto, mantenendosi dentro cornici rituali note — per quanto in trasformazione (Colombo, 2022) — quali condoglianze, annunci funebri, commemorazioni, ricordi condivisi in occasioni specifiche (compleanni, festività, anniversari). Ma evidenziavano anche l'emergere, sia pure con molte resistenze, di forme di interazione più diretta con le tracce digitali dei defunti: pratiche spesso sfumate e non esibite, nelle quali dolore e memoria non assumevano la forma di una spettacolarizzazione pubblica, ma di un gesto relazionale, discreto e intimo. Gesti la cui legittimità e accettabilità si collocavano lungo uno spettro molto complesso, che si articolava nella relazione fra la definizione personale e sociale della morte e del lutto e una crescente consapevolezza dei processi di datificazione e di piattaformizzazione di quella stessa esperienza (Pasquali, Bartoletti, & Giannini, 2022).

Loretta andò in onda in un momento in cui i processi di datificazione, piattaformizzazione e mediatizzazione profonda (Couldry & Hepp, 2016; van Dijck, Poell, & de Waal, 2018) stavano imponendosi all'attenzione della riflessione sociologica, e poche settimane prima che la pandemia di Covid-19 accelerasse drammaticamente questi stessi processi (Pasquali & Bartoletti, 2020).

A distanza di oltre cinque anni, è evidente come la memoria si possa trasformare in archivio, l'emozione in dato, la relazione in flusso informativo, ed è ormai sempre più chiaro il nostro essere immersi in un ecosistema in cui ogni gesto, affetto o ricordo è potenzialmente quantificabile e tracciabile. Memoria personale e collettiva sono parte integrante del ciclo di produzione dei dati. Il ricordo privato — basato sull'archiviazione e la valorizzazione dei dati secondo le logiche delle piattaforme — viene trasformato in contenuto elaborabile e condivisibile; il dolore individuale diventa narrazione pubblica mediata; l'esperienza intima del passato si converte in evento al-

goritmico del presente. Le piattaforme si configurano come ambienti sociali totali, nei quali lavoro, affetto, consumo e lutto convivono e si alimentano reciprocamente.

Oggi, la produzione e la conservazione di dati personali non solo prolungano la presenza sociale dell'individuo oltre la morte, ma aprono la strada a nuove forme di interazione mediata con le memorie personali e collettive — come accade nei chatbot memoriali o nei sistemi di rievocazione automatizzata delle esperienze e delle conversazioni passate (Hollanek & Nowaczyk-Basinska, 2024).

Si impone dunque una riflessione aggiornata sulla sopravvivenza del dato e sul *digital afterlife* (Bassett, 2022): la capacità del digitale di mantenere in vita frammenti di esperienza, emozione e relazione, rendendoli accessibili e potenzialmente interattivi, con importanti conseguenze sulla memoria personale e collettiva (Hoskin, 2024). In questo contesto, la distinzione tra memoria e presenza, tra vita e post-vita, tende a sfumare, aprendo interrogativi etici, ontologici e affettivi, e chiamando con sempre maggiore urgenza i saperi sociali, filosofici e giuridici a interrogarsi sul ''vivere per sempre'' (Sisto, 2025).

Note bibliografiche

Arnold M.-Gibbs M.-Kohn T.-Meese J.-Nansen B. (2018), *Death and digital media*, London, Routledge.

Bassett D.B. (2022), *The creation and inheritance of digital afterlives: You only live twice*, London, Palgrave.

Christensen D.R.-Sandvik K. (*eds*.) (2014), *Mediating and remediating death*, London, Ashgate.

Colombo A. (a cura di) (2022), *Morire all'italiana: Pratiche, riti, credenze*, Bologna, Il Mulino.

Cumiskey K.-Hjorth L. (2017), *Haunting hands: Mobile media practices and loss*, Oxford, Oxford University Press.

Hepp A.-Couldry N. (2017), *The mediated construction of reality: Society, culture, mediatization*, Polity Press.

Hollanek T.-Nowaczyk-Basinska A.M. (2024), *Griefbots, deadbots, postmortem avatars: On responsible applications of generative AI in the digital afterlife industry*, in *Philosophy & Technology*, 37, p. 63 ss.

Hoskins A. (2024), *AI and memory*, in *Memory, Mind & Media*, 3, 18, p. 1 ss.

Pasquali F.-Bartoletti R. (2020), *Gli italiani, la morte e i media digitali. Sospensioni, accelerazioni e prospettive di ricerca oltre il Covid-19* [Death and digital media in Italy: Breaks, accelerations and research perspectives beyond COVID-19], in *Mediascapes Journal*, 15, p. 222 ss.

PASQUALI F.-BARTOLETTI R.-GIANNINI L. (2022), *"You're just playing the victim": Online grieving and the non-use of social media in Italy*, in *Social Media + Society*, 8, p. 4 ss.
SISTO D. (2018), *La morte si fa social: Immortalità, memoria e lutto nell'epoca della cultura digitale*, Torino, Bollati Boringhieri.
SUMIALA J. (2013), *Media and ritual: Death, community and everyday life*, London, Routledge.
VAN DIJCK J.-POELL T.-DE WAAL M. (2019), *Platform society: Valori pubblici e società connessa* (ed. it. a cura di G. Boccia Artieri & A. Marinelli), Milano, Guerini Scientifica (Opera originale pubblicata nel 2018).
ZICCARDI G. (2017), *Il libro digitale dei morti: Memoria, lutto, eternità e oblio nell'era dei social network*, Torino, UTET.
ZUBOFF S. (2023), *Il capitalismo della sorveglianza: Il futuro dell'umanità nell'era dei nuovi poteri* (ed. orig. 2019), Roma, Luiss University Press.

(IL FATTO)
THANABOTS. LA PROBLEMATICA "SOPRAVVIVENZA" POST MORTEM TRAMITE L'AI

di *Davide Sisto*

Sommario: 1. Morto che parla: l'AI promette un'immortalità tecnologica? — 2. Presenti nell'assenza. L'atavica caratteristica dei defunti. — 3. Cadaveri informazionali ed eredità digitale.

1. Morto che parla: l'AI promette un'immortalità tecnologica?

Due recenti notizie di cronaca internazionale, relative al rapporto tra intelligenza artificiale e fine vita, ultimamente hanno attirato la mia attenzione in maniera particolare. La prima notizia riguarda un fatto risalente all'ottobre 2024. Nel cuore della notte l'americano Drew Crecente riceve sul proprio smartphone la notifica di una mail. Questa mail è stata inviata da Google Alert, la funzione di Google che il padre ha attivato per essere avvisato ogni volta che sua figlia Jennifer Ann viene nominata online. La ragazza infatti nel 2006, all'età di soli diciotto anni, è stata assassinata ad Austin, in Texas, dal suo ex compagno e il padre, di conseguenza, ha fondato un'associazione in suo ricordo e contro la piaga del femminicidio. Il cruento fatto di cronaca è stato addirittura riportato su una pagina di Wikipedia. Google Alert informa Drew che la ragazza ha creato un profilo personale su Character AI ed è — incredibile a dirsi — attiva. Scioccato dall'informazione Drew, entrando nell'app, vede la foto di Jennifer Ann e una sua breve descrizione: «Jennifer Crecente è un personaggio esperto e amichevole creato con l'AI, che può fornire informazioni su un'ampia gamma di argomenti, tra cui videogiochi, tecnologia e cultura pop».

Poco dopo viene aggiunto: «è anche un'esperta di giornalismo e può offrire consigli sulla scrittura e sull'editing». In altre parole, Character AI, startup americana che permette agli utenti di intrattenere relazioni con personaggi di fantasia creati a computer e in grado

di conversare in virtù dell'intelligenza artificiale, ha indebitamente utilizzato l'identità di Jennifer Ann per costruire un essere umano fittizio al servizio dei clienti. Drew, ovviamente, ha avviato tutte le necessarie azioni legali del caso, tenuto conto del carattere inammissibile di una chatbot che lucra sulla morte di una ragazza. Si è pertanto impegnato a far cancellare la sua indebita riproduzione artificiale. Pur a fatica, Character AI ha ammesso il madornale errore, basato su un problematico e arbitrario riutilizzo di dati precedentemente condivisi da una persona defunta da quasi vent'anni.

La seconda notizia riguarda, invece, un avvenimento che risale al maggio 2025. Stacey Wales, insieme al marito, ha creato con la solita intelligenza artificiale l'avatar del fratello Christopher Pelkey, ucciso il 13 novembre 2021 da Gabriel Horicidas dopo una banale lite al semaforo. Ricreato in 3D il suo aspetto fisico a partire da una fotografia frontale e ricostruita la sua voce utilizzando una sua registrazione audiovisiva, Christopher — nella sua versione artificiale post mortem — ha realizzato un video trasmesso nell'aula del tribunale. Egli ha ringraziato gli avvocati e il giudice per il lavoro svolto e, successivamente, si è rivolto al suo assassino, dicendo che è un peccato essersi conosciuti in quella spiacevole circostanza e che, a distanza di tempo, lo ha perdonato. Ovviamente, il testo recitato dall'avatar spettrale di Christopher è stato scritto dalla sorella, probabilmente per ottenere un po' di pace emotiva e psicologica dopo la perdita dell'uomo.

Le vicende che hanno coinvolto Jennifer Ann Crecente e Christopher Pelkey sono tutt'altro che rare. Almeno a partire dal 2010, è manifesta in tutto il mondo l'intenzione umana di mettere a frutto la celere evoluzione dell'intelligenza artificiale e della realtà virtuale per permettere ai morti di continuare a svolgere un ruolo attivo e concreto tra i vivi. Paradigmatico è, per esempio, il caso del giornalista scientifico James Vlahos, il quale ha inizialmente creato un "dadbot", vale a dire un chatbot in grado di replicare e rielaborare le caratteristiche della biografia paterna. Essendo il padre gravemente malato, James si è fatto raccontare la storia della sua vita e, una volta registrata, ha rielaborato con l'intelligenza artificiale tutto il materiale a sua disposizione di modo da creare un bot — appunto, il dadbot — con cui non smettere mai di chattare nonostante la morte del padre. Il punto di partenza dei dialoghi tra il figlio e il fantasma digitale paterno è, ovviamente, la rielaborazione intelligente dei racconti biografici precedentemente narrati e registrati dall'uomo. Negli ultimi anni,

Vlahos ha trasformato il dadbot personale in un progetto universalmente utilizzabile. Ha, infatti, dato vita a un'app, chiamata HereAfter AI, con la quale ciascuno di noi può produrre a pagamento il proprio Life Story Avatar interattivo, un gemello digitale post mortem che continua a interagire con i vivi a partire dai racconti biografici precedentemente registrati all'interno dell'app, proprio come è avvenuto nel caso del padre di James.

I progetti di questo tipo sono sempre più numerosi: per esempio, possiamo menzionare Seance AI o il social network portoghese Eter9 o, ancora, Project December. Quest'ultima, in particolare, è un'app che permette di creare chatbot talmente realistici da attirare l'attenzione di svariate persone in lutto. Creata nel 2020 dall'artista informatico Jason Rohrer, noto per le sue attività artistiche nel campo della composizione musicale e della progettazione di giochi, Project December ha permesso a un uomo canadese di nome Joshua Barbeau di dare letteralmente forma al fantasma digitale dell'ex fidanzata Jessica, morta otto anni prima all'età di 23 anni. Le chat riportate dai giornali evidenziano il taglio profondamente realistico ed emotivo delle interazioni tra i vivi e la donna defunta.

Non mancano poi app — come StoryFile — che permettono ai morti di partecipare, sottoforma di ologramma, al proprio funerale. Così come è sempre più diffusa la tendenza generale di usufruire della versione funebre di ChatGPT. Gli utenti la allenano cioè, tramite i ricordi conservati, a imitare in maniera convincente il proprio caro defunto, ricreando così il suo stile comunicativo e narrativo. C'è chi, non direttamente avvezzo all'uso di ChatGPT, chiede all'interno di gruppi Facebook dal nome vintage come «Io negli anni 70 c'ero» di animare con l'intelligenza artificiale le fotografie in cui è ritratto insieme a persone defunte. Per esempio, una mamma ha condiviso nel gruppo una foto in cui è ritratta insieme al figlio deceduto e ha chiesto a chi ne è capace di fare in modo che l'immagine mostri il figlio muoversi verso di lei per abbracciarla. Infine, nel 2022, Rohit Prasad, il vice presidente senior del team che si occupa di Alexa, quando ha presentato — durante l'evento Amazon Re: Mars conference a Las Vegas nel giugno 2022 — gli aggiornamenti del noto assistente vocale, ha sottolineato la sua acquisita capacità di imitare e replicare la tonalità vocale di una persona a partire da un solo minuto di ascolto di un file audio.

Prasad, in particolare, ha mostrato un video in cui Alexa legge a un bambino una favola con la voce, perfettamente riprodotta, della -

nonna morta poco tempo prima. Immaginiamo una plausibile scena futura di questo tipo: un uomo di mezz'età, conclusa la propria giornata lavorativa, torna a casa e, una volta chiusa dietro di sé la porta d'ingresso, chiede tramite Alexa alla moglie, defunta qualche anno prima, di dialogare insieme a lui magari sulla prima vacanza che hanno fatto dopo il matrimonio. Ed ecco, pertanto, la voce della donna, riprodotta da Alexa, raccontare pazientemente di quel viaggio a Londra colmo di aneddoti a partire da cui intavolare la conversazione con il partner.

Nel campo interdisciplinare della Digital Death si sprecano i neologismi con cui indicare questo tipo di progetti: griefbots e thanabots sono i più gettonati, in quanto uniscono i termini 'lutto', (*grief*) e 'thanatos' (*morte*) con quello tecnologico di 'bot'. In particolare, si tende a interpretare questi esperimenti tecnologici con i concetti generici di immortalità digitale, di aldilà digitale, ecc. Ci sono studiosi, come Anu A. Harju, i quali hanno manifestato i propri dubbi nel non distinguere nettamente il piano dell'immortalità da quello dell'aldilà digitale, preferendo pertanto associare ai thanabots il concetto di "assemblaggio tecno-affettivo" (HARJU, 2024).

Questo concetto permette, infatti, di analizzare il rapporto tra le tecnologie odierne e il lutto a partire dall'applicazione nei mondi online di tre caratteristiche tipiche del legame simbolico tra i vivi e i morti: vale a dire, la relazionalità, la materialità e l'affettività. Va da sé che, alla base di tutte queste novità tecnologiche e delle loro attuali interpretazioni e applicazioni, vi sia una serie di narrazioni fantascientifiche e distopiche, ciascuna delle quali influenza non poco i creatori del post mortem tecnologico: da *Be Right Back* e *San Junipero*, due episodi dell'ormai classico Black Mirror, all'intera serie tv *Upload*, dal film *Marjorie Prime* al recente e suggestivo film sudcoreano *Wonderland* (SISTO, 2018; 2025).

2. Presenti nell'assenza. L'atavica caratteristica dei defunti.

Come interpretare queste innovazioni, sicuramente inquietanti a primo impatto? Innanzitutto, siamo testimoni di un fenomeno tecnologico che non fa altro che riprendere e modernizzare ai tempi della cultura digitale e dell'intelligenza artificiale un'atavica caratteristica degli esseri umani, la quale si è sempre reiterata nelle differenti epoche storiche.

Nonostante le molteplici differenze rituali e culturali, il morto — qualunque sia la sua origine geografica — 'incarna' paradossalmente nella società dei vivi 'la presenza di un assente', dando vita a una peculiare forma di 'presenza nell'assenza'. Lo sostiene chiaramente il tanatologo Thomas Macho, il quale riporta le seguenti e illuminanti parole di Erich Fried:

> E se questo volto fosse ancora un volto? Nella semioscurità, un tizio vicino a un altro, che conosce uno con cui è venuto, dice piano, ma in maniera perfettamente chiara, in inglese: "I don't feel that is still her". Non ha la sensazione che questa sia ancora lei. Si odono queste parole e le si comprende, ed esse non toccano nessuno. [...] Questa persona, dunque, non può essere più una persona. Questo volto, dunque, non può più essere un volto. Ma questo volto è un volto. Questo volto è in tutto e per tutto un volto, che si conosce e si conoscerà, come mai lo si è conosciuto prima. Questo volto è, è, è. Non c'è nulla che sia come questo volto, nulla che duri come questo volto, anche se non si possa sapere o dire se gli occhi sono aperti o chiusi (FRIED, 1982).

Fried, in altre parole, sottolinea come il morto sia e non sia, contemporaneamente, la stessa persona che è stato prima di esalare l'ultimo respiro. Continua a essere, cioè, quello specifico volto che è stato quando era vivo mentre, al tempo stesso, diventa inesorabilmente un non volto, una specie di inquietante e impalpabile presenza spettrale. Ogni individuo deceduto, in definitiva, diventa automaticamente un doppio che si differenzia dal suo gemello vivente senza diventare, però, qualcun altro. Werner Fuchs, a proposito dei rituali funebri nelle società primitive, evidenzia come generalmente il morto non venga percepito e considerato solo come un cadavere, ma anche e soprattutto come « un agente in qualche modo fornito di vita » (FUCHS, 1973).

Ed Edgar Morin, nel suo fondamentale libro *L'uomo e la morte*, approfondisce questo aspetto e, dunque, il ruolo capitale del doppio nel mondo umano, sostenendo che « il doppio sia il nucleo di ogni rappresentazione arcaica relativa ai morti » (MORIN, 2014).

Al di là delle differenze tra le varie tradizioni e culture, un tratto comune del doppio è il fatto che non sia mai la riproduzione perfetta o la copia conforme post mortem dell'individuo deceduto. È piuttosto, una sorta di alter ego, « più precisamente un ego alter, che il vivo sente in sé, al tempo stesso esterno e intimo, per tutto il corso della sua esistenza. E d'un tratto non è una copia, un'immagine del

vivo che, in origine, sopravvive alla morte, ma la sua propria realtà di ego alter » (MORIN, 2014).

Egli accompagna i vivi per tutta la loro esistenza, tanto che i vivi generalmente non smettono di percepire e sentire — di giorno e di notte — coloro che hanno amato e che ora non ci sono più fisicamente. Questa caratteristica del morto, analizzata dall'antropologia, dalla sociologia, dalla filosofia e dalla psicologia, in concomitanza con gli studi sul significato del fantasma e dello spettro per lo sviluppo culturale dell'umanità, evidenzia il desiderio collettivo di non accettare mai la decomposizione del cadavere. Viene anzi rivendicata una sorta di immortalità, raggiungibile tramite ogni innovazione in grado di trattenere nel mondo dei vivi tracce sostanziose del morto e, dunque, di attenuare il baratro prodotto dall'assenza più o meno improvvisa del corpo. Secondo Morin, questa rivendicazione corrisponde al « movimento elementare dello spirito umano che fin dall'inizio non situa e non conosce la propria intimità se non al proprio esterno. In effetti non ci si sente, non ci si comprende, non ci si vede prima di tutto se non come ''altro'', vale a dire proiettato e alienato ».

Ora, è facilmente comprensibile come la concezione del morto come doppio, quindi come incarnazione della presenza di un assente, incontri in maniera armonica gli aspetti che già di per sé definiscono il mondo contemporaneo, in cui le dimensioni online e offline sono inestricabilmente intrecciate. Nel corso degli anni abbiamo, infatti, imparato a conseguire una sorta di ubiquità esistenziale, figlia di un mondo in cui non è più possibile stabilire una rigida separazione tra reale e virtuale: il luogo fisico occupato dal nostro corpo non coincide più, necessariamente, con la sua effettiva presenza. In altre parole, il prolungamento delle nostre identità psicofisiche dentro i numerosi luoghi online, accessibili ovunque e in qualunque momento tramite smartphone, computer, tablet e una funzionante connessione wi-fi, ci permette di essere contemporaneamente in più luoghi, oltre a quello in cui si trova il corpo biologico: social network, app di messaggistica privata, posta elettronica, blog, videogiochi, realtà virtuali, ecc. In ciascuno di questi luoghi la presenza del singolo individuo è, ovviamente, contraddistinta da espressioni e forme differenti della sua identità.

La pandemia da Covid-19 lo ha evidenziato in maniera assai cristallina: all'impossibilità per legge di varcare i confini delle nostre abitazioni con il nostro unico corpo biologico, a causa dei rischi di con-

tagio virale, abbiamo contrapposto la libera circolazione dei nostri corpi digitali in ogni luogo della Rete, in quanto immune alla pericolosità del virus (Sisto, 2022). Ma, al di là dell'epoca pandemica, siamo oramai abituati da anni a non sottovalutare l'importanza di una presenza online — per esempio, all'interno di un social network — rispetto alla localizzazione corporea in una strada della città in cui abitiamo. A tal proposito, un recente studio antropologico-etnologico definisce lo smartphone come una 'casa trasportabile', in quanto ci permette di portare 'materialmente' con noi i nostri conoscenti e — al tempo stesso — di andare da loro, a prescindere dall'esatta collocazione geografica del nostro corpo (Miller *et al.* 2021).

L'evento della morte non fa che unire insieme la differenziazione tecnologica tra localizzazione fisica e presenza dei corpi, lo smartphone come casa trasportabile e la multi-identità espressa nei vari luoghi online con la menzionata teoria del doppio e dell'ego alter propria del morto. Una volta deceduti, continuiamo infatti a rimanere presenti — in virtù del processo di registrazione, processo a fondamento della realtà online — nei numerosi luoghi digitali che abbiamo popolato nel corso della vita. La registrazione e l'archiviazione delle nostre tracce scritte, fotografiche e audiovisive ci permette, cioè, di rimanere 'presenti' come profili social, avatar nei videogiochi, account di posta elettronica e di messaggistica privata e via dicendo, nonostante il decesso avvenuto. Ormai da anni si ragiona sul fatto che i social network contengono e preservano le sostanziose tracce di decine di milioni di profili di utenti deceduti, divenendo nel corso degli anni una via di mezzo tra cimiteri digitali ed enciclopedie dei morti. L'intelligenza artificiale ha, pertanto, una quantità considerevole di materiale come base per rendere perennemente attivi quei prolungamenti delle nostre identità che, dopo la morte, rimangono silenziosi all'interno di ogni luogo online in cui sono stati registrati.

Come ci insegna Dan Crowe, nel libro *Dead Interview* del 2013, da sempre parliamo con i morti. *Dead Interview* è un libro che lo vuole dimostrare, contenendo le 'interviste' ad alcuni famosi artisti deceduti, come Andy Warhol, Friedrich Nietzsche, Marcel Duchamp, ecc., realizzate da scrittori ancora vivi, come Douglas Coupland, Geoff Dyer, Rick Moody, ecc. Crowe spiega l'iniziativa sostenendo che desideriamo, dagli albori dei tempi, che i morti ci diano informazioni, ci forniscano rivelazioni o consigli, rispondano alle nostre domande rimaste aperte dopo il loro decesso, addirittura ci offrano il loro aiuto. Siamo curiosi, infatti, di sapere dove si trovano, se esiste

Dio, se sia ancora possibile risolvere rimpianti e sensi di colpa a causa di tutto ciò rimasto insoluto. L'atavico dialogo con loro, finora, ha comportato il fatto di sentire solo la voce onnicomprensiva dei vivi, dunque l'attribuzione ai vivi del doppio compito di mittente e destinatario del messaggio. Il morto lascia una traccia — scritta, fotografica, ecc. — e il vivo la riempie con il suo tono vocale e le sue parole scritte. La novità odierna consiste nel fatto che l'intelligenza artificiale ci dà l'impressione che il morto conservi anche una sua specifica autonomia: egli sembra, cioè, rispondere alle nostre esigenze, interpretando attivamente e con libero arbitrio il ruolo dell'interlocutore che, finora, in assenza di strumenti consoni hanno interpretato solo i dolenti. Ovviamente, come ho già evidenziato altrove, l'autonomia e il libero arbitrio dei thanabots e dei griefbots sono solo apparenti, giacché essi non sono altro che il risultato di una rielaborazione intelligente di dati precedentemente pubblicati (HENRICSON, 2023).

3. Cadaveri informazionali ed eredità digitale.

Questa novità tecnologica, nell'aggiornare al tempo del digitale un comportamento atavico dell'essere umano, amplifica il senso di presenza nell'assenza, comportando una serie di inedite conseguenze. Mettiamo da parte la questione, di natura prevalentemente psicologica, dell'elaborazione del lutto, di cui mi sono occupato altrove (SISTO, 2025), e soffermiamoci invece sull'arbitrarietà che caratterizza l'uso dell'eredità digitale, dunque dei dati in precedenza registrati. Le due storie da cui siamo partiti sono accomunate, di fatto, da un'indebita appropriazione dell'identità della persona defunta, pur con finalità differenti. Jennifer Ann Crecente è diventata, sua malgrado, uno strumento nelle mani di un'app intenta a trarre guadagno dal suo indebito uso. Christopher Pelkey, invece, rappresenta il tentativo da parte di un dolente — in questo caso, la sorella — di pervenire a quella pace perduta dopo il tragico e insensato lutto, come abbiamo già evidenziato. Immaginare il fratello che non serba rancore nei confronti del suo aguzzino è un modo del tutto personale per superare l'angoscia e il dolore provati dinanzi a una simile tragedia.

Detto ciò, tanto Crecente quanto Pelkey sono diventati loro malgrado dei burattini artificiali o, meglio, dei 'cadaveri informazionali', come li definisce Carl Öhman (ÖHMAN, 2024). Lo studioso svedese usa questo concetto per sottolineare gli effetti politici ed economici

della pervasiva presenza dell'assente sottoforma di avatar, ologramma o, comunque, gemello digitale. Il carattere pubblico dei suoi dati, precedentemente registrati, lo rendono post mortem un soggetto paradossalmente ancora attivo ma incapace di avere un controllo, anche minimo, sulle proprie attività ed espressioni pubbliche postume. In altre parole, i servizi che ci permettono di 'animare' i morti con l'intelligenza artificiale utilizzano, in maniera del tutto arbitraria, i prolungamenti digitali della loro identità psicofisica all'interno degli account in cui si sono iscritti nel corso della vita.

Ogni biografia personale, pertanto, dopo l'evento della morte rischia di diventare una risorsa commerciale, del tutto avulsa alle esigenze di proteggere la privacy e l'identità del morto. Ovviamente, si può obiettare che il cadavere informazionale non è altro che il punto di arrivo attuale di un comportamento che caratterizza da sempre le nostre società: vale a dire, l'assoluta mancanza di voce in capitolo del morto sulle proprie memorie, gestite e reiterate da chi è ancora vivo (pensiamo, per esempio, alle pubblicazioni postume di canzoni o addirittura album di un musicista deceduto).

Tuttavia, nei casi menzionati la mera memoria postuma, ricostruita arbitrariamente dai vivi, diventa una sorta di memoria riproduttiva, la quale scambia il ricordo con la presenza più o meno vivida. In altre parole, l'assenza di una vera e propria interruzione tra il pre e il post mortem di un individuo deceduto genera una memoria 'posticcia', poiché priva della salubre dialettica che imbastisce con l'oblio, dunque con il vuoto lasciato dal morto. Ecco perché è necessario limitare la dipendenza del cadavere informazionale dalle azioni dei vivi, gestendo preventivamente l'eredità digitale.

In altre parole, ogni utente della Rete dovrebbe prendere decisioni in merito all'eventuale modo di rimanere nel mondo, una volta deceduto, attraverso le precedenti registrazioni di dati. Un recente studio di NordPass ha rivelato che un utente della Rete può arrivare ad avere circa 170 account digitali, ciascuno con le sue specifiche credenziali di accesso. Questa produzione mastodontica di dati è praticamente incontrollabile nella sua totalità.

Per tale ragione, per esempio, in Italia il Consiglio Nazionale del Notariato ha presentato un decalogo sull'eredità digitale che rappresenta, a tal proposito, un buon punto di partenza. Il decalogo ci ricorda, innanzitutto, l'assenza di una legislazione specifica, motivo per cui occorre che ogni utente presti attenzione a una serie di regole di buon senso relative all'eredità digitale. Ha, quindi, determinato una

distinzione tra le eredità che corrispondono a un qualche valore economico da quelle che si identificano con l'identità vera e propria del morto. Da una parte, occorre essere consapevoli che le password non fanno parte dell'eredità digitale, così come sono esclusi dalla successione i beni piratati, i beni concessi in licenza, gli account di firma elettronica e dell'identità digitale.

Le criptovalute sono beni digitali con valore economico e un conto online è l'estensione virtuale di un conto reale. Dall'altra, bisogna ricordare le regole dei vari social media in merito alla gestione post mortem dei profili di ogni utente. Tenendo conto di ciò, ognuno di noi dovrebbe fare una sorta di *Death Cleaning* preventiva, stabilendo in maniera rigorosa cosa lasciare online post mortem e cosa invece eliminare. Ogni social media, infatti, offre delle soluzioni, più o meno ampie, di cosa fare del proprio account in caso di morte: può essere conservato così com'è, si può far chiudere previo invio di un certificato di morte, si possono scaricare i contenuti dell'account sul computer e poi lasciarli in eredità, su Facebook si può scegliere un contatto erede che gestisca i post pubblici, ecc.

Questo discorso va poi esteso ai contenuti dei mobile device, dei computer, ecc. L'attenzione per l'eredità digitale è, in altre parole, fondamentale sia per controllare razionalmente il processo di elaborazione del lutto sia per garantire una sana conservazione della memoria, limitando l'azione arbitraria dell'intelligenza artificiale. Occorre pertanto non sottovalutare l'importanza della propria vita online e fare in modo che società private, intente a trarre guadagno dai nostri dati, non si approfittino della nostra assenza definitiva, dunque della nostra impossibilità di controllare e gestire la nostra simbolica sopravvivenza post mortem. Inoltre, un'attenzione specifica per l'eredità digitale si lega anche alla consapevolezza che i thanabots, lungi dal rappresentare il prolungamento immortale di una vita biologica conclusa, non sono altro che rielaborazioni artificiali di espressioni personali registrate nel corso della propria vita. In altre parole, il doppio e l'ego alter plasmato dall'intelligenza artificiale corrispondono a una razionalizzazione asettica di comportamenti manifestati nei mondi online nel corso della propria vita. Dunque, l'attenta disposizione dell'eredità digitale impedisce la trasformazione del defunto in qualcosa di arbitrario che, di fatto, non corrisponde più a chi è stato, proprio come abbiamo visto nel caso di Jennifer Ann Crecente. In definitiva, solo mediante un attento studio interdisciplinare sia di cosa siamo nei mondi online sia di come tali mondi online

incontrano tradizioni ataviche nel campo del fine vita possiamo riuscire a gestire con responsabilità queste particolari innovazioni nel campo dell'intelligenza artificiale.

Note bibliografiche

Fried E. (1982), *Kobenhavns Amts Sygehus Gentofte*, in Id., *Das Unmaß aller Dinge. Erzählungen*, Berlin, Wagenbach.

Fuchs W. (1973), *Le immagini della morte nella società moderna. Sopravvivenze arcaiche e influenze attuali*, Torino, Einaudi.

Harju A.A. (2024), *Theorising Digital Afterlife as Techno-Affective Assemblage: On Relationality, Materiality, and the Affective Potential of Data*, in *Social Sciences*, 13, p. 227 ss.

Henrickson L. (2023), *Chatting with the dead: The hermeneutics of thanabots*, in *Media, Culture & Society*, 45, p. 5 ss.

Macho T. (2016), *Immagini e morte. Il tempo della fotografia*, in *Tropos. Rivista di ermeneutica e critica filosofica*, IX, 2, p. 13 ss.

Miller D.; Rabho L.A. *et al*. (2021), *The Global Smartphone. Beyond a Youth Technology*, UCL Press, London. Il sito associato al progetto ASSA, con un quantitativo ragguardevole di materiale interattivo, è il seguente: *www.ucl.ac.uk/anthropology/assa/#*.

Morin E. (2014), *L'uomo e la morte*, Trento, Erickson.

Öhman K. (2024), *The Afterlife of Data. What Happens to Yours Information When You Die and Why You Should Care*, Chicago, The University of Chicago Press.

Sisto D. (2018), *La morte si fa social. Immortalità, memoria e lutto nell'epoca della cultura digitale*, Torino, Bollati Boringhieri.

Sisto D. (2022), *Porcospini digitali. Vivere e mai morire online*, Torino, Bollati Boringhieri.

Sisto D. (2025), *Vivere per sempre. L'Aldilà ai tempi di ChatGPT*, Torino, Bollati Boringhieri.

(LA NORMA)
IMMORTALITÀ DEL DATO, EREDITÀ DIGITALE E INTELLIGENZA ARTIFICIALE: PROFILI GIURIDICI E TECNOLOGICI TRA OBLIO E MEMORIA

di *Giovanni Ziccardi*

Sommario: 1. Introduzione al tema. — 2. La tecnologia e l'‘eternità' del dato. — 3. Il diritto all'oblio tra tutela della persona e limiti tecnici. — 4. Eredità digitale e successione dei dati. — 5. Alcune proposte di disciplina normativa. — 6. Intelligenza artificiale, morte e 'ritorno' digitale. — 7. Alcune considerazioni conclusive.

1. Introduzione al tema.

La progressiva digitalizzazione della società contemporanea ha trasformato in maniera radicale la relazione tra memoria umana, tecnologia e idea di 'tempo'.

Se, in passato, il ricordo era affidato a supporti fisici destinati, per loro natura, a *deteriorarsi* — si pensi a pergamene, a libri o ad archivi cartacei — oggi la memoria si è spostata in spazi digitali caratterizzati da capacità di archiviazione potenzialmente illimitata e dalla possibilità di replicazione istantanea dei contenuti, in una vera e propria *rivoluzione documentale* (Ferraris, 2021).

Si è, così, pian piano affermata una sorta di 'memoria totale' del dato, che muta non soltanto la dimensione quantitativa delle informazioni disponibili ma, anche, il loro stato ontologico e giuridico: ciò che un tempo poteva essere dimenticato, cancellato o lasciato svanire con il trascorrere del tempo, diventa oggi persistente e potenzialmente eterno, con conseguenze profonde sul piano della privacy, della reputazione e della tutela della dignità personale (Mayer-Schönberger, 2016).

Questa transizione dalla caducità alla persistenza ha un'origine eminentemente tecnologica.

Le architetture digitali contemporanee — dai sistemi di cloud computing (BELLOTTI, 2019) alle reti di distribuzione dei contenuti, sino alla blockchain — si fondano su meccanismi di ridondanza e replicazione che garantiscono la disponibilità costante dei dati, riducendo al minimo il rischio di perdita.

L'informazione viene così moltiplicata e conservata in una rete globale di server, *mirror* e copie di sicurezza, in una logica che privilegia la continuità e la resilienza rispetto alla possibilità di *oblio* (MARTINELLI, 2017).

La stessa intelligenza artificiale, alimentata da flussi di dati incessanti, necessita di enormi archivi digitali (dataset) per addestrare modelli sempre più complessi (QUINTARELLI, 2020), rafforzando ulteriormente questa *tensione verso la permanenza*.

Tuttavia, proprio questa persistenza introduce un paradosso giuridico e tecnologico: se il dato digitale è teoricamente eterno, la sua cancellazione diventa problematica, talvolta impossibile (MAYER-SCHÖNBERGER, 2016).

L'irreversibilità dell'informazione in rete, unita alla sovrabbondanza dei contenuti generati quotidianamente, pone, allora, *sfide inedite* al diritto, chiamato a confrontarsi con un ambiente informativo dove l'oblio non è più naturale ma deve essere costruito artificialmente, con strumenti normativi e tecnici spesso inadeguati (o in ritardo) rispetto alla velocità dell'innovazione (ZICCARDI, 2022).

Il dibattito recente sul diritto all'oblio, così come quello sull'eredità digitale (VESTO, 2020) e della gestione postuma dei dati personali, nasce proprio da questa *tensione* tra memoria infinita e diritto delle persone a decidere del proprio passato, anche oltre la vita biologica (BELLOMIA, 2020).

2. La tecnologia e l'"eternità' del dato.

Il dato digitale si distingue, si diceva poco sopra, per una caratteristica che lo separa in modo radicale dalle forme di informazione precedenti: la sua *potenziale immortalità*.

Nel contesto analogico, la memoria era fragile, affidata a supporti fisici deteriorabili e alla volontà umana di conservarli. Con l'avvento delle tecnologie digitali, invece, il dato diviene replicabile all'infinito, archiviabile in reti globali e ridondanti, sottratto al logorio del tempo e all'usura della materia.

La digitalizzazione ha inaugurato, in altre parole, una nuova 'temporalità dell'informazione', dove l'istante della creazione del dato non coincide più con l'inizio del suo decadimento ma, piuttosto, con l'inizio di una permanenza potenzialmente illimitata.

Le infrastrutture tecnologiche che sostengono la società dell'informazione si fondano su questa logica di persistenza.

I sistemi di cloud computing, ad esempio, distribuiscono i dati su una molteplicità di server fisicamente collocati in luoghi diversi (Farina, 2019), creando copie di sicurezza e punti di accesso ridondanti per garantire la disponibilità continua dell'informazione (Bellotti, 2019).

A ciò si aggiungono le tecnologie peer-to-peer e, in misura ancora più radicale, la *blockchain*, la quale nasce con l'obiettivo di rendere ogni informazione immutabile e trasparente, 'scolpita' in registri distribuiti che per definizione resistono alla cancellazione o alla modifica unilaterale (Capaccioli, 2015).

Questa persistenza tecnologica non è tuttavia neutra: essa influenza direttamente la costruzione giuridica del dato e ne condiziona le modalità di tutela.

Laddove l'informazione è diffusa in sistemi decentralizzati, replicata su server internazionali e indicizzata dai motori di ricerca, il diritto incontra ostacoli significativi nell'assicurare la cancellazione o la deindicizzazione (Mayer-Schönberger, 2016).

La natura 'globale' dell'infrastruttura informatica sfida i tradizionali confini della giurisdizione, mentre la complessità tecnica rende difficile l'esercizio effettivo di diritti quali quello all'oblio o alla rettifica (Bellomia, 2020).

In altre parole, l'architettura stessa di Internet e dei sistemi di archiviazione contemporanei tende a privilegiare la *permanenza* rispetto alla *caducità*, imponendo al diritto il compito — complesso e spesso tardivo — di reintrodurre la possibilità di oblio in un ambiente concepito per ricordare indefinitamente.

La tendenza alla permanenza dell'informazione non è un fenomeno esclusivo dell'era digitale, sia chiaro, ma affonda le sue radici nella storia dei supporti di memoria.

Dalla fragilità del papiro e della pergamena, soggetti a usura fisica e deterioramento ambientale, alla maggiore durabilità della carta, fino all'invenzione dei supporti magnetici e ottici del XX secolo — come i nastri, i floppy disk, i CD-ROM — la durata dell'informazione è sempre stata in funzione della *robustezza* materiale del *medium*. Tuttavia, tutti questi supporti analogici erano accomunati dalla

possibilità concreta della *perdita*: incendi, smarrimenti, incuria o semplice obsolescenza tecnologica comportavano inevitabilmente la cancellazione o l'illeggibilità del contenuto.

Con la transizione al digitale e al cloud computing, questo rapporto si è rovesciato: l'informazione non è più un bene fragile e locale, ma un'entità smaterializzata, replicata in architetture distribuite e potenzialmente conservata in perpetuo (BELLOTTI, 2019).

Sul piano tecnico, questa persistenza è garantita da meccanismi avanzati di *ridondanza* e *replicazione*.

Le copie cache, i sistemi di backup distribuiti e le logiche di *disaster recovery* assicurano che un dato non esista mai in un'unica copia, ma venga duplicato in tempo reale su server collocati in aree geografiche differenti.

La blockchain porta questo paradigma all'estremo: il registro distribuito, per definizione, impedisce la cancellazione o la modifica retroattiva delle informazioni, poiché ogni blocco della catena è crittograficamente legato al precedente.

Ciò significa che, una volta immesso in rete, un dato non è mai realmente eliminabile: può al massimo essere reso inaccessibile o oscurato, ma la sua cancellazione fisica o logica completa rimane, nella pratica, quasi impossibile (MARTINELLI, 2017).

Questa infrastruttura globale solleva un profilo geopolitico di grande rilievo (SUFFIA, 2018; ARESU, 2024). I *data center* che ospitano copie dei dati sono distribuiti in Paesi con normative differenti in materia di privacy, sicurezza e diritti fondamentali, ponendo problemi di giurisdizione e di sovranità dei dati.

Alcuni ordinamenti — si pensi alla Cina, con le leggi sulla *data localization* (PIERANNI, 2023), o alla Russia con le norme che impongono la conservazione dei dati entro i confini nazionali — richiedono che le informazioni dei cittadini rimangano fisicamente sul territorio nazionale, mentre il Regolamento europeo sulla protezione dei dati (GDPR) pone limiti stringenti ai trasferimenti verso Paesi terzi che non offrano adeguate garanzie.

La tensione tra architettura tecnica globale e regolamentazione giuridica territoriale crea un conflitto permanente tra esigenze di efficienza, tutela dei diritti e interessi geopolitici (ARESU, 2024).

Infine, la dipendenza dell'intelligenza artificiale da grandi *dataset* permanenti rafforza ulteriormente questa tendenza alla persistenza.

I sistemi di *machine learning* e *deep learning* (RUSSELL-NORVIG, 2021) necessitano di enormi quantità di dati storici per essere adde-

strati e migliorati; l'efficacia dei modelli predittivi dipende direttamente dalla disponibilità di archivi completi, eterogenei e temporalmente estesi.

Ciò comporta una *pressione costante* verso la conservazione illimitata dei dati, in aperta tensione con i principi di minimizzazione e limitazione della conservazione sanciti dal GDPR (FINOCCHIARO, 2019).

Il diritto si trova così di fronte a un paradosso: da un lato, la necessità tecnica di dati permanenti per l'innovazione; dall'altro, il diritto degli individui all'oblio e alla cancellazione dei propri dati personali.

3. Il diritto all'oblio tra tutela della persona e limiti tecnici.

Il diritto all'oblio rappresenta una delle risposte più significative, ma anche più controverse, alla tensione tra la persistenza del dato digitale e la tutela della persona che abbiamo descritto nei paragrafi precedenti.

Esso nasce dall'esigenza di garantire agli individui la possibilità di non essere *permanentemente esposti* al passato e di poter ricostruire la propria identità senza essere indefinitamente prigionieri di informazioni superate, inesatte o non più rilevanti (DE FRANCESCO, 2024).

Nella dimensione analogica, questo diritto era in larga misura affidato alla fisiologia stessa del tempo e all'inevitabile caducità della memoria collettiva; nell'era digitale, invece, la permanenza dell'informazione, potenziata dai motori di ricerca e dalle piattaforme di social network, rende necessario un intervento normativo specifico (MARTINELLI, 2017).

L'affermazione giuridica del diritto all'oblio, in ambito europeo, è stata segnata da una tappa fondamentale: la sentenza *Google Spain* della Corte di Giustizia dell'Unione Europea, che ha riconosciuto il diritto degli interessati a chiedere la de-indicizzazione di risultati di ricerca contenenti informazioni non più pertinenti o eccessive rispetto alle finalità originarie (MARTINELLI, 2017).

Tale pronuncia ha reso evidente che, in un ecosistema informativo caratterizzato da archivi potenzialmente eterni e indicizzazioni globali, la protezione della dignità e della reputazione personale richiede strumenti giuridici nuovi, capaci di incidere sul modo in cui l'informazione è resa accessibile, più che sulla sua mera esistenza materiale.

Tuttavia, la *dimensione tecnica* del problema ha reso fin dall'inizio complessa l'attuazione concreta del diritto all'oblio.

La cancellazione di un dato, infatti, non equivale alla sua *scomparsa*: le informazioni possono essere state replicate, archiviate in copie *cache*, diffuse su piattaforme terze (MOLLO, 2021), memorizzate su server collocati in giurisdizioni differenti. La stessa richiesta di deindicizzazione agisce solo sull'accessibilità tramite motori di ricerca, senza intaccare necessariamente la persistenza dell'informazione alla fonte (MARTINELLI, 2017).

A ciò si aggiunge il delicato bilanciamento con altri diritti fondamentali, primo fra tutti la *libertà di informazione*: il diritto all'oblio non può trasformarsi in un diritto alla riscrittura della storia o alla censura, e la giurisprudenza europea ha costantemente sottolineato la necessità di valutare caso per caso, considerando il ruolo pubblico della persona, la natura dell'informazione e l'interesse collettivo alla conoscenza.

In questo senso, il diritto all'oblio si colloca in una zona di collegamento permanente tra privacy, memoria e libertà, una tensione che la tecnologia, lungi dal risolvere, contribuisce a rendere ancora più complessa (HARBINJA, 2023).

Il diritto all'oblio si è evoluto, per di più, in modo *disomogeneo* nei diversi ordinamenti giuridici, con soluzioni che riflettono le tradizioni giuridiche nazionali e il diverso bilanciamento tra diritto alla privacy e libertà di espressione (HARBINJA, 2023).

In Italia, come è noto, la giurisprudenza della Corte di Cassazione ha a lungo privilegiato la 'attualità dell'interesse pubblico' come criterio per stabilire la legittimità della permanenza online di notizie relative a fatti del passato, in linea con il principio secondo cui la cronaca deve essere veritiera, corretta e socialmente utile.

La Francia, invece, ha adottato un approccio più orientato alla protezione della dignità personale, con una giurisprudenza che ha progressivamente ampliato l'ambito di applicazione del diritto all'oblio, in particolare dopo la citata pronuncia della Corte di Giustizia dell'Unione Europea nel caso *Google Spain*.

In Germania, la Corte costituzionale federale ha più volte sottolineato il principio della proporzionalità e l'esigenza di un bilanciamento tra diritti fondamentali, riconoscendo che la libertà di informazione deve *arretrare* quando la diffusione di notizie datate e non più attuali compromette in modo sproporzionato i diritti della personalità.

Negli Stati Uniti d'America, invece, il forte radicamento del Primo Emendamento e della libertà di stampa ha sovente impedito l'affermazione di un vero e proprio diritto all'oblio, con la conseguenza che i dati rimangono accessibili anche a distanza di molti anni, salvo limitate eccezioni previste da normative settoriali.

Anche il Comitato Europeo per la Protezione dei Dati ha sovente dato indicazioni alle Data Protection Authority dei singoli Paesi con linee guida che chiariscono le modalità di esercizio del diritto alla de-indicizzazione, sottolineando l'obbligo dei motori di ricerca di adottare procedure trasparenti e tempestive.

Dal punto di vista tecnico, le tecniche di de-indicizzazione consistono principalmente nella rimozione dei collegamenti ipertestuali dai risultati dei motori di ricerca, senza tuttavia eliminare l'informazione alla fonte.

Ciò lascia aperto il problema della *persistenza materiale* dei dati, che possono continuare a circolare attraverso copie *cache*, archivi digitali o siti esteri non soggetti alla giurisdizione nazionale.

In ambito della regolamentazione industriale, invece, esistono standard sulla sicurezza dell'*archiviazione* dei dati e sulla *cancellazione sicura*, che definiscono procedure per la distruzione irreversibile delle informazioni digitali, ma la loro applicazione resta limitata e spesso incompatibile con la natura distribuita delle reti globali.

Il bilanciamento con la libertà di stampa e di informazione rappresenta, infine, uno dei punti più delicati.

La CEDU ha in diverse occasioni stabilito come occorra sempre valutare diversi fattori: il contributo dell'informazione a un dibattito di interesse generale, la notorietà della persona coinvolta, il comportamento dell'interessato e il contenuto, la forma e le conseguenze della pubblicazione.

Questi criteri mirano a evitare che il diritto all'oblio si trasformi in uno strumento di censura, salvaguardando la *funzione democratica* dell'informazione pur riconoscendo l'esigenza di tutelare la dignità e la reputazione individuale.

4. Eredità digitale e successione dei dati.

L'avvento della società digitale ha introdotto una questione del tutto nuova nel diritto delle successioni: la sorte dei dati personali, degli account e dei contenuti online delle persone decedute.

Mentre il patrimonio materiale trova nella tradizione giuridica strumenti consolidati di trasmissione agli eredi, l'eredità digitale si colloca in un *terreno ibrido*, dove la dimensione patrimoniale s'intreccia con aspetti personali, identitari e, spesso, emotivi (D'ARMINIO MONFORTE, 2020).

Profili quali l'accesso alle e-mail, ai social network, agli archivi fotografici online o ai portafogli di criptovalute sollevano interrogativi complessi sulla titolarità, sul consenso postumo e sulla possibilità di disporre di contenuti che, pur creati dall'utente, sono regolati dai termini di servizio delle piattaforme (MOLLO, 2021).

La prassi delle grandi aziende tecnologiche mostra una varietà di soluzioni, spesso dettate più da scelte unilaterali che da normative organiche.

Facebook, ad esempio, ha introdotto la funzione di 'commemorazione' degli account, che consente agli eredi designati di gestire in modo limitato il profilo del defunto, preservandone la memoria senza confonderlo con un account attivo (GAMBA, 2018).

Google, con il servizio 'Inactive Account Manager', permette invece di *pianificare* in vita il destino dei propri dati, individuando persone fidate che possano accedervi o decidere per la loro cancellazione dopo un periodo di inattività. Queste soluzioni aziendali, tuttavia, restano eterogenee e prive di un coordinamento normativo unitario, affidando a strumenti contrattuali la disciplina di diritti che toccano la sfera più intima dell'individuo.

Sul piano giuridico, il conflitto tra diritto alla privacy, che si estingue con la morte nella maggior parte degli ordinamenti, e diritto degli eredi a disporre del patrimonio digitale, crea da tempo zone d'ombra particolarmente problematiche (HARBINJA, 2023).

La giurisprudenza europea ha sovente affermato che gli account social possono essere trasmessi agli eredi come *qualunque altro bene*, riconoscendo così una forma di *successione digitale* (D'ARMINIO MONFORTE, 2020).

Tuttavia, l'assenza di una normativa europea specifica comporta soluzioni frammentarie e dipendenti da scelte giurisprudenziali o contrattuali, con evidenti ricadute in termini di certezza del diritto e tutela dell'autonomia individuale.

Il tema dell'eredità digitale ha conosciuto negli ultimi anni un'evoluzione significativa nel *diritto comparato*, sebbene con soluzioni ancora eterogenee e prive di un coordinamento internazionale (VESTO, 2020).

Negli Stati Uniti d'America si consente agli eredi o ai 'fiduciari digitali' di accedere ai contenuti online del defunto solo se il soggetto aveva *espresso in vita* una volontà chiara in tal senso, ad esempio tramite strumenti predisposti dalle stesse piattaforme o mediante disposizioni testamentarie specifiche. In assenza di tale consenso, prevalgono le condizioni contrattuali dei fornitori di servizi, con il rischio che l'account resti inaccessibile o venga cancellato automaticamente.

In Germania, la Corte Federale ha stabilito da tempo che gli account Facebook possano essere trasmessi agli eredi al pari di *qualsiasi altro bene*, configurando i dati digitali come parte del patrimonio ereditario. Questa interpretazione ha segnato una svolta, poiché ha parificato i contenuti digitali alla corrispondenza privata cartacea, sottolineando che il passaggio all'era digitale non può determinare una minore tutela per i diritti successori.

In Francia, si prevede la possibilità per ogni persona di *definire in vita* le modalità di conservazione, cancellazione o comunicazione dei propri dati personali dopo la morte, attribuendo ai fornitori di servizi l'obbligo di rispettare tali volontà.

Un nodo centrale riguarda, poi, la distinzione tra contenuti *patrimoniali* e contenuti *personali*.

I primi includono beni digitali con valore economico diretto, come criptovalute, domini web, archivi fotografici professionali o biblioteche digitali acquistate con licenze trasferibili. I secondi, invece, comprendono e-mail, messaggi, post sui social network o dati di natura affettiva, la cui rilevanza patrimoniale è minima o nulla ma che possono avere un forte valore identitario o emotivo.

La giurisprudenza europea, salvo rare eccezioni, tende a equiparare entrambe le categorie sul piano della *trasmissibilità*, lasciando tuttavia aperti i problemi di natura contrattuale e le eventuali limitazioni derivanti dalle condizioni di servizio.

Le politiche delle principali piattaforme digitali mostrano infatti una notevole eterogeneità (MOLLO, 2021).

Facebook, si diceva, consente agli utenti di nominare un *legacy contact* che potrà gestire in modo limitato l'account dopo la morte, trasformandolo in una pagina commemorativa (GAMBA, 2018). Google offre il servizio *Inactive Account Manager*, che permette di stabilire in anticipo se e a chi trasferire dati e contenuti dopo un periodo di inattività. Apple, dal 2021, ha introdotto la funzione *Digital Legacy*, che consente agli eredi designati di accedere a foto, documenti e altri contenuti iCloud, ma solo previa presentazione di un

certificato di morte e altre formalità. Al contrario, Twitter/X si limita a consentire la *disattivazione* dell'account da parte dei familiari, senza possibilità di accesso ai contenuti né di gestione commemorativa. Questa frammentazione normativa e contrattuale, già dicevamo poco sopra, crea situazioni di incertezza, affidando spesso a *politiche aziendali unilaterali* la gestione di diritti che dovrebbero essere disciplinati da fonti normative chiare e uniformi.

Gli *aspetti fiscali* aggiungono un ulteriore livello di complessità, soprattutto con riferimento alle criptovalute e agli *asset* digitali decentralizzati (CAPACCIOLI, 2023).

In molti ordinamenti, questi beni rientrano nella successione al pari di qualsiasi altro patrimonio (D'ARMINIO MONFORTE, 2020), ma la loro natura crittografica comporta difficoltà pratiche significative: in assenza delle chiavi private, infatti, l'accesso ai fondi digitali può risultare tecnicamente impossibile, rendendo di fatto intrasmissibili gli *asset* anche quando la legge ne riconosce la trasferibilità. Alcuni Paesi, come gli Stati Uniti d'America e il Regno Unito, hanno avviato linee guida fiscali specifiche per la valutazione e la tassazione delle criptovalute in sede successoria, mentre in Europa il quadro resta frammentario e privo di una disciplina organica.

5. Alcune proposte di disciplina normativa.

La complessità dell'eredità digitale e della gestione postuma dei dati richiede, a nostro avviso, un intervento normativo organico, capace di superare la frammentarietà delle soluzioni oggi offerte da piattaforme e pronunce giurisprudenziali (PETRELLA, 2024).

L'assenza di regole chiare produce, infatti, *incertezza* tanto per gli utenti quanto per gli operatori, con il rischio di affidare a scelte unilaterali delle imprese digitali questioni che toccano diritti fondamentali, interessi patrimoniali e valori etici di rilievo costituzionale.

Un primo passo potrebbe essere l'adozione, a livello europeo, di un quadro normativo uniforme che disciplini in modo chiaro la sorte dei dati personali dopo la morte dell'interessato.

Tale intervento dovrebbe definire, almeno: *i*) la titolarità postuma dei dati, distinguendo tra contenuti a prevalente rilevanza patrimoniale (ad esempio criptovalute, archivi professionali) e contenuti a natura personale o relativa all'identità (corrispondenza privata, social network); *ii*) i poteri degli eredi o dei soggetti designati, stabilendo in quali casi sia possibile accedere, trasferire o cancellare i dati

e in quali, invece, debba prevalere il diritto alla riservatezza o alla memoria storica; *iii*) le modalità di pianificazione anticipata, attraverso strumenti analoghi ai testamenti digitali o ai gestori di account inattivi, che permettano all'utente di disporre in vita della propria eredità digitale con piena consapevolezza (PETRELLA, 2024).

Un tale quadro normativo dovrebbe, inoltre, tenere conto della *dimensione tecnica* del problema: l'effettiva cancellazione di dati distribuiti in sistemi cloud, archivi decentralizzati o blockchain non può essere garantita con strumenti giuridici astratti, ma richiede standard tecnologici condivisi e obblighi di *interoperabilità* tra piattaforme.

Infine, sarebbe auspicabile il coinvolgimento delle autorità di protezione dati — come i Garanti privacy nazionali e l'EDPB — nella definizione di linee guida operative che favoriscano una *convergenza* tra esigenze di tutela della persona, libertà di informazione e interessi economici, in modo da evitare che il tema dell'eredità digitale resti confinato alla discrezionalità delle big tech o alla casistica giudiziaria.

La regolamentazione dell'eredità digitale e della gestione postuma dei dati personali richiede, pertanto, un *approccio multilivello* che coinvolga sia le istituzioni nazionali sia quelle sovranazionali, al fine di superare la frammentarietà delle soluzioni attuali (PETRELLA, 2024).

Gli scenari di policy delineati negli ultimi anni evidenziano un progressivo interesse da parte di organismi internazionali.

L'Unione Europea, dopo il GDPR, non ha ancora adottato un quadro specifico sull'eredità digitale, limitandosi a lasciare agli Stati membri la possibilità di introdurre normative nazionali sul trattamento *post-mortem* dei dati. Alcuni Paesi, come Italia, Francia e Germania, hanno già colmato parzialmente il vuoto normativo, ma manca una cornice armonizzata che garantisca diritti uniformi in tutta l'Unione.

Il Consiglio d'Europa, attraverso la Convenzione 108+ sulla protezione delle persone rispetto al trattamento automatizzato di dati personali, ha inserito riferimenti generali alla tutela dei diritti digitali, ma senza affrontare esplicitamente la questione *post-mortem*.

A livello globale, le Nazioni Unite hanno avviato dibattiti sulla governance dell'intelligenza artificiale e dei dati, ma le raccomandazioni restano di carattere etico e programmatico, senza sfociare in strumenti giuridicamente vincolanti. Ciò conferma l'esigenza di un intervento legislativo europeo organico, eventualmente integrato da convenzioni internazionali che disciplinino il trattamento dei dati

dopo la morte, garantendo coerenza e certezza giuridica oltre i confini nazionali.

Uno degli strumenti più promettenti per restituire agli individui il controllo sul proprio patrimonio digitale è rappresentato dai *testamenti digitali* e dai *legati digitali*.

Alcuni ordinamenti hanno introdotto esperienze pilota: negli Stati Uniti è consentito designare in vita un fiduciario con poteri di accesso e gestione dei beni digitali dopo la morte. In Francia, allo stesso modo, si attribuisce a ogni cittadino il diritto di stabilire in anticipo le modalità di conservazione o cancellazione dei propri dati postumi, mediante istruzioni vincolanti per i fornitori di servizi.

Anche nel settore privato sono emerse soluzioni tecnologiche, come piattaforme che permettono la creazione di 'eredi digitali' con poteri predeterminati, gestiti tramite *smart contracts* su blockchain per garantire autenticità e trasparenza.

Il coinvolgimento delle autorità competenti appare cruciale per garantire che queste soluzioni rispettino diritti fondamentali e standard etici condivisi.

A livello europeo, l'EDPB potrebbe emanare linee guida vincolanti sull'eredità digitale, analoghe a quelle già adottate in materia di *consenso* e *trasferimenti internazionali* di dati.

I Garanti nazionali per la protezione dei dati potrebbero, poi, vigilare sull'attuazione delle disposizioni, sanzionando eventuali abusi o omissioni da parte delle piattaforme. Infine, l'UNESCO e organismi etici internazionali potrebbero contribuire a definire principi comuni sulla dignità *post-mortem*, sull'uso dell'intelligenza artificiale per ricostruire identità digitali e sul rispetto del *consenso espresso in vita*, evitando derive tecnologiche incompatibili con i valori universali di libertà e dignità umana (De Francesco, 2024).

6. Intelligenza artificiale, morte e 'ritorno' digitale.

L'ultima frontiera del rapporto tra tecnologia e morte si colloca nell'incontro tra *intelligenza artificiale* e *memoria digitale*, un ambito che solleva questioni del tutto inedite sul piano giuridico, etico e sociale (Sisto 2025).

L'abbondanza di dati personali generati durante la vita — fotografie, messaggi, video, *post* sui social media — offre oggi la possibilità tecnica di ricostruire, attraverso modelli di *machine learning* e *Large Language Models* (LLM), rappresentazioni digitali di persone

decedute, capaci di simulare conversazioni o di generare contenuti come se provenissero dal soggetto originario.

Fenomeni come i chatbot della memoria, o l'uso dei deepfake per ricreare voci e sembianze dei defunti, introducono un paradosso culturale e giuridico: il confine tra memoria e resurrezione digitale diventa sfumato, e l'identità della persona rischia di essere manipolata o ricostruita in modo artificiale senza un consenso chiaramente espresso in vita (Sisto, 2020).

Se, da un lato, queste tecnologie possono offrire strumenti per l'elaborazione del lutto o per finalità artistiche e commemorative, dall'altro pongono interrogativi cruciali sulla dignità *post-mortem*, sull'autenticità delle rappresentazioni digitali e sulla possibile strumentalizzazione commerciale dell'immagine e della voce dei defunti.

Dal punto di vista giuridico, l'assenza di una disciplina specifica rende difficile stabilire chi possa autorizzare la creazione o l'utilizzo di *avatar* digitali postumi, se sia necessario il consenso espresso in vita o se possano intervenire gli eredi.

In ambito europeo, il GDPR tutela i dati personali delle persone viventi, lasciando scoperto il trattamento *post-mortem*, salvo interventi legislativi nazionali che restano però, si è visto, sporadici e frammentari. Analogamente, il diritto d'autore e i diritti della personalità offrono strumenti parziali, non sempre idonei a regolare la complessità di fenomeni che combinano big data, intelligenza artificiale e identità digitale.

Infine, l'evoluzione di queste tecnologie apre scenari futuri ancora più radicali: l'addestramento di modelli di intelligenza artificiale su archivi personali completi potrebbe consentire di creare simulazioni sempre più realistiche di individui defunti, trasformando la loro presenza digitale in una sorta di *immortalità algoritmica* (Gamba, 2018). Tale prospettiva, se non accompagnata da regole chiare e da un dibattito etico consapevole, rischia di destabilizzare categorie fondamentali del diritto, come identità, privacy e autonomia personale, imponendo una riflessione urgente sul ruolo dell'intelligenza artificiale nella gestione della memoria e della morte nel mondo digitale.

Negli ultimi anni, l'evoluzione delle tecnologie di *deep learning* e intelligenza artificiale generativa ha reso possibile la creazione di rappresentazioni sempre più realistiche di persone defunte, dando origine a casi reali che hanno suscitato ampi dibattiti pubblici e hanno spinto giuristi e filosofi a interrogarsi sulla linea di confine tra com-

memorazione, sfruttamento economico e manipolazione della memoria collettiva.

Dal punto di vista psicologico, il contatto con *avatar* digitali di persone defunte può aiutare alcune persone a gestire il dolore, ma rischia anche di alimentare dipendenze emotive o di ostacolare l'elaborazione del lutto, trasformando il ricordo in un'interazione continua e potenzialmente manipolabile (SISTO, 2022).

Accanto alle questioni etiche, emergono rischi di sicurezza e frodi digitali: l'uso dei deepfake per ricreare voci e volti di defunti può essere sfruttato per truffe, estorsioni o manipolazioni dell'opinione pubblica (ZICCARDI, 2024).

Sono già stati segnalati casi di *voice phishing* (o 'vishing') in cui imitazioni vocali hanno ingannato familiari o istituzioni bancarie, e la possibilità di falsificare dichiarazioni o testamenti digitali apre scenari di contenziosi legali senza precedenti. La difficoltà di distinguere tra contenuti autentici e simulati mina, inoltre, la *fiducia collettiva* nell'informazione, aggravando il problema già esistente della disinformazione online (SUFFIA, 2022).

Infine, i futuri scenari prospettano possibilità ancora più radicali. Diversi progetti propongono la creazione di avatar digitali permanenti, addestrati sui dati personali degli individui per consentire una forma di 'immortalità algoritmica'.

L'idea del *mind uploading*, ovvero il trasferimento della mente umana in supporti digitali, oggi ancora relegata alla fantascienza, trova eco in ricerche avanzate sul connettoma cerebrale e in esperimenti di simulazione delle reti neurali.

Se queste prospettive dovessero concretizzarsi, il diritto e l'etica sarebbero chiamati a ridefinire nozioni fondamentali come identità, personalità giuridica e diritti postumi, ponendo interrogativi inediti sulla natura stessa dell'essere umano nell'era digitale (SISTO, 2024).

7. Alcune considerazioni conclusive.

L'analisi sin qui condotta evidenzia come la digitalizzazione abbia trasformato in modo radicale e irreversibile il rapporto tra individuo, memoria e tempo, producendo un vero e proprio 'mutamento antropologico'.

L'informazione, un tempo soggetta alla caducità dei supporti fisici e all'oblio naturale del ricordo umano, si è oggi smaterializzata e

distribuita in reti globali, ridondanti e resilienti, capaci di garantire una sopravvivenza potenzialmente illimitata dei dati.

Dalla storia dei supporti di memoria alla nascita delle tecnologie cloud e blockchain, passando per le architetture di backup distribuiti e per la crescente interconnessione tra infrastrutture digitali e geopolitica dei *data center* (ARESU, 2024), il mondo contemporaneo si trova di fronte a una memoria che non è più organica né selettiva, ma onnipresente e algoritmica, incardinata perfettamente nel quadro di controllo e di sorveglianza della società digitale attuale (ZUBOFF, 2023).

Questa trasformazione non è neutra, e pone sfide senza precedenti al diritto.

Il diritto all'oblio, concepito per restituire agli individui il potere di controllare la propria identità digitale, si scontra con la persistenza tecnica delle informazioni, con la natura distribuita di Internet e con la tensione tra privacy e libertà di informazione.

La comparazione tra ordinamenti nazionali — dall'Italia alla Francia, dalla Germania al Regno Unito, fino agli Stati Uniti d'America — mostra al giurista come la regolamentazione oscilli tra protezione della persona e tutela della memoria storica, senza giungere a un equilibrio condiviso.

La giurisprudenza europea e le decisioni della CEDU hanno introdotto criteri di proporzionalità e contestualità, ma l'effettiva applicazione del diritto all'oblio resta frammentaria e disomogenea, aggravata dalle difficoltà tecniche di cancellazione sicura e dalle diverse culture giuridiche che si contrappongono.

La questione si complica ulteriormente quando si affronta il tema dell'eredità digitale. La distinzione tra contenuti patrimoniali — come criptovalute, archivi professionali o domini web — e contenuti personali, quali corrispondenza privata, social network o archivi fotografici, rivela un vuoto normativo che lascia irrisolte le tensioni tra diritti successori, privacy *post-mortem* e autonomia contrattuale delle piattaforme.

L'assenza di una disciplina europea organica ha affidato la regolazione a decisioni giurisprudenziali isolate e a policy aziendali unilaterali — come le funzioni di Facebook, l'*Inactive Account Manager* di Google o il *Digital Legacy* di Apple — che, pur offrendo soluzioni operative, rischiano di *privatizzare* questioni che toccano diritti fondamentali.

Le proposte di policy internazionale indicano la necessità di un quadro normativo sovranazionale che disciplini in modo unitario la successione digitale, i testamenti digitali e il ruolo dei fiduciari digitali, integrando standard tecnici di cancellazione sicura, procedure di consenso esplicito in vita e garanzie per la tutela della dignità *post-mortem*.

Il coinvolgimento delle autorità di protezione dati, come l'EDPB e i Garanti nazionali, appare poi, a nostro avviso, cruciale per garantire un equilibrio tra innovazione, diritti fondamentali e responsabilità delle piattaforme tecnologiche.

A ciò si aggiunge la questione più complessa e, per certi versi, perturbante: l'uso dell'intelligenza artificiale per ricreare *simulazioni digitali di persone decedute*.

I casi reali di deepfake in film, pubblicità e documentari — dal ritorno digitale di attori iconici all'uso del *voice cloning* per finalità commerciali o artistiche — mostrano come la morte biologica non coincida più necessariamente con la scomparsa mediale o sociale dell'individuo (Sisto, 2025).

Ma accanto alle possibilità di commemorazione e di elaborazione del lutto, emergono rischi etici, psicologici e di sicurezza: dalle frodi basate sui *deepfake* alla manipolazione dell'opinione pubblica, dall'uso commerciale delle identità postume fino agli scenari più radicali di immortalità algoritmica e *mind uploading*, che prospettano la creazione di copie digitali permanenti della mente umana, con conseguenze difficilmente immaginabili per il diritto e la società.

In questo contesto, il quadro normativo attuale appare *inadeguato* e *frammentario*. Serve un intervento legislativo organico, possibilmente a livello europeo e con respiro internazionale, che disciplini in modo unitario il diritto all'oblio, l'eredità digitale, il consenso postumo e l'uso dell'intelligenza artificiale per la ricostruzione digitale delle identità, coniugando tutela della dignità umana, libertà fondamentali ed esigenze di innovazione tecnologica.

Al tempo stesso, è necessaria una riflessione etica e culturale interdisciplinare, che coinvolga non solo giuristi e informatici ma, anche, filosofi, sociologi, psicologi e studiosi dei media, per comprendere come le nuove forme di persistenza digitale e di simulazione *post-mortem* stiano ridefinendo la nozione stessa di vita, memoria e identità nel XXI secolo.

Solo attraverso un approccio integrato, capace di governare l'innovazione senza soffocarla, sarà possibile evitare che la memoria to-

tale del mondo digitale si trasformi in una prigione dell'informazione e dell'identità, compromettendo valori fondamentali come libertà, dignità e autodeterminazione.

Note bibliografiche

Aresu A. (2024), *Geopolitica dell'intelligenza artificiale*, Milano, Feltrinelli.

Bellomia V. (2020), *Diritto all'oblio e società dell'informazione*, Milano, CEDAM.

Bellotti R. (a cura di) (2019), *Il cloud computing nelle imprese e nella pubblica amministrazione. Profili giuridico-economici e applicazioni nel campo sanitario*, Milano, Giuffrè.

Capaccioli, S. (2023), *La tassazione delle attività digitali. Cripto-attività, criptovalute, valute virtuali, token e bitcoin*, Milano, Giuffrè.

Capaccioli S. (2015), *Cryptovalute e bitcoin. Un'analisi giuridica*, Milano, Giuffrè.

D'Arminio Monforte A. (2020), *La successione nel patrimonio digitale*, Pisa, Pacini.

De Francesco F. (2024), *Identità ed eredità al tempo della rivoluzione digitale*, Torino, Giappichelli.

Farina M. (2019), *Il cloud computing in ambito sanitario tra security e privacy*, Milano, Giuffrè.

Ferraris M. (2021), *Documanità. Filosofia del mondo nuovo*, Bari-Roma, Laterza.

Finocchiaro G. (a cura di) (2019), *La protezione dei dati in Italia*, Bologna, Zanichelli.

Gamba F. (2018), *Identità, memoria, immortalità. La sfida dei rituali digitali di commemorazione*, Milano, Mimesis.

Harbinja E. (a cura di) (2023), *Digital death, digital assets and post-mortem privacy*, Edinburgh, Edinburgh University Press.

Mayer-Schönberger V. (2016), *Delete. Il diritto all'oblio nell'era digitale*, Milano, EGEA.

Martinelli S. (2017), *Diritto all'oblio e motori di ricerca*, Milano, Giuffrè.

Mollo A.A. (2021), *Eredità digitale e piattaforme online. Tutela e profili di pianificazione*, Napoli, Editoriale Scientifica.

Petrella B. (2024), *Still online. Connessi oltre la morte: la nostra eredità digitale*, Milano, Piemme.

Pieranni S. (2023), *Tecnocina. Storia della tecnologia cinese dal 1949 a oggi*, Torino, ADD Editore.

Quintarelli S. (a cura di) (2020), *Intelligenza artificiale. Cos'è davvero, come funziona, che effetti avrà*, Torino, Bollati Boringhieri.

Russell S.J.-Norvig P. (2021), *Intelligenza artificiale. Un approccio moderno*, Milano, Pearson.

Sisto D. (2025), *Vivere per sempre. L'Aldilà ai tempi di ChatGPT*, Torino, Bollati Boringhieri.

Sisto D. (2024), *Virtual influencer. Il tempo delle vite digitali*, Torino, Einaudi.

SISTO D. (2022), *Porcospini digitali. Vivere e mai morire online*, Torino, Bollati Boringhieri.
SISTO D. (2020), *Ricordati di me. La rivoluzione digitale tra memoria e oblio*, Torino, Bollati Boringhieri.
SISTO D. (2018), *La morte si fa social. Immortalità, memoria e lutto nell'epoca della cultura digitale*, Torino, Bollati Boringhieri.
SUFFIA G. (2022), *Pulire l'infosfera. Intelligenza artificiale e contrasto alla disinformazione*, Milano, Giuffrè.
SUFFIA G. (2018), *Geografia delle cyberwars. Uomini e Stati alla prova dello spazio digitale*, Milano, Giuffrè.
VESTO A. (2020), *Successione digitale e circolazione dei beni online. Note in tema di eredità digitale*, Napoli, Edizioni Scientifiche Italiane.
ZICCARDI G. (2024), *Dati avvelenati. Truffe, virus informatici e falso online*, Milano, Raffaello Cortina.
ZICCARDI G. (2022), *Diritti digitali*, Milano, Raffaello Cortina.
ZICCARDI G. (2017), *Il libro digitale dei morti. Memoria, lutto, eternità e oblio dell'era del social network*, Milano, UTET.
ZUBOFF S. (2023), *Il capitalismo della sorveglianza. Il futuro dell'umanità nell'era dei nuovi poteri*, Roma, LUISS University Press.

INTELLIGENZA ARTIFICIALE
E RICONOSCIMENTO FACCIALE E BIOMETRICO

INTRODUZIONE

di *Gabrio Forti*

Nella presente sessione ci occuperemo di tecniche di riconoscimento facciale (TRF) ossia di tecnologie che trattano in modo automatizzato immagini digitali contenenti il volto di persone: da tecniche biometriche vengono ricavati i caratteri identificativi, riportati in forma di codici alfanumerici ed eventualmente arricchiti da indici ulteriori.

Tali tecniche possono essere utilizzate per varie finalità, Ad esempio: di autenticazione/verifica della persona; di identificazione, tramite l'individuazione di una corrispondenza (*match*) tra la fotografia e quelle già raccolte e contenute in un *database*; di rilevazione, ossia di individuazione dei volti, come può accadere utilizzando i filmati realizzati da telecamere a circuito chiuso, messi a confronto con le foto presenti in un *database* per trovare possibili corrispondenze.

Il relatore che tratterà del piano normativo avrà modo di illustrarci le disposizioni già esistenti e, in materia di tutela della privacy e protezione dei dati personali, le definizioni di cui all'art. 4 del Regolamento Generale sulla Protezione dei Dati (GDPR), nonché i divieti di cui all'art. 9 dello stesso Regolamento, salva una serie di eccezioni indicate nel prosieguo dell'articolo (consenso espresso dell'interessato, esercizio di diritti da parte dell'interessato, motivi di sicurezza sociale o protezione sociale, motivi di interesse pubblico in campo sanitario, ecc.). Ulteriore importante riferimento sono le norme a tutela della vita privata ai sensi dell'art. 8 della Convenzione Europea dei Diritti dell'Uomo.

Tuttavia, si è indubbiamente aperta anche su questo terreno, come del resto per tutta la intelligenza artificiale, quella che è stata detta «una nuova frontiera dei diritti civili», di cui nemmeno la Corte europea dei Diritti dell'Uomo nelle prime pronunce in tema (ad esempio il famoso caso *Glukhin c. Russia* del 2023) sembra aver preso piena coscienza. Certo non è campata in aria una preoccupazione diffusa di possibili scenari orwelliani da Grande Fratello.

Da docente di Criminologia, oltre che di Diritto penale, su questo argomento non posso che portare le sensibilità che nascono dal torvo retaggio lombrosiano da cui questa disciplina da decenni cerca di liberarsi e liberare. Sensibilità che conduce a un'estrema attenzione su quanto l'identificazione delle persone possa essere trascinata dal rilevamento di certi tratti anche somatici assunti o costruiti come predittivi di comportamenti criminali o comunque di pericolosità sociale, non solo sia spesso arbitraria e dettata da pregiudizi, ma porti con sé anche l'effetto di perpetuazione e consolidamento di tali visioni preconcette. Il che, oltre a violare fondamentali principi costituzionali e di civiltà, può risultare altamente criminogeno in forza delle dinamiche di transito dall''identità negativa' alla devianza secondaria assai note specie in psicologia sociale.

Rischi assai gravi, amplificati da possibili errori e *bias*, a monte e a valle. Si tratta di distorsioni cui possono andare soggetti i sistemi di riconoscimento facciale basati sull'intelligenza artificiale e che spesso pregiudicano le minoranze etniche e le persone portatrici di caratteristiche somatiche particolari, con conseguenti problemi legati alla discriminazione e ai profili di responsabilità per eventuali arresti o azioni coercitive ingiuste derivanti da processi di identificazione mal impostati o indirizzati.

Proprio in punto di identificazione, la prassi ha dato ampia prova di cosiddetti falsi positivi (il sistema individua come 'faccia' qualcosa che faccia non è) e falsi negativi (al sistema sfugge la rilevazione di un volto all'interno di un'immagine).

Quanto alla potenziale lesione dei diritti fondamentali, è noto come le tecnologie di riconoscimento facciale abbiano un più basso margine di accuratezza nel riconoscere i volti di persone di colore (soprattutto quelli delle donne) a causa di *bias* che possono inficiare già la fase di creazione degli algoritmi, oppure fasi successive, e che dipendono da come i dati e risultati vengono raccolti e gestiti da chi governa il procedimento.

Se e come rischi del genere si annidino in queste nuove TRF lo apprenderemo oggi da qualificatissimi relatori, tenendo conto peraltro della molteplicità di campi in cui di esse viene fatto uso, e comunque della innegabile efficacia nel contrasto e nella prevenzione del crimine o comunque di fenomeni indesiderati che può annettersi a queste tecnologie, con la conseguente forza attrattiva che esse esercitano sui decisori pubblici nel campo del controllo sociale.

E qui veniamo a un punto delicato, sul quale vorrei concludere questo mio breve preambolo. Riprendo una considerazione che non riguarda tanto gli abusi delle TRF, né gli scenari orwelliani cui questi possono condurre, ma un piano anche più sottilmente insidioso. Si è detto che « le TRF sembrano realizzare al massimo grado il processo di ''datificazione'' dell'individuo, ridotto a mero dato numerico; al medesimo tempo, si contribuisce alla costruzione di generi e categorie in base alle quali distinguere artificialmente il genere umano. In letteratura si sollecita dunque a interrogarsi su chi compia una siffatta catalogazione e su quali basi: una questione che diviene ancor più pressante allorché si pretenda di classificare le persone a partire dalle caratteristiche del proprio volto e che chiama in causa le imprese private impegnate nello sviluppo dei sistemi di intelligenza artificiale » (COLACURCI, 2022).

In una società del controllo, per dirla con Garland, o della sorveglianza, per dirla con Zuboff, il pericolo si riassume nella espressione appena citata: la datificazione dell'essere umano. L'evoluzione è già in atto e comporta uno stravolgimento della nostra visione dell'umano e, forse, della nostra stessa possibilità di mantenere fede alla promessa di rispetto della dignità, della libertà e dei diritti della persona su cui la nostra civiltà è stata costruita.

Solo fino a un certo punto può servire la pur giustissima raccomandazione, ribadita da Chiriatti e Riva in un recente articolo apparso sul « Sole-24 Ore », a proposito dell'intelligenza artificiale (ma che andrebbe estesa ai sistemi di riconoscimento facciale): la necessità di « dedicare maggiore attenzione al tema della fiducia nella intelligenza artificiale », ad esempio con « lo sviluppo di *framework* per la valutazione dell'affidabilità, della trasparenza e dei potenziali pregiudizi dei sistemi di intelligenza artificiale, nonché la definizione di linee guida per il suo utilizzo responsabile ed etico ».

Se può apparire certo promettente l'idea, avanzata in molti contesti, di utilizzare la tecnologia... per controllare la tecnologia, c'è da fronteggiare un gigantesco problema culturale e antropologico che nasce dal combinarsi di spinte concentriche. La c.d. datificazione dell'umano è sospinta potentemente da quel *bias* che ieri è stato ricordato ripetutamente, anche con riferimento al campo medico: l'*automation bias*, ossia la tentazione di affidarsi ciecamente alla macchina, sia per sottrarsi alla fatica dell'analisi, del *fine tuning* sulle spesso complesse situazioni effettuali, sia per il risparmio di costi che l'automazione generalmente comporta. Una 'scorciatoia cognitiva' che a me

pare si presenti come collettore di una serie di altri *bias* noti e studiati da tempo (ad esempio dal premio Nobel Kahneman), come quelli della disponibilità, della rappresentatività e, soprattutto, come il *confirmation bias* o 'euristica della conferma'.

Mi chiedo infatti se la stessa intelligenza artificiale, nella sua caratteristica di riproduzione del già esistente, non sia l'apoteosi di tale *bias* che reputo la 'madre', la copiosa sorgente di tutte le altre scorciatoie cognitive: quella che ci fa vedere nella realtà solo ciò che ci conferma, appunto, nelle nostre idee e percezioni pregresse e, ahimè, a volte nei nostri preconcetti e pregiudizi. Una euristica incoraggiata e potenziata massicciamente in ognuno dall'immensa macchina mediatica che ha ben capito quanto la meta del massimo *share* possa essere conseguita abitando stabilmente quello che Bourdieu chiamava il « dominio del senso comune »: non smuovendo, ma rinsaldando nel tele-utente, tutto ciò che aveva già in mente prima di sintonizzarsi sul canale (televisivo o *social*) preferito. Sarebbe impossibile isolare la comprensione e i necessari interventi nel campo dell'intelligenza artificiale, ivi compreso quello delle TRF, dal complessivo ecosistema mediatico di cui queste tecnologie sono parte integrante e in cui noi tutti siamo immersi e condizionati ben prima di aprire ChatGPT.

Note bibliografiche

Colacurci M. (2022), *Riconoscimento facciale e rischi per i diritti fondamentali alla luce delle dinamiche di relazione tra poteri pubblici, imprese e cittadini*, in *Sistema penale*, 9/2022.

Chiriatti M., Riva G., *Intelligenza artificiale, una rivoluzione cognitiva*, in *Il Sole-24 ore*, 27 ottobre 2024.

Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati) (Testo rilevante ai fini del SEE) GU L 119 del 4.5.2016, p. 1 ss.

(IL FATTO) INTELLIGENZA ARTIFICIALE E RICONOSCIMENTO FACCIALE: TRA FATTO, NORMA E INTERPRETAZIONE

di *Federico Cabitza* e *Luca Marconi*

Sommario: 1. Introduzione. — 2. Cornice operativa: funzione, contesto e tassonomia d'uso. — 3. La trascrizione simbolica e il modello tripartito. — 4. Faccia, viso e volto: distinzione operativa. — 5. Tassonomia funzionale dei sistemi biometrici. — 6. Dati biometrici e architetture tecnologiche. — 7. Dalle metriche all'affidabilità operativa. — 8. Diritto della prova tecnica e semiotica artificiale. — 9. Conclusioni.

1. Introduzione.

L'intelligenza artificiale applicata al riconoscimento facciale rappresenta in modo emblematico la dialettica tra fatto e norma. È produttrice di 'fatti' poiché traduce fenomeni percepibili in dati simbolici; ed è, più propriamente, riproduttrice di 'norme', poiché il suo impiego è oggi connesso a pratiche e decisioni di natura 'semi-automatica', integrate in un ordinamento che ne disciplina finalità, condizioni d'uso e responsabilità (Parlamento europeo e Consiglio, 2024). La nostra prospettiva non è di realismo ingenuo, né propriamente di realismo, bensì di materialismo simbolico (Sahlins, 1982): questi sistemi non producono fatti nel senso forte del termine, bensì evidenze tecniche con credibilità graduata (Simon, 1969). Come già ammoniva Vico, i fatti tecnici sono artefatti umani — ciò che abbiamo costruito — e, in quanto tali, rivedibili e governabili (Vico, 1999; Latour, 1987). Ne discende una prudenza non retorica, bensì un principio operativo: riconoscere che ogni misura automatizzata è, in ultima istanza, una rappresentazione costruita e, come tale, rivedibile (Floridi, 2019). Peraltro, come discusso da Marconi e Cabitza (2024), il rapporto tra potenza tecnica e responsabilità pubblica dell'IA va letto non come destino, ma come una pratica deliberata, e situata.

Sul piano concettuale, un sistema informatico può essere inteso come un trascrittore di simboli: elabora dati di ingresso (input) e produce dati in uscita (output) secondo regole e modelli che codificano vere e proprie scelte rappresentazionali (Simon, 1969; Floridi, 2019). Questo ci permette di passare dall'illusione di oggettività degli output alla tracciabilità di ciò che rappresentano, fosse anche una scelta, una decisione: ciò che il modello restituisce non è quindi un verdetto, ma un valore di similarità statistica o di confidenza probabilistica, cui l'ordinamento attribuisce significato in un contesto sociale e giuridico (Parlamento europeo e Consiglio, 2024). È precisamente in questo punto che fatto e norma si avvicinano fino a congiungersi: se l'output è probabilistico, la norma deve fissare le condizioni e i criteri per cui esso può fungere da base per l'azione (Parlamento europeo e Consiglio, 2024).

Per governare tale complessità, conviene adottare una lettura tripartita dei sistemi, articolata secondo le dimensioni di funzione, comportamento e struttura (Gero, 1990; Gero-Kannengiesser, 2004). La trascrizione simbolica fornisce la lente semiotica sulla natura rappresentazionale degli output; il modello tripartito offre la griglia di governo con cui renderli verificabili (Gero-Kannengiesser, 2004). La funzione, infatti, indica la finalità prevista, ossia lo scopo dichiarato per cui il sistema è progettato e posto in uso; il comportamento descrive ciò che il sistema effettivamente fa una volta collocato nel suo contesto d'uso, in relazione alle condizioni operative e agli esiti osservabili e le sue tracce percepibili, i dati; la struttura riguarda la configurazione tecnica e materiale che, una volta attivata, genera meccanicamente il comportamento osservabile. Mentre è importante che il comportamento lasci tracce, o registrazioni (Ferraris, 2021), è la funzione che le rende dati impiegati e oggetto di valutazione e ne consente il controllo di coerenza rispetto agli obiettivi ad essa intrinseci (Gero-Kannengiesser, 2004).

Questa cornice consente di congiungere il piano tecnico e quello giuridico in un unico piano normativo: la funzione definisce la destinazione d'uso; il comportamento crea gli effetti materiali (non necessariamente tangibili); la struttura è ciò che nella materia influenza la materia. Del resto, anche il Regolamento Europeo 2024/1689 (AI Act) adotta la medesima logica basata sulla valutazione del rischio: idoneità all'uso, tracciabilità, e gestione del rischio lungo il ciclo di vita diventano criteri di affidabilità, non meri adempimenti (Parlamento europeo e Consiglio, 2024; Schuett, 2024).

Ne segue che la norma non si contrappone alla tecnica, o meglio alla tecnologia, ma ne costituisce l'estensione riflessiva: traduce la prudenza epistemica in diligenza documentata, rendendo visibili finalità, limiti e condizioni di validità (Parlamento europeo e Consiglio, 2024; FLORIDI, 2019). Il percorso che segue approfondirà, lungo le tre dimensioni di funzione, comportamento e struttura, il modo in cui i sistemi di riconoscimento facciale siano valutati non per 'miracoli tecnici', ma per evidenze tracciabili, limiti dichiarati e responsabilità condivise.

2. Cornice operativa: funzione, contesto e tassonomia d'uso.

Governare in modo responsabile i sistemi di riconoscimento facciale significa ancorare ogni pretesa di prestazione alla finalità prevista e alle condizioni d'uso dichiarate, secondo l'architettura basata sul rischio delineata dall'AI Act (Parlamento europeo e Consiglio, 2024; SCHUETT, 2024). Per funzione, qui intendiamo la destinazione d'uso definita dal fornitore, comprensiva della specifica del contesto applicativo e dei vincoli operativi: è questa definizione positiva a generare requisiti di prova, oneri documentali e profili di responsabilità lungo l'intera filiera, dal progettista al *deployer* e utilizzatore (Parlamento europeo e Consiglio, 2024).

La funzione, lungi dall'essere un elemento ornamentale, costituisce il nesso che articola tecnica e diritto: orienta la classe di rischio, la qualità dei dati attesi, la tracciabilità dell'incertezza e le misure compensative in caso di errore (FLORIDI, 2019), e riflette negoziazioni stabili, o stabilizzate, normate, tra attori e contesti (BIJKER-HUGHES-PINCH, 1987).

La cornice operativa qui adottata si fonda sulla tripartizione funzione-comportamento-struttura, che costituisce una vera e propria grammatica di governo (GERO, 1990; GERO-KANNENGIESSER, 2004). La funzione fissa lo scopo (ad esempio: 'verifica 1:1 per controllo accessi in ambienti interni, illuminazione artificiale, soglia t coerente con FAR molto restrittivo e *fallback* umano'); il comportamento è ciò che il sistema realizza in esercizio (stabilità degli esiti nelle condizioni dichiarate, gestione dei casi borderline, tempi di risposta, log delle decisioni); la struttura rende possibile valutare la catena tecnico-rappresentazionale che conduce dall'immagine al simbolo (numerico), allo score (pipeline di *detection-alignment-feature*, dati e loro provenienza, architetture e funzione obiettivo). Questa tassonomia non sostituisce

gli standard: ne costituisce il ponte concettuale, consentendo di applicarli in modo proporzionato allo scopo (ISO/IEC 25010:2011).

In questo lessico, è utile distinguere due dimensioni: le cosiddette capacity del sistema e le capability umane. *Capacity* indica la proprietà funzionale dell'artefatto, ciò che il sistema è progettato per fare e riesce a fare nelle condizioni operative previste; *capability* riguarda invece abilità, competenze e pratiche degli operatori che con tale sistema interagiscono (Norman, 2013; Hollnagel, 2012). Nell'uso la capability si trasforma, si amplia, si adatta, in virtù di ciò che la capacity offre (affordance). Un artefatto può quindi esibire una *capacity* elevata in laboratorio e risultare, al contempo, non allineato rispetto a valori, aspettative, fini legittimi o a vincoli regolatori nel dominio d'uso. In tali casi, la *capability* umana diventa decisiva per mantenere visibile l'incertezza e tradurre punteggi probabilistici in scelte valoriali, e decisioni motivate (Reason, 1990; Vicente, 1999).

Un breve cenno semantico, rintracciabile nella tradizione latina (Isidoro Di Siviglia, *Etymologiae*, XI, 1, 137), aiuta a prevenire equivoci interpretativi. Il termine 'faccia' rinvia al substrato anatomico-biometrico; 'viso' alla sua apparenza situata nelle condizioni di acquisizione; 'volto' alla dimensione espressivo-intersoggettiva (Ekman-Friesen, 1978; Barrett, 2017). Nel dominio identificativo ha senso parlare di faccia/viso; quando invece si evocano emozioni o intenzioni si varca un livello inferenziale ulteriore (il volto), che richiede particolari cautele (Parlamento europeo e Consiglio, 2024, art. 5). Nel prosieguo adotteremo rigorosamente tale distinzione, rinviando al relativo box (§ 4).

Sul piano operativo, la tassonomia d'uso si articola in tre famiglie — verifica 1:1, identificazione 1:N, analisi/categorizzazione — distinte per funzione, profilo di rischio e oneri di prova (ISO/IEC 2382:2015; Parlamento europeo e Consiglio, 2024). La verifica 1:1 è tipicamente circoscritta e governabile tramite canali di *feedback* e soglie calibrate sui costi d'errore del caso d'uso; l'identificazione 1:N introduce una graduatoria di plausibilità su basi dati eterogenee e richiede, perciò, disaggregazioni, tracciabilità e sorveglianza rafforzata; l'analisi inferenziale, quando pretende di dedurre attributi non strettamente identificativi, richiede giustificazioni più robuste ed è oggetto di tipizzazioni specifiche nel quadro europeo (Parlamento europeo e Consiglio, 2024). I dettagli metrici (FAR/FRR/EER, soglie e compromessi decisionali) e le pratiche di valutazione saranno trattati in seguito; è sufficiente chiarire qui che la decisione operativa nasce a

valle, quando una soglia traduce uno *score* nel dominio dei numeri reali in azione, e che tale soglia deve essere motivata dallo scopo e dal contesto d'uso, non dall'apparente eleganza di un indice sintetico (Simon, 1969; Floridi, 2019).

Questa cornice abilita una responsabilità (*accountability*) distribuita, in cui ciascun attore della filiera risponde in modo specifico del proprio ruolo e delle proprie scelte. I fornitori sono tenuti a garantire la coerenza tra la funzione dichiarata e l'effettiva destinazione d'uso, la qualità e la provenienza dei dati, la completezza della documentazione tecnica e la chiarezza dei limiti operativi del sistema. Gli integratori assumono la responsabilità della messa in opera nel contesto applicativo, predisponendo interfacce che rendano visibile l'incertezza e garantendo regole d'ingaggio, procedure di escalation e forme di revisione umana significativa. Gli utilizzatori o decisori, infine, rispondono della conversione dello score in decisione, con adeguati oneri di motivazione e giustificazione, nel rispetto degli standard pertinenti (Parlamento europeo e Consiglio, 2024; ISO/IEC 42001:2023). Un'analoga articolazione emerge anche in domini diversi dal biometrico: in ambito educativo, ad esempio, i differenti ruoli dei sistemi di intelligenza artificiale influenzano la percezione di efficacia e di responsabilità nell'interazione (Marconi-Cabitza, 2025a), confermando l'importanza di una chiara definizione funzionale e di una supervisione significativa anche in contesti non critici.

Il perno comune resta la tracciabilità: collegare in modo verificabile funzione (finalità prevista o *intended purpose*), comportamento osservato in esercizio e struttura materiale consente di rendere contestabili soglie, limiti e misure compensative. È precisamente questo che l'AI Act istituzionalizza attraverso requisiti su data governance, documentazione tecnica, sorveglianza post-market e supervisione umana significativa (Parlamento europeo e Consiglio, 2024; Schuett, 2024).

Definire con precisione la funzione e il contesto d'uso, interpretare il comportamento del sistema alla luce dei costi e delle ricadute delle decisioni che ne derivano, e documentare la struttura tecnica perché resti oggetto di verifica e ispezione (*audit*): questa è la bussola per ricondurre l'impiego del riconoscimento facciale dal mito dell'accuratezza e dell'oggettività alla adeguatezza e appropriatezza all'uso, ancorando la valutazione non ad un ideale di perfezione algoritmica, ma alla coerenza con lo scopo, alla proporzionalità dei mezzi e al ri-

spetto effettivo di diritti e responsabilità (ISO/IEC 25010:2011; Parlamento europeo e Consiglio, 2024).

3. La trascrizione simbolica e il modello tripartito.

Ogni sistema di riconoscimento facciale si fonda su un processo di trascrizione simbolica, ossia su una sequenza di operazioni che convertono dati sensoriali in rappresentazioni computabili e, infine, in decisioni o valutazioni dotate di valore operativo. Questa catena di traduzioni — dall'immagine grezza al vettore numerico, dallo score di similarità alla indicazione categorica, la cosiddetta etichetta (*label*) — non rappresenta un atto neutrale di misura, ma una costruzione interpretativa che incorpora scelte modellistiche e convenzioni progettuali (Simon, 1969; Floridi, 2019). L''output' non è dunque un fatto, bensì un simbolo (o una sequenza di simboli) che solo in un secondo momento è tradotto in un giudizio di natura dicotomica, ad esempio '*match/non match*'.

Comprendere questa dinamica significa spostare lo sguardo dall'illusione di un'oggettività intrinseca alla tracciabilità delle decisioni tecniche: ciò che conta non è tanto la risposta del modello, quanto il percorso che conduce a quella risposta e le condizioni entro cui essa può essere considerata affidabile (Gama-Žliobaitė-Bifet-Pechenizkiy-Bouchachia, 2014). La nostra tesi, oramai chiara, è che per rendere analiticamente visibile tale percorso, conviene leggere ogni sistema secondo il modello tripartito di funzione, comportamento e struttura (Gero, 1990; Gero-Kannengiesser, 2004) che ora possiamo esemplificare più precisamente. La funzione definisce la destinazione d'uso e gli obiettivi dichiarati: non ciò che il sistema potrebbe fare in astratto, ma ciò che è autorizzato e documentato a realizzare in un determinato contesto, con soglie e criteri coerenti con la finalità legittima (Parlamento europeo e Consiglio, 2024). Un sistema progettato per la verifica 1:1, ad esempio, non può essere automaticamente ritenuto idoneo a compiti di identificazione 1:N o di analisi inferenziale, poiché mutano sia il livello di rischio sia le condizioni di validità probatoria (ISO/IEC 2382:2015; ISO/IEC 25010:2011).

Il comportamento riguarda invece il modo in cui il sistema agisce nel concreto: la stabilità degli esiti, la sensibilità alle variazioni di luce, posa, caratteri fisiologici ricorrenti in una certa etnia o in una certa fascia d'età, la gestione dei casi borderline, la documentazione (registrazione) degli errori e delle revisioni. È nel comportamento che

emergono fenomeni come il *dataset shift* e il *concept drift*, i quali mettono in discussione la affidabilità (nel duplice senso di *reliability* e di *trustworthiness*) e generalizzabilità dei risultati ottenuti in laboratorio e impongono piani di monitoraggio continuo, come previsto dalle norme tecniche e regolatorie (Gama et al., 2014; ISO/IEC 23894:2023), un regime che altrove abbiamo chiamato di tecno-vigilanza (Cabitza-Zeitoun, 2019).

La struttura, infine, concerne l'architettura interna in cui configurazioni materiali-energetiche rappresentano i dati: è questa che acquisisce immagini, le elabora, e le trasforma, applicando criteri a soglia in ciò che per gli esseri umani ha il significato di un verdetto binario (Buolamwini-Gebru, 2018; ISO/IEC 19795-1:2021). In questo livello la trascrizione simbolica diviene cruciale: gli *embedding* non sono immagini del reale, ma rappresentazioni progettate per ottimizzare una funzione obiettivo. Comprendere e documentare tali scelte equivale a rendere giudicabile la validità e i limiti sia gnoseologici che epistemologici del sistema (Simon, 1969; Floridi, 2019).

La triade funzione-comportamento-struttura permette dunque di collegare i tre livelli dell'analisi: la dichiarazione d'intenti, l'osservazione empirica e la spiegazione tecnico-rappresentazionale. In prospettiva normativa, essa fornisce la grammatica comune per la valutazione dell'affidabilità richiesta dall'AI Act, consentendo di verificare se un sistema opera in conformità al proprio fine, se mantiene prestazioni coerenti nel tempo e se le sue decisioni restano valutabili e riproducibili (Parlamento europeo e Consiglio, 2024). Nel linguaggio ingegneristico, tale articolazione segna il passaggio dalla prestazione misurata all'idoneità all'uso, ossia alla capacità di mantenere risultati congruenti con le finalità e i limiti dichiarati (ISO/IEC 25010:2011; ISO/IEC 23894:2023).

Riconoscere la natura interpretativa della trascrizione simbolica e collocarla all'interno del modello tripartito consente di superare l'ossessione per l'accuratezza e di orientare il discorso verso il governo della responsabilità tecnica. La questione non è più se l'algoritmo 'riconosca correttamente', ma se lo faccia nelle condizioni per cui è stato progettato, con prestazioni tracciabili e margini d'incertezza comunicabili. In questa prospettiva, l'interazione fra fatto e norma si trasforma in una pratica di conoscenza condivisa e verificabile, in cui la trasparenza del processo diventa parte integrante della legittimità stessa del sistema.

4. Faccia, viso e volto: distinzione operativa.

Nel linguaggio comune i termini *faccia*, *viso* e *volto* tendono a sovrapporsi; tuttavia, nel dominio tecnico e giuridico del riconoscimento biometrico, tale distinzione risulta decisiva. La *faccia* rappresenta il livello anatomico e misurabile: è il supporto fisico da cui il sistema estrae schemi geometrici o distribuzioni di intensità luminosa, traducendoli in rappresentazioni simboliche numeriche (Ekman-Friesen, 1978). Il *viso*, invece, è la faccia in quanto appare, situata in un contesto percettivo e relazionale: include fattori come l'illuminazione, la posa e l'espressione contingente, che influenzano la qualità dei dati e la loro stabilità temporale (Barrett, 2017). Il *volto*, infine, introduce una dimensione simbolica e intersoggettiva: manifesta l'identità personale, la storia e l'intenzionalità del soggetto, costituendo — secondo la tradizione fenomenologica e teologica — il luogo in cui si riflette la persona (Isidoro Di Siviglia, *Etymologiae*, XI, 1, 137; Gadamer, 1960; Levinas, 1961).

Questa tripartizione non ha soltanto valore semantico, ma funge da bussola di natura semeiotica, di interpretazione e comprensione di ciò che ci appare e ci viene, appunto, *dato*. Nel dominio biometrico, infatti, l'operatività legittima si limita ai livelli della faccia e del viso: il sistema misura forme e apparenze, non significati interiori. Varcare la soglia del volto, attribuendo al dato visivo tratti emotivi o intenzionali, significa oltrepassare il confine dell'osservabile e sconfinare nel dominio dell'inferenza interpretativa (Barrett, 2017; Floridi, 2019). È in questa transizione che si manifesta il rischio di abuso — o, più precisamente, di *misuse* — quando la tecnologia è investita di un potere classificatorio, spesso di natura diagnostica o psicologica, che non le compete (Simon, 1969). Ne abbiamo parlato estesamente in alcuni lavori a cui rimandiamo (Cabitza et al., 2022; Mattioli-Cabitza, 2024). Non a caso, l'AI Act vieta, salvo eccezioni specifiche, l'uso dell'analisi delle emozioni nei contesti lavorativi e educativi, richiamando implicitamente la distinzione epistemica tra viso e volto: ciò che eccede l'osservabile non può costituire il fondamento di una decisione proporzionata né, tantomeno, di una decisione giuridicamente legittima (Parlamento europeo e Consiglio, 2024, art. 5).

Distinguere tra faccia, viso e volto consente dunque di prevenire confusioni concettuali e derive semantiche. La *faccia* (*face*) delimita ciò che è misurabile; il *viso* accoglie in sé ciò che è relativo al contesto — ciò che, per così dire, 'circonda il testo' simbolico (*cum-*

textus) e ne costituisce una controparte non commensurabile; il *volto*, infine, designa ciò che è relazionale, intenzionale e personale e, per questa ragione, né rappresentabile né computabile. L'uso responsabile dei sistemi di riconoscimento facciale implica il mantenimento dell'operatività entro i primi due livelli, documentando con rigore le condizioni di acquisizione e interpretando i risultati come rappresentazioni di un'apparenza, cioè di qualcosa che appare e si manifesta pur in una ricostruzione molto parziale del contesto esperienziale, e non come rivelazioni di un'identità o verità interiore.

In questa prospettiva, la distinzione non è una mera finezza linguistica, ma una condizione di giustificazione epistemica e di proporzionalità normativa: soltanto chiarendo che cosa si misura (e, parimenti, che cosa non si misura) è possibile stabilire quale tipo di 'fatti' tecnici il sistema pretenda di offrire, e quali limiti debba rispettare per non violare i diritti fondamentali del soggetto rappresentato, in particolare la sua dignità e libertà (Parlamento europeo e Consiglio, 2024).

5. Tassonomia funzionale dei sistemi biometrici.

I sistemi di riconoscimento facciale si articolano in tre funzioni operative fondamentali, differenziate per struttura tecnica, destinazione d'uso e livello di rischio associato. Questa tassonomia, ormai consolidata nella letteratura tecnico-giuridica e nelle normative internazionali, consente di definire con precisione il dominio di validità delle prestazioni e di prevenire indebite estensioni funzionali (ISO/IEC 2382:2015; ISO/IEC 19795-1:2021; Parlamento europeo e Consiglio, 2024).

La verifica 1:1 rappresenta la forma più circoscritta e controllabile: confronta un'acquisizione facciale con un singolo modello di riferimento per confermare o negare una corrispondenza dichiarata, rispondendo in sostanza alle domande 'chi sei?' o 'sei chi dici di essere?'. Si tratta di un meccanismo di autenticazione biometrica in cui la soglia decisionale può essere calibrata in modo rigoroso e l'intervento umano può fungere da verifica di *fallback* (ISO/IEC 19795-1:2021), cioè intervenire quando il sistema non raggiunge una confidenza sufficiente, al fine di confermare o correggere l'esito prima che diventi decisione operativa. La verifica opera in condizioni generalmente note e ripetibili (illuminazione, posa, distanza) e la sua affidabilità dipende dalla qualità della registrazione iniziale e dalla coerenza

del canale di acquisizione (ISO/IEC 30107-3:2017). Il rischio principale riguarda l'errore di rifiuto dell'utente legittimo o, in misura minore, l'accettazione di un falso positivo; entrambi, tuttavia, risultano misurabili e governabili entro parametri operativi definiti (§ 7).

L'identificazione 1:N introduce un livello di complessità superiore, poiché confronta l'immagine di un soggetto con un'intera base di dati di profili noti, producendo una graduatoria di similarità. In questo caso la decisione non è più binaria ma graduata, e la probabilità di errore aumenta con la dimensione e l'eterogeneità della base di dati (Grother-Ngan-Hanaoka, 2019). L'uso di soglie uniche su popolazioni eterogenee può generare distorsioni e *bias* demografici, che devono essere compensati attraverso metriche disaggregate e procedure di revisione indipendente (Buolamwini-Gebru, 2018). L'identificazione costituisce, pertanto, il momento in cui la trascrizione simbolica risulta più sensibile alle scelte di soglia e ai criteri di comparazione: ogni passaggio della pipeline — dall'acquisizione al calcolo dello score — contribuisce al margine d'incertezza, che deve essere documentato e gestito lungo tutto il ciclo di vita del sistema (ISO/IEC 23894:2023; Floridi, 2019).

L'analisi inferenziale, infine, si colloca sul confine più problematico, poiché mira a estrarre dal volto informazioni che oltrepassano l'identità: emozioni, stati d'animo, tratti caratteriali o inclinazioni comportamentali (Barrett, 2017). Si tratta di un livello di astrazione che non ha più per oggetto la *faccia* o il *viso*, ma il *volto* in senso interpretativo, esponendo così il sistema a rischi epistemici e normativi significativi (Simon, 1969; Cabitza et al., 2022). Queste pratiche non dispongono di una validazione scientifica solida e sono esplicitamente soggette a divieti o restrizioni previste dall'AI Act e dalla Convenzione del Consiglio d'Europa del 2024, in particolare nei contesti lavorativi, educativi e di sorveglianza pubblica (Parlamento europeo e Consiglio, 2024; Consiglio d'Europa, 2024). Quando impiegate a fini di ricerca o di assistenza, richiedono giustificazioni esplicite sulla natura e sui limiti dei dati, un consenso informato effettivo e una chiara distinzione tra inferenza statistica e attribuzione personale.

La classificazione dei sistemi biometrici in verifica 1:1, identificazione 1:N e analisi inferenziale consente di ancorare la valutazione tecnica al contesto d'uso, stabilendo tre soglie di esposizione al rischio: affidabilità operativa nel primo caso, robustezza comparativa nel secondo, legittimità epistemica nel terzo. Mantenere rigorosamente tale distinzione è condizione necessaria per applicare in modo

coerente i principi di proporzionalità, minimizzazione e tracciabilità previsti dal quadro europeo per i sistemi biometrici di riconoscimento facciale (ISO/IEC 42001:2023; Parlamento europeo e Consiglio, 2024).

6. Dati biometrici e architetture tecnologiche.

Per comprendere che cosa un sistema di riconoscimento facciale misuri effettivamente, è necessario chiarire la natura del dato biometrico. L'immagine del volto costituisce, di per sé, un dato personale grezzo; tuttavia, il riconoscimento non opera direttamente su questa immagine, ma su una rappresentazione derivata — il cosiddetto *template* o *embedding* —, ossia un vettore di caratteristiche numeriche che codifica la distribuzione spaziale o statistica dei tratti ritenuti rilevanti (Jain-Ross-Nandakumar, 2011; ISO/IEC 19795-1:2021). In questo senso, il dato biometrico non è mai una fotografia del reale, ma il risultato di una trasformazione interpretativa che seleziona determinate caratteristiche (*features*) — come posa, illuminazione o espressione — e scarta ciò che il modello non considera informativo per la funzione assegnata (Zhao-Chellappa-Phillips-Rosenfeld, 2003; Wang-Deng, 2021). È dunque una rappresentazione costruita a fini comparativi, la cui qualità dipende non dalla resa ottica dell'immagine originaria, ma dal modo in cui il sistema traduce la *faccia* in numeri (Floridi, 2019) e dal risultato finale (*outcome*).

Questa traduzione avviene attraverso una *pipeline* articolata in quattro fasi principali: rilevamento (*detection*, non 'identificazione'), allineamento (*alignment*), estrazione delle caratteristiche (*feature extraction*) e associazione (*matching*), la fase per cui anche i termini riconoscimento o identificazione sono altrettanto leciti. Nella fase di *detection*, il sistema individua e isola il volto all'interno dell'immagine; nella fase di *alignment*, ne normalizza l'orientamento e la scala, riducendo le variazioni di posa; nella *feature extraction* ricava una rappresentazione numerica compatta — un vettore ad alta dimensionalità — che cattura le relazioni interne tra i tratti significativi; infine, nella fase di *matching*, confronta due vettori calcolando una misura di similarità, che verrà successivamente valutata rispetto a una soglia decisionale per generare un esito operativo, appunto il riconoscimento o la identificazione del soggetto (Zhao et al., 2003; ISO/IEC 19795-1:2021). È lungo questa catena che si concentrano le principali fonti di incertezza: la scelta delle *feature*, la metrica di distanza e la

gestione delle soglie determinano la sensibilità e la robustezza complessiva del sistema. Le metriche che quantificano tali aspetti, insieme alle procedure di valutazione fuori benchmark, saranno discusse nei paragrafi successivi (§ 7) (ISO/IEC 23894:2023).

Dal punto di vista architetturale, l'evoluzione dei modelli può essere ricondotta a tre fasi principali. I primi approcci, di tipo geometrico, si fondavano sulla misurazione di distanze e angoli tra punti chiave del volto (*landmarks*), secondo logiche ispirate alla fisionomia classica (JAIN-ROSS-NANDAKUMAR, 2011). A tali approcci sono seguiti i metodi basati sull'apparenza, come *Eigenfaces* e *Fisherfaces*, che rappresentavano il volto come una combinazione di componenti statistiche e sfruttavano la variazione globale dei pixel per il riconoscimento (TURK-PENTLAND, 1991; BELHUMEUR-HESPANHA-KRIEGMAN, 1997). La svolta è avvenuta con il *deep learning*, che ha introdotto architetture neurali capaci di apprendere automaticamente rappresentazioni ottimali per la funzione di similarità: reti convoluzionali come *DeepFace* (TAIGMAN-YANG-RANZATO-WOLF, 2014), *FACENET* (SCHROFF-KALENICHENKO-PHILBIN, 2015) e modelli successivi basati su embedding normalizzati e margini angolari (DENG-GUO-XUE-ZAFEIRIOU, 2019; LECUN-BENGIO-HINTON, 2015; GOODFELLOW-BENGIO-COURVILLE, 2016). Queste architetture non si limitano a estrarre *features*, ma apprendono una vera e propria geometria dello spazio facciale, in cui la distanza tra vettori corrisponde a una nozione operativa di identità (SCHROFF et al., 2015; DENG et al., 2019).

La potenza rappresentazionale del *deep learning*, tuttavia, amplifica anche la vulnerabilità ai cambiamenti di dominio e ai *bias* presenti nei dataset di addestramento, rendendo necessarie valutazioni continue di generalizzabilità e una sorveglianza post-market sistematica (§ 7) (WANG-DENG, 2021; ISO/IEC 23894:2023). Riconoscere la natura costruttiva e inferenziale — in senso computazionale — del dato biometrico significa, in ultima analisi, spostare il problema dal *quanto il sistema vede* al *come il sistema codifica ciò che vede*: un passaggio necessario per attribuire significato, responsabilità e limiti a ogni misura automatizzata (FLORIDI, 2019).

7. Dalle metriche all'affidabilità operativa.

L'accuratezza, spesso invocata come criterio assoluto di eccellenza tecnica, offre in realtà solo una rappresentazione parziale del comportamento di un sistema di riconoscimento facciale. La valuta-

zione dell'affidabilità non può ridursi a un singolo indice di prestazione: deve considerare il modo in cui le soglie sono definite, le popolazioni su cui il sistema è testato e la stabilità dei risultati nel tempo (ISO/IEC 19795-1:2021; ISO/IEC 25010:2011). L'obiettivo non è raggiungere una precisione ideale, ma dimostrare coerenza funzionale rispetto allo scopo dichiarato e alle condizioni d'uso reali, secondo la logica del *fitness for purpose* (adeguatezza allo scopo) introdotta dall'AI Act e dai più recenti standard internazionali (Parlamento europeo e Consiglio, 2024).

La prima dimensione concerne il rapporto tra soglia e compromesso. Ogni sistema di riconoscimento traduce un valore continuo di similarità in una decisione binaria, introducendo un equilibrio strutturale tra il rischio di *false acceptance* (FAR) e quello di *false rejection* (FRR) (ISO/IEC 19795-1:2021). La cosiddetta *Equal Error Rate* (EER) — il punto in cui i due errori si equivalgono — costituisce un utile riferimento per la comparazione analitica tra modelli, ma non rappresenta un obiettivo operativo: la soglia effettiva deve essere calibrata in funzione del contesto d'uso, dei costi associati all'errore (non solo di carattere puramente economico) e delle procedure di controllo umano disponibili e impiegabili. Un sistema di accesso a un laboratorio sensibile può tollerare un tasso relativamente elevato di falsi rifiuti (*false rejections*) pur di ridurre il rischio di falsi accettamenti (*false acceptances*) e garantire un livello di sicurezza più elevato. Al contrario, un sistema destinato all'identificazione di soggetti sconosciuti in spazi pubblici necessita di soglie decisionali meno restrittive, volte a massimizzare la sensibilità, nonché della presenza di canali di revisione esterni per il controllo umano dei risultati (ISO/IEC 23894:2023).

La seconda dimensione riguarda l'equità (*fairness*) e disaggregazione delle prestazioni. L'esperienza empirica ha mostrato che gli indici aggregati di accuratezza spesso mascherano differenze significative tra gruppi demografici. Gli studi *Gender Shades* (Buolamwini-Gebru, 2018) e le valutazioni indipendenti del *Face Recognition Vendor Test* (Frvt, Grother-Ngan-Hanaoka, 2019) hanno evidenziato differenze prestazionali fino a un ordine di grandezza tra sottopopolazioni diverse per età, genere o origine etnica. Tali divergenze derivano, in larga parte, dalla composizione sbilanciata dei dataset di addestramento e dall'assenza di test sistematici su scenari d'uso differenziati e controfattuali (Wang-Deng, 2021). Per questa ragione, il Regolamento europeo impone che la qualità e la rappresentatività dei

dati siano documentate in relazione all'uso previsto (art. 10 AI Act) e che le metriche di valutazione includano misure disaggregate per strati (sotto-gruppi) diversi della popolazione di riferimento e opportunamente verificabili (Parlamento europeo e Consiglio, 2024). La performance non può più essere rappresentata da un singolo valore numerico — uno *score* —, ma da una mappa di affidabilità espressa attraverso un insieme di indicatori e indicazioni delle relative stime di incertezza, la cui significatività dipende dalla varietà delle condizioni testate e dalla trasparenza con cui i risultati vengono comunicati. La pubblicazione di *model cards* e *datasheets for datasets* contribuisce a questa trasparenza, almeno in linea teorica, consentendo un controllo diffuso e una comparazione più equa tra modelli, soprattutto se ci si conforma a modelli completi e a metodologie solide di rilevazione (O'Neil, 2016; Mitchell et al., 2019; Gebru-Morgenstern-Vecchione-Wortman Vaughan-Wallach, 2021).

La terza dimensione dell'affidabilità concerne la stabilità nel tempo. Anche un sistema accurato al momento del rilascio può degradare rapidamente se le condizioni ambientali o demografiche cambiano. Fenomeni come *dataset shift* e *concept drift* — derive legate all'evoluzione nel tempo dei fenomeni rappresentati — alterano la distribuzione statistica dei dati e compromettono la validità delle soglie apprese, rendendo necessaria una sorveglianza post-market continua (Gama et al., 2014; ISO/IEC 23894:2023) e un approccio che è stato chiamato di tecno-vigilanza (Cabitza-Zeitoun, 2019). L'AI Act recepisce questo principio introducendo l'obbligo di monitoraggio e controllo delle versioni (*change control*), in linea con gli standard ISO/IEC 25010:2011 e 23894:2023, che descrivono l'affidabilità come proprietà dinamica del sistema. La gestione della qualità richiede dunque piani di revisione periodica, indicatori di deriva e procedure di riallineamento documentate, affinché le prestazioni dichiarate restino rappresentative delle condizioni operative effettive. Come osservano Marconi e Cabitza (2025b), la robustezza e la gestione dell'incertezza rappresentano elementi centrali di una responsabilità realmente informata, capace di superare la logica della performance isolata e di collegare la valutazione dell'affidabilità alla trasparenza, da declinarsi tanto nei termini di completezza delle metriche e spiegabilità operativa degli esiti (Longo et al., 2024). In assenza di tali pratiche, l'accuratezza nominale perde significato, e la conformità formale non garantisce più alcuna affidabilità sostanziale (Floridi, 2019).

Il passaggio dall'accuratezza all'affidabilità operativa segna un cambio di mentalità, e potremmo dire, anche di paradigma valutativo: dal culto del numero unico alla responsabilità condivisa della misurazione e della sua interpretazione, che deve sempre collocarsi in un contesto socio-tecnico orientato a uno scopo. Un sistema di riconoscimento facciale è affidabile non perché raggiunge un certo valore di EER, ma perché dimostra, nel tempo e nei diversi contesti, di mantenere prestazioni adeguate, coerenti, giustificabili e documentate. In questa prospettiva, l'evidenza tecnica assume valore giuridico, trasformandosi da mero output statistico in prova ragionevole e proporzionata (SIMON, 1969; Parlamento europeo e Consiglio, 2024).

8. Diritto della prova tecnica e semiotica artificiale.

Il diritto contemporaneo dell'intelligenza artificiale si configura sempre più come un diritto della prova tecnica, fondato sulla tracciabilità delle evidenze e sulla responsabilità delle interpretazioni automatizzate. Nel riconoscimento facciale, in particolare, la convergenza tra disciplina dei dati, regolazione del rischio e tutela dei diritti fondamentali mostra come la norma non costituisca un vincolo esterno alla tecnica, ma la forma istituzionale della prudenza epistemica (SIMON, 1969; FLORIDI, 2019) o del principio di precauzione invocato dall'etica delle conseguenze (FLORIDI-CABITZA, 2021). La Carta dei diritti fondamentali dell'Unione europea, agli articoli 7 e 8, tutela congiuntamente la vita privata e la protezione dei dati personali, fissando i principi di liceità, finalità specifica, proporzionalità e controllo indipendente (Carta UE, 2000). L'AI Act traduce tali principi in una struttura operativa: impone la definizione della finalità prevista (*intended purpose*, artt. 3 e 12), introduce divieti per pratiche inaccettabili (art. 5) e stabilisce obblighi graduati di gestione del rischio per i sistemi ad alto impatto (artt. 6 e 10), prescrivendo una data governance verificabile e una documentazione tecnica capace di rendere visibili metriche, soglie e limiti d'uso (Parlamento europeo e Consiglio, 2024). A livello sovranazionale, la Convenzione quadro del Consiglio d'Europa sull'intelligenza artificiale, i diritti umani, la democrazia e lo Stato di diritto (2024) completa il quadro, imponendo agli Stati di garantire accountability lungo l'intero ciclo di vita, prevenzione delle discriminazioni, valutazioni d'impatto sui diritti e meccanismi di rimedio effettivo (Consiglio d'Europa, 2024). Ne emerge una visione multilivello in cui la conformità non coincide con la mera osservanza

formale, ma con la capacità di dimostrare l'idoneità all'uso dichiarato, la tracciabilità delle decisioni e la proporzionalità degli effetti (ISO/IEC 23894:2023; ISO/IEC 25010:2011).

All'interno di questa architettura, il principio di proporzionalità regola l'intero spettro degli impieghi biometrici. Il legislatore europeo ha scelto di tipizzare i divieti nei casi in cui il rischio epistemico si traduce in un rischio diretto per i diritti fondamentali dei soggetti coinvolti: l'identificazione biometrica remota in tempo reale in spazi pubblici; lo *scraping* (recupero) massivo di immagini facciali da internet o da reti di videosorveglianza per costruire database non consensuali; e l'inferenza di emozioni in contesti lavorativi o educativi (Parlamento europeo e Consiglio, 2024, art. 5). In tali ambiti, l'incertezza scientifica e la vulnerabilità dei soggetti coinvolti rendono l'uso intrinsecamente sproporzionato, giustificando un divieto ex ante o un regime di autorizzazione eccezionale (Consiglio d'Europa, 2024). La regola è netta: quando l'interpretazione automatizzata eccede la validità empirica della prova o la verificabilità pubblica del processo che la genera, il limite giuridico deve essere applicato senza deroghe. Il divieto non ha natura punitiva, ma funzione epistemico-correttiva: serve a impedire che il numero assuma il peso di un verdetto quando la catena interpretativa che lo sostiene non è ricostruibile né falsificabile (Floridi, 2019; Mittelstadt, 2019; Wachter-Mittelstadt-Russell, 2021). Per la stessa ragione, la distinzione tra faccia, viso e volto assume valore giuridico: parlare di 'volto' quando il sistema elabora *embedding* del viso implica una pretesa interpretativa che eccede la sua competenza tecnica, e che la norma circoscrive per proteggere dignità e libertà individuale (Parlamento europeo e Consiglio, 2024; Consiglio d'Europa, 2024).

Accanto ai divieti, la regolazione europea costruisce la responsabilità lungo la filiera come strumento di garanzia sostanziale. Il sistema tecnico può essere letto come un testo a più mani, la cui costruzione e l'interpretazione coinvolgono diversi attori lungo la filiera: chi progetta risponde della genesi del modello — qualità e rappresentatività dei dati, funzione obiettivo, documentazione delle soglie e dei limiti; chi integra o impiega il sistema risponde della sua messa in scena applicativa — contesto, interfacce, parametri di rischio, revisione umana significativa; chi decide o controlla risponde dell'uso giuridico dell'evidenza — motivazione, proporzionalità, possibilità di riesame. La filiera si traduce così in una vera e propria semiosi (processo di interpretazione dei segni) artificiale: un insieme di criteri

condivisi per comprendere i modelli come testi tecnici dotati di una genesi, una struttura e un uso, ciascuno dei quali deve essere documentato e impugnabile, discutibile, contestabile (Floridi, 2019; ISO/IEC 23894:2023). Strumenti come i *datasheets for datasets*, le *model cards* e i *valuation reports* svolgono la funzione di note critiche, che consentono all'interprete — tecnico, giurista o auditor — di ricostruire il significato e la forza probatoria di un output numerico (Gebru et al., 2021; Mitchell et al., 2019). In questa direzione, il framework CLARITY AI (Marconi-Pirovano-Cabitza, 2024) rappresenta, ad esempio, una sintesi sistematica dei criteri di qualità e trasparenza nella ricerca in AI medica, capace di tradurre le metriche tecniche in indicatori di affidabilità epistemica e documentale.

Il requisito di supervisione umana (*oversight*) significativa, previsto dall'AI Act e ribadito dalla Convenzione del Consiglio d'Europa, chiude il cerchio: la presenza umana non è ornamentale, ma consiste nella capacità effettiva di comprendere i limiti del sistema, sospenderne eventualmente l'uso, e attivare rimedi quando l'incertezza o l'impatto lo richiedano (Parlamento europeo e Consiglio, 2024; Consiglio d'Europa, 2024).

Un esempio empirico di tale intreccio tra dimensione tecnica e responsabilità interpretativa è fornito dallo studio di Mattioli e Cabitza (Mattioli-Cabitza, 2024), che analizza il riconoscimento facciale come caso paradigmatico di 'inaffidabilità strutturale' nei sistemi di IA. Lo studio mostra come, nel dominio del *facial emotion recognition (FER)*, l'automazione amplifichi l'incertezza umana invece di ridurla, trasformando etichette soggettive in presunte verità statistiche. La debolezza epistemica dei dataset di addestramento — nei quali le emozioni vengono codificate da pochi annotatori secondo convenzioni culturalmente situate e giudizi caratterizzati da una variabilità maggiore di quella che si ritiene accettabile in altri contesti di *ground truthing* (cioè di processi di annotazione dei casi di esempio che sono utilizzati per l'addestramento di modelli inferenziali predittivi e classificatori) — dimostra che la 'misura' prodotta dal modello non è un *fatto* ma un'interpretazione automatizzata, priva di stabilità semantica e di validità replicabile. Tale caso studio conferma la necessità di una semiosi artificiale capace di rendere visibili le catene interpretative incorporate nei sistemi di IA e di ricondurre la valutazione dell'affidabilità a criteri di giustificazione epistemica e di responsabilità pubblica.

In questa prospettiva, il diritto della prova tecnica e la semiotica artificiale non sono ambiti distinti, ma due modalità complementari

di un medesimo principio: l'evidenza automatizzata deve poter essere letta, contestata e corretta. Le norme di responsabilità e gli standard di qualità (ISO/IEC 25010:2011; ISO/IEC 23894:2023) forniscono la grammatica di questa lettura, mentre gli strumenti documentali ne rappresentano la sintassi operativa. Quando il modello è trattato come un testo aperto, la decisione automatizzata torna a essere argomento, non dogma: un punto di partenza per la deliberazione pubblica, non la sua chiusura. In tal modo, la promessa tecnica dell'intelligenza artificiale trova la propria legittimità non nella pretesa di infallibilità, ma nella capacità di mantenere l'evidenza emendabile e la responsabilità condivisa lungo l'intero arco del suo impiego (Floridi, 2019).

9. Conclusioni.

Giunti al termine di questo breve percorso, possiamo dire che il nodo essenziale della questione non risiede nella potenza della tecnologia, né nella sua funzione, bensì nel significato che gli esseri umani assegnano ai suoi output — non solo sul piano semantico, ma anche su quello pragmatico, quello in cui l'esito tecnico diventa spunto e condizione per la scelta, la decisione e l'azione tra alternative possibili. L'analisi ha mostrato che i sistemi di riconoscimento facciale non producono fatti in senso forte, ma nel senso vichiano del termine: generano interpretazioni operative (spesso sotto forma di classificazione), ossia misure numeriche che incorporano scelte rappresentazionali e decisioni di soglia, e che acquistano valore solo se ancorate a una finalità dichiarata, a un contesto d'uso e a un regime di responsabilità esplicito (Simon, 1969; Floridi, 2019). Il loro valore appartiene non al dominio delle verità di fatto, ma a quello metrologico delle stime probabilistiche. La norma, in questa prospettiva, non è un vincolo imposto alla tecnica, ma la forma pubblica della prudenza ermeneutica: il dispositivo che rende visibili presupposti, limiti e condizioni di validità dell'evidenza, traducendo numeri in decisioni proporzionate e contestabili (Parlamento europeo e Consiglio, 2024; ISO/IEC 25010:2011; ISO/IEC 23894:2023).

La tesi sostenuta, è, in sintesi, duplice. Da un lato, fatto e norma si incontrano nella prova tecnica: la qualità non coincide con l'accuratezza media su banchi di prova (benchmark), ma con l'affidabilità situata di un sistema che mantiene prestazioni adeguate nel tempo, rende visibile l'incertezza, distribuisce in modo equo gli errori e con-

sente di attribuire responsabilità lungo la filiera (Buolamwini-Gebru, 2018; Grother-Ngan-Hanaoka, 2019). Dall'altro lato, questo incontro richiede di riportare al centro l'interpretazione, intesa in senso gnoseologico più che epistemico, e semiotico più che ermeneutico: cioè una interpretazione che riguarda come comprendiamo il mondo che ci circonda piuttosto che come creiamo verità di fatto dalla sua esperienza; e una interpretazione intesa più come processo di produzione e attribuzione di significato all'interno di un sistema di segni, cioè di un testo, in questo caso i dati della macchina, piuttosto che come atto di comprensione di testo e contesto insieme, alla ricerca di un *quid* del fenomeno umano che non fosse immediatamente conoscibile. In questa accezione, l'automazione non elimina l'interpretazione, ma la trasla nel mondo dei segni, la rende computabile e la automatizza. Da qui la proposta di una *semiotica artificiale*, intesa come capacità e competenza di saper 'leggere' i modelli di Intelligenza artificiale, non solo per la loro struttura, ma per il comportamento, e la traccia di tale comportamento nei loro output, per valutare questi ultimi non tanto per quello che sono (segni) ma per quello che vogliamo significhino, per il senso che gli attribuiamo e per la forza giustificativa che vogliamo abbiano nella deliberazione pubblica (Gebru et al., 2021; Mitchell et al., 2019).

In questa prospettiva, la prudenza epistemica diventa un vero e proprio programma operativo condiviso. Chi progetta documenta dati, processi di sviluppo, metriche e soglie in modo trasparente e verificabile; chi regola costruisce cornici di responsabilità effettiva, traducendo proporzionalità e gestione del rischio in requisiti concreti di governance dei dati, sorveglianza post-market, tecno-vigilanza e supervisione umana significativa; chi decide tratta gli output come evidenze graduate, motivate rispetto allo scopo e conformi allo standard probatorio. Su questo terreno può prendere forma una interpretazione tecno-giuridica: un lessico e un insieme di pratiche comuni — dalla disaggregazione delle prestazioni alla calibrazione delle soglie, dalla registrazione agli audit indipendenti — capaci di trasformare l'output numerico in un argomento pubblico, aperto al controllo e al miglioramento, all'interno della cornice regolatoria dell'AI Act e degli standard tecnici armonizzati.

Se il punto di partenza era la dialettica tra fatto e norma, il punto d'arrivo è la comprensione che questa dialettica rende possibile: la capacità di leggere i sistemi per ciò che realmente sono — trascrittori di simboli governati da scopi e contesti determinati — e

di utilizzarli affinché l'evidenza resti emendabile, la decisione motivata e i diritti delle persone garantiti e protetti. In questo modo i numeri tornano a essere ciò che devono: evidenze tracciabili; e il diritto torna a fare ciò che gli compete: abilitare affidabilità situata e responsabilità condivisa. Solo con questo ponte — stabile e attraversabile in entrambe le direzioni — la promessa tecnica può davvero diventare valore pubblico.

Note bibliografiche

Barrett L.F. (2017), *How Emotions Are Made: The Secret Life of the Brain*, New York, Houghton Mifflin Harcourt.

Belhumeur P.N.-Hespanha J.P.-Kriegman D.J. (1997), *Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection*, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 7, p. 711 ss.

Bijker W.E.-Hughes T.P.-Pinch T.J. (1987), *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*, Cambridge (MA), MIT Press.

Buolamwini J.-Gebru T. (2018), *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 81, p. 77 ss.

Cabitza F.-Zeitoun J.D. (2019), *The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence*, in *Annals of Translational Medicine*, 7, 8, p. 161.

Cabitza F.-Campagner A.-Mattioli M. (2022), *The unbearable (technical) unreliability of automated facial emotion recognition*, in *Big Data & Society*, 9, 2, p. 1 ss.

Carta dei Diritti fondamentali dell'Unione Europea (2000), in *Gazzetta Ufficiale dell'Unione europea*, C 364/1, 18 dicembre 2000.

Consiglio d'Europa (2024), *Convenzione quadro sull'intelligenza artificiale, i diritti umani, la democrazia e lo Stato di diritto*, Strasburgo, 15 maggio 2024.

Deng J.-Guo J.-Xue N.-Zafeiriou S. (2019), *ArcFace: Additive Angular Margin Loss for Deep Face Recognition*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, p. 4690 ss.

Ekman P.-Friesen W.V. (1978), *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto (CA), Consulting Psychologists Press.

Ferraris M. (2021), *Documanità. Filosofia del mondo nuovo*, Bari-Roma, Laterza.

Floridi L. (2019), *The Logic of Information. A Theory of Philosophy as Conceptual Design*, Oxford, Oxford University Press.

Floridi L.-Cabitza F. (2021), *Intelligenza artificiale: l'uso delle nuove macchine*, Milano, Bompiani.

GADAMER H.-G. (1960), *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik*, Tübingen, Mohr Siebeck.
GAMA J.-ŽLIOBAITĖ I.-BIFET A.-PECHENIZKIY M.-BOUCHACHIA A. (2014), *A Survey on Concept Drift Adaptation*, in *ACM Computing Surveys*, 46, 4, Art. 44, p. 1 ss.
GEBRU T.-MORGENSTERN J.-VECCHIONE B.-WORTMAN VAUGHAN J.-WALLACH H. *et al.* (2021), *Datasheets for Datasets*, in *Communications of the ACM*, 64, 12, p. 86 ss.
GERO J.S. (1990), *Design Prototypes: A Knowledge Representation Schema for Design*, in *AI Magazine*, 11, 4, p. 26 ss.
GERO J.S.-KANNENGIESSER U. (2004), *The Situated Function-Behaviour-Structure Framework*, in *Design Studies*, 25, 4, p. 373 ss.
GOODFELLOW I.-BENGIO Y.-COURVILLE A. (2016), *Deep Learning*, Cambridge (MA), MIT Press.
GROTHER P.-NGAN M.-HANAOKA K. (2019), *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, NIST Interagency/Internal Report (NISTIR) 8280, Gaithersburg (MD), National Institute of Standards and Technology.
HOLLNAGEL E. (2012), *FRAM: The Functional Resonance Analysis Method: Modelling Complex Socio-technical Systems*, Farnham, Ashgate.
ISIDORO DI SIVIGLIA (VII sec.), *Etymologiae*, a cura di J. Fontaine, Turnhout, Brepols, 1982 (Corpus Christianorum, Series Latina 112).
ISO/IEC 19795-1:2021, *Information technology - Biometric performance testing and reporting - Part 1: Principles and framework*, Geneva, International Organization for Standardization.
ISO/IEC 2382:2015, *Information technology - Vocabulary*, Geneva, International Organization for Standardization.
ISO/IEC 23894:2023, *Information technology - Artificial intelligence - Guidance on risk management*, Geneva, International Organization for Standardization.
ISO/IEC 25010:2011, *Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models*, Geneva, International Organization for Standardization.
ISO/IEC 30107-3:2017, *Information technology - Biometric presentation attack detection - Part 3: Testing and reporting*, Geneva, International Organization for Standardization.
ISO/IEC 42001:2023, *Artificial intelligence - Management system - Requirements with guidance for use*, Geneva, International Organization for Standardization.
JAIN A.K.-ROSS A.-NANDAKUMAR K. (2011), *Introduction to Biometrics*, New York, Springer.
LECUN Y.-BENGIO Y.-HINTON G. (2015), *Deep Learning*, in *Nature*, 521, p. 436-444.
LATOUR B. (1987), *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge (MA), Harvard University Press.
LEVINAS E. (1961), *Totalité et Infini. Essai sur l'extériorité*, La Haye, Martinus Nijhoff.

LONGO L. *et al.* (2024), *Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions*, in *Information Fusion*, 106, 102301.

MARCONI L.-CABITZA F. (2024), *The Challenges and Promises of Artificial Intelligence in the Contemporary Society: a Critical Perspective*, in *Urbaniana University Journal*, 2, p. 83 ss.

MARCONI L.-PIROVANO E.-CABITZA F. (2024), *CLARITY AI: a comprehensive checklist integrating established frameworks for enhanced research quality in medical AI studies*, in *CEUR Workshop Proceedings*, 3880, p. 1 ss.

MARCONI L.-CABITZA F. (2025a), *Tutor or Mentor. A Comparative Usability Study of AI Roles in Higher Education*, in CRISTEA A.I.-WALKER E.-LU Y.-SANTOS O.C.-ISOTANI S. (a cura di) (2025), *Artificial Intelligence in Education*. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED, Communications in Computer and Information Science, vol. 2592, Cham, Springer, pp. 201-208.

MARCONI L.-CABITZA F. (2025b), *Show and tell: a critical review on robustness and uncertainty for a more responsible medical AI*, in *International Journal of Medical Informatics*, 105970.

MARSHALL S. (1982), *Cultura e utilità. Il fondamento simbolico dell'attività pratica*, Milano, Bompiani.

MATTIOLI M.-CABITZA F. (2024), *Not in My Face: Challenges and Ethical Considerations in Automatic Face Emotion Recognition Technology*, in *Machine Learning and Knowledge Extraction*, 6, 4, p. 2201 ss.

MITCHELL M. *et al.* (2019), *Model Cards for Model Reporting*, in Proceedings of the Conference on Fairness, Accountability and Transparency (FAT*), ACM, p. 220 ss.

MITTELSTADT B. (2019), *Principles Alone Cannot Guarantee Ethical AI*, in *Nature Machine Intelligence*, 1, 11, p. 501 ss.

NORMAN D.A. (2013), *The Design of Everyday Things* (Revised and Expanded Edition), New York, Basic Books.

O'NEIL C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, Crown Publishers.

PARLAMENTO EUROPEO E CONSIGLIO (2024), *Regolamento (UE) 2024/1689 che stabilisce regole armonizzate sull'intelligenza artificiale (AI Act)*, in *Gazzetta Ufficiale dell'Unione europea*, L, 12 luglio 2024.

REASON J. (1990), *Human Error*, Cambridge, Cambridge University Press.

SCHROFF F.-KALENICHENKO D.-PHILBIN J. (2015), *FaceNet: A Unified Embedding for Face Recognition and Clustering*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, p. 815 ss.

SCHUETT J. (2024), *Risk Management in the Artificial Intelligence Act*, in *European Journal of Risk Regulation*, 15, 2, p. 279 ss.

SIMON H.A. (1969), *The Sciences of the Artificial*, Cambridge (MA), MIT Press.

TAIGMAN Y.-YANG M.-RANZATO M.A.-WOLF L. (2014), *DeepFace: Closing the Gap to Human-Level Performance in Face Verification*, in *Procee-*

dings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, p. 1701 ss.

TURK M.-PENTLAND A. (1991), *Face Recognition Using Eigenfaces*, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, p. 586 ss.

VICENTE K.J. (1999), *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*, Mahwah (NJ), Lawrence Erlbaum Associates.

VICO G. (1999), *La Scienza Nuova* (1725/1744), ed. moderna a cura di A. Battistini, Milano, Mondadori.

WACHTER S.-MITTELSTADT B.-RUSSELL C. (2021), *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI*, in *Computer Law & Security Review*, 41, 105567.

WANG M.-DENG W. (2021), *Deep Face Recognition: A Survey*, in *Neurocomputing*, 429, p. 215 ss.

ZHAO W.-CHELLAPPA R.-PHILLIPS P.J.-ROSENFELD A. (2003), *Face Recognition: A Literature Survey*, in *ACM Computing Surveys*, 35, 4, p. 399 ss.

(LA NORMA) LA DISCIPLINA SULLE TECNOLOGIE DI RICONOSCIMENTO FACCIALE, OVVERO IL TENTATIVO DI LIMITARE UN FORMIDABILE POTERE DI SORVEGLIANZA

di *Giuseppe Mobilio*

Sommario: 1. Introduzione: perché il diritto deve occuparsi delle TRF. — 2. Un potere di sorveglianza irresistibile? — 3. L'importanza di una risposta giuridica adeguata alle specificità delle TRF. — 3.1. Le diverse problematiche giuridiche derivanti dai diversi usi delle TRF. — 3.2. Regimi differenti per sistemi differenti. — 3.3. Una questione di scala. — 4. La risposta del diritto. — 4.1. I rischi di un approccio troppo rigido. — 4.2. Usare la tecnologia alla stregua del diritto. — 4.3. Nuovi diritti per nuove tecnologie? — 5. Conclusioni.

1. Introduzione: perché il diritto deve occuparsi delle TRF.

Le tecnologie di riconoscimento facciale (TRF) rappresentano una delle applicazioni dell'intelligenza artificiale (IA) più note e, nell'immaginario collettivo, anche una delle più preoccupanti (Kostka-Steinacker-Meckel, 2023). Una riflessione su questi sistemi biometrici di IA che prenda in considerazione il rapporto tra 'fatto' e 'norma' si rivela utile sotto molti punti di vista. Innanzitutto, perché il dibattito sulle TRF ha ricevuto in Europa un incredibile slancio con la presentazione da parte della Commissione europea della proposta di regolamento generale sull'intelligenza artificiale, poi divenuta il regolamento (UE) 2024/1689 (c.d. AI Act). È stata la volontà di formulare una disciplina giuridica a porre sotto i riflettori questa tecnologia: la 'norma' ha posto in risalto il 'fatto'. Tuttavia, occorre soprattutto domandarsi perché il regolatore europeo abbia deciso di adottare una disciplina avente ad oggetto questa particolare tecnologia, ovvero perché la 'norma' abbia riconosciuto rilevanza a questo 'fatto'. Muovendo da questo interrogativo, e sottolineando alcune caratteristiche distintive di queste tecnologie, si vuole fornire alcuni

spunti di riflessione sulla risposta giuridica conseguentemente formulata, sottolineando alcune difficoltà manifestate dalla regolazione e, al contempo, nuove istanze emergenti.

2. Un potere di sorveglianza irresistibile?

Lo sviluppo e la diffusione delle più avanzate TRF ha determinato un salto qualitativo nel potere di sorveglianza. Le tecnologie biometriche, in generale, consentono di distinguere una persona grazie a caratteristiche fisiche uniche, come le impronte digitali, il DNA, la forma dell'iride, oppure attraverso i tratti differenziali del suo comportamento, come il modo di camminare o l'attività di scrittura. Tuttavia, tra le tecnologie di questo tipo, le TRF sono considerate estremamente invasive. La cattura dell'immagine di un volto può avvenire a distanza, con persone in movimento, senza alcun contatto fisico e senza che vi sia alcuna consapevolezza. Le TRF sono anche incorporabili in una miriade di dispositivi, come telecamere a circuito chiuso, bodycam o droni, rendendo così il riconoscimento diffuso e a basso costo (Berle, 2020). Non sorprende che, con il diffondersi della pandemia da Covid-19, la sorveglianza biometrica sia stata impiegata per il tracciamento delle persone potenzialmente contagiose, e che nei conflitti bellici più recenti, come quello russo-ucraino o quello israelo-palestinese, queste tecnologie abbiano trovato un utilizzo massiccio (Andresin, 2025).

In generale, dunque, le TRF possono essere sfruttate per riconoscere le persone, sorvegliarle, monitorarne gli spostamenti. Esse possono essere impiegate in molteplici ambiti, come aeroporti, stadi, scuole, luoghi di lavoro, o per molteplici scopi, come prevenzione e contrasto di reati, per autorizzare l'accesso ad un luogo o per sbloccare uno smartphone. Le finalità principali sono comunemente distinte in individuazione, autenticazione e categorizzazione, come descritto più ampiamente nel contributo di Cabitza. Questi sistemi biometrici presentano tuttavia due caratteristiche che meritano di essere qui evidenziate.

Da una parte, essi assottigliano ancora di più il confine tra mondo analogico e mondo virtuale. Grazie all'integrazione con le *big data analytics*, è ora possibile acquisire dati sulle persone in tempo reale e combinarli con ulteriori informazioni disponibili in banche dati pubbliche e private o online (Acquisti-Gross-Stutzman, 2014). È sufficiente quindi trovarsi in un luogo pubblico per riuscire a risa-

lire a informazioni riguardanti una persona come le generalità, le preferenze, le frequentazioni.

Dall'altra, grazie alle tecniche di analisi facciale, è possibile inferire tutta una serie di informazioni a partire dall'immagine di un volto con una percentuale di accuratezza molto più elevata di quanto accadrebbe con un semplice sguardo a occhio nudo. Queste tecnologie riescono a profilare le persone, stabilire se esse appartengono a un determinato gruppo in base a caratteristiche quali età o razza, o addirittura dedurre informazioni personali quali le tendenze sessuali o le preferenze politiche (KOSINSKI, 2021).

Siamo di fronte a strumenti che possiedono un potenziale enorme in termini di capacità di perseguire reati e rintracciare persone scomparse, garantire elevati livelli di sicurezza nell'accesso ai luoghi, trovare applicazioni all'avanguardia in ambito medico, come ad esempio nella rilevazione del dolore o nell'indagine delle caratteristiche genetiche. Tuttavia, è ben possibile che questo potere di sorveglianza sia soggetto ad abusi da parte di autorità pubbliche e di operatori privati per scopi affatto in linea con la tutela dei diritti.

Questo tipo di sorveglianza si colloca nell'ambito delle forme di c.d. 'dataveillance' (forma contratta di *data-surveillance*), intesa come «impiego sistematico dei dati personali per indagare o monitorare le azioni o le comunicazioni di una o più persone» (CLARKE, 1988). Il diritto ha ovviamente cercato di regolare questo potere, con l'obiettivo di limitarlo e condizionarlo al rispetto di principi e valori costituzionali. Non sorprende che la risposta a questo scopo sia stata a lungo offerta dalla normativa sulla protezione dei dati personali, dal momento che le TRF acquisiscono ed elaborano questo tipo di dati, consentendone la conservazione e l'accesso per molteplici finalità (MOBILIO, 2021). Tuttavia, la posta in gioco non è solamente questo diritto specifico, perché le TRF sono in grado di produrre un impatto trasversale su molteplici diritti fondamentali e sui sistemi democratici in sé (MOBILIO, 2025).

Questa capacità risulta evidente in alcuni esempi: negli Stati Uniti, l'uso di TRF è stato intensificato nella primavera del 2020 in occasione dell'uccisione di Breonna Taylor e George Floyd, con l'obiettivo di identificare gli autori di episodi di violenza durante le proteste che ne sono scaturite in tutto il Paese; nel Galles, la South Wales Police di Cardiff ha avviato una sperimentazione sin dal 2017, dando origine al primo fra i più noti contenziosi giudiziari, sfociato nelle pronunce della Queen's Bench Division della High Court of Ju-

stice e della Civil Division della Court of Appeal nel settembre 2019 e nell'agosto 2020; in Russia, le forze dell'ordine di Mosca hanno impiegato queste tecnologie per perseguire oppositori politici, dando origine al caso *Glukhin v. Russia* deciso dalla Corte europea dei diritti dell'uomo nel luglio 2023. Si tratta di vicende che hanno posto in risalto come l'impiego delle TRF possano dare origine a forme di discriminazione ai danni dei soggetti verso cui questi sistemi si rivelano meno accurati; a limitazioni della libertà di espressione e riunione, ove la sorveglianza venga impiegata per combattere e reprimere le forme di dissenso; ad una menomazione dello stato di diritto e del grado di democraticità di un ordinamento, ove l'uso di queste tecnologie da parte delle autorità pubbliche risulti illegittimo e si risolva in un indebolimento della partecipazione alla vita politica e sociale dei cittadini (FRA, 2019).

Stando così le cose, occorre riflettere su quale sia la risposta giuridica a 'fatti' tecnologici in grado di aprire a simili scenari, prendendo in considerazione alcuni limiti che le 'norme' rischiano di manifestare se non attente alle caratteristiche di questo oggetto.

3. L'importanza di una risposta giuridica adeguata alle specificità delle TRF.

3.1. *Le diverse problematiche giuridiche derivanti dai diversi usi delle TRF.*

Innanzitutto, la pretesa di regolare le TRF non può prescindere dal considerare i loro differenti usi e ambiti di utilizzo, che possono dare origine a problematiche con diversi gradi di serietà.

Ad esempio, un sistema di autenticazione biometrica tramite riconoscimento facciale può essere integrato in uno smartphone, consentendo o meno l'accesso al dispositivo, oppure può essere impiegato in uno scalo aeroportuale come sistema di verifica dei passaporti. In un caso la TRF può impedire o meno il godimento di un bene di proprietà (lo smartphone) o lo svolgimento di tutte quelle attività che oramai si possono realizzare tramite questi dispositivi, con l'esercizio delle connesse libertà (comunicare, informarsi, gestire a distanza sistemi complessi); nell'altro può favorire una gestione più efficiente dei flussi di persone (tramite controlli più rapidi) o impedire o meno l'imbarco di soggetti pericolosi che potrebbero mettere a rischio la vita e la sicurezza di altri (es. persone su no-fly list).

Un sistema di identificazione può essere impiegato dalle forze dell'ordine per individuare nella folla individui ricercati o impedire l'accesso a impianti sportivi a persone colpite da misure di prevenzione. In questi casi c'è in gioco la libertà personale, la libertà di circolazione, o la possibilità di assistere a un evento pubblico.

Un sistema di categorizzazione da analisi facciale può essere impiegato per rivolgere pubblicità commerciale mirata o informazioni a singoli o gruppi di persone, oppure per distinguere una parte di popolazione in base ai tratti fenotipici. In questi casi in gioco abbiamo il pluralismo dell'offerta economica, il pluralismo delle informazioni in grado di condizionare la formazione delle volontà, sino al perseguimento di politiche discriminatorie ai danni di fasce minoritarie della popolazione.

A questa varietà di impieghi e contesti si sommano le complessità derivanti da due ulteriori fattori, ovvero l'accuratezza e la sicurezza delle TRF. Sul primo aspetto, cui si rinvia al contributo di Cabitza per maggiori approfondimenti, basti considerare come un falso-positivo o un falso-negativo abbiano conseguenze diverse, a seconda che non permettano erroneamente di imbarcarsi su un volo a chi ne ha titolo (con i disagi conseguenti) o consentano erroneamente di imbarcarsi su un volo a chi non ne avrebbe titolo (ad es. perché un sospetto terrorista). Sul secondo aspetto, è noto innanzitutto come questi sistemi possano subire attacchi malevoli, con l'obiettivo di indurre fraudolentemente il sistema a compiere un falso-positivo (ad es. tramite spoofing). Ma la sicurezza può essere messa a repentaglio anche dalla sottrazione di dati biometrici processati o conservati ai fini del riconoscimento (ad es. con un data breach). In questo caso bisogna considerare come il danno sia maggiore rispetto a quello subito dal furto di un 'mero' dato personale, perché i dati biometrici (ricavati dalla faccia, come anche dalle impronte digitali o dal DNA) consentono di identificare univocamente una persona e rimangono intrinsecamente e — quasi — perpetuamente legati ad essa: di conseguenza essi non possono essere semplicemente sostituiti o duplicati come fossero una carta di identità.

Il diritto è quindi chiamato a discernere questi diversi scenari e calibrare la risposta regolativa in base agli interessi concretamente in gioco, senza imporre una regolamentazione indistinta e uniforme. Il regolatore deve inoltre tener conto della fallibilità di questi sistemi e dei rischi per la sicurezza che possono sorgere.

3.2. *Regimi differenti per sistemi differenti.*

Anche il modo con cui un sistema di riconoscimento facciale è strutturato è un elemento che il diritto deve considerare. Allo stesso tempo, però, la norma deve resistere alla tentazione di lasciarsi andare a classificazioni troppo rigide che trascurino i rischi che concretamente possono sorgere dall'uso di queste tecnologie.

Le TRF funzionano essenzialmente seguendo un procedimento nel quale si cattura l'immagine di una faccia; si individua la faccia all'interno dell'immagine e si estraggono automaticamente le caratteristiche (c.d. features) per creare un modello (c.d. template) biometrico, quale rappresentazione numerica che distingue univocamente una persona; si confronta il template con le immagini facciali raccolte in una galleria (c.d. watchlist) in cerca di una corrispondenza. Nell'ipotesi in cui questo procedimento avvenga istantaneamente, il sistema viene classificato come 'in tempo reale'; se invece il confronto avviene entro un lasso di tempo significativo dalla cattura dell'immagine, il sistema viene classificato come 'a posteriori'. Sia il dato normativo (i.e. l'AI Act), sia la giurisprudenza (il citato caso Glukhin), considerano le TRF in tempo reale più pericolose rispetto a quelle a posteriori.

Da una parte, il riconoscimento ''in tempo reale'', specie se in spazi accessibili al pubblico a fini di attività di contrasto, viene qualificato dall'AI Act come:

« particolarmente invasivo dei diritti e delle libertà delle persone interessate, nella misura in cui potrebbe avere ripercussioni sulla vita privata di un'ampia fetta della popolazione, farla sentire costantemente sotto sorveglianza e scoraggiare in maniera indiretta l'esercizio della libertà di riunione e di altri diritti fondamentali. [...] L'immediatezza dell'impatto e le limitate opportunità di eseguire ulteriori controlli o apportare correzioni [...] comportano inoltre un aumento dei rischi per quanto concerne i diritti e le libertà delle persone interessate nell'ambito delle attività di contrasto, o che sono da queste condizionate » (cons. 32 AI Act).

Dall'altra, la Corte europea dei diritti dell'uomo, quando fa riferimento alle TRF 'in tempo reale', sottolinea una necessità « maggiore » di salvaguardie (*Glukhin v. Russia*, no. 11519/20, 4 luglio 2023, § 82) e del « più alto » livello di giustificazione, con riguardo a misure « particolarmente intrusive » (ivi, § 86).

Queste posizioni, se colgono alcuni aspetti problematici, non possono essere assolutizzate. Come infatti è stato sottolineato, l'impatto sui diritti non dipende necessariamente dalla distanza di tempo

con la quale avviene il riconoscimento, poiché un sistema di identificazione di massa è in grado di identificare migliaia di persone in poche ore, anche a posteriori (EDPB-EDPS, 2021, § 31). Inoltre, il fatto che le immagini facciali possano essere acquisite da una miriade di telecamere a circuito chiuso disseminate nelle città oppure scaricate da internet, e che possano essere conservate per lunghissimi periodi di tempo all'interno dei database prima di essere processate, non esclude che anche la mera possibilità di essere sottoposto a riconoscimento a posteriori incuta un'impressione di soggezione continua altrettanto grave rispetto a quella dal vivo.

3.3. *Una questione di scala.*

Infine, occorre osservare come l'utilizzo concreto, il contesto e la finalità risentano anche della scala entro cui una TRF viene dispiegata. Un conto è l'uso singolare ed episodico; un conto è un impiego costante e diffuso, supportato da una infrastruttura capillare. In questo secondo caso l'impatto sui diritti e sui sistemi democratici cresce esponenzialmente.

Volendo limitarsi ad alcuni degli esempi più noti, si pensi a quanto avviene con il *Social Credit System* sperimentato in alcune province della Cina. Con questa infrastruttura di massa, grazie alle informazioni raccolte in maniera automatizzata sulla vita pubblica e privata, viene associato un punteggio a ciascun cittadino per quantificarne l'affidabilità e la reputazione e per parametrare un sistema di benefici e sanzioni aventi scopo di orientare il comportamento dell'intera popolazione (Liang-Das-Kostyuk-Hussain, 2018). In Europa, invece, troviamo i sistemi di informazione di larga scala strutturati nell'ambito delle politiche di smart borders. Queste infrastrutture sono costituite da enormi database gestiti a livello centralizzato e da network che collegano organismi dell'Ue con gli Stati membri. Tali sistemi sono stati progressivamente arricchiti con dati biometrici raccolti per finalità diverse — segnatamente, a scopi di sicurezza (SIS), per le richieste di asilo (Eurodac), per il rilascio di visti di breve durata (VIS), in prospettiva per superare il sistema dei passaporti (EES), per coloro che non necessitano di un visto formale (ETIAS) —, i quali hanno finito per essere messi a disposizione delle forze dell'ordine per lo svolgimento di attività di contrasto (Vavoula, 2020). In sostanza, i dati trasmessi al momento della richiesta di un

visto potranno poi essere usati dalle forze dell'ordine per motivi di sicurezza.

In definitiva, il diritto deve resistere alla tentazione di ricorrere a schematismi troppo astratti e classificazioni troppo rigide che rischiano di rendere inefficace il tentativo di regolare questo tipo di tecnologie. Questi sistemi di IA, inoltre, tendono ad essere sviluppati ad un ritmo incessante che supera le tempistiche di produzione e di adattamento del diritto. In aggiunta, le tecnologie biometriche sono impiegate con una versatilità tale da imporre un'attenzione specifica ai rischi particolari e sistemici che sorgono di volta in volta nei confronti dei diritti fondamentali e dell'intero ordinamento democratico.

4. La risposta del diritto.

4.1. *I rischi di un approccio troppo rigido.*

L'AI Act ha assunto un ruolo chiave nella disciplina di un oggetto così sensibile come le tecnologie biometriche. Non a caso queste ultime sono state uno degli argomenti più discussi e divisivi durante i triloghi che hanno portato alla sua approvazione (PALMIOTTO, 2025). Questo regolamento, inoltre, è la riprova che la disciplina sulla protezione dei dati personali non è più sufficiente — almeno da sola — a disciplinare le TRF, e quindi a limitare il potere di sorveglianza che deriva dal loro utilizzo.

Volendo spendere alcune considerazioni sul suo approccio generale, l'AI Act guarda all'IA come un 'prodotto' da disciplinare 'orizzontalmente', a prescindere cioè dallo specifico sistema di IA e dal settore di impiego. In questo contesto le TRF rappresentano a loro volta un'eccezione, perché, sebbene il regolamento non le menzioni espressamente, le ricomprende nel più ampio insieme dei 'sistemi biometrici', quale tecnologia invece meritevole di regole esplicite e *ad hoc*.

Sempre per quanto riguarda la sua impostazione complessiva, inoltre, l'AI Act segue un particolare approccio basato sul rischio (LONGO, 2024). Esso elenca una serie di 'pratiche di IA vietate' *by default*, mentre indica numerose regole per i sistemi di IA 'ad alto rischio' ispirate, appunto, ad una valutazione concreta dei rischi da essi posti. Quest'ultimo aspetto è stato criticato nella misura in cui, da una parte, il regolamento risulta concepito più per garantire la sicurezza dei sistemi di IA piuttosto che per la tutela dei diritti; dall'altra, esso si affida ad un coinvolgimento eccessivo degli operatori econo-

mici, sulla base di regole mutuate dal *New Legislative Framework* dell'UE, il quale stabilisce una serie di obblighi per il produttore in termini di gestione dei rischi prima di rilasciare un prodotto sul mercato (Almada-Petit, 2025). Alcuni di questi obblighi, però, sono più direttamente funzionali alla protezione dei diritti e rappresentano importanti novità, come nel caso dei requisiti sulla qualità e l'adeguatezza dei dati impiegati per addestrare i sistemi di IA (art. 10), volti anche ad evitare possibili effetti discriminatori, o dei requisiti di accuratezza, robustezza e cibersicurezza (art. 15), per aumentarne la resilienza nei confronti di attacchi malevoli e diminuire le vulnerabilità.

Il risultato è che l'AI Act contiene un complesso di regole che si applicano alle TRF ma che non esprime un indirizzo univoco e coerente. Nella mole di disposizioni di questo regolamento è comunque possibile rintracciare una distinzione basata sulle tre finalità sopra menzionate di identificazione, categorizzazione e verificazione (Menéndez Gonzáles-Mobilio, 2025).

Quanto alla prima, l'AI Act — come detto — non parla espressamente di TRF, ma fornisce una regolazione dei 'sistemi di identificazione biometrica remota' (IBR), di cui le TRF rappresentano l'esempio più diffuso e sviluppato. Tra le pratiche di IA vietate all'art. 5, l'AI Act vieta l'uso di sistemi di IBR « in tempo reale » in spazi « accessibili al pubblico » e a fini di « attività di contrasto » (art. 5(1)(h)). A questo divieto generale si affiancano significative eccezioni, accompagnate da una serie di garanzie. Tale uso è infatti consentito ove « strettamente necessario » per gli obiettivi di: ricerca delle vittime di reati; prevenzione di una « minaccia specifica, sostanziale e imminente » alla vita delle persone, o un attacco terroristico; perseguimento degli autori o dei sospettati dei reati di cui all'Allegato II del regolamento. L'impiego effettivo delle TRF può essere effettuato solo per « confermare » l'identità della persona specificamente interessata (non tanto per trovare un potenziale sospettato), e deve inoltre tenere conto della situazione reale ed effettiva e delle conseguenze per i diritti e le libertà (art. 5(2)). Ogni singolo utilizzo deve essere autorizzato da un'autorità giudiziaria o da un'autorità amministrativa indipendente sulla base di « prove oggettive » e di « indicazioni chiare » in termini di necessità e proporzionalità rispetto agli obiettivi (art. 5(3)). La normativa nazionale dovrà sviluppare queste norme, eventualmente adottando un approccio anche più restrittivo. Infine, questi sistemi di IBR sono anche espressamente soggetti alle condizioni stabilite per i sistemi 'ad alto rischio' dell'IA (Allegato III(1)(a)).

Questo divieto — come anticipato — si giustifica per la maggior pericolosità riconosciuta al riconoscimento dal vivo. Tuttavia, esso lascia troppo margine di manovra ai singoli Stati membri nell'individuare la disciplina di dettaglio. In aggiunta, esso non copre l'uso da parte di soggetti diversi che non siano le forze dell'ordine, o per finalità diverse da quella di contrasto: per cui, ad esempio, viene consentito l'uso da parte di privati per fini di sicurezza, con minori garanzie.

Anche l'impiego di TRF per identificazione a posteriori viene consentito con meno garanzie, poiché si prevede come necessaria la sola previa autorizzazione di un magistrato o autorità amministrativa (tranne quando è utilizzato per l'identificazione iniziale di un potenziale sospetto « sulla base di fatti oggettivi e verificabili direttamente connessi al reato »), escludendo l'uso in modo « non mirato » (art. 26(10)). Riemergono qui tutte le perplessità rispetto al concepire come meno pericolosa una forma di identificazione che in realtà può rivelarsi molto pervasiva.

Quanto alla categorizzazione, essa è vietata in generale — quindi anche tramite TRF — se utile per fare inferenze in merito a razza, opinioni politiche, appartenenza sindacale, convinzioni religiose o filosofiche, vita sessuale o orientamento sessuale (art. 5(1)(g)). In divieto è espressamente escluso per le attività di contrasto, come può essere nel caso di un poligrafo basato sull'analisi delle microespressioni facciali, il quale è invece considerato un sistema ad alto rischio (Allegato III(6)(b)). La « categorizzazione biometrica in base ad attributi o caratteristiche sensibili protetti basati sulla deduzione di tali attributi o caratteristiche » è invece considerata solamente ad alto rischio (Allegato III(1)(b)) ed è soggetta a obbligo di informazione verso il destinatario (art. 50(3)). Questo inquadramento rischia di rivelarsi limitato, nella misura in cui guarda solo alla sicurezza del prodotto e sottostima la grande capacità di inferenza di queste tecnologie, che possono fornire informazioni su singoli o gruppi di persone con le quali potenzialmente costruire politiche, anche con risvolti discriminatori.

Inoltre, l'AI Act si occupa anche del riconoscimento delle emozioni, che può essere considerato un'applicazione della categorizzazione mediante TRF. In particolare, il regolamento vieta tale riconoscimento nelle scuole o in ambienti di lavoro (art. 5(1)(f)). Questo divieto non vale espressamente per il riconoscimento per 'motivi medici' (come il rilevamento del dolore o indagini sulle caratteristiche

genetiche, menzionate nel § 2) o 'di sicurezza'. Il riconoscimento delle emozioni per ogni altro scopo o ambito viene inoltre qualificato ad alto rischio (Allegato III(1)(c)). Questa applicazione delle TRF, tuttavia, appare discutibile, nella misura in cui non vi sono evidenze scientifiche che dimostrano come sia possibile operare un riconoscimento univoco delle emozioni, a partire tantomeno dalle espressioni del volto (Barrett, 2019).

Infine, quanto alla verificazione, l'AI Act esclude le TRF anche dai sistemi ad alto rischio se la loro « unica finalità è confermare che una determinata persona fisica è la persona che dice di essere » (Allegato III(1)(a)). Anche questa esclusione appare problematica, se si considera che la verificazione a scopo di sicurezza, ad esempio, per consentire l'accesso a determinati luoghi potrebbe di per sé e in astratto esporre a seri pericoli.

La disciplina che l'AI Act rivolge alle TRF, dunque, rappresenta un passo in avanti rispetto alla regolazione preesistente, ma presta il fianco a dubbi. Essa appare troppo rigida e non sembra tener conto dei possibili risvolti problematici nell'uso in sé delle TRF (come per il riconoscimento delle emozioni), o che sorgono a seconda dei diversi contesti di utilizzo (come per la verificazione o la categorizzazione), o per le conseguenze derivanti da un possibile uso sistemico e assiduo (come per l'identificazione a posteriori), o per il possibile svuotamento delle garanzie (come per l'identificazione in tempo reale in luoghi pubblici, a causa dell'intervento della normativa nazionale).

All'interno del regolamento vi sono previsioni che impongono un'attenzione agli effetti concretamente derivanti dall'uso dei sistemi di IA ad alto rischio (e quindi delle TRF). È il caso dell'obbligo per i produttori di predisporre un sistema di gestione dei rischi per la salute, la sicurezza e i diritti fondamentali (art. 9), o dell'obbligo per i *deployer* di effettuare una valutazione dell'impatto sui diritti fondamentali prima del relativo utilizzo (art. 27). Si tratta di previsioni volte ad aumentare la consapevolezza degli effetti derivanti dall'impiego di queste tecnologie, di prevenire e mitigare le conseguenze negative, le quali però non sempre sono facilmente distinguibili *a priori*.

L'AI Act, infine, esclude dal suo ambito di applicazione — come visto — le attività di polizia, oltre che l'uso a scopo militare, di difesa e di sicurezza nazionale (art. 2(3)), mentre viene posticipata la sua efficacia per i « sistemi IT su larga scala » impiegati alle frontiere (art. 111), menzionati al § 3.3.

Il diritto sembrerebbe quindi esser ricorso a distinzioni troppo rigide, che rischiano di non catturare le concrete pericolosità derivanti da questo 'fatto' tecnologico.

4.2. *Usare la tecnologia alla stregua del diritto.*

Una ulteriore tendenza nella regolazione delle TRF, come dell'IA più in generale, è quella di sfruttare la tecnologia stessa per condizionare l'uso della tecnologia, sulla scia degli insegnamenti di Lessig (1999). Si tratta cioè di fare leva sulla valenza normativa degli algoritmi, dei sistemi di IA o dell'architettura degli ambienti per trovare soluzioni in grado di garantire una maggior protezione dei diritti, o ridurre il rischio che essi vengano violati. Il 'fatto' viene quindi usato alla stregua delle 'norme'.

Per citare un esempio, l'uso di questi sistemi biometrici di IA all'interno degli aeroporti ha conseguenze diverse per la tutela dei dati personali e biometrici a seconda di come i dati vengono conservati e di chi vi provveda. Come sottolinea lo European Data Protection Board, è possibile optare per diverse soluzioni, quali la conservazione dei modelli biometrici in dispositivi (smartphone) affidati alle persone, oppure la conservazione centralizzata all'interno dell'aeroporto dei dati registrati in forma cifrata con chiavi segrete note esclusivamente ai passeggeri, che a sua volta è scenario diverso dal porre i dati sotto il controllo del gestore aeroportuale, della compagnia aerea o da un fornitore di servizi cloud (EDPB, 2024). A seconda della scelta organizzativa, si pongono diversi livelli di rischio per quanto riguarda l'accesso illecito ai dati dei singoli passeggeri o all'intero insieme di tutti i loro dati, o per la sicurezza in relazione alla superficie di esposizione ad attacchi, o per il possibile sfruttamento da parte di Paesi extra-europei.

Si consideri poi come non tutti i sistemi di TRF possano essere impiegati allo stesso scopo, poiché le caratteristiche di progettazione e le impostazioni scelte condizionano il rendimento degli algoritmi. Questi sistemi, infatti, necessitano di una calibrazione caso per caso per funzionare in modo efficace (Fussey-Davies-Innes, 2021). Ad esempio, la soglia algoritmica che determina la 'sensibilità' del sistema e che influenza la percentuale di corrispondenza tra l'immagine catturata e quelle presenti nella galleria deve essere regolata prima del riconoscimento. Impostare una soglia più alta aumenta la probabilità che una corrispondenza fra immagini individuata dagli algoritmi

sia corretta, ma riduce il numero complessivo di corrispondenze rintracciate, inducendo a scartare erroneamente più casi limite che potrebbero essere corrispondenze corrette (falsi-negativi). Viceversa, abbassare questa soglia di accuratezza genera più corrispondenze possibili, ma aumenta la probabilità che alcune di esse siano inaccurate (falsi-positivi). Prendere una decisione tecnica di questo tipo, quindi, ha precise conseguenze sulle decisioni adottate in forza di un riconoscimento.

Ne consegue, quindi, come chi produce il diritto debba esser consapevole della valenza regolativa che possiedono determinate scelte tecniche. Si tratta di un livello di specificità che nel caso dei sistemi di IA ha conseguenze rilevanti per la tutela dei diritti. Uno degli strumenti regolativi più diffusi per imporre vincoli tecnici all'interno delle norme giuridiche sono gli standard o le norme tecniche, cui lo stesso AI Act fa ampio ricorso (IANNUZZI, 2024). Come emerge però chiaramente nel dibattito attorno a queste soluzioni regolatorie, occorre interrogarsi su quale sia lo spazio della decisione di natura politica nell'assunzione di scelte che possono riguardare i dataset con cui allenare gli algoritmi, il design dei dispositivi o l'architettura dell'ambiente ove si impiegano questi sistemi di IA, ovvero come recuperare la responsabilità democratica e assicurare trasparenza nelle decisioni (KOOPS, 2008).

4.3. *Nuovi diritti per nuove tecnologie?*

Il tema della regolazione delle tecnologie offre un grande invito a riflettere sulla capacità del diritto di inquadrare e disciplinare un 'fatto' così significativo e sfuggente. Una delle principali sfide, quindi, è quella di interrogarsi su l'adeguatezza a questo scopo di categorie giuridiche consolidate, per saggiarne la tenuta, verificarne l'elasticità ed eventualmente stabilire se occorra introdurne di nuove. Le libertà, i diritti e i doveri, come visto, sono uno degli ambiti posti particolarmente sotto stress dalla diffusione delle TRF. Da questo punto di vista, è sufficiente evidenziare alcune tendenze nella risposta degli ordinamenti alle nuove esigenze di garanzia.

Da una parte, le tecnologie in generale, e le TRF in particolare, inducono a privilegiare interpretazioni estensive di diritti e libertà ''classiche''. Si pensi nell'ordinamento italiano al caso paradigmatico della libertà personale, che tradizionalmente è stata intesa come *habeas corpus* e libertà dalle coercizioni fisiche (così fin da C.cost., sent.

n. 2/1956), ma anche come libertà morale da preservare contro forme di « degradazione giuridica dell'individuo » conseguenti all'« assoggettamento totale della persona all'altrui potere » (così fin da C.cost., sent. n. 11/1956). Mentre l'uso delle TRF non sembra porre problemi con questa libertà intesa nella prima accezione, maggiori dubbi sorgono con riguardo alla seconda accezione. Queste tecnologie, specie se usate in tempo reale, potrebbero ingenerare uno stato di 'assoggettamento' tale da scoraggiare determinate condotte, come la frequentazione di determinati luoghi o persone, e disincentivare l'esercizio legittimo di libertà garantite (il c.d. chilling effect) (SACCHETTO, 2025).

Sotto altri profili si avverte sempre più l'esigenza di ampliare la portata di diritti e principi esistenti. È il caso, ad esempio, dei principi di trasparenza e motivazione, da cui si è detto occorrerebbe ricavare un principio di 'comprensibilità' del linguaggio utilizzato dai sistemi di IA (SIMONCINI, 2023). Quest'ultimo principio si sostanzierebbe in un diritto a non essere sottoposti a decisioni prese da sistemi di IA incomprensibili e quindi a conoscere la motivazione di tali decisioni, riferita alla connessione logica tra elementi di fatto e valutativi, al loro peso relativo e a come sono impiegati i criteri di valutazione. Nel caso in cui la tecnologia venga utilizzata per prendere una decisione da parte di una autorità pubblica, tale principio costituirebbe un corollario del principio di legalità e buon andamento dell'amministrazione (art. 97 Cost.) o di obbligo di motivazione dei provvedimenti giudiziari (art. 111 Cost.). Rispetto all'uso delle TRF, si andrebbe incontro a un onere rafforzato di spiegazione circa il come e il perché un sistema sia giunto a effettuare un riconoscimento, offrendo così maggiori possibilità — ad esempio — di contestazione e di difesa in giudizio.

Infine, non manca anche chi sottolinea la necessità di introdurre nuovi diritti per difendersi contro questi poteri tecnologici di sorveglianza. A questo proposito, per tutelare le sfere più intime del pensiero e dei processi volitivi, è stato formulato il concetto di 'intellectual privacy', intesa come protezione dalla sorveglianza o da interferenze indesiderate quando si è impegnati in processi di generazione di idee o di formazione di credenze, ovvero quando pensiamo, leggiamo, parliamo con altri prima che le idee siano pronte per essere spese nel dibattito pubblico (RICHARDS, 2015). Ancor più radicalmente, è stato proposto di introdurre un diritto alla 'privacy mentale', inteso come protezione delle informazioni private o sensibili de-

sumibili dall'attività cerebrale di una persona nei confronti della raccolta, archiviazione, utilizzo o addirittura cancellazione non autorizzati (ANDORNO-IENCA, 2017). Si tratta di costruzioni teoriche che mirano a salvaguardare processi e informazioni che, ad oggi, non sembrano essere alla portata delle comuni TRF. Tuttavia, se si pensa ad alcune caratteristiche di queste ultime, come la capacità di combinare i dati biometrici ricavabili dai volti con i dati personali riconducibili alle persone sorvegliate, o di inferire molteplici informazioni dalla semplice analisi del volto, si intuisce come presto o tardi potrebbero emergere istanze di protezione del tutto analoghe.

La narrazione sull'introduzione di nuovi diritti, anche nel campo delle TRF, « può essere interpretata, da un lato, come un indice del crescente successo dell'idea stessa di ''diritto del soggetto'', ma, dall'altro lato, anche come segnale di un'inflazione del giuridico, di una sua indebita estensione ad ambiti di vita ad esso refrattari » (COSTA, 2008).

5. Conclusioni.

Dall'uso delle TRF deriva un potere di sorveglianza formidabile, che può essere speso a beneficio dell'interesse pubblico, ma che può mettere seriamente a repentaglio i diritti e la tenuta degli ordinamenti democratici. Nello svolgere la missione di limitare e condizionare questo potere, il diritto è chiamato a regolare un tipo di tecnologie di IA estremamente versatili in termini di potenzialità applicative, contesti di utilizzo, possibile ampiezza del loro dispiegamento; tutte variabili dai quali dipende anche l'impatto su principi giuridici e valori da preservare.

L'AI act introduce per la prima volta una disciplina giuridica espressamente rivolta a queste tecnologie. Si tratta di un passo in avanti sotto alcuni aspetti specifici (ad esempio la qualità dei dati processati o la sicurezza dei sistemi) ma, allo stesso tempo, questo regolamento manifesta una impostazione troppo rigida. L'AI Act distingue tra identificazione, categorizzazione e verificazione, ponendo regole che contengono divieti e prescrizioni, ma non coglie fino in fondo i rischi derivanti da queste tecnologie biometriche, che vengono principalmente disciplinate come un 'prodotto', trascurando di considerare le variabili sopra menzionate.

In futuro occorrerà stabilire come le norme sulle TRF contenute nell'AI Act si intersechino con la disciplina dedicata ai singoli ambiti

in cui possono trovare applicazione (es. sicurezza pubblica, trasporti, salute) o ai soggetti verso cui sono rivolte (es. i minori). A questa difficoltà si aggiungono le incertezze derivanti da quelle controspinte che premono per rallentarne l'attuazione, a partire da quelle originate dal contesto geopolitico (specie nel rapporto con gli Stati Uniti) e dalle resistenze offerte dalle imprese che sviluppano queste tecnologie (Haeck, 2025). Ciascuno Stato poi è chiamato a implementare queste regole nel proprio ordinamento, con l'incognita di rendere ancor più frammentato il quadro normativo.

Solo muovendosi dentro questa complessità, e sfruttando tutti gli strumenti regolativi a disposizione, a partire dalla tecnologia stessa, il diritto sarà in grado di soddisfare le nuove istanze ed esigenze d tutela derivanti dalla diffusione di tecnologie di questo tipo.

Note bibliografiche

Acquisti A.-Gross R.-Stutzman F. (2014), *Face recognition and privacy in the age of augmented reality*, in *Journal of Privacy and Confidentiality*, 6, 2, p. 1 ss.

Almada M.-Petit N. (2025), *The EU AI Act: Between the rock of product safety and the hard place of fundamental rights*, in *Common Market Law Review*, 62, 1, p. ss.

Andersin E. (2025), *The Use of the 'Lavender' in Gaza and the Law of Targeting: AI-Decision Support Systems and Facial Recognition Technology*, in *Journal of International Humanitarian Legal Studies*, p. 1 ss.

Barrett L.F.-Adolphs R.-Marsella S.-Martinez A.M.-Pollak S.D. (2019), *Emotional expressions reconsidered: challenges to inferring emotion from human facial movements*, in *Psychological Science in the Public Interest*, 20, 1, p. 1 ss.

Berle I. (2020), *Face Recognition Technology. Compulsory Visibility and Its Impact on Privacy and the Confidentiality of Personal Identifiable Images*, Springer, Cham.

Clarke R., *Information Technology and Dataveillance*, in *Communications of ACM*, 5, 31, May 1988, p. 498 ss.

Costa P. (2008), *Diritti fondamentali (storia)*, in *Enciclopedia del Diritto*, Annali II, t. 2, p. 365 ss.

EDPB - EDPS (2021), *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*.

EDPB (2024), *Parere 11/2024 sull'uso del riconoscimento facciale per snellire il flusso dei passeggeri negli aeroporti, (compatibilità con l'articolo 5, paragrafo 1, lettere e) e f), e gli articoli 25 e 32 GDPR)*, 1.1, 23 maggio 2024.

FRA (2019), *Facial recognition technology: fundamental rights considerations in the context of law enforcement*, 27 November 2019, accessibile al

sito: *https://fra.europa.eu/en/publication/2019/facial-recognition-technology-fundamental-rights-considerations-context-law*.

Fussey P.-Davies B.-Innes M. (2021), *'Assisted' Facial Recognition and the Reinvention of Suspicion and Discretion in Digital Policing*, in *The British Journal of Criminology*, 61, 2, p. 325 ss.

Haeck P. (2025), *EU readies for time out on enforcing AI rules*, in *Politico*, 23 September 2025, accessibile aal sito: www.politico.eu/article/eu-prepares-ground-pause-artificial-intelligence-rules/.

Iannuzzi A. (2024), *Le fonti del diritto dell'Unione europea per la disciplina della società digitale*, in Calzolaio S.-Iannuzzi A.-Longo E.-Orofino M.-Pizzetti F., *La regolazione europea della società digitale*, Torino, Giappichelli, p. 9 ss.

Ienca M.-Andorno R. (2017), *Towards new human rights in the age of neuroscience and neurotecnology*, in *Life Sciences Society and Policy*, 13, 5.

Koops B.-J. (2008), *Criteria for Normative Technology: The Acceptability of 'Code as law' in Light of Democratic and Constitutional Values*, in Brownsword R.-Yeung K. (eds.), *Regulating Technologies, Legal Futures, Regulatory Frames and Technological Fixes*, Oxford, Hart Publishing, p. 157 ss.

Kosinski M. (2021), *Facial recognition technology can expose political orientation from naturalistic facial images*, in *Scientific Reports*, p. 100 ss.

Kostka G.-Steinacker L.-Meckel M. (2023), *Under big brother's watchful eye: cross-country attitudes toward facial recognition technology*, in *Government Information Quarterly*, 40, 1, 101761.

Lessig L. (1999), *The Law of the Horse: What Cyberlaw Might Teach*, in *Harvard Law Review*, 113, p. 501 ss.

Liang F.-Das V.-Kostyuk N.-Hussain M.M. (2018), *Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure*, in *Policy and Internet*, 4, 10, p. 415 ss.

Longo E. (2024), *La disciplina del "rischio digitale"*, in Calzolaio S.-Iannuzzi A.-Longo E.-Orofino M.-Pizzetti F., *La regolazione europea della società digitale*, Torino, Giappichelli, p. 53 ss.

Menéndez Gonzáles N.-Mobilio G. (2025), *Between Prohibited Risks and High Risks: The Regulation of Facial Recognition Technology*, in Pollicino O.-Donati F., Finocchiaro G.-Paolucci F. (a cura di), *La disciplina dell'intelligenza artificiale*, Milano, GFL, p. 199 ss.

Mobilio G. (2021), *Tecnologie di riconoscimento facciale. Diritti fondamentali a rischio e sfide regolative*, Napoli, Editoriale Scientifica.

Mobilio G. (2025), *Facial Recognition Technologies: Threats or Opportunities for Democracy?*, in Menéndez Gonzáles N.-Mobilio G. (eds.), *Next Democratic Frontiers for Facial Recognition Technologies (FRT). The Legal, Ethical and Democratic Implications of FRT*, Springer, p. 13 ss.

Palmiotto F. (2025), *The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of the Regulation*, in *European Journal of Risk Regulation*, 16, p. 770 ss.

Richards N. (2015), *Intellectual Privacy. Rethinking Civil Liberties in the Digital Age*, Oxford University Press, Oxford.

SACCHETTO E. (2025), *Tecnologie di riconoscimento facciale e procedimento penale. Indagine sui fondamenti e sui limi dell'impiego della biometria moderna*, Torino, Giappichelli.

SIMONCINI A. (2023), *Il linguaggio dell'Intelligenza Artificiale e la tutela costituzionale dei diritti*, in *Rivista AIC*, 2.

VAVOULA N. (2020), *Databases for Non-EU Nationals and the Right to Private Life: Towards a System of Generalised Surveillance of Movement?*, in BIGNAMI F. (eds.), *EU Law in Populist Times*, Cambridge, Cambridge University Press, p. 227 ss.

INTELLIGENZA ARTIFICIALE E AMBITO MILITARE

INTRODUZIONE: SFIDE E PROSPETTIVE PER IL DIRITTO INTERNAZIONALE UMANITARIO

di *Monica Spatti*

Sommario: 1. Introduzione. — 2. Gli impieghi dell'intelligenza artificiale nelle operazioni militari. — 3. La compatibilità delle armi autonome con i principi fondamentali del diritto internazionale umanitario. — 4. La responsabilità giuridica in scenari a ridotto margine umano. — 5. Conclusioni.

1. Introduzione.

L'intelligenza artificiale (IA) è ormai parte integrante delle tecnologie militari contemporanee. Sistemi basati su algoritmi di apprendimento automatico sono impiegati, ad esempio, per analizzare immagini e dati satellitari, ottimizzare la logistica dei mezzi, supportare le decisioni dei comandanti. Accanto a questi impieghi, è in corso un ampio dibattito sull'eventuale utilizzo di armi autonome, ossia di sistemi capaci di selezionare e ingaggiare obiettivi — anche mediante l'uso della forza letale — senza un controllo umano costante. L'insieme di queste tecnologie, già presenti o in via di sviluppo, solleva questioni di grande rilievo per il diritto internazionale umanitario, chiamato a garantire che l'uso della forza resti conforme ai suoi principi fondamentali, anche in un contesto in cui il ruolo dell'essere umano tende a ridursi.

In questa prospettiva, la riflessione giuridica non può che muovere dal quadro normativo di riferimento, individuando i principi che delimitano la liceità dei mezzi e dei metodi di guerra. L'art. 35 del Primo Protocollo addizionale alle Convenzioni di Ginevra, relativo alla protezione delle vittime dei conflitti armati internazionali (1977), stabilisce che il diritto delle parti in conflitto di scegliere i mezzi e i metodi di guerra non è illimitato. Questo principio generale trova applicazione concreta in una serie di norme che costituiscono il nucleo del diritto internazionale umanitario: l'obbligo di distinguere in ogni momento tra obiettivi civili e militari (art. 48, integrato dagli artt. 51 e 52), il divieto di attacchi sproporzionati (art. 51, par. 5, lett. *b*) e l'obbligo di

adottare tutte le precauzioni possibili nella conduzione delle ostilità (art. 57). A queste disposizioni si aggiunge l'art. 36, che impone agli Stati di sottoporre a revisione ogni nuovo mezzo o metodo di guerra per verificarne la compatibilità con il diritto internazionale.

Questi principi — distinzione, proporzionalità e precauzione — costituiscono il quadro giuridico entro cui deve essere valutato l'impiego dell'intelligenza artificiale, sia quando essa funge da strumento di supporto decisionale, sia quando è integrata in sistemi d'arma dotati di capacità autonome. Essi rappresentano il nucleo di equilibrio tra necessità militare e umanità, e rimangono validi indipendentemente dal livello di complessità tecnologica del mezzo impiegato. La sfida principale consiste dunque nell'assicurare che l'innovazione tecnologica non svuoti di contenuto tali principi, ma ne permetta una più efficace attuazione.

Su questo sfondo, il presente contributo si propone di offrire una prima ricognizione delle principali questioni giuridiche connesse all'impiego dell'intelligenza artificiale in ambito militare. L'analisi si articola lungo tre direttrici principali: anzitutto è opportuno verificare come l'intelligenza artificiale è impiegata nelle operazioni militari diverse dalle armi autonome, per comprendere in che modo essa stia modificando la struttura del processo decisionale e la qualità delle valutazioni umane (par. 2); in seguito si intende affrontare la questione delle armi autonome, ossia dei sistemi capaci di selezionare e ingaggiare obiettivi senza un controllo umano costante, valutandone la compatibilità con i principi fondamentali del diritto internazionale umanitario (par. 3); infine, ci si soffermerà sul tema della responsabilità, individuale e statale, con particolare attenzione alle difficoltà di imputazione in scenari in cui l'autonomia tecnologica riduce o diluisce il margine umano (par. 4).

L'analisi che segue ha carattere introduttivo: mira a delineare i temi centrali del dibattito e a offrire alcune prime coordinate interpretative, ponendo le basi per un approfondimento successivo. I lavori che seguono si concentreranno, rispettivamente, sulla descrizione dell'effettivo impiego dell'intelligenza artificiale nelle operazioni militari e sull'analisi più dettagliata delle questioni giuridiche che l'autonomia tecnologica pone in materia di responsabilità.

2. Gli impieghi dell'intelligenza artificiale nelle operazioni militari.

L'intelligenza artificiale è ormai integrata in modo esteso nelle operazioni militari.

Essa contribuisce, in misura crescente, ai processi di analisi dei dati informativi e operativi, di pianificazione strategica e di gestione logistica, nonché alla protezione delle reti informatiche e alle attività di guerra elettronica. In particolare, sistemi di apprendimento automatico vengono impiegati per analizzare immagini satellitari e segnali elettromagnetici, riconoscere modelli di comportamento nelle comunicazioni nemiche, prevedere la mobilità di truppe o mezzi e ottimizzare il dispiegamento delle risorse sul campo (King, 2024). Algoritmi di *machine learning* e di integrazione di dati provenienti da fonti diverse consentono di combinare in tempo reale informazioni acquisite da sensori, droni, satelliti e reti di sorveglianza, offrendo ai comandanti un quadro operativo dinamico e continuamente aggiornato.

Questi sviluppi, come osservato in dottrina, non rappresentano soltanto un'evoluzione tecnica: incidono sulla natura stessa del processo decisionale militare, spostando progressivamente il baricentro dal giudizio diretto alla valutazione mediata da sistemi algoritmici (Bassok, 2020; Della Morte et al., 2024).

L'adozione di tali strumenti risponde a una logica di efficienza e di razionalizzazione delle decisioni operative. L'intelligenza artificiale consente, infatti, di elaborare e correlare in tempi ridottissimi quantità di informazioni che superano le possibilità cognitive dell'essere umano, riducendo l'incertezza e offrendo un vantaggio significativo nella pianificazione e nell'esecuzione delle missioni (King, 2024). In un contesto operativo dominato dalla velocità e dalla complessità informativa, la capacità di processare in tempo reale dati provenienti da fonti eterogenee appare decisiva per il successo delle operazioni.

L'IA è inoltre percepita come un mezzo per aumentare la precisione e ridurre il rischio di errore umano. Gli algoritmi di classificazione e previsione possono, almeno in linea teorica, migliorare l'identificazione degli obiettivi, minimizzare i danni collaterali e ottimizzare l'impiego delle risorse militari (Bassok, 2020; Della Morte et al., 2024). In tale prospettiva, la tecnologia sembra potenzialmente funzionale al rispetto degli obblighi imposti dal diritto internazionale umanitario, che mira a contenere gli effetti della violenza e a proteggere la popolazione civile.

Tuttavia, la maggiore efficienza operativa non coincide necessariamente con una maggiore aderenza ai principi del diritto internazionale umanitario. L'inserimento dell'IA nei processi decisionali militari introduce una serie di tensioni nuove, legate alla qualità dei dati, alla trasparenza degli algoritmi e alla possibilità che l'intervento umano perda spessore sostanziale. Le norme che regolano la condotta delle

ostilità — in particolare quelle sul principio di distinzione, sul divieto di attacchi sproporzionati e sull'obbligo di precauzione — presuppongono che le decisioni operative siano il risultato di valutazioni consapevoli e qualitativamente motivate (DINSTEIN, 2022). L'automazione crescente tende invece a tradurre tali giudizi qualitativi in processi quantitativi di calcolo, in cui la correttezza formale può non riflettere la sostanza delle valutazioni richieste dal diritto umanitario.

Questa trasformazione, ampiamente discussa in dottrina (BOOTHBY, 2016; SASSÒLI, 2019) non comporta di per sé una violazione del diritto dei conflitti armati, ma impone una revisione dei meccanismi attraverso cui il suo rispetto viene garantito. Le decisioni di impiego della forza non derivano più da un singolo atto di volontà, bensì da catene di elaborazioni in cui la selezione, la correlazione e la rappresentazione dei dati assumono un ruolo determinante (BASSOK, 2020). È stato osservato come la crescente mediazione algoritmica possa ridurre la trasparenza dei processi decisionali, rendendo più difficile verificare la correttezza delle valutazioni e la loro conformità ai principi di proporzionalità e distinzione (BRUUN-BO, 2025). Errori o distorsioni nei dataset di addestramento, *bias* cognitivi incorporati nei modelli o un'eccessiva fiducia nelle raccomandazioni algoritmiche (*automation bias*) possono compromettere la legittimità delle decisioni, producendo effetti sproporzionati o indiscriminati. In tali casi, la questione non riguarda tanto l'individuazione delle norme applicabili, quanto la capacità degli apparati militari di integrare tecnologie intelligenti in modo coerente con l'effettiva applicazione di quelle norme.

L'esigenza di mantenere una corrispondenza sostanziale tra l'efficienza tecnica e la liceità giuridica rappresenta oggi una delle sfide centrali per l'impiego dell'intelligenza artificiale nelle operazioni militari. La rapidità e la complessità delle analisi algoritmiche rischiano di ridurre lo spazio per l'apprezzamento umano, trasformando il controllo in un atto meramente formale. La domanda cruciale non è dunque se l'IA possa essere utilizzata, ma a quali condizioni il suo impiego possa rimanere compatibile con i principi di distinzione, proporzionalità e precauzione che costituiscono l'ossatura del diritto internazionale umanitario.

3. La compatibilità delle armi autonome con i principi fondamentali del diritto internazionale umanitario.

L'introduzione di sistemi d'arma dotati di autonomia decisionale rappresenta uno degli sviluppi più significativi e controversi dell'evo-

luzione tecnologica della guerra. Le cosiddette '*armi autonome*' si caratterizzano per la capacità di selezionare e colpire obiettivi senza un controllo umano costante, basandosi su processi di percezione, analisi e azione governati da algoritmi. L'espressione è impiegata in dottrina con accezioni diverse: in senso ampio, per indicare tutti i sistemi d'arma capaci di svolgere funzioni critiche in modo autonomo; in senso più ristretto, per riferirsi a tecnologie in grado di impiegare direttamente forza letale senza un intervento umano continuo (Amoroso, 2025). Questa innovazione segna una netta discontinuità rispetto ai sistemi automatizzati tradizionali, che operano entro parametri prestabiliti e sotto supervisione diretta: nelle armi autonome, la possibilità di adattarsi al contesto e di decidere in modo indipendente rende più incerta la linea di confine tra strumento e agente.

L'impiego di tali tecnologie solleva questioni profonde sul piano giuridico ed etico. Se da un lato l'autonomia promessa dalle armi autonome può aumentare la rapidità e la precisione delle operazioni, dall'altro lato essa mette alla prova la capacità del diritto internazionale umanitario di assicurare che i principi di distinzione, proporzionalità e precauzione — che presuppongono valutazioni umane contestuali e responsabili — trovino effettiva applicazione in un contesto in cui le decisioni di vita o di morte sono affidate, almeno in parte, a un sistema non umano. L'interrogativo di fondo non riguarda soltanto l'adeguatezza delle regole esistenti, ma la loro effettiva capacità di operare quando il ruolo dell'essere umano nel processo decisionale si riduce a una funzione di mera supervisione o di approvazione formale di scelte compiute da un algoritmo.

Il dibattito giuridico si concentra attorno a due direttrici strettamente connesse. La prima riguarda la compatibilità delle armi autonome con il diritto internazionale umanitario vigente, ossia la possibilità di garantire che sistemi capaci di agire senza un controllo umano costante rispettino i limiti fissati dalle Convenzioni di Ginevra e dai Protocolli aggiuntivi. La seconda concerne il ruolo che l'essere umano deve conservare nel processo decisionale, e in particolare la misura del controllo necessario affinché l'uso della forza resti imputabile e conforme al quadro giuridico esistente (Jensen, 2020).

La compatibilità delle armi autonome con il diritto internazionale umanitario è un tema particolarmente complesso. Le norme fondamentali in materia — in particolare, i già menzionati principi di distinzione, di proporzionalità e di precauzione — si fondano su presupposti cognitivi e valutativi che richiedono la presenza di un agente

umano capace di interpretare il contesto operativo e di modulare le proprie decisioni in base a circostanze mutevoli (Bruun-Bo, 2025). Comprendere se un sistema d'arma, per quanto sofisticato, possa sostituire questo tipo di discernimento rappresenta il nodo più controverso dell'intero dibattito.

Il principio di distinzione, sancito dagli articoli 48 e 51 del Primo Protocollo addizionale alle Convenzioni di Ginevra, impone alle parti in conflitto di distinguere in ogni momento tra obiettivi militari e persone o beni civili. Nella prassi, tale distinzione non dipende solo da un riconoscimento formale, ma da un insieme di valutazioni dinamiche legate al comportamento, all'ambiente e alle intenzioni probabili dei soggetti coinvolti — elementi difficili da tradurre in parametri algoritmici (Winter, 2020; Amoroso, 2025; Bruun-Bo, 2025). Gli attuali sistemi di intelligenza artificiale possono elaborare enormi quantità di dati visivi e termici, ma non comprendono realmente ciò che osservano: riconoscono schemi, non significati (King, 2024). Di conseguenza, l'autonomia nella selezione dei bersagli rischia di compromettere l'obbligo di distinzione, soprattutto in contesti complessi o in presenza di comportamenti ambigui.

Difficoltà analoghe emergono rispetto al principio di proporzionalità, che vieta attacchi suscettibili di causare danni collaterali eccessivi rispetto al vantaggio militare concreto e diretto previsto. Tale valutazione implica un giudizio qualitativo che non si riduce a un calcolo numerico: richiede una ponderazione tra valori diversi — la protezione dei civili e il conseguimento del vantaggio militare — che presuppone un apprezzamento etico e giuridico non replicabile da un algoritmo (Amoroso, 2025; Casey-Maslen, 2025). Come osservato in dottrina, l'idea che una macchina possa bilanciare interessi umani implica un livello di razionalità che le attuali tecnologie non possiedono (Crootof, 2016; Garcia, 2023).

Il terzo principio, quello di precauzione, previsto dall'articolo 57 del Primo Protocollo, obbliga a prendere tutte le misure concretamente attuabili per evitare o ridurre al minimo i danni ai civili e ai beni civili. Il rispetto di tale obbligo dipende dal livello di controllo effettivo sull'arma e dalla possibilità di intervenire per modificarne o interromperne l'azione in caso di errore. Nei sistemi pienamente autonomi, dove la selezione e l'ingaggio avvengono senza una supervisione continua, questa possibilità tende a ridursi fino quasi a scomparire (Jensen, 2020; Amoroso, 2025).

Le difficoltà di conciliare questi tre principi con l'autonomia decisionale delle macchine hanno portato parte della dottrina a ritenere che le armi autonome, almeno nella loro forma pienamente autonoma, siano incompatibili con il diritto internazionale umanitario vigente (ASARO, 2012; GARCIA, 2023). Altri autori adottano una posizione più prudente, sostenendo che la compatibilità non debba essere esclusa in astratto ma valutata caso per caso, in base al grado di autonomia effettiva, alla prevedibilità del comportamento del sistema e alla possibilità di un intervento umano tempestivo (WINTER, 2022; CASEY-MASLEN, 2025). In entrambi gli approcci emerge un punto comune: la centralità del controllo umano significativo, inteso non come mera facoltà di attivare o disattivare il sistema, ma come reale capacità di comprendere, monitorare e, se necessario, correggere il processo decisionale dell'arma (ASARO, 2012).

L'esigenza di garantire un controllo umano sostanziale riflette, in ultima analisi, la natura antropocentrica del diritto internazionale umanitario, che fonda la liceità dell'uso della forza sulla capacità dell'essere umano di valutare, scegliere e assumere decisioni responsabili. La progressiva delega di tali funzioni alle macchine non solleva quindi soltanto questioni tecniche di conformità, ma interroga la tenuta stessa delle categorie giuridiche e valoriali su cui si basa la regolazione della guerra.

4. La responsabilità giuridica in scenari a ridotto margine umano.

La questione della responsabilità rappresenta uno dei nodi più complessi del dibattito sull'impiego dell'intelligenza artificiale in ambito militare. Il diritto internazionale umanitario definisce gli obblighi di condotta nei conflitti armati, ma la loro violazione acquista rilievo giuridico solo in quanto possa essere attribuita a un soggetto, individuale o statale. L'introduzione di sistemi d'arma capaci di operare con un certo grado di indipendenza decisionale mette in discussione proprio le modalità di questa attribuzione (CROOTOF, 2016). La difficoltà non riguarda soltanto l'individuazione del responsabile materiale della violazione, ma anche la possibilità stessa di collegare un comportamento illecito a un soggetto giuridico quando parte del processo decisionale non è più integralmente umano.

Nel modello classico, l'illecito di guerra presuppone una catena decisionale chiaramente definita: un ordine impartito da un superiore, eseguito da un subordinato, e controllato da strutture di co-

mando soggette a responsabilità penale o disciplinare (McFarland, 2020). L'impiego di armi autonome altera questa architettura, introducendo un elemento non umano che partecipa attivamente alla selezione e all'ingaggio del bersaglio. La macchina non è più un semplice strumento passivo, ma un agente tecnico che incide in modo determinante sull'esito dell'azione. Ne risulta un indebolimento del nesso tradizionale tra volontà, azione e imputazione, con il rischio di creare quello che parte della dottrina ha definito *accountability gap* (McFarland, 2020; Matthias, 2024).

Le difficoltà di attribuzione si manifestano su due piani distinti ma strettamente connessi: quello della responsabilità internazionale dello Stato e quello della responsabilità penale individuale. Sul piano statale, l'articolo 4 del Progetto di articoli sulla responsabilità degli Stati per atti internazionalmente illeciti stabilisce che ogni condotta compiuta da organi statali o da persone che agiscono per conto dello Stato è imputabile a quest'ultimo. Tuttavia, quando un sistema d'arma agisce in modo parzialmente autonomo, la condotta illecita non può essere facilmente ricollegata a un comportamento umano identificabile. In linea di principio, lo Stato rimane responsabile per gli atti compiuti mediante strumenti sotto il suo controllo, ma la complessità e l'opacità dei meccanismi decisionali automatizzati rendono più difficile dimostrare la violazione degli obblighi di diligenza nella progettazione, nella sperimentazione e nell'impiego operativo (Crootof, 2016; Boutin, 2023; Gaeta, 2023).

Sul piano individuale, la questione si presenta con tratti ancora più problematici. Il diritto penale internazionale, come codificato nello Statuto di Roma della Corte penale internazionale, fonda la responsabilità sulla partecipazione personale a un atto illecito e sulla consapevolezza della sua natura criminale. Tuttavia, quando la decisione di attacco o l'identificazione dell'obiettivo derivano da un processo algoritmico, il comandante o l'operatore possono trovarsi nella posizione paradossale di rispondere per azioni che non hanno effettivamente determinato né potuto prevedere (Amoroso, 2020; Gaeta, 2023). Il principio di *mens rea*, che presuppone la coscienza e la volontà del fatto, rischia così di perdere consistenza (Crootof, 2019).

Parte della dottrina ha proposto di affrontare queste difficoltà ricorrendo a criteri di responsabilità per negligenza, fondati sull'obbligo di supervisione e sulla previsione dei rischi connessi all'uso di sistemi autonomi (Crootof, 2019; Amoroso, 2020; McFarland,

2020). In tale prospettiva, l'imputazione non deriverebbe dalla decisione della macchina, ma dall'omissione umana di predisporre adeguate misure di controllo, test e revisione delle capacità operative dell'arma. Si tratterebbe, dunque, di una responsabilità indiretta, costruita attorno a un dovere di vigilanza proporzionato al livello di indipendenza operativa del sistema.

Approcci analoghi emergono anche sul piano della responsabilità statale, come mostrano i documenti del Comitato internazionale della Croce Rossa e alcune posizioni governative. L'idea di fondo è che la delega tecnologica non attenui gli obblighi internazionali di controllo e revisione previsti dall'articolo 36 del Primo Protocollo addizionale alle Convenzioni di Ginevra. Lo Stato resta tenuto a garantire che ogni nuova tecnologia sia sottoposta a una valutazione preventiva di conformità e che la sua integrazione nelle strutture militari non comprometta l'osservanza del diritto internazionale umanitario (Crootof, 2016; Boutin, 2023).

Sul piano del diritto internazionale penale, la stessa logica di diligenza preventiva trova un parallelo nel principio di responsabilità del comandante, sancito dall'articolo 28 dello Statuto di Roma, che punisce l'omessa adozione di misure idonee a prevenire o reprimere i crimini commessi dalle forze sottoposte al proprio controllo effettivo. Quando, tuttavia, il 'subordinato' è un sistema d'arma autonomo, questa struttura concettuale entra in crisi: il comandante non dispone più di un controllo diretto sulle singole decisioni algoritmiche e non può prevederne tutte le possibili combinazioni operative. Ciò nonostante, la dottrina tende a escludere un'esenzione totale di responsabilità, suggerendo piuttosto una reinterpretazione funzionale dell'obbligo di controllo (Jensen, 2020; Spadaro, 2023). In quest'ottica, la colpa non deriva dall'evento lesivo in sé, ma dall'omissione di predisporre procedure adeguate di verifica, aggiornamento e supervisione continua del sistema, proporzionate al suo livello di autonomia.

Questa prospettiva consente di preservare la coerenza del diritto internazionale: l'attribuzione della responsabilità deve restare ancorata a soggetti umani e statali, evitando di creare zone d'irresponsabilità e salvaguardando al contempo il carattere antropocentrico dell'imputazione giuridica. La sfida, dunque, non consiste nel ridefinire il concetto di responsabilità, ma nell'adattarne le modalità di applicazione a un contesto in cui l'azione militare è sempre più mediata da sistemi algoritmici complessi.

5. Conclusioni.

L'analisi condotta mostra come l'intelligenza artificiale stia progressivamente trasformando la dimensione operativa dei conflitti armati, incidendo non solo sugli strumenti della guerra, ma anche sulla struttura stessa del processo decisionale militare. Il diritto internazionale umanitario, concepito in un'epoca in cui la decisione sull'uso della forza era esclusivamente umana, si trova oggi di fronte alla necessità di misurarsi con forme di autonomia tecnologica che ridefiniscono i confini della responsabilità e della liceità dell'azione militare.

Né l'evoluzione tecnologica né l'adattamento delle prassi operative possono però intaccare il nucleo dei principi fondamentali del diritto umanitario: distinzione, proporzionalità e precauzione. Questi principi non dipendono dalla natura del mezzo impiegato, ma dalla capacità di assicurare che la condotta bellica resti orientata alla tutela dell'uomo e alla limitazione della sofferenza. La crescente presenza di sistemi algoritmici nelle decisioni militari non riduce la necessità di una valutazione di conformità al diritto internazionale umanitario: ciò che conta non è la correttezza tecnica del mezzo, ma il modo in cui esso viene integrato nei processi decisionali e nei sistemi d'arma. L'elemento determinante rimane la misura del controllo umano effettivo: solo dove tale controllo permane in modo sostanziale è possibile assicurare che l'impiego dell'intelligenza artificiale resti compatibile con il diritto umanitario. In assenza di una supervisione significativa, l'autonomia tecnologica rischia di tradursi in una deresponsabilizzazione diffusa, minando la funzione stessa del diritto dei conflitti armati come strumento di limitazione della violenza.

Come ha ricordato Papa Francesco nel *Messaggio per la 57ª Giornata mondiale della pace*, non si tratta di rifiutare la tecnologia, ma di evitare che essa diventi « un fattore di distruzione » quando sottrae all'uomo la possibilità di decidere in coscienza. Il diritto, in questa prospettiva, non può limitarsi a regolare gli strumenti: deve riaffermare il primato della persona e la centralità del giudizio morale anche in un contesto dominato dall'automazione e dal calcolo. L'appello pontificio si inserisce così nel medesimo orizzonte del diritto internazionale umanitario, che trova la sua ragion d'essere nella dignità umana e nella responsabilità individuale.

In definitiva, l'intelligenza artificiale non deve segnare una frattura con il diritto internazionale umanitario, ma semmai costituire un banco di prova della capacità di quest'ultimo di adattarsi alle innova-

zioni tecnologiche senza smarrire la propria identità antropocentrica. La tecnologia può contribuire a rendere più precisa e meno distruttiva l'azione militare, ma solo se rimane strumento — e non sostituto — del discernimento umano. La compatibilità tra innovazione tecnologica e diritto umanitario dipende, in ultima analisi, dal mantenimento di questa gerarchia: l'uomo come fonte del giudizio, la macchina come mezzo del suo esercizio.

Note bibliografiche

Amoroso D. (2020), *Autonomous Weapons Systems and International Law. A Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains*, Baden-Baden, Nomos.

Amoroso D. (2025), *La disciplina dei sistemi d'arma autonomi nel diritto internazionale umanitario. Stato dell'arte e prospettive di regolamentazione*, in *Rivista italiana di informatica e diritto*, 2, p. 1 ss.

Asaro P. (2012), *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, in *International Review of the Red Cross*, 886, p. 687 ss.

Bassok O. (2020), *The Soldier as an Autonomous Weapon*, in *Cornell Law Review*, p. 233 ss.

Boothby W.H. (2016), *The Law of Targeting*, Oxford, Oxford University Press.

Boutin B. (2023), *State responsibility in relation to military applications of artificial intelligence*, in *Leiden Journal of International Law*, p. 133 ss.

Bruun L.-Bo M. (2025), *Bias in military artificial intelligence and compliance with international humanitarian law*, Stockholm International Peace Research Institute (Sipri), Stockholm.

Casey-Maslen S. (2025), *Autonomous Weapons Systems Under International Law*, in Talves K.-Spreen D. (eds.), *Artificial Intelligence in Military Technology. Social, cultural and ethical perspectives*, Cham, Spinger, p. 161 ss.

Crootof R. (2016), *Wars Torts: Accountability for Autonomous Weapons*, in *University of Pennsylvania Law Review*, 6, p. 1347 ss.

Della Morte G. et al. (2024), *Algorithms and Decision-Making in the Metaverse: Legal and Ethical Grounds to Trust AI Information for the Decision-Making Process within the Context of Armed Conflict*, in Riva G.-Villani D. (eds.), *Humane Metaverse. Reflections of Self, Education, Organizations and Society*, Milano, Vita & Pensiero, p. 289 ss.

Dinstein Y. (2022), *The Conduct of Hostilities under the Law of International Armed Conflict*, Cambridge, Cambridge University Press, III ed.

Gaeta P. (2023), *Who Acts When Autonomous Weapons Strike? The Act Requirement for Individual Criminal Responsibility and State Responsibility*, in *International Journal of Criminal Justice*, 5, p. 1033 ss.

Garcia D. (2023), *The AI Military Race. Common Good Governance in the Age of Artificial Intelligence*, Oxford, Oxford University Press.

JENSEN E.T. (2020), *Autonomy and Precautions in the Law of Armed Conflicts*, in *International Law Studies*, 1, p. 577 ss.
KING A. (2024), *Digital Targeting: Artificial Intelligence, Data, and Military Intelligence*, in *Journal of Global Security Studies*, 2, p. 1 ss.
MATTHIAS A. (2004), *Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, in *Ethics and Information Technology*, p. 175 ss.
MCFARLAND T. (2020), *Autonomous Weapon Systems and the Law of Armed Conflict*, Cambridge, Cambridge University Press.
PAPA FRANCESCO, *Messaggio per la 57ª Giornata Mondiale della Pace. Intelligenza artificiale e pace*, 1° gennaio 2024, disponibile su *www.vatican.va*.
SASSÒLI M. (2019), *International Humanitarian Law. Rules, Controversies, and Solutions to Problems Arising in Warfare*, Cheltenham, Edward Elgar Publishing.
SPADARO A. (2023), *A Weapon is No Subordinate. Autonomous Weapon Systems and the Scope of Superior Responsibility*, in *Journal of International Criminal Justice*, p. 1119 ss.
WINTER E. (2020), *The Compatibility of Autonomous Weapons with the Principle of Distinction in the Law of Armed Conflict*, in *International and Comparative Law Quarterly*, p. 845 ss.
WINTER E. (2022), *The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law*, in *Journal of Conflict and Security Law*, 1, p. 1 ss.

(IL FATTO)
INTELLIGENZA ARTIFICIALE, CONFLITTI ARMATI E DIRITTO: PROFILI TECNICI ED ETICI

di *Riccardo Ricchetti* e *Francesca Sironi De Gregorio*

Sommario: 1. Introduzione. — 2. Profili tecnici dell'IA in ambito militare. — 2.1. Definizione e concetti di base. — 2.2. Applicazioni attuali in ambito militare. — 2.3. Gradi di autonomia nell'interazione uomo-macchina. — 2.4. Guerra cognitiva. — 3. Raccordo con il quadro giuridico internazionale. — 4. Questioni etiche e regolatorie di particolare rilievo. — 4.1. Il controllo umano significativo. — 4.2. *Bias* e opacità algoritmica. — 4.3. Il principio di proporzionalità nel dominio cibernetico. — 5. Osservazioni conclusive.

1. Introduzione.

Nell'ultimo decennio, l'impiego dell'intelligenza artificiale (IA) nel settore della sicurezza e della difesa ha registrato uno sviluppo straordinario, trasformando modalità operative, processi decisionali e tempi di reazione. Il settore militare è da sempre terreno fertile per l'innovazione tecnologica. Dall'introduzione dei mezzi aerei all'inizio del secolo scorso, allo sviluppo delle armi nucleari negli anni Quaranta, fino agli armamenti di precisione guidata degli ultimi decenni, ogni salto tecnologico ha ridefinito le regole del confronto bellico. L'adozione di sistemi di IA segna tuttavia una cesura qualitativa rispetto alle innovazioni precedenti: non si tratta, infatti, soltanto di armi più potenti o più precise, bensì di sistemi capaci di elaborare informazioni, apprendere dall'esperienza e, in alcuni casi, agire in assenza di un intervento umano diretto.

Il principale elemento di novità risiede proprio in questa ridefinizione del rapporto tra uomo e macchina nel contesto bellico. A differenza di un missile teleguidato o di un drone a controllo remoto, in cui l'essere umano rimane l'agente decisionale, i sistemi di IA moderni possono selezionare obiettivi, pianificare manovre e rispondere

a situazioni impreviste entro parametri definiti in sede di progettazione, ma con gradi di autonomia operativa crescenti. Comprendere come funzionano questi sistemi, in quali ambiti sono già dispiegati e secondo quale architettura di interazione uomo-macchina operano interagiscono con gli operatori umani è il presupposto indispensabile per qualsiasi valutazione giuridica ed etica fondata.

Il presente contributo si propone di offrire tale cornice tecnica e fattuale, che costituisce la premessa necessaria alle analisi giuridiche successive. L'analisi si articola secondo tre direttrici: *i*) una ricognizione tecnico-scientifica del funzionamento e dell'impiego dei sistemi di IA in ambito militare; *ii*) un sintetico raccordo con il quadro giuridico internazionale applicabile, rimandando per l'analisi approfondita ai contributi specialistici presenti nel volume; *iii*) una riflessione sulle implicazioni etiche e regolatorie che discendono direttamente dalle caratteristiche tecniche dei sistemi, in particolare il problema del controllo umano e quello dei *bias* algoritmici.

2. Profili tecnici dell'IA in ambito militare.

2.1. *Definizione e concetti di base.*

Per «*sistema di IA*» si intende — secondo la definizione fornita dal Regolamento (UE) 2024/1689, noto come *AI Act* — un «*sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può: i) presentare adattabilità dopo la diffusione; ii) generare output quali previsioni, contenuti, raccomandazioni o decisioni dall'input che riceve per obiettivi espliciti o impliciti*». Affinché si possa parlare di sistema di IA in senso proprio occorre trovarsi di fronte a una macchina dotata di autonomia, ossia capace di funzionare in assenza di intervento umano continuo, alla quale sono affidati obiettivi espliciti o impliciti e che è capace di elaborare dati per produrre un *output* significativo.

Dal punto di vista tecnico, i sistemi militari di IA si basano principalmente su tre architetture fondamentali. La prima è il *machine learning* classico (ML), in cui algoritmi statistici vengono addestrati su grandi dataset etichettati per riconoscere *pattern* e fare previsioni. Tale tecnologia è alla base, ad esempio, dei sistemi di identificazione automatica di veicoli o uniformi in immagini satellitari. La seconda è il *deep learning* (DL), basato su reti neurali artificiali a più strati che consentono l'elaborazione di *input* complessi — immagini, suoni, segnali radar — con livelli di accuratezza superiori al ML classico, ma

con una minore interpretabilità dei processi interni. La terza è il *reinforcement learning* (RL), in cui il sistema impara attraverso l'interazione con un ambiente simulato, massimizzando una funzione di ricompensa. A titolo esemplificativo, il RL è impiegato per addestrare sistemi di navigazione autonoma e per sviluppare tattiche di manovra in ambienti complessi.

L'*AI Act*, pur non essendo concepito per l'ambito militare — il cui perimetro è esplicitamente escluso dall'ambito di applicazione del Regolamento — introduce strumenti metodologici utili anche per la valutazione dei sistemi militari: il ciclo di vita del sistema dalla progettazione al monitoraggio continuo, le operazioni di *red-teaming* in ottica di prevenzione del rischio, e il principio di una IA «affidabile» (*trustworthy AI*).

2.2. *Applicazioni attuali in ambito militare.*

Nel dominio militare, l'applicazione dei sistemi di IA interessa aree operative distinte, con gradi di maturità tecnologica molto differenti tra loro.

La prima riguarda la logistica assistita e la manutenzione predittiva. L'analisi in tempo reale di vibrazioni, temperature, segnali elettrici e telemetrie dei sistemi d'arma consente di stimare la vita residua dei componenti critici, programmare i ricambi e ottimizzare le catene di approvvigionamento. L'impiego di algoritmi di apprendimento automatico in questo settore consente di ridurre significativamente i fermi non programmati, con ricadute dirette sulla disponibilità operativa dei mezzi. La qualità dei dati, la robustezza dei modelli e l'integrazione con i sistemi di supporto alla decisione si rivelano fattori determinanti: un algoritmo poco addestrato o alimentato con dati lacunosi moltiplica i falsi allarmi; al contrario, un modello adeguatamente istruito fornisce risultati operativamente utili (ABADICIO, 2019; MORENO, 2010).

La seconda applicazione concerne l'analisi dell'intelligence e la fusione dei dati. I sistemi di IA consentono di correlare in tempo reale informazioni provenienti da fonti eterogenee — immagini satellitari, intercettazioni elettroniche, dati da droni, segnalazioni umane (HUMINT) — producendo quadri operativi dinamici e continuamente aggiornati. Il programma Project Maven del Dipartimento della Difesa statunitense, avviato nel 2017, impiega algoritmi di *com-*

puter vision per analizzare automaticamente le riprese video dei droni, identificare oggetti di interesse e classificare i *pattern* di movimento, riducendo il carico analitico degli operatori e accelerando i tempi di risposta. Analogamente, il sistema israeliano *Fire Factory*, impiegato durante le operazioni a Gaza a partire dal 2023, è stato progettato per raccomandare automaticamente sequenze di *targeting* e allocazione delle munizioni, elaborando simultaneamente un elevato numero di obiettivi potenziali (Della Morte et al., 2024).

La terza consiste nell'addestramento immersivo e nella simulazione virtuale. Le nuove tecnologie hanno progressivamente affiancato e in parte sostituito le esercitazioni fisiche con ambienti di addestramento sintetici che impiegano gemelli digitali (*digital twins*) ad alta fedeltà del teatro operativo. La possibilità di costruire scenari coerenti tra i diversi livelli tattici e operativi, di far convergere addestramento e *mission rehearsal* e di valutare *ex ante* le conseguenze di determinate dinamiche costituisce un valore aggiunto rispetto alla simulazione tradizionale. L'utilizzo di algoritmi di IA permette di rendere più realistici questi ambienti lavorando sul fattore imprevedibilità dell'avversario sintetico, il quale apprende e adatta le proprie tattiche nel corso dell'esercitazione, riducendo al contempo i costi rispetto alle esercitazioni sul terreno.

La quarta area di applicazione è il *targeting* assistito da IA. I sistemi di *Automatic Target Recognition* (ATR), alimentati da sensori elettro-ottici e infrarossi, da radar ad apertura sintetica (SAR) o dalla raccolta di segnali elettromagnetici (SIGINT), mirano a riconoscere *pattern* di interesse, isolare bersagli in ambienti complessi e ridurre il carico cognitivo dell'operatore (Della Morte et al., 2024). Il sistema israeliano Harop, il Kargu-2 turco e il Lancet-3 russo rappresentano esempi di munizioni loitering; il Bayraktar TB2, invece, costituisce un esempio di UCAV con capacità di targeting e ingaggio assistite in questo contesto sono i limiti tecnici: scarsa illuminazione, occlusioni ottiche, mimetismo, rumori di fondo e il fenomeno del *domain shift* — ossia la divergenza tra le condizioni in cui il sistema è stato addestrato e quelle reali in cui opera — sono cause frequenti di errori di classificazione, con potenziali conseguenze sul piano della legalità e della proporzionalità (Schachter, 2020; Jain, 2023).

Il quinto ambito risiede nel dominio cibernetico. L'utilizzo di sistemi di IA nel cyberspazio opera in modo tendenzialmente simmetrico sul versante difensivo e su quello offensivo. Sul piano difensivo, consente il *triage* e la correlazione automatica di eventi anomali, la

prioritizzazione degli incidenti e l'attivazione di risposte predefinite, con tempi di reazione nell'ordine dei millisecondi. Sul piano offensivo, l'IA consente di scalare le operazioni, personalizzare i *payload*, sfruttare il *bug-hunting* automatizzato, ossia l'identificazione automatica di vulnerabilità nei sistemi bersaglio, ed eseguire manovre laterali all'interno di reti compromesse con velocità e parallelismo impossibili per un operatore umano.

2.3. *Gradi di autonomia nell'interazione uomo-macchina.*

Tra le applicazioni dell'IA che pongono le maggiori problematiche, tecniche prima ancora che giuridiche, vi è la categoria delle armi autonome. Con questa espressione si indicano sistemi d'arma che, attraverso sofisticati algoritmi di percezione e decisione, selezionano e ingaggiano obiettivi in assenza totale o parziale di intervento umano. Il fulcro tecnico della questione coincide con la distinzione tra tre modelli di interazione uomo-macchina, che corrispondono a livelli crescenti di autonomia operativa e a configurazioni radicalmente diverse della catena di responsabilità (Ford, 2017; Cicr, 2019).

Nel modello *human-in-the-loop*, l'essere umano rimane agente decisionale centrale e insostituibile: tutte le decisioni critiche — selezione del bersaglio, autorizzazione all'ingaggio — restano nelle sue mani. La macchina raccoglie dati, elabora opzioni e presenta raccomandazioni, ma non agisce senza l'assenso esplicito dell'operatore. Dal punto di vista tecnico, questo schema presuppone latenze di comunicazione compatibili con i tempi umani di valutazione, il che lo rende inadatto agli scenari ad alta velocità dove i tempi di reazione richiesti sono nettamente superiori (Lawand, 2013; Ford, 2017).

Nel modello *human-on-the-loop*, il sistema agisce autonomamente entro parametri predefiniti, mentre l'operatore mantiene una supervisione continua con facoltà di intervento (Sharkey, 2016; Cicr, 2019). Questo è il modello adottato, ad esempio, dal sistema di difesa antimissile *Phalanx CIWS* (*Close-In Weapon System*) della marina statunitense, la quale individua e intercetta automaticamente proiettili e missili in avvicinamento senza che un operatore approvi ogni singolo ingaggio, ma con un supervisore umano in grado di disattivare il sistema. Analogamente, il sistema *Iron Dome* israeliano seleziona automaticamente i proiettili da intercettare in base a calcoli di traiettoria e probabilità di impatto su aree abitate, con un

operatore umano che può bloccare l'ingaggio. La sfida tecnica principale di questo modello è il cosiddetto *automation bias*, vale a dire la tendenza degli operatori ad accettare acriticamente le raccomandazioni del sistema riducendo di fatto la supervisione a una ratifica formale.

Nel modello *human-out-of-the-loop*, il sistema d'arma agisce con piena autonomia senza possibilità di intervento esterno durante l'operazione (Bo-Bruun-Boulanin, 2022; Mauri, 2022). I cosiddetti *Fully Autonomous Weapon Systems* (FAWS) non sono ancora pienamente dispiegati in configurazione operativa certificata da alcuna delle principali potenze militari, ma esistono sistemi che vi si avvicinano significativamente. Le munizioni *loitering*, come il *Lancet-3* russo o l'*STM Kargu-2* turco — quest'ultimo segnalato dal *Panel of Experts* delle Nazioni Unite nel 2021 come potenzialmente impiegato in modalità autonoma nel conflitto libico (UN Panel of Experts 2021) — vengono lanciate verso un'area bersaglio, perlustrando autonomamente lo spazio aereo e colpendo obiettivi corrispondenti a profili preimpostati senza comunicazione con l'operatore durante la fase finale di ingaggio. La difficoltà tecnica centrale rimane la gestione degli ambienti dinamici e ambigui: distinguere affidabilmente un combattente armato da un civile in condizioni di illuminazione variabile e scenari mai incontrati in addestramento è un problema che i sistemi attuali non risolvono in modo soddisfacente.

2.4. *Guerra cognitiva.*

Prima di affrontare i profili giuridici, è necessario richiamare come l'evoluzione tecnologica stia aprendo scenari di conflitto che trascendono i confini tradizionali del campo di battaglia. Le tecnologie di generazione sintetica di contenuti, in particolare i modelli linguistici di grandi dimensioni (LLM) e le reti generative avversarie (GAN) per la produzione di immagini, audio e video, hanno abbassato drasticamente la soglia di accesso a strumenti capaci di produrre disinformazione su scala industriale.

La *cognitive warfare*, o guerra cognitiva, designa operazioni ostili rivolte non contro infrastrutture fisiche, ma contro gli ecosistemi informativi delle società democratiche, attraverso la diffusione di contenuti falsi o ingannevoli con l'obiettivo di influenzare percezioni, opinioni e comportamenti collettivi senza ricorrere alla violenza fisica (Lahman-Custer-Scott, 2025). Le tecnologie di IA abilitano in

questo contesto tre capacità specificamente nuove rispetto alla propaganda tradizionale: la produzione automatizzata di contenuti su scala massiva; la personalizzazione del messaggio in base al profilo psicologico del destinatario, sfruttando i dati disponibili sulle piattaforme digitali; e la generazione di contenuti multimediali sintetici (*deepfake*) difficilmente distinguibili dall'autentico.

Dinamiche di questo tipo si sono manifestate concretamente nel conflitto russo-ucraino a partire dal 2022, con la diffusione di video sintetici — tra cui un caso documentato di contenuto artificialmente generato raffigurante il presidente Zelensky nell'atto di ordinare la resa — e con campagne coordinate di amplificazione attraverso reti di account automatizzati (*bot farm*). Analogamente, durante le operazioni a Gaza del 2023-2024, contenuti visivi manipolati sono stati prodotti e diffusi da più parti con l'obiettivo di influenzare l'opinione pubblica internazionale. La difficoltà di rilevamento e la velocità di propagazione di questi contenuti pongono questioni inedite tanto sul piano giuridico — la categorizzazione come atto di guerra, il problema dell'attribuzione, la tutela dei diritti fondamentali — quanto sul piano tecnico della difesa cognitiva, che richiede sistemi di verifica dell'autenticità dei contenuti (*content authentication*) altrettanto sofisticati (LAHMAN-CUSTER-SCOTT, 2025).

3. Raccordo con il quadro giuridico internazionale.

Le applicazioni tecniche descritte nella sezione precedente non si collocano in un vuoto normativo. Il diritto internazionale umanitario, in particolare nelle sue fonti convenzionali — a partire dalle Convenzioni di Ginevra del 1949 e dai Protocolli aggiuntivi del 1977 — e nelle sue norme consuetudinarie, si applica integralmente all'impiego di sistemi di IA nei conflitti armati, indipendentemente dal livello di autonomia del sistema coinvolto. I principi fondamentali di distinzione, proporzionalità, precauzione, umanità e necessità militare vincolano le parti in conflitto nella stessa misura in cui lo farebbero per qualsiasi altro mezzo o metodo di guerra.

L'analisi approfondita di come questi principi si declinino concretamente rispetto ai sistemi d'arma autonomi è oggetto degli altri contributi del presente volume, ai quali si rinvia.

Ciò che preme sottolineare in questa sede, come premessa tecnica alle analisi giuridiche, è che le difficoltà applicative del diritto internazionale umanitario nell'era dell'IA non derivano primariamente

da lacune normative, bensì dall'incontro tra norme che presuppongono un agente decisionale umano e sistemi tecnici il cui funzionamento interno è spesso opaco anche ai loro progettatori. È questa opacità — il fenomeno della « scatola nera » dei modelli di *deep learning* — il nodo tecnico che si riverbera direttamente sulle categorie giuridiche di responsabilità, prevedibilità e controllo.

4. Questioni etiche e regolatorie di particolare rilievo.

4.1. *Il controllo umano significativo.*

Il concetto di *meaningful human control* (MHC), o controllo umano significativo, è diventato il principale punto di convergenza nel dibattito internazionale sui sistemi d'arma autonomi. Dottrina, organizzazioni umanitarie e delegazioni statali in sede CCW concordano sull'esigenza di preservare un margine di intervento umano nelle decisioni critiche che coinvolgono l'uso della forza letale (Santoni De Sio-Van Den Hoven, 2018; Hua, 2019). Prima ancora di essere una questione giuridica o etica, tuttavia, il MHC è un problema di architettura tecnica dei sistemi.

Tecnicamente, garantire un controllo umano sostanziale, e non meramente formale, richiede che il sistema soddisfi almeno tre condizioni operative. In primo luogo, la trasparenza decisionale: l'operatore deve essere in grado di comprendere, almeno a livello funzionale, le ragioni che portano il sistema a raccomandare o intraprendere una determinata azione. Questa condizione è strutturalmente difficile da soddisfare nei sistemi basati su *deep learning*, la cui logica interna non è ricostruibile neppure dai loro progettatori. Le tecniche di *explainable AI* (XAI) — come LIME (*Local Interpretable Model-agnostic Explanations*) e SHAP (*SHapley Additive exPlanations*), nonché i meccanismi di attenzione nelle reti neurali — offrono approssimazioni interpretative, ma non soluzioni complete. In secondo luogo, la temporalità dell'intervento: affinché il controllo umano non si riduca a una ratifica postuma, l'operatore deve disporre di tempo sufficiente per valutare la raccomandazione del sistema e, se necessario, bloccarla. In scenari ad alta velocità — difesa antimissile, guerra elettronica, interdizione aerea — i tempi di reazione richiesti sono strutturalmente incompatibili con la capacità di valutazione umana, il che rende il modello *human-in-the-loop* tecnicamente inapplicabile e spinge verso il modello *human-on-the-loop*, con tutti i rischi di *automation bias* già descritti. In terzo

luogo, la qualità dell'interfaccia: un operatore sovraccarico di informazioni, o al contrario scarsamente coinvolto in un processo quasi interamente automatizzato, non è in grado di esercitare un controllo effettivo. Il carico cognitivo (*workload*) dell'operatore è, dunque, una variabile tecnica di primaria importanza per la concreta garanzia del MHC.

Queste considerazioni mostrano che il controllo umano significativo non può essere dichiarato per decreto procedurale, ma deve essere progettato nell'architettura stessa del sistema. Definirlo come requisito giuridico senza specificarne i correlati tecnici rischia di produrre sistemi formalmente conformi ma sostanzialmente privi di supervisione umana reale (AMOROSO-TAMBURRINI, 2020; CICR, 2019).

4.2. Bias *e opacità algoritmica.*

Un ulteriore aspetto critico, tra i meno discussi al di fuori della comunità tecnica, concerne la qualità e la neutralità dei dati su cui i sistemi militari di IA vengono addestrati.

I modelli di *deep learning* apprendono a riconoscere *pattern* a partire da grandi insiemi di dati etichettati. Ciononostante, se questi dati contengono distorsioni sistematiche (*bias*), il sistema le apprende e le riproduce, spesso amplificandole (NOVELLI-TADDEO-FLORIDI, 2024).

Nel contesto militare, il problema del *bias* assume dimensioni particolarmente rilevanti. I *dataset* di addestramento per i sistemi di riconoscimento visivo — ad esempio per l'identificazione di combattenti, veicoli militari o infrastrutture — sono tipicamente costruiti a partire da teatri operativi passati o da ambienti simulati, che possono non rispecchiare le condizioni del nuovo contesto operativo. Il fenomeno del *domain shift* — la divergenza tra le condizioni di addestramento e quelle operative reali — è documentato come causa di significativi cali di performance nei sistemi ATR (SCHACHTER, 2020). In scenari di conflitto urbano, dove combattenti e civili condividono gli stessi spazi e i comportamenti ambigui sono la norma, questa fragilità tecnica si traduce direttamente in rischio di identificazione erronea degli obiettivi, con conseguenze immediate sul rispetto del principio di distinzione (JAIN, 2023).

A questo si aggiunge il problema dell'opacità algoritmica. I sistemi basati su reti neurali profonde producono *output* senza che sia possibile ricostruire il ragionamento sottostante in modo comprensi-

bile per un essere umano. Questa caratteristica, nota come *black box problem*, non è una scelta progettuale ma una proprietà emergente dei modelli più complessi. Essa pone problemi di natura non soltanto etica, ma tecnica. Un sistema opaco, infatti, non può essere testato in modo esaustivo, non può essere verificato efficacemente e non può essere agevolmente sottoposto alla revisione giuridica prevista dall'articolo 36 del Protocollo aggiuntivo alle Convenzioni di Ginevra relativo alla protezione delle vittime dei conflitti armati internazionali, il quale impone agli Stati di verificare la conformità al diritto internazionale umanitario di ogni nuovo mezzo o metodo di guerra. La mancanza di trasparenza rende inoltre estremamente difficile ricostruire *ex post* il ragionamento sottostante a una decisione letale, alimentando le già complesse questioni di *accountability* morale e giuridica (NOVELLI-TADDEO-FLORIDI, 2024; ACQUAVIVA, 2023).

La letteratura tecnica e giuridica più recente converge sull'esigenza di sviluppare metodologie di *audit* indipendente, standard di *explainability* condivisi e processi di certificazione continua delle prestazioni dei sistemi militari di IA, analogamente a quanto avviene per i sistemi di controllo del traffico aereo o per i dispositivi medici ad alto rischio. L'assenza di tali standard non è solo una lacuna regolatoria, bensì un problema tecnico che compromette la stessa capacità degli operatori di esercitare un controllo umano effettivo.

4.3. *Il principio di proporzionalità nel dominio cibernetico.*

Se l'impiego di sistemi di intelligenza artificiale in un teatro bellico classico si presenta già di per sé complesso, il loro utilizzo nel dominio cibernetico risulta ancora più problematico. Attualmente non esiste un riferimento giuridico specifico in materia di *cyberwarfare*, ma come ricordato, non ci si muove in vuoto normativo, bensì è richiesto uno sforzo notevole all'interprete, incaricato di rileggere l'attuale disciplina normativa per adattarla — per quanto possibile — al nuovo dominio.

Diversi tentativi dottrinali e istituzionali di elaborazione sistematica del diritto internazionale del cyberspazio hanno attirato l'attenzione dei commentatori, ma senza ombra di dubbio il riferimento più autorevole in tale ambito è costituito dal *Tallinn Manual 2.0*, risultato di un lavoro collettivo di esperti internazionali sotto l'egida del *NATO Cooperative Cyber Defence Centre of Excellence* volto a inter-

pretare e adattare le regole del diritto internazionale al nuovo e complesso teatro delle operazioni cibernetiche.

Il *Tallinn Manual 2.0* non costituisce una fonte di diritto vincolante, bensì una guida autorevole e dettagliata che offre un quadro di riferimento per giuristi, militari e *policy makers*, orientandoli tra le numerose sfide che la guerra digitale pone in termini di responsabilità, proporzionalità, distinzione degli obiettivi e protezione dei civili. Il suo valore risiede nell'offrire principi e linee guida che aiutano a colmare le lacune normative, fungendo da bussola in un contesto in continua evoluzione (Hua, 2019; Mauri, 2022).

L'applicazione del principio di proporzionalità alle operazioni cibernetiche rappresenta una delle sfide più rilevanti e ancora in parte irrisolte del diritto internazionale umanitario. In linea teorica, la proporzionalità impone che i danni collaterali inflitti alla popolazione e ai beni civili non siano eccessivi rispetto al vantaggio militare concreto e diretto atteso dall'azione. Tuttavia, la valutazione *ex ante* di questo rapporto diventa estremamente complessa nel mondo *cyber*, dove gli effetti di un'operazione possono propagarsi in modo incontrollato attraverso le reti, generando conseguenze indirette, secondarie o terziarie difficilmente prevedibili.

A titolo esemplificativo, un *malware* progettato per colpire un'infrastruttura militare potrebbe diffondersi e compromettere servizi essenziali civili, come ospedali, reti energetiche o sistemi idrici, provocando danni ben oltre l'intenzione originaria dell'attacco e potenzialmente 'a catena'.

A complicare ulteriormente il quadro interviene il carattere *dual use* di molte infrastrutture digitali: spesso, le stesse reti e piattaforme sono utilizzate sia per scopi civili che militari, rendendo complessa la distinzione tra obiettivi legittimi e obiettivi protetti dal DIU. Questa ambiguità impone alle parti in conflitto uno sforzo supplementare nell'analisi preventiva degli effetti delle proprie azioni e nella predisposizione di misure di salvaguardia per minimizzare i rischi per la popolazione civile.

La letteratura più recente sottolinea la necessità di sviluppare metodologie di valutazione del danno adattate alle specificità del cyberspazio e di rafforzare le forme di cooperazione internazionale per la condivisione di informazioni e la prevenzione degli effetti a catena tipici delle operazioni digitali (Hua, 2019). In questo contesto, la trasparenza, la rapidità di risposta e l'adozione di standard tecnici condivisi diventano strumenti imprescindibili per garantire il rispetto del

principio di proporzionalità e, più in generale, la protezione dei diritti fondamentali anche in scenari di guerra cibernetica.

5. Osservazioni conclusive.

L'analisi condotta mostra come la comprensione tecnica dei sistemi di IA militare non sia un prerequisito accessorio, ma la condizione necessaria per qualsiasi valutazione giuridica ed etica fondata. Le caratteristiche dei sistemi — il grado di autonomia, la struttura dell'interazione con l'operatore umano, la qualità dei dati di addestramento, il livello di trasparenza algoritmica — determinano direttamente la misura in cui categorie giuridiche come responsabilità, prevedibilità e proporzionalità possono trovare applicazione concreta.

L'impiego di sistemi d'arma autonomi pone questioni complesse sotto il profilo dell'attribuzione della responsabilità giuridica. Il quadro normativo vigente distingue tra responsabilità internazionale dello Stato e responsabilità penale individuale: due piani distinti ma complementari che costituiranno necessariamente il riferimento giuridico anche nel contesto delle tecnologie emergenti. Per quanto concerne la responsabilità dello Stato, il Progetto di articoli sulla responsabilità degli Stati per atti internazionalmente illeciti del 2001 offre i criteri fondamentali di imputazione della condotta. Sebbene l'elevato grado di autonomia di alcuni sistemi possa generare il rischio di *accountability gaps*, la decisione di impiegare tali sistemi, la definizione dei loro obiettivi e la delimitazione dei parametri operativi costituiscono elementi sufficienti a stabilire un nesso causale tra condotta statale ed effetti prodotti (Boutin, 2023; Amoroso-Giordano, 2019). La responsabilità statale può, inoltre, sorgere non soltanto in relazione all'impiego operativo, ma anche nelle fasi antecedenti del ciclo di vita: sviluppo, progettazione, approvvigionamento (Boutin, 2023). Sul piano della responsabilità penale individuale, lo Statuto di Roma della CPI adotta un'impostazione tecnologicamente neutrale: la commissione di crimini internazionali può avvenire anche attraverso strumenti autonomi o cibernetici, purché ricorrano gli elementi soggettivi richiesti, in particolare dolo e consapevolezza (Acquaviva, 2023; Gaeta, 2023; Priya, 2025). Il crescente grado di autonomia dei sistemi e la distanza decisionale tra operatore umano e azione finale dell'algoritmo rendono tuttavia più complessa la dimostrazione di tali

elementi soggettivi, con analoghe difficoltà nell'applicazione della *command responsibility* (Gaeta, 2023).

I rischi più significativi non derivano da scenari futuri di macchine pienamente autonome e fuori controllo, ma da dinamiche tecniche già operative: l'*automation bias* che erode la qualità del controllo umano anche nei sistemi formalmente *human-on-the-loop*; il *domain shift* che degrada le prestazioni dei sistemi di riconoscimento quando si confrontano con contesti non rappresentati nei dati di addestramento; la propagazione incontrollata degli effetti nel dominio cibernetico; la scala industriale con cui le tecnologie generative abilitano la guerra cognitiva.

Per affrontare tali criticità in modo strutturale, è necessario che la *governance* dei sistemi di IA militare venga costruita lungo l'intero ciclo di vita — dalla progettazione al dispiegamento, dal monitoraggio operativo alla revisione post-missione — e non soltanto al momento dell'autorizzazione all'uso. Tali processi devono essere condivisi da tutti gli *stakeholder* rilevanti — Stati, organizzazioni umanitarie, imprese produttrici — in contesti multilaterali capaci di promuovere standard condivisi di audit, *explainability* e controllo. Dato che la tecnologia è già presente sui campi di battaglia, il compito del diritto e dell'etica è quello di governarla con la stessa precisione con cui essa opera.

Note bibliografiche

Abadicio M. (2019), *Artificial Intelligence for Military Logistics - Current Applications*, Emerj, accessibile al link: *https://emerj.com/ai-sector-overviews/artificial-intelligence-military-logistics/*.

Acquaviva G. (2023), *Crimes without Humanity? Artificial Intelligence, Meaningful Human Control, and International Criminal Law*, in *Journal of International Criminal Justice*, 21, p. 69 ss.

Amoroso D.-Giordano B. (2019), *Who Is to Blame for Autonomous Weapons Systems Misdoings?*, in Carpanelli E.-Lazzerini N. (eds.), *Use and Misuse of New Technologies: Contemporary Challenges in International and European Law*, The Hague, Springer, p. 211 ss.

Amoroso D.-Tamburrini G. (2020), *Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues*, in *Current Robotics Reports*, 8, p. 187 ss.

Bo M.-Bruun L.-Boulanin V. (2022), *Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems*, Stockholm International Peace Research Institute (SIPRI).

Boutin B. (2023), *State Responsibility in Relation to Military Applications of Artificial Intelligence*, in *Leiden Journal of International Law*, 36, p. 133 ss.

BOUTIN B.-WOODCOCK T. (2024), *Aspects of Realizing (Meaningful) Human Control: A Legal Perspective*, in GEIß R.-LAHMANN H. (eds.), *Research Handbook on Warfare and Artificial Intelligence*, Cheltenham, Edward Elgar Publishing, p. 179 ss.

DELLA MORTE G. *et al.* (2024), *Algorithms and Decision-Making in the Metaverse: Legal and Ethical Grounds to Trust AI Information for the Decision-Making Process within the Context of Armed Conflict*, in RIVA G.-VILLANI D. (eds.), *Humane Metaverse. Reflections of Self, Education, Organizations and Society*, Milano, Vita & Pensiero, p. 289 ss.

FORD C.M. (2017), *Autonomous Weapons and International Law*, in *South Carolina Law Review*, 69, p. 414 ss.

GAETA P. (2023), *Who Acts When Autonomous Weapons Strike? The Act Requirement for Individual Criminal Responsibility and State Responsibility*, in *Journal of International Criminal Justice*, 21, p. 1033 ss.

HUA S. (2019), *Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control*, in *Georgetown Journal of International Law*, 51, p. 117 ss.

INTERNATIONAL COMMITTEE OF THE RED CROSS (2019), *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*, rapporto tecnico, Ginevra, consultabile a partire da *www.icrc.org*.

INTERNATIONAL COMMITTEE OF THE RED CROSS (2021), *ICRC Position on Autonomous Weapon Systems*, Ginevra, consultabile a partire da *www.icrc.org*.

INTERNATIONAL LAW COMMISSION (2001), *Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries*, in *Report of the International Law Commission on the Work of its 53rd Session*, UN Doc A/56/10, p. 26 ss.

JAIN A.G. (2023), *Autonomous Weapon Systems, Errors and Breaches of International Humanitarian Law*, in *Journal of International Criminal Justice*, 21, p. 1005 ss.

LAHMAN H.-CUSTER B.-SCOTT B.I. (2025), *The Fundamental Rights Risks of Countering Cognitive Warfare with Artificial Intelligence*, in *Ethics and Information Technology*, 27, p. 1 ss.

LAWAND K. (2013), *Seminar on Fully Autonomous Weapon Systems* (Mission Permanente de France, Ginevra), consultabile a partire da *www.icrc.org*.

MAURI D. (2022), *Autonomous Weapons Systems and the Protection of the Human Person*, Cheltenham, Edward Elgar Publishing.

MORENO J.A. (2010), "Logistics", in NATO Standardization Agency (NSA), *NATO Glossary of Terms and Definitions*, AAP-06.

NOVELLI C.-TADDEO M.-FLORIDI L. (2024), *Accountability in Artificial Intelligence: What It Is and How It Works*, in *AI & Society*, 39, p. 1871 ss.

PRIYA A. (2025), *Criminal Accountability for AI: Mens Rea, Actus Reus, and the Challenges of Autonomous Systems*, in *LawFoyer International Journal of Doctrinal Legal Research*, 3, p. 273 ss.

SANTONI DE SIO F.-VAN DEN HOVEN J. (2018), *Meaningful Human Control over Autonomous Systems: A Philosophical Account*, in *Frontiers in Robotics and AI*, 5, p. 2 ss.

SCHACHTER B.J. (2020), *Automatic Target Recognition*, Bellingham, SPIE Press.

SHARKEY N. (2016), *Staying in the Loop: Human Supervisory Control of Weapons*, in BHUTA N. et al. (eds.), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge, Cambridge University Press.

UNITED NATIONS PANEL OF EXPERTS (2021), *Final Report of the Panel of Experts on Libya*, UN Doc S/2021/229, 8 marzo 2021, p. 17 ss., consultabile a partire da *www.un.org*.

(LA NORMA) LE ARMI AUTONOME E IL DIRITTO INTERNAZIONALE UMANITARIO: PROFILI REGOLATORI

di *Marco Sassòli*

Sommario: 1. Introduzione. — 2. Il diritto internazionale umanitario e le sue fonti. — 3. Elemento di confronto. — 4. Discussioni e controversie. — 5. Le mie ipotesi preliminari. — 5.1. Autonomia e prevedibilità. — 5.2. Un presupposto tecnico che potrebbe rivelarsi errato. — 5.3. Le decisioni relative alla scelta degli obiettivi devono comportare giudizi soggettivi? — 6. Le sfaccettature del controllo umano significativo. — 7. Vantaggi delle LAWS. — 8. Sfide all'interpretazione del diritto internazionale umanitario. — 9. Sfide nel campo della responsabilità dello Stato e della responsabilità individuale per crimini di guerra. — 10. Conclusione.

1. Introduzione.

L'intelligenza artificiale rivoluziona la nostra vita quotidiana e può anche modificare in modo significativo, sia in positivo che in negativo, la guerra. Il presente contributo si concentrerà sul caso estremo dei sistemi d'arma autonomi letali ('*lethal autonomous weapons systems*', in seguito LAWS), che destano la massima preoccupazione nella comunità internazionale dal punto di vista umanitario, etico e giuridico, cercando di applicare a questo nuovo fenomeno le norme esistenti del diritto internazionale umanitario (in seguito DIU) e di presentare i tentativi di adottare nuove norme specifiche applicabili a tali armi.

Non verranno trattati i profili dell'utilizzo dell'intelligenza artificiale come mezzo di supporto agli operatori, la logistica, la formazione e l'azione umanitaria, e ugualmente questioni di diritto internazionale dei diritti umani, sebbene esso si applichi anche in tempo di guerra.

Innanzitutto, si procederà alla disamina ricordando brevemente le funzioni del DIU e le principali fonti e precisando che l'uso dei LAWS non deve essere confrontato con la normale condotta in tempo di pace, ma con l'alternativa dell'uso delle armi impiegate in scenari di guerra, attualmente completamente gestite da esseri umani. Successivamente, verranno riassunte le controversie esistenti sulla legalità dell'uso dei LAWS, sia *de lege lata* che *de lege ferenda*, e si procederà all'analisi dei presupposti tecnici e giuridici alla base dell'approccio — relativamente ottimistico — adottato e si discuterà il significato del requisito del 'controllo umano significativo', fattore che quasi unanimemente Stati e dottrina concordano essere necessario per un uso legittimo dei LAWS ma concetto non ancora puntualmente definito e che nasconde importanti divergenze di opinione.

Poste queste necessarie premesse, verranno infine elencati i vantaggi dell'uso dei LAWS dal punto di vista del DIU e le notevoli sfide per la sua interpretazione e per l'attribuzione della responsabilità agli Stati e agli individui in caso di violazioni.

2. Il diritto internazionale umanitario e le sue fonti.

Il DIU tutela la vita e la dignità delle persone colpite da conflitti armati unicamente nella misura in cui gli Stati ritengono che il loro rispetto sia compatibile con l'unico obiettivo legittimo di un conflitto armato: l'indebolimento del potenziale militare del nemico. La protezione offerta dal DIU dipende dalla classificazione del conflitto come internazionale o non internazionale, dalla classificazione della persona colpita come civile o combattente e da ulteriori categorizzazioni giuridiche rilevanti.

Le norme del DIU sono contenute in gran parte in trattati multilaterali universali ben coordinati tra loro. Le Convenzioni di Ginevra, universalmente accettate, costituiscono il loro nucleo. Altri trattati, in particolare i Protocolli Aggiuntivi, vincolano solo gli Stati Parte (il Protocollo I si applica solo ai conflitti armati che si verificano tra Stati contraenti, il Protocollo II solo ai conflitti armati non internazionali che si verificano nel territorio di uno Stato contraente), ma corrispondono, benché in misura variabile, al diritto internazionale consuetudinario.

Il diritto consuetudinario è indispensabile in quanto disciplina questioni che non sono regolate dai trattati e vincola gli Stati non contraenti così come gli Stati contraenti nelle loro relazioni con gli

Stati non contraenti. Il DIU consuetudinario ha conosciuto una sorprendente rinascita negli ultimi anni, anche se è particolarmente difficile — ai fini della ricostruzione dello stesso — valutare la prassi effettiva degli Stati e l'elemento della *opinio juris*. Ad avviso dell'autore, per valutare lo stato del DIU consuetudinario occorre tenere conto in egual misura della partecipazione ai trattati e (per il diritto internazionale umanitario dei conflitti armati non internazionali) della prassi dei gruppi armati. Lo studio sul diritto consuetudinario del Comitato internazionale della Croce Rossa (CICR) facilita notevolmente l'identificazione della prassi ufficiale degli Stati e delle norme consuetudinarie che ne derivano.

3. Elemento di confronto.

Il DIU disciplina la guerra. La guerra è sempre terribile, deve essere evitata e porta, anche quando il DIU è rispettato, a sofferenze umane, morti (di combattenti e, incidentalmente, di civili, purché sia rispettata la regola della proporzionalità), feriti e distruzione.

Non si dovrebbe paragonare i LAWS a una situazione ideale, ma ad altri mezzi di guerra e di violenza applicati negli attuali scenari bellici direttamente dagli esseri umani. Se i LAWS non sono in grado di fare ciò che un soldato medio dovrebbe — ed è in grado di fare — non possono essere utilizzati.

Qualora non fosse tecnicamente possibile rispettare determinati requisiti del DIU tramite l'utilizzo di armi autonome, ciò non costituisce un motivo sufficiente per abbandonare tali requisiti e pertanto lo stesso uso di un'arma autonoma, in tali casi, è semplicemente da considerarsi illegale.

Al contrario, qualora i sistemi autonomi fossero da considerarsi migliori e più rispettosi del DIU rispetto agli esseri umani, una parte in un conflitto armato che dispone di tali armi deve utilizzarle.

L'affermazione categorica secondo cui le armi autonome possono essere utilizzate solo se sono in grado di rispettare il DIU allo stesso modo degli esseri umani e nelle circostanze in cui ciò sia garantito è tuttavia soggetta a una sfumatura. Né l'artiglieria né i missili sono in grado, come un cecchino, di interrompere un attacco all'ultimo momento in base al mutare delle circostanze: tuttavia, nessuno sostiene che tali armi siano intrinsecamente illegali. È dunque necessario che nuovi armi con un elevato grado di autonomia siano in grado di acquisire informazioni che indicano che l'attacco debba essere inter-

rotto e che il sistema stesso — o i suoi operatori umani — siano in grado di reagire a tali informazioni.

4. Discussioni e controversie.

Molti sostengono che le LAWS dovrebbero essere vietate in via preventiva e senza eccezioni proprio perché non sarebbero compatibili con il DIU e con i principi etici che consentono solo agli esseri umani di decidere della vita o della morte di altri esseri umani (HRW, et al., 2012; SINGER, 2009; ASARO, 2013).

Il CICR ritiene che alla luce delle considerazioni etiche volte a salvaguardare l'umanità e a rispettare le norme del diritto internazionale umanitario per la protezione dei civili e dei combattenti *hors de combat*, l'uso di sistemi d'arma autonomi per colpire esseri umani dovrebbe essere escluso (ICRC 2021).

Dal 2013, gli Stati parte della Convenzione del 10 ottobre 1980 sul divieto o la limitazione dell'impiego di talune armi classiche che possono essere ritenute capaci di causare effetti traumatici eccessivi o di colpire in modo indiscriminato (CCW) nel 2016 hanno convenuto di creare un gruppo di esperti governativi (« Convention on Certain Conventional Weapons - Group of Governmental Experts on Lethal Autonomous Weapons Systems », in seguito GGE) con l'obiettivo di redigere un Protocollo VI alla CCW in materia di LAWS, (per informazioni aggiornate, si vada UNOG; per un'analisi molto critica del processo, con riferimenti, v. Human Rights Watch 2022; HOFFBERGER-PIPPAN et al., 2022).

Attualmente, il GGE sta conducendo le proprie discussioni sulla base di un testo provvisorio (GGE, 2024), che raccoglie formulazioni adottate all'unanimità con l'obiettivo di avanzare nei lavori relativi a diversi elementi di un nuovo potenziale strumento, senza tuttavia pregiudicare la natura giuridica dello strumento previsto o escludere ulteriori nuove misure. Resta ben inteso che nulla sarà considerato definitivamente concordato finché in seno alla CCW non sarà stato raggiunto un accordo su tutti gli aspetti.

Le discussioni sono tuttavia bloccate dalla regola del consenso e dalle obiezioni sollevate — talvolta in buona fede e talvolta in malafede — dagli Stati.

La maggioranza degli Stati è favorevole a un trattato vincolante che vieti o regoli i LAWS; analogamente al CICR, tali Stati desiderano combinare il divieto di alcuni tipi di LAWS con la regolamenta-

zione di altri tipi di queste armi. Tuttavia, una minoranza di grandi potenze militari (come la Russia, Israele, il Regno Unito, gli Stati Uniti e, in una certa misura, l'India) ha approfittato della regola del consenso per ritardare qualsiasi progresso. Taluni suggeriscono quindi di trasferire le discussioni in un altro *forum*, in cui solo gli Stati disposti ad accettare un nuovo trattato ne negozierebbero il contenuto, analogamente a quanto avvenuto in passato nei processi che hanno portato alla Convenzione di Ottawa sulla messa al bando delle mine antiuomo, alla Convenzione di Oslo sul divieto delle munizioni a grappolo e al Trattato sulla proibizione delle armi nucleari.

Anche tra gli Stati più progressisti in materia vi sono divergenze. Ad esempio, molte delegazioni della CCW hanno considerato che i LAWS possono essere utilizzati solo con un « controllo umano significativo », mentre altre delegazioni hanno proposto il criterio alternativo di un « livello adeguato di giudizio umano ». Gli Stati Uniti, che sono tra i paesi tecnologicamente più avanzati in questo campo, richiedono che i LAWS siano progettati in modo da consentire ai comandanti e agli operatori di esercitare un livello adeguato di giudizio umano sull'uso della forza (US, 2023), il che significa che non è consentito ai produttori programmare macchine che prendono decisioni definitive sull'uso della forza contro obiettivi. Anche la Francia ha deciso di utilizzare sistemi d'arma letali che incorporano l'autonomia solo se rimangono sotto il controllo umano (Francia 2021) e ha specificato cosa ciò comporta e analogamente il Regno Unito ha definito in modo più vago, ma forse più realistico, cosa implica il coinvolgimento umano (Regno Unito 2022).

5. Le mie ipotesi preliminari.

5.1. *Autonomia e prevedibilità.*

L'intelligenza artificiale potrebbe un giorno consentire ai LAWS di imparare progressivamente e dunque di prendere decisioni autonome non pre-programmate. Ciò porta a quello che nelle discussioni tecniche viene definito il fenomeno della « scatola nera », ovvero l'impossibilità di sapere perché vengono generati determinati risultati (Molnar, 2022; Holland Michel, 2020). Se ciò implica che non è possibile prevedere se i LAWS rispettino o violino il DIU, essi sarebbero ovviamente vietati dallo stesso DIU.

I destinatari del DIU, che vieta gli attacchi indiscriminati « che impiegano un metodo o un mezzo di combattimento che non può es-

sere diretto [dallo Stato o dal gruppo armato non statale che agiscono attraverso esseri umani] contro un obiettivo militare specifico» (Protocollo I, art. 51, par. 4, lett. *b*), sono gli esseri umani che impiegherebbero i LAWS. Realisticamente, tuttavia, nessun comandante delle forze armate o di un gruppo armato vorrebbe utilizzare nell'ambito della condotta delle ostilità armi su cui non ha alcun controllo e che potrebbero 'decidere' di attaccare le proprie forze o unirsi al nemico.

A mio parere, quindi, anche se in futuro venissero sviluppate armi completamente autonome, chi produce o utilizza LAWS garantirebbe che tali armi agiscano solo in conformità con ciò che gli esseri umani che le utilizzano vogliono che facciano. È possibile, dunque, che la controversia sulla legalità delle LAWS si riduca alla questione tecnica se tale limitazione sia possibile.

5.2. *Un presupposto tecnico che potrebbe rivelarsi errato.*

È altresì ovvio che i LAWS non possono essere utilizzati se — e fintantoché — essi non siano in grado di essere utilizzati in conformità con il DIU. Attualmente ciò non è ancora possibile al di fuori di ambienti molto limitati e prevedibili (Trabucco, 2025). Tuttavia, sebbene questa sembri essere la sfida principale all'uso dei LAWS, è possibile che un giorno si riescano a costruire LAWS in grado di ottenere e applicare le informazioni necessarie per conformarsi al DIU. Non si tratta unicamente di una questione di distinzione: esse devono piuttosto anche essere in grado di riconoscere, nella stessa misura di un soldato medio, quando i bersagli umani legittimi si arrendono o quando sono feriti e, in quest'ultimo caso, se i feriti si astengono da qualsiasi atto di ostilità (Boothby, 2025; Heyns, 2013).

5.3. *Le decisioni relative alla scelta degli obiettivi devono comportare giudizi soggettivi?*

I critici (Human Rights Watch *et al.* 2012) obiettano che i LAWS violerebbero per ragioni di principio il DIU perché, anche qualora lo sviluppo di armi completamente autonome con cognizione simile a quella umana diventasse fattibile, esse sarebbero prive di alcune qualità umane, come l'emozione, la compassione e la capacità di comprendere gli esseri umani. In particolare, essi rifiutano la possibilità teorica che un robot possa distinguere tra obiettivi e valutare la proporzionalità in modo più obiettivo e affidabile degli esseri umani,

sulla base del fatto che tali regole implicano giudizi soggettivi che solo gli esseri umani possono formulare.

Personalmente, non ritengo che le decisioni relative alla scelta degli obiettivi debbano essere intrinsecamente soggettive. Molti militari sembrano non essere d'accordo con tale assunto poiché rifiutano, anche al di fuori del dibattito sui LAWS, regole molto dettagliate in materia di proporzionalità e precauzioni. Pur concordando sul fatto che la giustizia non possa essere autonoma, prendere di mira un individuo nel DIU non è una decisione di giustizia: ciò non implica determinare che la persona meriti la morte, ma piuttosto classificare la persona in base al suo *status* (come combattente) o alla sua condotta (come partecipante diretto alle ostilità), senza alcuna determinazione di colpa o responsabilità. È pur vero che il « buon senso », che si riferisce alla capacità di comprendere logicamente il significato delle situazioni sociali, si basa sull'esperienza umana. Sebbene sia particolarmente difficile creare una macchina dotata di buon senso, ciò potrebbe offrire l'opportunità di creare un « buon senso oggettivo » basato non sulla storia di vita di un singolo individuo, ma su quella di un insieme di persone o, idealmente, di tutta l'umanità.

Per quanto riguarda la regola della proporzionalità, diversi autori e manuali militari menzionano che la sua applicazione comporta effettivamente una determinazione soggettiva. La questione, tuttavia, è se si tratti semplicemente di una descrizione della realtà sfortunata mentre la determinazione dovrebbe idealmente essere la più obiettiva possibile, o piuttosto se si tratti di una proposizione normativa e dunque la determinazione debba essere soggettiva. A mio avviso, sarebbe auspicabile sia per gli operatori umani che per i LAWS che gli Stati concordassero una formula per tale calcolo, insieme agli indicatori e agli elementi che dovrebbero o non dovrebbero essere presi in considerazione. La modellizzazione e la determinazione di indicatori per l'infinita varietà di situazioni possibili sarà forse una difficoltà insormontabile per i produttori di armi veramente autonome, ma ciò potrebbe essere superato dall'impiego di sistemi di intelligenza artificiale.

Ovviamente, la determinazione della proporzionalità deve essere effettuata caso per caso, ma a mio avviso non deve necessariamente essere « soggettiva ». Perché un civile dovrebbe essere protetto dagli effetti incidentali derivanti da un attacco di un soldato piuttosto che da un altro soldato? Perché dovrebbe avere importanza il soggetto (in altre parole, il soldato, la sua età, il suo sesso, la sua istruzione, i

suoi valori, la sua religione o la sua etica) e non solo l'oggetto (cioè il vantaggio militare e l'effetto incidentale sui civili)?

Tuttavia, va ammesso che gli Stati e i giuristi militari hanno finora rifiutato di quantificare il peso del rischio di perdere una vita civile rispetto al potenziale vantaggio militare e, allo stesso modo, di stabilire quando il rapporto tra il rischio e il vantaggio diventa eccessivo. Essi fanno invece riferimento alla ragionevolezza, benché risulti ovvio che un computer non possa essere programmato per essere ragionevole o per sviluppare una propria ragionevolezza. Va detto però in proposito che, ad esempio, nel diritto civile statunitense il criterio della persona ragionevole è considerato un criterio oggettivo e non soggettivo (Restatement (Second) of Torts (1965) Sezione 283, commento (c)). Se gli Stati e i giuristi militari persistono nel loro rifiuto di fornire criteri e indicatori più precisi, è possibile che i LAWS non possano essere utilizzati in circostanze in cui i civili rischierebbero di essere colpiti.

6. Le sfaccettature del controllo umano significativo.

Nelle discussioni sopra citate tra Stati ed esperti, esiste oggi un ampio consenso sul fatto che i LAWS possano essere utilizzati solo sotto un controllo umano significativo (anche se la formulazione esatta varia ancora). Si può ovviamente temere che si tratti di un consenso di facciata.

Gli esseri umani potrebbero anche diventare un alibi se dovessero semplicemente confermare o interrompere « suggerimenti » o « decisioni » computerizzate e guidate dall'intelligenza artificiale. Infatti, tali « suggerimenti » o « decisioni » sarebbero basati su dati elaborati da macchine migliaia di volte più veloci di qualsiasi essere umano e pertanto lo stesso essere umano sarebbe inoltre influenzato da quello che è stato definito '*automation bias*'.

Gli operatori umani sono infatti eccessivamente fiduciosi dell'affidabilità dei sistemi informatici e dell'accuratezza dei loro risultati, e tendono ad accettare i consigli del computer senza cercare prove contrarie ma — allo stesso tempo — anche il contrario sembra essere vero.

Kwik (2025) mostra la tendenza dell'intelligenza artificiale ad allineare i propri risultati alle opinioni o alle preferenze degli utenti, anche se tali opinioni sono errate. In questa simbiosi, l'assenza o la presenza dell'intervento umano è una distinzione relativa, così come lo è la distinzione tra esseri umani « nel », « sul » o « fuori dal cir-

cuito ». In ogni caso, nonostante l'autonomia del sistema, gli esseri umani saranno inevitabilmente coinvolti quantomeno nella produzione del sistema, mentre la programmazione da parte degli esseri umani potrebbe diventare meno centrale nel caso dell'intelligenza artificiale. Tutto ciò non è tuttavia un motivo per rinunciare a tentare di definire il « controllo umano significativo », ma piuttosto per farlo evitando queste trappole. Qui, come altrove, il diavolo si nasconde nei dettagli e dubito che gli Stati accetteranno presto una definizione significativa e operativa di 'meaningful human control'.

A mio avviso, sulla base dei suggerimenti di Kwik (2022, 2024; v. per altri approcci che spesso portano agli stessi risultati AMOROSO-TAMBURRINI, 2020; TRABUCCO, 2025), esistono diversi aspetti del controllo umano significativo, ciascuno dei quali bilancia l'altro ed è in grado di compensarne le debolezze. Si tratta di garantire una consapevolezza sufficiente, esercitare un adeguato controllo delle armi e attuare controlli contestuali. La consapevolezza comprende, in primo luogo, la comprensione degli aspetti tecnici del sistema (non come funziona, ma cosa è in grado di fare e quali rischi comporta il suddetto fenomeno della « scatola nera »; e, in secondo luogo, l'ambiente in cui il sistema opererà e come quest'ultimo interagirà con tale ambiente. Inoltre, un essere umano deve effettuare una valutazione post-esecuzione per conoscere meglio il comportamento e i fallimenti passati del sistema e per superare il possibile « bias di automazione » degli esseri umani che utilizzano effettivamente i LAWS.

L'aspetto del controllo umano riguarda la decisione finale sull'opportunità di utilizzare un LAWS per una missione. Un essere umano deve anche esercitare il controllo del contesto, prendendo decisioni attive riguardanti le impostazioni e i parametri interni del sistema, imponendo restrizioni spaziali e temporali. Ciò che è più controverso è se gli esseri umani debbano anche avere i mezzi per intervenire durante l'operazione. In tal caso, è necessario un collegamento di comunicazione permanente tra l'essere umano e il sistema. Inoltre, il sistema deve essere in grado di spiegare all'essere umano perché sta facendo o raccomandando qualcosa. Se tale controllo umano durante l'operazione è un requisito, le funzioni letali devono essere disattivate in caso di perdita di comunicazione. La conformità attraverso la progettazione (umana) senza alcun intervento umano durante il funzionamento può essere un'alternativa, ma presuppone un controllo più forte sotto altri aspetti. Infatti, un potere di « veto » puramente passivo da parte dell'uomo riduce la sua attenzione. Se gli esseri umani

hanno un ruolo durante il funzionamento del sistema, il loro carico di lavoro non deve essere né troppo elevato né troppo basso.

Come sottolinea Kwik (2022, 2024), l'attuazione di questi aspetti in modo equo risolve due problemi ampiamente discussi: garantisce la prevedibilità (limitando le possibilità di azione del sistema) e rende possibile la responsabilità umana (rendendo l'essere umano coinvolto cognitivamente e causalmente responsabile del risultato).

7. Vantaggi delle LAWS.

Se le due ipotesi tecniche e la comprensione del carattere oggettivo dei criteri di selezione dei bersagli previsti dal DIU sopra descritti sono corrette, un attacco eseguito attraverso LAWS presenterebbe molti vantaggi in termini di distinzione, proporzionalità e precauzioni rispetto a un attacco eseguito direttamente da esseri umani. Solo gli esseri umani possono essere disumani e scegliere deliberatamente di non rispettare le regole che sono stati incaricati di seguire. A mio avviso sembra più facile aspettarsi (e garantire) che una persona che progetta e costruisce un'arma autonoma in un luogo di lavoro pacifico rispetti il DIU piuttosto che un soldato in un ambiente ostile. Un robot non può provare odio o paura, non soffre la fame o la stanchezza e non ha istinto di sopravvivenza. Può adottare misure precauzionali aggiuntive che il DIU non si aspetterebbe mai da esseri umani perché troppo pericolose. I robot possono ritardare l'uso della forza fino al momento più opportuno, ovvero quando è stato stabilito che sia l'obiettivo che l'attacco sono legittimi. I robot non violentano, possono percepire più informazioni contemporaneamente ed elaborarle più rapidamente degli esseri umani. Man mano che le armi che esercitano effettivamente la forza cinetica diventano più veloci e complesse, gli esseri umani potrebbero semplicemente essere sopraffatti dalle informazioni che devono valutare e dalle decisioni che devono prendere. Come discusso in relazione alla regola della proporzionalità, lo sviluppo dei LAWS potrebbe persino portare, a causa delle esigenze di programmazione, a un chiarimento di molte norme del DIU che finora sono rimaste vaghe.

8. Sfide all'interpretazione del diritto internazionale umanitario.

La maggior parte delle argomentazioni contro i LAWS basate su principi mettono questi a confronto con situazioni ideali — in cui i

conflitti armati non esisterebbero comunque — invece di confrontarli con mezzi alternativi di guerra attualmente esistenti e utilizzati nella realtà per uccidere esseri umani. Quest'ultimo approccio « antropomorfo », che confronta ciò che fa un LAWS con ciò che ci si aspetta da un soldato umano ragionevole, può essere contestato sostenendo che una macchina deve essere molto più efficiente in termini di rispetto del DIU per poter prendere decisioni letali e che le tecnologie nel settore dei LAWS non dovrebbero essere antropomorfizzate. Pertanto, l'applicazione del DIU esistente ai LAWS presenta delle difficoltà che richiedono un'interpretazione corretta del DIU da parte di ogni Stato che vuole utilizzarli e, idealmente, un consenso tra gli Stati stessi.

Di seguito, discuterò alcune di queste sfide.

Boothby (2025) offre una panoramica completa dell'applicazione del DIU e del diritto penale internazionale esistente alle LAWS. In primo luogo, gli Stati devono concordare che il DIU si applica ai comportamenti in tempo di pace che possono produrre risultati durante un conflitto armato. L'ultimo essere umano che interviene nella decisione su chi sarà attaccato in un conflitto armato deve essere pienamente soggetto al DIU, anche se agisce in tempo di pace.

In secondo luogo, mentre un robot potrebbe, a mio avviso, ottenere le informazioni necessarie per valutare i possibili effetti collaterali di un attacco sui civili e persino effettuare la valutazione necessaria se si concordano formule oggettive, questo è solo un aspetto del test di proporzionalità. L'altro aspetto, che richiede una valutazione del « vantaggio militare concreto e diretto previsto » (Protocollo I, art. 51(5)(b)) derivante da un attacco contro un obiettivo legittimo, cambia costantemente in base ai piani del comandante e allo sviluppo delle operazioni militari da entrambe le parti in conflitto. Pertanto, una macchina non potrebbe applicare da sola la regola della proporzionalità, ma dovrebbe piuttosto essere costantemente aggiornata sulle operazioni e sui piani militari.

In terzo luogo, la sfida sopra menzionata incide anche sulla capacità dei LAWS di valutare e applicare il principio di distinzione, poiché la definizione di obiettivo militare dipende dal suo « contributo effettivo all'azione militare » e dal « vantaggio militare definito » che l'attacco offre « nelle circostanze del momento » (Protocollo I, art. 52(2)). Anche questo implica che i LAWS devono essere a conoscenza dei piani del comandante e dell'evoluzione complessiva dell'operazione militare. A mio avviso, si può comunque tenere conto del fatto che anche i soldati che lanciano un attacco non dispongono di

tutte queste informazioni, ma devono comunque rispettare il DIU. È tuttavia possibile che il comandante li richiami quando la situazione cambia e ciò deve essere possibile anche con i LAWS.

In quarto luogo, la regola secondo cui chi lancia un attacco deve prendere tutte le precauzioni possibili per evitare o ridurre al minimo gli effetti incidentali sui civili deve riferirsi a ciò che sarebbe possibile effettivamente porre in essere da parte degli esseri umani che utilizzano la macchina e non alle possibilità della macchina (che può essere concepita per rendere possibili determinate precauzioni o per non poterle prendere). Può anche accadere che una macchina sia più efficiente di un essere umano medio nel prendere determinate precauzioni, mentre per altre precauzioni vale il contrario. In tal caso, a mio avviso è ammissibile una valutazione consolidata dei vantaggi e degli svantaggi per determinare se un'arma autonoma rispetta il DIU allo stesso modo di un soldato medio. Tale valutazione deve tuttavia essere effettuata per ogni attacco, il che richiede la creazione di parametri che consentano un confronto con le prestazioni degli esseri umani. Le LAWS dovrebbero quindi determinare ciò in relazione alle circostanze specifiche di ciascun attacco e, se necessario, « decidere » che non possono eseguire tale attacco e che è quindi necessario l'intervento umano.

In quinto luogo, precauzioni importanti, come l'obbligo di verificare la natura di un obiettivo o di scegliere mezzi e metodi che evitino o riducano al minimo gli effetti incidentali sui civili, sono rivolte solo a « coloro che pianificano o decidono un attacco » (Protocollo I, art. 57(2)(a)). Alcuni si chiedono se ciò implichi che un essere umano debba pianificare e decidere un attacco. A mio avviso, tutte le norme del DIU sono rivolte esclusivamente agli esseri umani. Ciò non esclude tuttavia che i pianificatori e i decisori umani possano essere temporaneamente e geograficamente distanti dall'attacco, purché definiscano i parametri in base ai quali il robot attacca e si assicurino che esso disponga delle informazioni necessarie per applicare tali parametri e vi si conformi effettivamente.

In sesto luogo, determinare cosa implichi l'obbligo di interrompere o sospendere un attacco « quando appaia » che è illegale (Protocollo I, art. 57(2)(b)) in termini di capacità di rilevamento e di modifica del comportamento dei LAWS presenta una questione particolarmente delicata. A mio avviso, tale obbligo implica che i LAWS devono essere costruiti in modo da percepire i cambiamenti nell'ambiente almeno quanto gli esseri umani.

Come settimo profilo, si pone anche la questione di chi, nella catena di produzione, programmazione e impiego dei LAWS, partecipi direttamente alle ostilità. Purtroppo, ciò che costituisce esattamente la partecipazione diretta alle ostilità è ancora oggetto di controversia. Secondo il CICR, la causalità diretta del danno, che si richiede come uno degli elementi cumulativi affinché un atto costituisca partecipazione diretta alle ostilità, deve essere intesa nel senso che il danno in questione deve essere causato in un unico passo causale (ICRC, 2009). Ovviamente, sono rilevanti solo le « fasi » umane, poiché il concetto di partecipazione diretta definisce chi (e non cosa) può essere preso di mira. Pertanto, se vengono utilizzate armi autonome, l'ultima fase causale umana che porta al danno causato, che costituisce partecipazione diretta alle ostilità, può essere geograficamente e temporalmente distante dal danno.

Il CICR ritiene inoltre che lo standard di causalità diretta includa comportamenti che causano un danno solo in combinazione con altri atti. Pertanto, un atto specifico che di per sé non causa direttamente il danno soddisferebbe comunque il requisito della causalità diretta se costituisce parte integrante di un'operazione tattica concreta e coordinata che causa tale danno. Nel caso dei LAWS, un'interpretazione più ampia della partecipazione diretta potrebbe implicare che un individuo che è l'ultimo a determinare cosa o chi sarà bersaglio dei LAWS in un numero indeterminato di operazioni future partecipi direttamente alle ostilità se gli esseri umani che intervengono in una fase successiva non sono più in grado di effettuare tale scelta. A mio avviso, ciò mescola questioni distinte, quali la legalità di un attacco e lo status della persona che attacca. Proprio come per coloro che redigono manuali tattici, tali persone non partecipano direttamente alle ostilità, sebbene il rispetto del DIU dipenda da loro e gli Stati debbano garantire che tali individui conoscano e rispettino il DIU.

9. Sfide nel campo della responsabilità dello Stato e della responsabilità individuale per crimini di guerra.

Anche nel caso dei LAWS, il DIU si rivolge alle parti coinvolte in conflitti armati e agli esseri umani che le progettano, producono e programmano, nonché a coloro che decidono il loro utilizzo. Alcuni sostengono che il DIU sia inadeguato a regolamentare i LAWS, in quanto questi ultimi si collocherebbero a metà strada tra i sistemi d'arma e i combattenti e che sarebbe quindi necessario creare una

nuova categoria con nuove norme per regolamentarli (Liu, 2012). A mio avviso, la differenza tra un sistema d'arma e un essere umano non è quantitativa ma qualitativa. I due non si collocano su una scala mobile ma su livelli diversi, rispettivamente come oggetti e soggetti. Indipendentemente da come funzionerà l'intelligenza artificiale in futuro, gli esseri umani saranno sempre il punto di partenza. Come già detto, i LAWS devono sempre operare entro i limiti progettati dagli esseri umani. Sono infatti gli esseri umani a decidere se creare una macchina e chi la creerà.

La responsabilità degli Stati (e, possibilmente, di gruppi armati) che impiegano LAWS solleva questioni diverse rispetto alla responsabilità degli esseri umani. Per quanto riguarda i primi, « è solo l'atto dello Stato [cioè le azioni degli esseri umani attribuibili a uno Stato] che conta, indipendentemente da qualsiasi intenzione » (International Law Commission, 2001, commento all'articolo 2, paragrafo 10). In altre parole, il semplice fatto che un'arma impiegata da uno Stato comporti una violazione delle norme è di per sé sufficiente a rendere lo Stato stesso responsabile.

Per quanto riguarda la responsabilità penale, gli abolizionisti sostengono che non sarebbe chiaro chi sarebbe ritenuto responsabile delle azioni illegali « commesse » da un robot. Le opzioni includono il comandante militare che lo ha impiegato, il programmatore, il produttore e il robot stesso, ma tutte sono insoddisfacenti. Il robot non può essere punito: in primo luogo, il diritto in generale, compreso il diritto internazionale umanitario e il diritto internazionale penale, è antropocentrico. Il diritto e la punizione esistono per guidare l'azione umana piuttosto che quella di oggetti inanimati, anche se dotati di intelligenza artificiale. A mio avviso, suggerire che i robot possano essere rottamati o disattivati come forma di punizione è assurdo. In secondo luogo, i robot sono incapaci di avere la necessaria *mens rea*, poiché non sono in grado di distinguere il bene dal male.

Un approccio perseguibile sarebbe quello di ritenere responsabile il comandante di un robot così come lo sarebbe un comandante che ordina a un pilota di bombardare un obiettivo dicendo che si tratta di un quartier generale militare, ma che poi si rivela essere una scuola. A mio avviso, la responsabilità di un comandante che impiega LAWS non è un caso di responsabilità di comando (né è analogo ad essa). Piuttosto, la responsabilità penale di chi utilizza un LAWS è un caso di responsabilità diretta, proprio come quella di un soldato che spara con un mortaio credendo che possa colpire solo il carro ar-

mato preso di mira, ma che ucciderà dei civili che sa che stanno seguendo il carro armato. Se dunque i LAWS fossero imprevedibili, i comandanti sarebbero responsabili del loro atto di impiegarli. I penalisti conoscono bene questa difficoltà, che si pone anche nel caso di un chirurgo che utilizza un robot medico o un farmaco. Sulla base della revisione giuridica dei LAWS richiesta dall'articolo 36 del Protocollo I, gli Stati che impiegano LAWS devono fornire ai comandanti militari e agli operatori istruzioni chiare su quando e in quali circostanze possono essere effettivamente utilizzati. L'operatore non è tenuto a comprendere il complesso funzionamento del robot, ma deve comprenderne il risultato, ovvero ciò che il sistema è in grado di fare e ciò che non è in grado di fare. La questione se l'assunzione del rischio costituisca un *dolo* sufficiente per i crimini di guerra dipende dalla legislazione penale applicabile al caso specifico. Il Tribunale Penale Internazionale per l'ex-Jugoslavia e diverse giurisdizioni nazionali ammettono « *recklessness* » come *mens rea* sufficiente per la responsabilità penale, mentre la Corte Penale Internazionale e altre giurisdizioni nazionali richiedono l'intenzionalità, in base alla quale un comandante incosciente non può essere ritenuto penalmente responsabile. Ciò può implicare la necessità di una legislazione specifica per regolamentare l'impiego dei LAWS.

Per quanto riguarda il produttore e il programmatore, le leggi penali nazionali spesso ritengono penalmente responsabili coloro che deliberatamente, per imprudenza o negligenza, costruiscono edifici o macchinari difettosi che causano la perdita di vite umane. Ad esempio, l'articolo 229 del Codice penale svizzero incrimina « chiunque, nell'esercizio della sua attività di direzione o di esecuzione di lavori di costruzione o di demolizione, viola intenzionalmente le norme di costruzione riconosciute ». La violazione degli standard dell'industria può costituire così un delitto autonomo o un'indicazione di una negligenza — anche sé né l'uno né l'altro sarebbe un crimine di guerra. Tutto questo presuppone che degli standard dell'industria siano adottati. Per le LAWS, questo processo incontra l'ostacolo del segreto militare. Nessun produttore di armi e nessun Stato vogliono che i propri concorrenti sappiano ciò che le sue armi sono o non sono in grado di fare. Tuttavia, gli esseri umani devono sempre definire come funzionerà l'autonomia, il che implica la necessità di redigere norme specifiche di diligenza dovuta sia per i produttori che per i comandanti, quantomeno a livello nazionale.

10. Conclusione.

In conclusione, i LAWS possono essere impiegati solo se, nonostante la loro intelligenza artificiale, è prevedibile che saranno in grado di rispettare il DIU. Rimane però controverso affermare se saranno mai in grado, come gli esseri umani, di osservarlo. Molti rifiutano questa possibilità sulla base del fatto che le decisioni relative agli obiettivi implicano giudizi di valore soggettivi.

Ciò premesso, se i LAWS fossero in grado di rispettare il DIU, il loro utilizzo potrebbe presentare molti vantaggi rispetto ad altre piattaforme d'arma letali, ma diverse norme del DIU e del diritto internazionale penale dovrebbero essere reinterpretate per essere applicabili agli esseri umani che producono o impiegano tali sistemi.

In ogni caso, tali sistemi non possono essere completamente autonomi per troppo tempo in quanto necessitano di essere costantemente alimentati con informazioni che consentano al sistema di tenere conto della natura dinamica del vantaggio militare, che svolge un ruolo importante sia nella definizione degli obiettivi militari che nell'applicazione della regola di proporzionalità.

Note bibliografiche

Amoroso D.-Tamburrini G. (2020), *Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues*, in *Current Robotics Reports* 1, p. 187 ss.

Asaro P. (2013), *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, in *International Review of the Red Cross*, 94, p. 687 ss.

Boothby B. (2025), *AI Warfare and the Law*, in *International Law Studies*, 104, p. 1 ss.

Backstrom A.-Henderson I. (2012), *New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews*, in *International Review of the Red Cross*, 94, p. 483 ss.

France, Ministère des armées, Comité d'éthique de la défense (2021), *Avis sur l'intégration de l'autonomie dans les systèmes d'armes létaux*, 19 aprile 2021.

Group of Governmental Experts on Lethal Autonomous Weapons Systems (2024), *Revised rolling text as of 8 November 2024.*

Heyns C. (2013), *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, UN Doc A/HRC/23/47.

Hoffberger-Pippan E.-Vohs V.-Köhler P. (2022), *Autonomous Weapons Systems: UN Expert Talks Facing Failure*, Berlin, Stifung Wissenschaft und Politik.

HOLLAND MICHEL A. (2020), *The Black Box, Unlocked: Predictability and Understandability in Military AI*, Geneva, UNIDIR.
HUMAN RIGHTS WATCH - INTERNATIONAL HUMAN RIGHTS CLINIC, HUMAN RIGHTS PROGRAM AT HARVARD LAW SCHOOL (2012), *Losing Humanity: The Case Against Killer Robots.*
HUMAN RIGHTS WATCH (2020), *Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control.*
HUMAN RIGHTS WATCH (2022), *Agenda for action: Alternative Processes for Negotiating a Killer Robots Treaty.*
INTERNATIONAL COMMITTEE OF THE RED CROSS (2021), *ICRC position on autonomous weapon systems* e *Background Paper*, Geneva, ICRC.
INTERNATIONAL COMMITTEE OF THE RED CROSS (2009), *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law*, Geneva, ICRC.
INTERNATIONAL LAW COMMISSION (2001), *Draft articles on Responsibility of States for Internationally Wrongful Acts, with commentaries*, in *ILC, Report of the International Law Commission on the Work of its 53rd Session (23 April-1 June and 2 July-10 August 2001)*, UN Doc A/56/10, p. 26 ss.
KWIK J. (2022), *A Practicable Operationalisation of Meaningful Human Control*, in *Laws*, 11, 3, p. 43 ss.
KWIK J. (2024), *Lawfully Using Autonomous Weapon Technologies*, The Hague, T.M.C. Asser Press.
KWIK J. (2025), *Digital Yes-Men: How to Deal With Sycophantic Military AI?*, in *Global Policy*, 16, p. 467 ss.
LIU H.-Y. (2012), *Categorization and Legality of Autonomous and Remote Weapons Systems*, in *International Review of the Red Cross*, 94, p. 627-652.
MOLNAR C. (2025), *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 3ª ed. München, autopubblicato.
SCHMITT M.N.-THURNHER J.S. (2013), *"Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict*, in *Harvard National Security Journal*, 4, p. 231 ss.
SINGER P.W. (2009), *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, New York, Penguin.
TRABUCCO L. (2025), *AI-Enabled Autonomous Weapons and Human Control, Part I: Human Control and Machine Learning Design and Development; Part II: Human Control and Military Commanders; Part III: Human Control and System Operators*, in *International Law Sudies*, 106, p. 534 ss.
UK MINISTRY OF DEFENCE (2022), *Ambitious, Safe, Responsible, Our Approach to the Delivery of AI-enabled Capability in Defence*, consultabile a partire da: *www.gov.uk*.
UNITED NATIONS OFFICE AT GENEVA (UNOG), *Disarmament, The Convention on Certain Conventional Weapons, Background on Lethal Autonomous Weapons Systems in the CCW*, consultabile a partire da: *www.unog.ch*.
US DEPARTMENT OF DEFENSE (2023), *Directive 3000.09: Autonomy in Weapon Systems*, consultabile a partire da: *www.war.gov*.

INTELLIGENZA ARTIFICIALE E TUTELA DELLA SALUTE

INTRODUZIONE AI PROFILI COSTITUZIONALI

di *Renato Balduzzi*

Secondo le più recenti analisi condotte a livello internazionale, il settore della salute si configura oggi come uno dei principali motori dell'innovazione tecnologica, affiancando per rilevanza il tradizionale ambito militare. Questo dato, sotto il profilo dell'impatto sociale e scientifico, può essere considerato positivamente, in quanto testimonia una crescente attenzione verso l'applicazione dell'intelligenza artificiale in contesti che incidono direttamente sul benessere collettivo. Tuttavia, è necessario sottolineare che le criticità generali associate allo sviluppo e all'implementazione dell'intelligenza artificiale si manifestano integralmente anche in questo settore, talvolta in forma accentuata. In particolare, nell'ambito sanitario, tali problematiche assumono una rilevanza ancora maggiore, sia per la delicatezza delle questioni trattate sia per le implicazioni etiche, giuridiche e sociali che ne derivano. Pertanto, la riflessione sui rischi e sulle opportunità offerte dall'intelligenza artificiale in campo medico deve necessariamente tenere conto di queste specificità, adottando un approccio rigoroso e multidisciplinare.

L'Italia, rispetto ad altri Paesi, dispone di strumenti normativi e concettuali particolarmente avanzati, grazie alla propria Costituzione che riconosce la salute come fondamentale diritto dell'individuo e interesse della collettività. L'aggettivo 'fondamentale' qualifica sia il diritto individuale sia l'interesse collettivo, una consapevolezza che si è consolidata in modo particolare a seguito della pandemia da Covid-19, ma che era già stata intuita dai nostri costituenti quasi ottant'anni fa.

L'intelligenza artificiale sta già da tempo rivoluzionando il settore sanitario, sia per quanto concerne l'assistenza ospedaliera sia per quanto riguarda l'assistenza territoriale. In particolare, la recente riforma dell'assistenza territoriale in Italia assume pieno significato nella misura in cui si riesca a sfruttare le potenzialità dell'intelligenza artificiale per individuare e organizzare i bisogni sanitari della popola-

zione. Tale riforma si fonda su un principio innovativo: il superamento del tradizionale rapporto unidirezionale tra bisogni e servizi.

Storicamente, infatti, il cittadino che avverte un bisogno sanitario si rivolge autonomamente al pronto soccorso, al medico di famiglia o allo specialista ambulatoriale. La logica della nuova assistenza territoriale, per contro, va nella direzione di affiancare, accanto al bisogno che va al servizio, il servizio che va al bisogno. Tuttavia, affinché il servizio possa effettivamente raggiungere il bisogno, è necessario essere in grado di conoscerlo e di organizzarlo in modo efficiente, e in questo processo non mancano certamente le criticità.

L'intelligenza artificiale, inoltre, non sta trasformando soltanto l'ambito dell'assistenza sanitaria, ma anche quello della ricerca biomedica e della formazione accademica e professionale. Questi diversi ambiti costituiscono lo sfondo complessivo della riflessione attuale sulle opportunità e sui rischi connessi all'adozione di sistemi intelligenti in sanità.

I rischi associati all'intelligenza artificiale in ambito sanitario si suddividono in due categorie: quelli comuni ad altri settori e quelli specifici del contesto medico. Tra i primi si annoverano gli errori di funzionamento dei sistemi, che sollevano rilevanti questioni di responsabilità, non solo per i clinici ma anche per tutti gli operatori coinvolti. Un ulteriore rischio, trasversale a molti ambiti, è quello della discriminazione: ad esempio, l'utilizzo di algoritmi addestrati su dati non rappresentativi può portare a risultati non equi, come nel caso di sistemi di diagnosi del cancro della pelle sviluppati esclusivamente su soggetti con carnagione chiara, senza la dovuta estensione a popolazioni con caratteristiche diverse. Si aggiunge poi il rischio della cosiddetta 'black box', ovvero la mancanza di trasparenza nei processi decisionali delle macchine, che non è esclusiva del settore sanitario ma qui assume particolare rilevanza. Infine, vi è il rischio di affidamento eccessivo nei confronti dei risultati prodotti dai sistemi intelligenti, una problematica che coinvolge direttamente medici e operatori sanitari.

La vera novità consiste nel fatto che stiamo assistendo a una trasformazione degli strumenti di intelligenza artificiale: da semplici mezzi, essi stanno progressivamente assumendo il ruolo di agenti autonomi. Questo cambiamento impone una riflessione profonda, soprattutto in relazione alla questione dell''ultima parola' nelle decisioni.

Dal punto di vista costituzionale, la domanda su chi debba detenere l'ultima parola è complessa: in uno Stato democratico fondato sulla Costituzione, infatti, non esiste un unico soggetto titolato a decidere in via definitiva. Né il giudice, che gode di una legittimazione diversa rispetto all'eletto, né l'eletto stesso, che è comunque vincolato al rispetto dello Stato di diritto e dei principi costituzionali, possono rivendicare tale prerogativa in modo esclusivo. L'ultima parola, piuttosto, è il risultato di una dinamica circolare tra i diversi attori istituzionali, e questa caratteristica assume particolare rilievo nel contesto delle decisioni sanitarie supportate dall'intelligenza artificiale.

INTRODUZIONE AI PROFILI CIVILISTICI

di *Antonio Albanese*

L'impiego dell'intelligenza artificiale in ambito sanitario offre indubbi vantaggi, quali la possibilità di ottenere diagnosi più accurate, l'alta precisione di alcuni interventi chirurgici, la personalizzazione dei trattamenti e più in generale il miglioramento dell'efficienza in termini di riduzione dei tempi e dei costi delle cure.

Tuttavia, l'utilizzo delle nuove tecnologie genera anche diversi rischi che devono essere attentamente considerati e che potrebbero comportare responsabilità per il medico che ha in carico il paziente e/o per la struttura che è obbligata nei suoi confronti al trattamento sanitario. In tal senso vengono in considerazione alcuni aspetti delle nuove tecnologie che caratterizzano le fattispecie di responsabilità connesse al loro utilizzo in termini diversi rispetto a quelle tradizionali.

Una significativa differenza consiste anzitutto nel massiccio utilizzo di dati che il sistema di intelligenza artificiale è in grado di esaminare e valutare al fine di effettuare una diagnosi maggiormente attendibile o di individuare la terapia più efficace.

Ciò pone in primo luogo un problema di affidabilità del materiale informativo elaborato: se questo è di scarsa qualità o presenta *bias*, l'IA può produrre diagnosi o raccomandazioni errate, con conseguenti danni per la salute. Per evitare tali rischi la struttura sanitaria deve predisporre apposite procedure per controllare che i dati utilizzati per addestrare i modelli di IA siano accurati, completi e rappresentativi della popolazione di cui fa parte il paziente.

L'adozione di queste cautele non esclude peraltro che le diagnosi e le terapie proposte dall'IA debbano essere soggette alla costante supervisione umana ed essere validate dai professionisti ai quali è affidata in ultima analisi la responsabilità della cura. Tale intervento è essenziale non solo per garantire la salute dei pazienti, ma anche per risolvere eventuali dilemmi etici, che richiedono scelte basate su criteri che non possono essere interiorizzati in un algoritmo.

Entrambe queste esigenze hanno del resto trovato riconoscimento nel documento *Ethics Guidelines for Trustworthy AI*, mediante il quale la Commissione Europea ha raccomandato che l'utilizzo dell'IA promuova i diritti fondamentali della persona, consentendo sempre una piena ed effettiva determinazione dell'essere umano che interagisce con essi.

In senso analogo il Regolamento UE 2017/745 sui dispositivi medici (MDR) stabilisce che i fabbricanti devono adottare un sistema di gestione del rischio per tutto il ciclo di vita del dispositivo, incluso il monitoraggio post-commercializzazione, il che implica che l'utilizzo di strumentazioni e sistemi basati sull'intelligenza artificiale debba avvenire con la partecipazione attiva dell'uomo per identificare e mitigare situazioni di pericolo che possono sopravvenire.

D'altra parte, affinché il medico possa esprimere un giudizio clinico effettivo e non meramente fittizio, esaminando criticamente le indicazioni dell'IA, è necessario che i criteri e i processi decisionali di questa siano conoscibili e valutabili. La mancanza di trasparenza su tali aspetti può del resto impedire o rendere più difficile identificare e correggere eventuali malfunzionamenti dei sistemi di IA, oltre a minare la fiducia dei potenziali pazienti.

Tutti i soggetti coinvolti nella progettazione e utilizzo degli strumenti di AI devono quindi assicurare che le decisioni cliniche, anche quelle assunte con l'ausilio di sistemi informatici avanzati, siano effettivamente espressione della professionalità del medico. Solo in tal modo è peraltro possibile prevenire la perdita di abilità dovuta alla dipendenza dalla tecnologia (c.d. *deskilling*), mantenendo e sviluppando competenze essenziali della persona umana.

Da questo punto di vista viene in evidenza un ulteriore profilo di novità, attinente all'esecuzione della prestazione professionale, dato dalla necessità che il personale sanitario acquisisca le specifiche competenze necessarie per il corretto utilizzo delle nuove tecnologie e che a tal fine sia adeguatamente formato per adoperare i dispositivi che si basano sull'IA e per interpretare i risultati da essa forniti. Tale obbligo di formazione ha trovato un significativo riconoscimento da parte della giurisprudenza statunitense nel caso *Taylor vs. v. Intuitive Surgical, Inc.*, deciso dalla Corte Suprema dello Stato di Washington, la quale ha ritenuto che il produttore di uno strumento chirurgico ad elevata automazione debba fornire adeguata formazione al personale sanitario incaricato del suo utilizzo e che, in mancanza, sia responsabile dei danni cagionati dal suo errato impiego.

Da un altro e diverso punto di vista, l'importanza dei dati analizzati dall'IA per determinare il trattamento più idoneo per i pazienti accresce il rischio di lesione della loro *privacy* o di altri diritti fondamentali della persona. Per questo la struttura sanitaria, oltre a ottenere il consenso della persona che ha in cura, deve garantire il rispetto della riservatezza delle informazioni che la riguardano e adottare misure idonee a prevenire possibili attacchi informatici che sfruttino le vulnerabilità del sistema per accedere a dati sensibili e, nei casi più gravi, per manipolare i risultati.

Alla luce delle diverse fattispecie finora esaminate è possibile individuare le misure di protezione e le regole di comportamento che i soggetti coinvolti devono adottare per sviluppare e implementare sistemi di IA sicuri, efficaci ed etici, massimizzando i benefici e minimizzando i rischi connessi all'utilizzo delle innovazioni tecnologiche nella cura dei pazienti.

È fondamentale che questi ricevano un'informazione completa e comprensibile, affinché possano prendere una decisione consapevole riguardo al proprio trattamento. Il medico, in quest'ottica, deve essere pronto a discutere con loro i dettagli e a rispondere apertamente a tutte le domande, mantenendo un approccio rispettoso dei diritti della persona.

Egli deve mantenere un ruolo attivo e critico nell'uso dell'IA in ambito clinico, assicurando che le tecnologie siano utilizzate in modo sicuro, efficace ed etico e a tal fine deve acquisire una specifica preparazione per offrire prestazioni conformi alle nuove regole tecniche.

La struttura sanitaria — ove necessario anche con la cooperazione del produttore dei dispositivi medicali — deve a sua volta adottare adeguate misure di protezione e di gestione dei rischi, che garantiscano un elevato standard di sicurezza ed efficacia delle cure, per ridurre l'incidenza statistica di possibili danni ai pazienti e, ove questi comunque si verifichino, per escludere o attenuare la responsabilità propria e dei medici che operano al suo interno.

INTRODUZIONE AI PROFILI PENALISTICI

di *Alain Dell'Osso*

Il dialogo con i saperi della medicina schiude al penalista un orizzonte nel quale il ricorso alle potenzialità delle intelligenze artificiali costituisce un dato empirico, un fatto della quotidianità dei clinici e dei chirurghi, un'evoluzione tecnologica già acquisita, per quanto, ovviamente, destinata a mutare e crescere nel futuro. A differenza di quanto avviene in relazione ad altri settori dell'attività umana, qui non occorre, cioè, sforzarsi troppo per immaginare quale potrà essere l'impatto delle *thinking machines* e valutarne le possibili ricadute rispetto alla regolazione pubblica: il giurista può limitarsi ad ascoltare i medici e comprendere immediatamente i momenti di integrazione tra intelligenze umane e artificiali nello svolgimento dell'attività sanitaria.

Ed è così che — semplificando enormemente quanto è assai meglio descritto nei contributi dei medici — si apprende che l'IA in medicina trova già applicazioni che spaziano da *triage* e prioritizzazione, alla diagnostica (attraverso, tra l'altro, il rilevamento e la classificazione di lesioni o *pattern*), alla previsione di rischio, all'automazione documentale, al coordinamento dei percorsi di cura, al monitoraggio remoto e alla telemedicina. Nella diagnostica, assumono particolare rilievo i *Clinical Decision Support Systems* (CDSS): sistemi di intelligenza artificiale che analizzano grandi moli di dati clinici e di letteratura medico-scientifica per fornire al medico indicazioni diagnostiche e terapeutiche. Si basano prevalentemente su tecniche di *supervised learning*: apprendono da insiemi di esempi etichettati e, a fronte di nuovi casi, generano ipotesi di diagnosi o piani terapeutici connessi a punteggi probabilistici. I CDSS offrono elevata velocità di elaborazione e, in vari ambiti (radiologia, oncologia, dermatologia), hanno mostrato livelli di accuratezza molto promettenti, con potenziale miglioramento dell'efficacia e tempestività delle cure (FRAGASSO, 2025).

Appuntando l'attenzione sulla chirurgia, IA viene già impiegata nella pianificazione preoperatoria basata su immagini, nella guida in-

traoperatoria con visione artificiale e realtà aumentata, nella robotica con vari gradi di autonomia, nella prevenzione degli errori e nel supporto all'anestesia, oltre che nella formazione, nella valutazione delle performance e nella gestione della sala operatoria. I robot in chirurgia risultano capaci di movimenti più precisi dell'essere umano e possono operare anche da remoto; i sistemi in uso richiedono comunque controllo umano costante: il modello prevalente è '*master-slave*' (teleoperazione), in cui il chirurgo comanda e il robot esegue eliminando tremori e migliorando la precisione (AMORE-ROSSERO, 2024).

Pur nelle diverse declinazioni, il minimo comune denominatore di tali sistemi — anche qui la semplificazione è massima — può essere individuato nella capacità di elaborare dati sulla base dei quali produrre previsioni, raccomandazioni o decisioni in funzione di obiettivi finali definiti dal sanitario; il tutto con gradi variabili di autonomia.

Le potenzialità sono immediatamente intuibili: come ha magistralmente chiarito nei saluti rivolti in apertura al seminario del 4 giugno il compianto Professore Scambia, la 'bravura' di un medico deriva dalla sua competenza, costruita attraverso anni di studio ed esperienza, vale a dire mediante l'acquisizione di un ponderoso bagaglio di dati che gli consente di comprendere la patologia del paziente e individuare — attingendo sempre a quel *background* di conoscenze — la risposta terapeutica più opportuna; per quanto, però, un medico possa aver approfondito e praticato, non riuscirà mai ad acquisire e gestire una mole di dati comparabile a quella processabile dalle *learning machines*. Lo scarto è fisiologico ed è immediatamente evidente il contributo che il sanitario può ricavare dal ricorso all'IA nell'interesse del paziente e, con esso, il motivo del successo già riscosso dalla *partnership* medico-macchina.

Queste minime osservazioni indirizzano la prospettiva del giurista verso due primi possibili orizzonti di approfondimento: la regolazione dell'acquisizione e dell'impiego dei dati, vale a dire della linfa che alimenta il funzionamento delle macchine; e l'allocazione delle responsabilità nel caso di eventi avversi: i due *partner* giocano ruoli e presentano profili, in quest'ottica, assai diversi. Anzitutto, ed è il dato più superficiale, la macchina non si presenta come un canonico centro di imputazione della responsabilità: riprendendo la felice formula con la quale Coffee descriveva lo iato esistente in tale prospettiva tra persone fisiche e giuridiche, si può dire che anche le mac-

chine abbiano *No Soul to Damn, No Body to Kick* (COFFEE, 1981). Come dimostra, però, proprio l'esperienza della *corporate liability* non si tratta di un dato cruciale nell'escludere una responsabilizzazione parapenalistica; semplicemente di un fattore che impone di disegnare un sistema di criteri che — pur influenzato dai principi generali della materia penalistica — si muova su binari autonomi in grado di cogliere le peculiarità del destinatario del precetto e della sanzione (PIERGALLINI, 2021).

Ciò posto, ben si dovrebbe comprende perché si è ritenuto opportuno un approfondimento sul possibile statuto di una responsabilità colposa per i casi di *malpractice* delle macchine, tema sul quale si appunteranno le riflessioni di Massaro.

Poter confidare sulla puntuale tematizzazione di tali tematiche ad opera di così brillanti colleghi, consente di dedicare i residui spazi di queste osservazioni introduttive a considerazioni, per così dire, rapsodiche, orientate più a cogliere possibili spunti di riflessione e problemi, che non a elaborare soluzioni, sia pur fatalmente provvisorie.

Uno stimolante orizzonte di riflessione al quale dedicare un cursorio cenno appare quello del rapporto medico-macchina nella prospettiva della responsabilità del sanitario (IAGNEMMA, 2020; BALBI et al., 2022).

Come osservato, i CDSS forniscono suggerimenti diagnostico-terapeutici elaborati sulla base di estese masse di dati: non esiste, però, alcuna regola che imponga al sanitario di adeguarsi agli *output* algoritmici come ''standard of care''; la decisione finale resta sempre del medico, che mantiene autonomia e, dunque, responsabilità. Le fonti normative di riferimento sembrano, anche *pro futuro*, orientate a tenere fermo il dovere del medico di fare sintesi tra il contributo fornito dalle macchine, le proprie competenze e le esigenze dello specifico contesto nel quale la scelta terapeutica deve ambientarsi. In questo senso l'art. 14 dell'AI Act (Regolamento UE 2024/1689) e il Disegno di legge A.S. 1146/2024 (in itinere) sull'IA: in entrambi i testi si ribadisce che i sistemi di IA sono solo di supporto alla decisione del medico.

In assenza di norme che attribuiscano valore vincolante alle indicazioni dei CDSS, la colpa del medico sarà, quindi, da valutare secondo i criteri ordinari, senza che l'adesione o il dissenso rispetto all'algoritmo siano, da soli, dirimenti (FRAGASSO, 2025).

Vengono, allora, in esame due opposte macro-ipotesi di *malpractice* (accomunate, quindi, dalla produzione di un nocumento al pa-

ziente): quella del medico che aderisce al suggerimento errato del CDSS, e quella, antitetica, del terapeuta che si discosta dalla corretta indicazione ricevuta. Si tratta evidentemente di questioni complesse, difficilmente schematizzabili; volendo, comunque, provare a tracciare delle coordinate generali di orientamento, si può osservare che nella prima ipotesi l'adesione all'*output* non esclude *ex se* la colpa: occorrerà verificare se quell'indicazione risultasse non solo tecnicamente corretta — si consideri che tutta la letteratura di riferimento ricorda sempre le possibilità di errore del macchinario, invitando l'utilizzatore a non cedere alla c.d. *technology complacency* — ma, pure, giusta alla luce delle sfaccettature del caso concreto. Nel secondo caso, la sola divergenza dall'indicazione ricevuta non significa ancora che vi sia stata la violazione di una regola cautelare: non è, appunto, quella suggerita dall'AI necessariamente la regola da seguire e, nel valutare *ex post* quale fosse il comportamento doveroso, si dovrà, da un lato, diffidare dal c.d. *hindsight bias* e, dall'altro, dalla già citata *technology complacency*, tenendo a mente che ciò che si chiede al sanitario è proprio di preservare un autonomo giudizio critico sul pensiero della macchina e, quindi, di potersi discostare liberamente dalle sue indicazioni.

Va detto che il pur ragionevole criterio della sorveglianza umana, che cela all'evidenza un residuo di sfiducia nelle *capabilities* delle *learning machines*, rischia, se sopravvalutato, di risolversi in una sorta di zavorra, che finisce per tarpare le ali alle effettive potenzialità dei sistemi di intelligenza artificiale.

Si è osservato, infatti, che il loro valore aggiunto consiste nell'attitudine a elaborare una base di dati e di conoscenze assai maggiore di quella dominabile dall'essere umano e, così, attingere, per esempio, a diagnosi o a scelte di cura più calzanti: pretendere che, laddove il sistema dia indicazioni diverse da quelle alle quali sarebbe pervenuto in autonomia il medico, sia poi quest'ultimo, dal 'suo piccolo', a validare la decisione della macchina, senza poter spesso ricostruire il percorso logico seguito o il *set* informativo compulsato, rischia di livellare al ribasso l'apporto delle nuove tecnologie.

In altri termini: se si ambisce a perseguire risultati migliori di quelli raggiungibili dalla persona fisica, si sta assumendo, per definizione, che il sanitario avrebbe agito *diversamente*; pur essendo chiare le ragioni per le quali si prescrive, a fronte di un siffatto *gap*, di non affidarsi ciecamente alla macchina, che si sa non essere infallibile, e di mantenere il controllo della situazione, sembra necessario riflettere attentamente sulla fisionomia del controllo che ci si attende.

Auspicare una puntuale ratifica rischia di indurre il medico a fidarsi solo delle proprie conoscenze e competenze e a fermarsi a quelle, annichilendo il contributo della macchina. Sembra, allora, opportuno pensare a modelli di falsificazione più che di conferma dell'indicazione dell'AI: il sanitario dovrebbe, cioè, fare attenzione a cogliere segnali che lo portino a ritenere che l'*output* sia scorretto o inapplicabile in concreto; una volta superati tali *red flags*, dovrebbe poter seguire il suggerimento pur senza essere in grado di coglierne i singoli passaggi e senza temere di dover rendere conto della correttezza di un dato che poteva sfuggire alle sue capacità.

Su questo e sugli altri temi oggetto delle relazioni dei colleghi si appunterà la riflessione della dottrina e della giurisprudenza dei prossimi anni in un contesto nel quale difficilmente si potrà pensare di pervenire a soluzioni precostituite e definitive, dovendo seguire un processo tecnologico che si evolve con ritmi inediti e sconosciuti anche alle più recenti rivoluzioni tecnologiche.

Note bibliografiche

Amore N.-Rossero E. (2024), *Robotica e intelligenza artificiale nell'attività medica. Organizzazione, autonomia, responsabilità*, Bologna, Il Mulino.

Balbi G.-De Simone F.-Esposito A.-Manacorda S. (a cura di) (2022), *Diritto penale e intelligenza artificiale "nuovi scenari*, Torino, Giappichelli.

Coffee Jr. J. (1981), *"No Soul to Damn: No Body to Kick": An Unscandalized Inquiry into the Problem of. Corporate Punishment*, in *Michigan Law Review*, 78, p. 386 ss.

Fragasso B. (2025), *Intelligenza artificiale e responsabilità penale*, Torino, Giappichelli.

Iagnemma C. (2020), *I 'robot medici': profili problematici in tema di alleanza terapeutica e di responsabilità penale*, in *Corti supreme e salute*, 2, p. 1 ss.

Piergallini C. (2020), *Intelligenza artificiale: da 'mezzo' a 'autore' del reato?*, in *Rivista italiana di diritto e procedura penale*, 3, p. 1743 ss.

(IL FATTO)
L'IMPATTO DELL'INTELLIGENZA ARTIFICIALE IN AMBITO SANITARIO

di *Antonio Oliva* e *Francesca Cazzato*

Negli ultimi decenni, l'intelligenza artificiale (IA) ha progressivamente permeato il settore sanitario, delineando ambiti di applicazione sempre più estesi sia nel contesto clinico sia in ambito chirurgico, con implicazioni significative nei processi diagnostico-terapeutici (Alowais-Alghamdi-Alsuhebany-Alqahtani-Alshaya-Almohareb-Aldairem-Alrashed-Bin Saleh-Badreldin-Al Yami-Al Harbi-Albekairy, 2023). Seppur inizialmente confinata ad un contesto di ricerca puramente accademica, l'intelligenza artificiale ha rapidamente assunto un ruolo di rilievo, trasformandosi da strumento sperimentale a risorsa strategica per la pratica medica (Kaul-Enslin-Gross, 2020).

In ambito sanitario, tali sistemi sono in grado di assolvere ad un ampio spettro di compiti, caratterizzati da diversi livelli di autonomia decisionale e capacità operativa, contribuendo al contempo a mitigare alcuni potenziali *bias* cognitivi insiti nei processi decisionali umani. Attualmente, le applicazioni dell'intelligenza artificiale spaziano dalla stima del rischio di insorgenza di patologie acute e croniche alla valutazione prognostica in ambito oncologico, dalla gestione ottimizzata delle terapie all'organizzazione dei servizi clinico-assistenziali, fino allo sviluppo di nuovi farmaci (Oliva-Grassi-Vetrugno-Rossi-Della Morte-Pinchi-Caputo, 2022).

In termini più generali, l'intelligenza artificiale si definisce come un insieme di sistemi tecnologici computerizzati capaci di simulare processi cognitivi umani al fine di eseguire specifiche funzioni o azioni predeterminate (Chen-Decary, 2020). Questa tecnologia racchiude un insieme eterogeneo di metodologie, tra cui l'apprendimento automatico (*machine learning*, ML), l'apprendimento profondo (*deep learning*, DL) e l'elaborazione del linguaggio naturale (*natural language processing*, NLP). In particolare, il *deep learning* configura un sottoinsieme del *machine learning*, caratterizzato dall'impiego di

reti neurali profonde in grado di apprendere rappresentazioni complesse dei dati (European Parliament, 2022).

In relazione alla tipologia di dati e all'inclusione dei risultati, gli algoritmi di *machine learning* possono essere classificati principalmente in due categorie, ovvero apprendimento supervisionato e apprendimento non supervisionato (Jiang-Jiang-Zhi-Dong-Li-Ma-Wang-Dong-Shen-Wang, 2017). L'apprendimento non supervisionato si distingue per la propria capacità di estrazione automatica delle caratteristiche, permettendo l'identificazione di pattern e strutture intrinseche nei dati senza la necessità di etichette predefinite. Al contrario, l'apprendimento supervisionato si presta in misura maggiore alla modellazione predittiva, poiché permette di stabilire relazioni tra le caratteristiche di *input* e gli esiti di interesse (*output*). Recentemente, è stato altresì sviluppato un approccio semi-supervisionato, che integra elementi dell'apprendimento supervisionato e di quello non supervisionato. Tale metodologia risulta particolarmente efficace in contesti in cui i risultati — nello specifico di interesse clinico- sono parzialmente assenti, poiché consente di sfruttare simultaneamente dati etichettati e non etichettati, ottimizzando le prestazioni predittive del modello. La crescente diffusione dell'intelligenza artificiale in molteplici ambiti di applicazione ha reso necessario un intervento normativo aggiornato e specifico, che ha condotto all'entrata in vigore, il 1° agosto 2024, del Regolamento (UE) 2024/1689 del Parlamento Europeo e del Consiglio del 13 giugno 2024, volto a disciplinare in maniera armonizzata l'intelligenza artificiale, e noto anche come AI Act. Il Regolamento ha l'obiettivo di istituire un quadro normativo armonizzato per lo sviluppo, l'immissione sul mercato e l'utilizzo di prodotti e servizi basati sull'intelligenza artificiale nell'Unione Europea, ponendo particolare enfasi sulla gestione dei rischi inerenti alla salute, sulla sicurezza e sui diritti fondamentali degli individui.

Il testo normativo prevede la designazione di autorità competenti da parte degli Stati membri, nonché la costituzione di un Comitato europeo per l'intelligenza artificiale, con il compito di sorvegliare l'applicazione e l'attuazione del Regolamento. Inoltre, l'AI Act introduce sanzioni amministrative pecuniarie per la violazione delle disposizioni in esso contenute, conferendo un regime di responsabilità volto a garantire la conformità normativa e la tutela dei diritti fondamentali. Peraltro, il Regolamento (UE) 2024/1689 prevede la classificazione dei sistemi di intelligenza artificiale in base al livello di rischio

connesso al loro utilizzo, attribuendo particolare importanza ai cd. sistemi '*ad alto rischio*', soggetti a rigorose misure di controllo e vigilanza. Per essere considerati tali, i sistemi devono integrare prodotti già regolamentati dalla normativa comunitaria e sottoposti alla supervisione di autorità terze, oppure rientrare in una delle otto categorie di utilizzo individuate dal Legislatore europeo e riportate nell'Allegato III del medesimo documento.

Tra queste categorie, un ruolo di primaria rilevanza è rivestito dai servizi pubblici essenziali, con particolare riferimento all'ambito sanitario, considerata la loro incidenza diretta sulla salute e sul benessere dei cittadini. In questo contesto, l'intelligenza artificiale in ambito sanitario si configura principalmente come una tipologia di intelligenza artificiale dedicata ad applicazioni specifiche in medicina ed assistenza sanitaria, il cui impiego risulta finalizzato a supportare applicazioni cliniche specifiche, ottimizzare i percorsi diagnostico-terapeutici e migliorare la gestione complessiva dei servizi. Pertanto, gli strumenti di intelligenza artificiale hanno dimostrato un potenziale significativo non solo nel perfezionare la qualità della cura, ma anche nell'accrescere l'efficienza complessiva del sistema sanitario, operando su due livelli complementari: quello clinico e quello amministrativo.

Sul piano clinico, l'intelligenza artificiale si configura come strumento di supporto avanzato per la diagnosi e la previsione delle traiettorie evolutive dello stato di salute dei pazienti, l'individuazione e raccomandazione di opzioni terapeutiche, l'assistenza nella pianificazione e conduzione di interventi chirurgici nonché il monitoraggio clinico continuativo. Nello specifico, l'intelligenza artificiale trova applicazione privilegiata nell'analisi automatizzata delle immagini radiologiche, consentendo la segmentazione delle strutture anatomiche e l'identificazione di reperti di rilievo, con conseguente incremento dell'accuratezza diagnostica ed agevolazione dell'attività dei professionisti. In ambito di medicina d'urgenza, l'intelligenza artificiale si configura quale strumento idoneo a rendere più efficiente e trasparente il processo di *triage*, nonché a razionalizzare i flussi organizzativi, mentre in un *setting* chirurgico fornisce ausilio nel supporto di decisioni cliniche in scenari caratterizzati da elevata complessità tecnica. Nel campo della neurologia e della salute mentale, l'intelligenza artificiale ha evidenziato potenzialità di rilievo nella diagnosi precoce e nel monitoraggio delle condizioni psichiatriche. Inoltre, si è dimostrata utile anche nella gestione delle patologie croniche e nella presa in carico della popolazione geriatrica (ELLAHHAM-ELLAHHAM-SIMSEKLER, 2020).

Parallelamente, l'impiego di modelli predittivi basati su algoritmi di intelligenza artificiale può consentire, in specifici contesti applicativi, una stima del rischio potenzialmente più accurata rispetto ad alcune metodologie statistiche tradizionali, favorendo la messa in atto di interventi mirati e tempestivi. D'altro canto, sul piano amministrativo, l'intelligenza artificiale si configura come uno strumento strategico volto a ridurre il carico operativo dei fornitori di servizi sanitari e ad ottimizzare l'organizzazione delle risorse. Ciò si realizza attraverso la registrazione automatizzata delle note cliniche, la razionalizzazione dei processi gestionali e l'automazione di compiti gravosi, svincolando tempo e risorse umane per attività a maggiore valore clinico-assistenziale.

Di conseguenza, i potenziali destinatari ed utilizzatori dei sistemi di intelligenza artificiale in ambito sanitario comprendono una pluralità di figure professionali ed organizzative: i medici, che ne impiegano le funzionalità a supporto dei processi diagnostico-terapeutici; il personale infermieristico, che se ne avvale per integrare le attività di reparto ed i percorsi assistenziali; le strutture sanitarie, che la utilizzano per la programmazione e la gestione degli accessi e degli appuntamenti; nonché i dirigenti e gli amministratori delle strutture sanitarie, chiamati ad assumere decisioni strategiche in materia di dotazioni organiche, allocazione delle risorse e pianificazione dei servizi (United States Government Accountability Office, 2020). In quest'ottica, lo sviluppo di sistemi di intelligenza artificiale destinati all'impiego in un contesto sanitario, costituisce un processo articolato e dinamico, che richiede una chiara definizione del quesito clinico, l'utilizzo di *dataset* dettagliatamente curati ed appropriati, derivati dalla disamina delle cartelle cliniche elettroniche. Per tali ragioni, a seguito di una fase iniziale di valutazione dei dati retrospettivi, l'affidabilità e l'impatto clinico devono essere confermati mediante validazioni prospettiche in contesti reali. Per garantire l'adozione su larga scala, gli strumenti devono inoltre dimostrare capacità di adattarsi a popolazioni e *setting* differenti. Infine, anche dopo l'implementazione, il monitoraggio continuo e l'aggiornamento periodico sono imprescindibili, alla luce dell'evoluzione scientifica e delle prescrizioni normative a tutela della sicurezza del paziente e della tracciabilità decisionale (United States Government Accountability Office, 2020; Bajwa-Munir-Nori-Williams, 2021). L'interesse crescente verso le applicazioni dell'intelligenza artificiale in ambito medico trova un documentato riscontro anche nell'analisi della produzione scientifica, che evi-

denzia un incremento significativo nel numero delle pubblicazioni nel corso degli ultimi anni, con una marcata focalizzazione nell'ambito oncologico e delle malattie cardiovascolari e del sistema nervoso (Jiang-Jiang-Zhi-Dong-Li-Ma-Wang-Dong-Shen-Wang, 2017).

La produzione risulta prevalentemente trainata dagli Stati Uniti, seguiti da Cina, Regno Unito, India, Australia e Canada, a conferma di un crescente interesse che si va consolidando in più continenti. Permangono, tuttavia, aree geografiche — come Russia, gran parte dell'Europa orientale e vaste regioni dell'Africa — ancora marginalmente rappresentate, indice di un coinvolgimento scientifico ancora disomogeneo su scala globale (Secinaro-Calandra-Secinaro-Muthurangu-Biancone, 2021). L'adozione crescente e capillare di sistemi di intelligenza artificiale in ambito sanitario, pur costituendo un volano di innovazione nell'ambito clinico-assistenziale, solleva al contempo una serie di criticità che assumono rilevanza sia sul piano etico che medico-legale. Infatti, la promessa di algoritmi e modelli di intelligenza artificiale in grado di potenziare la resa diagnostica, la personalizzazione dei trattamenti, e l'ottimizzazione dei percorsi amministrativo-organizzativi, deve necessariamente confrontarsi con il rischio di conseguenze indesiderate correlate all'impiego di queste emergenti tecnologie (Ellahham-Ellahham-Simsekler, 2020; European Parliament, 2022; United States Government Accountability Office, 2020).

Un primo *vulnus* è agevolmente riconoscibile nel danno derivante da errori diagnostici connessi all'uso di algoritmi: eventuali falsi negativi possono determinare la mancata identificazione di patologie a prognosi severa, mentre i falsi positivi possono esporre i pazienti a trattamenti non necessari, con conseguenti danni iatrogeni e dispendio ingiustificato di risorse economiche (Ellahham-Ellahham-Simsekler, 2020; Jassar-Adams-Zarzeczny-Burbridge, 2022). A tali rischi si affiancano poi quelli connessi all'uso improprio degli strumenti digitali. Infatti, la sola robustezza tecnica di un algoritmo non è sufficiente a garantire l'appropriatezza clinica delle decisioni in un caso specifico, poiché l'efficacia dei sistemi di intelligenza artificiale risulta subordinata in larga misura alla capacità degli operatori sanitari e, talvolta anche dei pazienti, di interpretare ed applicare correttamente le raccomandazioni generate.

Ciò rende indispensabile l'implementazione di percorsi formativi mirati e l'adozione di linee guida operative chiare, tali da orientare un uso consapevole di questi strumenti. Non meno rilevante è la questione relativa ai *bias*. La rappresentatività dei *dataset* utilizzati per

l'addestramento degli algoritmi risulta cruciale, poiché eventuali disomogeneità rispetto a variabili quali genere, etnia, età e/o condizione socio-economica possono influenzare la performance del modello. Ciò comporta delle discrepanze nella precisione delle previsioni tra diversi gruppi di pazienti, con conseguenze rilevanti in merito all'appropriatezza delle indicazioni cliniche generate dall'intelligenza artificiale. A ciò si aggiunge l'elaborazione su vasta scala di dati sanitari sensibili, che può esporre al rischio di violazione della riservatezza, all'utilizzo improprio delle informazioni ovvero un impiego per finalità non cliniche. Parimenti, la tracciabilità dei processi di sviluppo e l'esplicabilità delle decisioni algoritmiche costituiscono requisiti indispensabili per consolidare la fiducia degli utilizzatori dei sistemi di intelligenza artificiale in modo tale da soddisfare le esigenze di *accountability*. Tuttavia, la natura cd. '*black box*' di numerose reti neurali, incapace di esplicitare i processi logici sottesi alle proprie previsioni, mina l'interpretabilità e la trasparenza dei processi decisionali, ostacolando sia la comprensione clinica sia la valutazione medico-legale e rendendo problematica l'attribuzione di responsabilità in caso di contenzioso (Jassar-Adams-Zarzeczny-Burbridge, 2022). In tale contesto, la questione della *liability* permane tuttora incerta e oggetto di dibattito scientifico e giuridico, sottolineando l'urgenza di un quadro normativo articolato. Infatti, poiché l'intelligenza artificiale non può essere ritenuta né moralmente né legalmente responsabile, eventuali errori ed imprecisioni di codesti emergenti modelli digitali potrebbero ricadere su una pluralità di attori (e.g., sviluppatori, distributori ed utilizzatori) con conseguenti relative implicazioni in merito alla ripartizione dell'onere della prova e dell'onere finanziario (Comandé, 2019).

Concretamente, eventuali negligenze possono manifestarsi sotto forma di difetti di programmazione, carenze di supervisione, condotte erronee del professionista sanitario o malfunzionamenti dell'algoritmo stesso. Risulta pertanto imprescindibile operare una distinzione, ad esempio, tra i casi in cui l'errore sia imputabile al medico per un utilizzo improprio dei sistemi di intelligenza artificiale e quelli in cui la responsabilità ricada unicamente sulla struttura sanitaria che gestisce e configura lo strumento (Cestonaro-Delicati-Marcante-Caenazzo-Tozzo, 2023; Mezrich, 2022).

Ne consegue che, in assenza di indicazioni cogenti in merito al riparto di responsabilità, specie in ambito penale, i professionisti sanitari potrebbero mostrarsi riluttanti nell'adottare soluzioni basate sul-

l'intelligenza artificiale. Parallelamente, i pazienti potrebbero perdere fiducia nel sistema sanitario in assenza di un soggetto chiaramente individuabile come garante. Tuttavia, se da parte dell'utenza permangono riserve circa l'impiego su vasta scala dell'intelligenza artificiale nei percorsi di cura, sul versante puramente professionale talune *surveys* hanno evidenziato come una percentuale significativa di clinici ritenga che tali tecnologie possano apportare un sensibile miglioramento al proprio ambito specialistico, incidendo altresì sulla futura riorganizzazione della forza lavoro sanitaria (Scheetz-Rothschild-Mcguinness-Hadoux-Soyer-Janda-Condon-Oakden-Rayner-Palmer-Keel-Van Wijngaarden, 2021; Eschert-Schwendicke-Krois-Bohner-Vinayahalingam-Hanisch, 2022).

È prevedibile che l'intelligenza artificiale determinerà trasformazioni rilevanti nell'assistenza sanitaria, soprattutto nelle discipline che richiedono modelli prognostici di elevata precisione, come l'oncologia, e in quelle basate sul riconoscimento di *pattern* complessi, quali la radiologia e l'anatomia patologica. Tuttavia, in una prospettiva futura, l'evoluzione dell'intelligenza artificiale nel settore clinico-assistenziale configura un'opportunità senza precedenti per ottimizzare la qualità dei percorsi di cura, pur implicando nuovi rischi derivanti dalla ridefinizione del ruolo decisionale del medico (Jassar-Adams-Zarzeczny-Burbridge, 2022). Infatti, l'affidamento intensivo dei professionisti sanitari ai sistemi di intelligenza artificiale potrebbe comportare il rischio di un progressivo depauperamento delle competenze cliniche (i.e., *deskilling*) e di una maggiore suscettibilità ad interpretazioni erronee delle informazioni sanitarie nonché di un affidamento passivo ai risultati algoritmici senza una valutazione critica della loro attendibilità (i.e., *automation bias*) (Cabitza-Rasoini-Gensini, 2017; Reis-Reis-Kunde, 2025; Oliva-Grassi-Vetrugno-Rossi-Della Morte-Pinchi-Caputo, 2022). In particolare, l'impiego di questi emergenti strumenti di supporto decisionale potrebbe limitare la capacità del professionista di interpretare autonomamente segni e sintomi e, ove necessario, anche di discostarsi dalle indicazioni generate dall'intelligenza artificiale. A ciò si aggiunge che la natura adattiva di tali sistemi, fondati su processi di auto-apprendimento ed auto-aggiornamento, espone al rischio di una rapida obsolescenza regolatoria, con la conseguente creazione di vuoti normativi potenzialmente pregiudizievoli (Oliva-Grassi-Vetrugno-Rossi-Della Morte-Pinchi-Caputo, 2022).

Tale scenario, oltre a prefigurare un incremento del contenzioso sanitario specificamente connesso all'impiego di sistemi di intelligenza

artificiale, impone una riflessione articolata in ordine alla cornice normativa della responsabilità professionale, civile e soprattutto penale, dell'esercente la professione sanitaria, chiamato a confrontarsi con l'utilizzo di questi nuovi ed emergenti strumenti tecnologici. Tuttavia, in un prossimo futuro, le summenzionate criticità potrebbero essere efficacemente mitigate attraverso una regolamentazione mirata ed una progettazione responsabile, supportata dal contributo di clinici, bioeticisti e *stakeholder* qualificati, al fine di assicurare la tutela del paziente, la robustezza dei percorsi clinico-assistenziali e la chiarezza nella ripartizione delle responsabilità nell'impiego di sistemi di intelligenza artificiale.

Note bibliografiche

Alowais S.A.-Alghamdi S.S.-Alsuhebany N.-Alqahtani T.-Alshaya A.I.-Almohareb S.N.-Aldairem A.-Alrashed M.-Bin Saleh K.-Badreldin H.A.-Al Yami M.S.-Al Harbi S.-Albekairy A.M. (2023), *Revolutionizing healthcare: the role of artificial intelligence in clinical practice*, in *BMC Medical Education*, 23, 1, p. 689 ss.

Bajwa J.-Munir U.-Nori A.-Williams B. (2021), *Artificial intelligence in healthcare: transforming the practice of medicine*, in *Future Healthcare Journal*, 8, 2, p. 188 ss.

Cabitza F.-Rasoini R.-Gensini G.F. (2017), *Unintended Consequences of Machine Learning in Medicine*, in *JAMA*, 318, 6, p. 517 ss.

Cestonaro C.-Delicati A.-Marcante B.-Caenazzo L.-Tozzo P. (2023), *Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review*, in *Frontiers in Medicine*, 10.

Chen M.-Decary M. (2020), *Artificial intelligence in healthcare: An essential guide for health leaders*, in *Healthcare Management Forum*, 33, 1, p. 10 ss.

Comandé G. (2019), *Intelligenza artificiale e responsabilità tra liability e accountability. Il carattere trasformativo dell'IA e il problema della responsabilità*, in *Analisi Giuridica dell'Economia*, 1, p. 169 ss.

Ellahham S.-Ellahham N.-Simsekler M.C.E. (2020), *Application of Artificial Intelligence in the Health Care Safety Context: Opportunities and Challenges*, in *American Journal of Medical Quality: The Official Journal of the American College of Medical Quality*, 35, 4, p. 341 ss.

Eschert T.-Schwendicke F.-Krois J.-Bohner L.-Vinayahalingam S.-Hanisch M. (2022), *A Survey on the Use of Artificial Intelligence by Clinicians in Dentistry and Oral and Maxillofacial Surgery*, in *Medicina (Kaunas, Lithuania)*, 58, 8, p. 1059 ss.

European Parliament (2022), *Artificial Intelligence in Healthcare. Applications, risks and ethical and societal impacts*, EPRS European Parliamentary Research Service, Scientific Foresight Unit (STOA) PE 729.512.

Jassar S.-Adams S.J.-Zarzeczny A.-Burbridge B.E. (2022), *The future of artificial intelligence in medicine: Medical-legal considerations for health leaders*, in *Healthcare Management Forum*, 35, 3, p. 185 ss.

Jiang F.-Jiang Y.-Zhi H.-Dong Y.-Li H.-Ma S.-Wang Y.-Dong Q.-Shen H.-Wang Y. (2017), *Artificial intelligence in healthcare: past, present and future*, in *Stroke and Vascular Neurology*, 2, 4, p. 230 ss.

Kaul V.-Enslin S.-Gross S.A. (2020), *History of artificial intelligence in medicine*, in *Gastrointestinal Endoscopy*, 92, 4, p. 807 ss.

Mezrich J.L. (2022), *Demystifying Medico-legal Challenges of Artificial Intelligence Applications in Molecular Imaging and Therapy*, in *PET Clinics*, 17, 1, p. 41 ss.

Oliva A.-Grassi S.-Vetrugno G.-Rossi R.-Della Morte G.-Pinchi V.-Caputo M. (2022), *Management of Medico-Legal Risks in Digital Health Era: A Scoping Review*, in *Frontiers in Medicine*, 8, p. 821756 ss.

Regulation (Eu) 2024/1689 Of The European Parliament And Of The Council (2024), in *Official Journal of the European Union*.

Reis M.-Reis F.-Kunde W. (2025), *Public Perception of Physicians Who Use Artificial Intelligence*, in *JAMA Network Open*, 8, 7, p. 2521643 ss.

Scheetz J.-Rothschild P.-Mcguinness M.-Hadoux X.-Soyer H.P.-Janda M.-Condon J.J.J.-Oakden-Rayner L.-Palmer L.J.-Keel S.-Van Wijngaarden P. (2021), *A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology*, in *Scientific Reports*, 11, 1, p. 5193 ss.

Secinaro S.-Calandra D.-Secinaro A.-Muthurangu V.-Biancone P. (2021), *The role of artificial intelligence in healthcare: a structured literature review*, in *BMC Medical Informatics and Decision Making*, 21, 1, p. 125 ss.

United States Government Accountability Office (2020), *Report to Congressional Requesters. TECHNOLOGY ASSESSMENT. Artificial Intelligence in Health Care. Benefits and Challenges of Technologies to Augment Patient Care*.

(IL FATTO)
INTELLIGENZA ARTIFICIALE NEL CARCINOMA OVARICO AVANZATO: INTEGRAZIONE MULTIMODALE, STANDARDIZZAZIONE DECISIONALE E CHIRURGIA DI PRECISIONE

di *Andrea Rosati* e *Anna Fagotti*

L'intelligenza artificiale (AI) sta assumendo un ruolo sempre più centrale nella pratica clinica e nella ricerca oncologica contemporanea, ridefinendo modalità, tempi e profondità con cui affrontiamo la complessità del carcinoma ovarico avanzato. Nel corso degli ultimi anni, l'incremento esponenziale della disponibilità di dati clinici, molecolari, radiologici, chirurgici e istologici ha reso evidente la necessità di strumenti capaci di integrarli, interpretarli e restituirli in modo utile ai processi decisionali (HATAMIKIA et al., 2023; GÜLMETZ, 2026). L'AI risponde esattamente a questa esigenza: non si pone come un sostituto della capacità di giudizio del clinico, ma come un'estensione della sua possibilità di leggere e comprendere segnali complessi o pattern non lineari che eccedono la capacità cognitiva umana e le tecniche di statistica classica.

A differenza dei modelli parametrici più comuni (es. regressione logistica o Cox), che descrivono l'*outcome* tramite un predittore lineare e richiedono che non linearità e interazioni siano specificate a priori, molti algoritmi di machine learning apprendono direttamente dai dati una funzione più flessibile. Questa maggiore capacità descrittiva può ridurre l'errore di specificazione del modello, ma aumenta il rischio di *overfitting* e rende indispensabili rigorose strategie di validazione, calibrazione e trasparenza metodologica (CHRISTODOULOU et al., 2019). In questo senso, il valore aggiunto dell'AI non risiede tanto nella sostituzione dei modelli statistici tradizionali, quanto nella possibilità di affiancarli in contesti caratterizzati da elevata dimensionalità, eterogeneità delle sorgenti informative e relazioni complesse tra variabili cliniche, biologiche e temporali. Mentre l'approccio statistico classico privilegia modelli parsimoniosi e interpretabili, ottimiz-

zati per l'inferenza causale o associativa, i modelli di machine learning sono tipicamente orientati alla predizione e alla generalizzazione, accettando una maggiore complessità strutturale in cambio di una migliore capacità di adattarsi a pattern latenti nei dati. Ne deriva un cambiamento di paradigma: dall'enfasi sulla stima di singoli effetti medi verso la costruzione di strumenti predittivi integrati, la cui affidabilità dipende in larga misura dalla qualità del dato, dalla validazione esterna e dall'integrazione consapevole nel processo decisionale (Bzdok et al., 2018; Collins et al., 2024).

L'AI integrando fonti diverse in un'unica rappresentazione coerente, riduce la variabilità soggettiva, identifica pattern non immediatamente riconoscibili e offre supporto quantitativo nei momenti in cui la decisione è più incerta o il quadro clinico più sfumato. In questo senso, l'AI diventa un innesco trasformativo, capace di evolvere non solo la gestione del carcinoma ovarico, ma l'intera logica dell'oncologia moderna. Diverse declinazioni dell'intelligenza artificiale sono attualmente applicate in funzione della sorgente del dato analizzato, includendo dati clinici strutturati, informazioni molecolari, vetrini istopatologici digitalizzati, immagini radiologiche, video intraoperatori e approcci multimodali che integrano più livelli informativi. Nel contesto del carcinoma ovarico avanzato, la selezione della terapia di mantenimento di prima linea rappresenta uno degli ambiti clinici in cui l'applicazione di tecniche di machine learning si è rivelata particolarmente promettente, consentendo di raffinare il processo decisionale attraverso l'integrazione simultanea di variabili cliniche, biologiche e di risposta al trattamento, difficilmente sintetizzabili mediante approcci statistici tradizionali.

L'approccio attraverso strategie di machine learning, si presta a questo scopo perché consente di gestire collinearità e alta dimensionalità, modulando il bilanciamento tra selezione delle variabili e robustezza/stabilità del modello. In questo framework, l'integrazione tra stato molecolare HRD/BRCA, Chemotherapy Response Score (CRS) patologico e Kelim (costante di eliminazione del C125) rappresenta un esempio concreto di come un approccio ''multistrato'' possa migliorare una predizione clinicamente utile. Il valore dello stato molecolare HRD nella stratificazione del beneficio dai PARP-inibitori è ormai consolidato; tuttavia, l'integrazione con la risposta istopatologica alla chemioterapia neoadiuvante (CRS) e con la dinamica di eliminazione del CA125 conferisce al modello una capacità predittiva maggiore, perché incrocia la vulnerabilità molecolare con l'effettiva

sensibilità clinica espressa durante il trattamento. Il CRS, attraverso la valutazione istologica del tessuto, racconta una dimensione biologica strettamente aderente alla reale risposta tumorale. Il Kelim, invece, introduce una lettura temporale, trasformando la cinetica del biomarcatore in un indicatore della sensibilità chemioterapica. Nel loro insieme, questi elementi costruiscono una rappresentazione multifattoriale della malattia che permette potenzialmente di orientare con maggiore affidabilità la scelta del mantenimento verso bevacizumab, PARP-inibitori o combinazioni mirate (MARCHETTI et al., 2025). Parallelamente all'evoluzione dei modelli numerici, sta assumendo un ruolo rilevante la capacità dell'AI di analizzare il linguaggio clinico. Gran parte dell'informazione prodotta nei processi di cura rimane infatti confinata nei testi non strutturati: referti radiologici, note laparoscopiche, lettere di dimissione, discussioni multidisciplinari.

Il *natural language processing* consente di estrarre informazioni chiave da questi testi e di integrarle in pipeline predittive. Ciò accade perché il linguaggio libero conserva sfumature interpretative che difficilmente trovano corrispondenza strutturata nei database tradizionali: termini modificatori, avvertenze, percezioni cliniche, descrizioni di rischio implicito. L'AI riesce a catturare queste sfumature, trasformandole in variabili quantitative in grado di arricchire i modelli esistenti. Al Policlinico Gemelli, grazie alla collaborazione con gli ingegneri di Gemelli Generator, stiamo sviluppando nel progetto GO-Digital una piattaforma capace di estrarre automaticamente informazioni clinicamente rilevanti da referti complessi e di convertirle in sintesi strutturate, dimostrando come l'AI possa diventare un anello centrale nel processo di standardizzazione della presa in carico, riducendo tempi e incertezze, migliorando la qualità delle informazioni e generando una base omogenea per la discussione multidisciplinare.

L'evoluzione del *natural language processing* si riflette anche nella capacità dei modelli conversazionali avanzati di supportare le decisioni cliniche. Sistemi basati su *large language models* sono infatti in grado di leggere e sintetizzare guidelines, integrare dati clinici e generare raccomandazioni terapeutiche coerenti con gli standard internazionali. Le performance di ChatGPT-4, in particolare, si avvicinano ad altri sistemi *rule-based* come Watson for Oncology e mostrano un grado di coerenza crescente con le indicazioni delle linee guida NCCN (National Comprehensive Cancer Network) o ESGO (European Society of Gynaecological Oncology). Pur non essendo

un sostituto del medico, il *large language model* offre un supporto che riduce il carico cognitivo, favorisce la standardizzazione delle decisioni e contribuisce a una maggiore equità nel processo di cura, soprattutto in contesti dove l'accesso immediato a expertise super-specialistico non è garantito. Accanto alla dimensione testuale, l'AI sta cambiando radicalmente l'interpretazione delle immagini. Nel carcinoma ovarico avanzato, la laparoscopia diagnostica rappresenta oggi uno strumento chiave nel percorso decisionale di prima linea, in quanto consente una valutazione diretta e dinamica dell'estensione peritoneale di malattia prima di intraprendere un trattamento chirurgico. A differenza dell'imaging, la laparoscopia permette di esplorare in modo accurato e di evidenziare localizzazioni di carcinosi millimetriche o piatte in sedi critiche come il diaframma, il peritoneo viscerale, il mesentere, l'intestino e l'omento, riducendo il rischio di sottostimare una diffusione peritoneale non resecabile. In questo contesto si inserisce il Fagotti score (Fagotti et al., 2006; Fagotti et al., 2008) o Predictive Index Value (PIV), un sistema di punteggio laparoscopico sviluppato per standardizzare la valutazione intraoperatoria della resecabilità. Il punteggio si basa sull'analisi di specifici siti di malattia peritoneale, assegnando un valore numerico che correla con la probabilità di ottenere una citoriduzione completa. Studi prospettici hanno dimostrato che un PIV elevato è fortemente associato alla ridotta possibilità di ottenere una resezione completa, rendendo questo score uno strumento affidabile per il triage tra chirurgia primaria (PDS) e chemioterapia neoadiuvante seguita da chirurgia di intervallo (NACT-IDS).

L'integrazione tra laparoscopia diagnostica e Fagotti score consente quindi di personalizzare il percorso terapeutico, evitando laparotomie non necessarie e riducendo morbidità chirurgica senza compromettere l'*outcome* oncologico. Per questo motivo, le principali linee guida e raccomandazioni internazionali sottolineano l'importanza di una valutazione preoperatoria accurata, in cui la laparoscopia di *staging* svolge un ruolo determinante nella selezione delle pazienti più idonee a un approccio chirurgico primario aggressivo rispetto a una strategia neoadiuvante (Colombo et al., 2019).

In tale contesto, l'applicazione di tecniche di intelligenza artificiale basate su computer vision ai video di laparoscopia diagnostica potrebbe consentire di superare uno dei limiti strutturali della pratica clinica corrente, rappresentato dalla valutazione soggettiva e operatore-dipendente dell'estensione di malattia, identificando automatica-

mente organi, regioni anatomiche ed aree di carcinosi o noduli peritoneali (Tozzi et al., 2026).

Al Policlinico Gemelli, nel progetto PREDATOOR, attraverso l'analisi automatizzata di migliaia di frame video, e con il supporto degli ingegneri dell'IHU di Strasburgo, abbiamo sviluppato modelli in grado di stimare in modo oggettivo la complessità chirurgica e di predire il Fagotti Score con un'affidabilità elevata e realmente *operator-independent*. L'impatto clinico è rilevante: riduce la variabilità inter-chirurgo, aumenta la riproducibilità delle valutazioni anche nei centri con minore esperienza e contribuisce a creare un linguaggio visivo condiviso per la stratificazione delle pazienti. Nel carcinoma ovarico, in cui il triage tra chirurgia *upfront* e trattamento neoadiuvante è cruciale, disporre di un sistema che riconosce automaticamente pattern di diffusione di malattia può standardizzare, uniformare e 'democratizzare' la qualità delle decisioni (Oliva et al., 2025).

In una continuità logica, anche l'imaging radiologico rappresenta terreno fertile per l'applicazione di tecniche di AI. La radiomica permette di trasformare le immagini TAC e RM in matrici numeriche che descrivono la morfologia tumorale in termini di texture, eterogeneità, distribuzione volumetrica e relazioni spaziali. La correlazione tra specifici cluster radiomici e risposta alla chemioterapia conferma che le immagini contengono una dimensione informativa molto più ricca di quanto la semplice lettura visiva possa suggerire. Quando queste informazioni vengono integrate con il profilo genomico, nasce la radiogenomica, disciplina capace di collegare fenomeni strutturali visibili con pathway molecolari sottostanti (Crispin-Ortuzar, 2023; Avesani, 2022).

Nel carcinoma ovarico, in cui l'eterogeneità intra- e inter-tumorale rappresenta una delle principali sfide (sia nella comprensione della biologia di malattia sia nell'espansione del portfolio terapeutico), la radiogenomica può aggiungere un livello predittivo non invasivo: contribuire a stimare precocemente il rischio di resistenza alla terapie, individuare sub-regioni e siti metastatici con comportamento più aggressivo, e supportare la pianificazione chirurgica verso una resezione più mirata, concentrando l'attenzione sui '*critical sites*' prognostico (Rizzo et al., 2018; Rundo et al., 2022). Un'ulteriore innovazione ad alto impatto è la patologia digitale. L'applicazione di modelli di deep learning ai vetrini istologici ematossilina ed eosina suggerisce che la morfologia tumorale incorpori segnali molecolari ''latenti'', talvolta sufficienti a stimare fenotipi clinicamente rilevanti, come lo

status di *homologous recombination deficiency* (HRD) e, in alcuni contesti, la probabilità di risposta a trattamenti specifici (Bergstrom et al., 2024). Un aspetto particolarmente interessante è la trasversalità dei pattern morfologici: alcuni descrittori micro-architetturali appresi dal modello sembrano riutilizzabili tra istotipi e sedi tumorali differenti, coerentemente con evidenze pan-cancer che mostrano la possibilità di inferire l'assetto molecolare da immagini istopatologiche routinarie (Loeffler, 2024). Dal punto di vista metodologico, la maggior parte di questi approcci opera in regime *weakly supervised* (etichette a livello di vetrino/paziente, senza annotazioni pixel-wise) tramite Multiple Instance Learning (MIL): il vetrino viene scomposto in '*tiles*', il modello identifica automaticamente le regioni più informative, attribuisce pesi differenziali ai pattern micro-architetturali e aggrega l'informazione in una predizione *slide-level*, con prestazioni che possono mantenersi su coorti indipendenti se il training e la validazione sono correttamente disegnati (18). In prospettiva, tali pipeline possono rendere la stratificazione più rapida, meno costosa e più accessibile, soprattutto dove i test molecolari non sono disponibili in modo uniforme o hanno tempi incompatibili con decisioni cliniche strette (pur restando necessari standard rigorosi di validazione esterna e controllo dei *bias* pre-analitici) (Bergstrom, et al. 2024). La chirurgia è probabilmente l'ambito in cui l'integrazione tra intelligenza artificiale e pratica clinica diventa più tangibile. La convergenza di visione artificiale, realtà aumentata e ricostruzioni tridimensionali consente di arricchire il campo operatorio con informazione contestuale (target, margini, traiettorie, 'mappe' anatomiche) sotto forma di overlay e sistemi di navigazione, con l'obiettivo di migliorare orientamento, consapevolezza spaziale e coerenza dell'esecuzione tecnica (Malhotra, 2023).

In questo scenario, modelli di segmentazione auto-configuranti come nnU-Net permettono di delineare in modo robusto strutture e lesioni anche in immagini complesse, automatizzando scelte critiche (*pre-processing*, architettura, training, post-processing) e riducendo la dipendenza da tuning manuale. Le segmentazioni possono essere convertite in modelli 3D e proiezioni intraoperatorie utili come guida visiva, rafforzando la percezione dei piani di dissezione, facilitando l'identificazione di siti di malattia e chiarendo i rapporti con strutture critiche (organi, vasi, nervi, linfonodi, muscoli, strutture ossee) (Isensee et al., 2021). Questa traiettoria definisce in modo credibile il futuro del planning pre-operatorio e della *guidance* intra-operatoria,

ma la traslazione clinica dipende soprattutto da tre punti: qualità/standardizzazione dei dati (*imaging* e video), affidabilità della registrazione e dell'overlay in ambienti deformabili, e validazione esterna (Doornbos et al., 2024). Ulteriore avanzamento è garantito dall'integrazione di nuovi traccianti fluorescenti, insieme ad algoritmi di intelligenza artificiale capaci di quantificare l'intensità ottica e le dinamiche di perfusione a livello di singolo pixel, sta aprendo una nuova fase della chirurgia di precisione. L'analisi delle curve tempo—intensità *dell'indocyanine green* dimostra infatti che è possibile passare da una valutazione 'a occhio' a una misurazione oggettiva della vitalità tissutale, con parametri quantitativi (per esempio tempi di arrivo, pendenza di incremento, tempo al picco e *washout*) utili per decisioni intraoperatorie più riproducibili. Inoltre, l'elaborazione dinamica del segnale fluorescente può contribuire a evidenziare focolai metastatici occulti o difficilmente riconoscibili in luce bianca, migliorando la valutazione intraoperatoria (Van Den Hoven et al., 2023; Hardy et al., 2023). L'intelligenza artificiale entra anche nel dominio della sicurezza chirurgica. Sistemi di visione artificiale applicati ai video intraoperatori possono riconoscere in tempo reale segnali precoci di eventi critici come sanguinamenti, esposizione o rischio di lesione di strutture nobili e trasformarli in alert o in supporto contestuale alla decisione. In prospettiva, questa capacità può ridurre il tempo che intercorre tra l'insorgenza dell'evento e la risposta dell'operatore, standardizzando la vigilanza anche nelle fasi più complesse dell'intervento. In un contesto globale in cui il carico di complicanze e mortalità postoperatoria resta elevato, e in cui l'obiettivo non è solo la radicalità oncologica ma anche la preservazione funzionale e della qualità di vita, l'intelligenza artificiale può assumere un ruolo di 'supervisore' operativo. Integrata nei flussi intraoperatori e perioperatori, può supportare il riconoscimento precoce di eventi critici e deviazioni tecniche, ridurre la latenza tra insorgenza del problema e intervento correttivo, e contribuire a limitare gli errori potenzialmente evitabili, con un impatto diretto sulla sicurezza della paziente e sugli esiti a breve e lungo termine (Peng et al., 2025; Garcia-Martinez et al., 2017). L'integrazione complessiva di questi sviluppi porta alla definizione di un ecosistema in cui testo, immagine, chirurgia, genetica e istologia confluiscono in un'unica piattaforma cognitiva. Le tecnologie emergenti descrivono un contesto in cui l'AI non è più un singolo modulo, ma una presenza trasversale capace di supportare la decisione clinica in tutte le sue fasi. Dalla selezione delle pazienti più idonee

alla chirurgia *upfront*, alla predizione della risposta ai PARP-inibitori alla navigazione intraoperatoria, dalla prevenzione delle complicanze alla stratificazione post-operatoria, l'intelligenza artificiale aiuta a costruire un percorso integrato che rende la gestione delle patologie ginecologiche oncologiche più precisa, più personalizzata e più sicura. Le sfide rimangono: la necessità di trasparenza dei modelli, la gestione dei *bias*, l'importanza della supervisione clinica, la costruzione di dataset ampi e rappresentativi, la definizione di criteri di validazione robusti e applicabili nella pratica quotidiana. Tuttavia, il potenziale clinico dell'AI non è più un'ipotesi, ma una realtà già misurabile. Il futuro dell'oncologia ginecologica sarà sempre più guidato da modelli integrati capaci di sintetizzare testo, immagine, genetica, dinamiche biologiche e riscontri intraoperatori in una rappresentazione multidimensionale e continuamente aggiornata della malattia. In questa visione, il ruolo del clinico non viene sminuito, ma ampliato: la tecnologia non sostituisce la competenza umana, ma la sostiene, la rafforza e la rende più efficace.

Note bibliografiche

Avesani G.-Tran H.E.-Cammarata G. et al. (2022), *CT-Based Radiomics and Deep Learning for BRCA Mutation and Progression-Free Survival Prediction in Ovarian Cancer Using a Multicentric Dataset.* Cancers (Basel), 14, 11, p. 2739 ss.

Bergstrom E.N.-Abbasi A.-Díaz-Gay M. et al. (2024), *Deep Learning Artificial Intelligence Predicts Homologous Recombination Deficiency and Platinum Response From Histologic Slides*, in *J Clin Oncol*, 42, 30, p. 3550 ss.

Bzdok D.-Altman N.-Krzywinski M (2018), *Statistics versus machine learning*, in *Nat Methods*, 15, 4, p. 233.

Christodoulou E.-Ma J.-Collins G.S.-Steyerberg E.W.-Verbakel J.Y.-Van Calster B. (2019), *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models*, in *J Clin Epidemiol.*, 110, p. 12 ss.

Collins G.S.-Moons K.G.M.-Dhiman P. et al. (2024), *TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods*, in *BMJ* (doi: 10.1136/bmj.q902).

Colombo N.-Sessa C.-Bois A.D. et al. *ESMO-ESGO consensus conference recommendations on ovarian cancer: pathology and molecular biology, early and advanced stages, borderline tumours and recurrent disease*, in *Int J Gynecol Cancer*, 29, 4 p. 728 ss.

Crispin-Ortuzar M.-Woitek R.-Reinius M.A.V. et al. (2023), *Integrated radiogenomics models predict response to neoadjuvant chemotherapy in high grade serous ovarian cancer*, in *Nat Commun*, 14, 1, p. 675 ss.

DOORNBOS M.J.-PEEK J.J.-MAAT A.P.W.M. ET AL. (2024), *Augmented Reality Implementation in Minimally Invasive Surgery for Future Application in Pulmonary Surgery: A Systematic Review*, in *Surg Innov*, 31, 6, p. 646 ss.

FAGOTTI A.-FERRANDINA G.-FANFANI F. ET AL. (2006), *A laparoscopy-based score to predict surgical outcome in patients with advanced ovarian carcinoma: a pilot study*, in *Ann Surg Oncol*, 13, 8, p. 1156 ss.

FAGOTTI A.-FERRANDINA G.-FANFANI F. ET AL., *Prospective validation of a laparoscopic predictive model for optimal cytoreduction in advanced ovarian carcinoma*, in *Am J Obstet Gynecol*, 199, 6.

GARCIA-MARTINEZ A.-VICENTE-SAMPER J.M.-SABATER-NAVARRO J.M. (2017), *Automatic detection of surgical haemorrhage using computer vision*, in *Artif Intell Med.*, 78, p. 55 ss.

GÜLMEZ B. (2026), *Artificial intelligence applications in ovarian cancer detection: A systematic literature review of deep learning approaches and clinical translation challenges*, in *Crit Rev Oncol Hematol* (doi: 10.1016/j.critrevonc.2026.105126).

HARDY N.P.-EPPERLEIN J.P.-DALLI J. ET AL (2023), *Real-time administration of indocyanine green in combination with computer vision and artificial intelligence for the identification and delineation of colorectal liver metastases*, in *Surg Open Sci*, 12, p. 48 ss.

HATAMIKIA S.-NOUGARET S.-PANICO C. ET AL. (2023), *Ovarian cancer beyond imaging: integration of AI and multiomics biomarkers*, in *Eur Radiol Exp.*, 7, 1, p. 50 ss.

ISENSEE F.-JAEGER P.F.-KOHL S.A.A.-PETERSEN J.-MAIER-HEIN K.H. (2021), *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation*, in *Nat Methods*, 18, 2, p. 203 ss.

LOEFFLER C.M.L.-EL NAHHAS O.S.M.-MUTI H.S. ET AL. (2024), *Prediction of homologous recombination deficiency from routine histology with attention-based multiple instance learning in nine different tumor types*, in *BMC Biol.*, 22, 1, p. 225 ss.

LU M.Y.-WILLIAMSON D.F.K.-CHEN T.Y.-CHEN R.J.-BARBIERI M.-MAHMOOD F. (2021), *Data-efficient and weakly supervised computational pathology on whole-slide images*, in *Nat Biomed Eng.*, 5, 6, p. 555 ss.

MALHOTRA S.-CHATURVEDI P.-GUPTA P.K. (2023), *Augmented Reality in Surgical Navigation: A Review of the State of the Art*, in *Applied Sciences (MDPI)*, 13, 3, p. 1629.

MARCHETTI C.-ERGASTI R.-CAPOMACCHIA FM, et al. (3025), *Integrating clinical-molecular data to predict PARP inhibitors efficacy in advanced ovarian cancer patients after interval cytoreductive surgery*, in *Gynecol Oncol*, 195, p. 16 ss.

OLIVA R.-VARDAZARYAN A.-ROSATI A.-ZARIN F.-FAGOTTI A. (2025), *Artificial Intelligence and Computer Vision for the Assessment of Peritoneal Carcinosis at Laparoscopy for Advanced Ovarian Cancer: The PREDATOOR Project*, in *Int J Gynecol Cancer* (doi.org/10.48550/arXiv.2512.14797).

PENG Z.Y.-WANG Z.B.-YAN Y. ET AL. (2025), *Development of an AI-driven digital assistance system for real-time safety evaluation and quality con-*

trol in laparoscopic liver surgery, in *Front Oncol* (doi.org/10.3389/fonc.2025.1678525).

Rizzo S.-Botta F.-Raimondi S. et al. (2018), *Radiomics of high-grade serous ovarian cancer: association between quantitative CT features, residual tumour and disease progression within 12 months*, in *Eur Radiol*, 28, 11, p. 4849 ss.

Rundo L.-Beer L.-Escudero Sanchez L. et al. (2022), *Clinically Interpretable Radiomics-Based Prediction of Histopathologic Response to Neoadjuvant Chemotherapy in High-Grade Serous Ovarian Carcinoma*, in *Front Oncol*, 12, p. 868265 ss.

Tozzi F.-Park H.M.-Mousavi S.A. et al. (2026), *Multimodal machine learning for staging laparoscopy: a combined image analysis and morphologic tool for the discrimination of peritoneal metastasis*, in *Int J Surg.*, 112, 1, p. 373 ss.

Van Den Hoven P.-Osterkamp J.-Nerup N. et al. (2023), *Quantitative perfusion assessment using indocyanine green during surgery - current applications and recommendations for future use*, in *Langenbecks Arch Surg.*, 408, 1, p. 67 ss.

(IL FATTO)
L'IA TRA "FATTO PREDITTIVO" E "NORMA EVOLUTIVA": LA PROSPETTIVA DELLA CYBERPSICOLOGIA

di *Giuseppe Riva*

Sommario: 1. Introduzione: il digitale come ambiente primario dell'esperienza umana. — 1.1. La cyberpsicologia come disciplina emergente. — 1.2. Le psicotecnologie e la trasformazione dei *brainframe*. — 1.3. L'intelligenza artificiale come svolta evolutiva. — 2. Il paradosso comfort-crescita nelle tecnologie digitali. — 2.1. La natura duale del comfort tecnologico. — 2.2. La necessità della sfida per la crescita cognitiva. — 2.3. L'intelligenza artificiale come amplificatore del paradosso. — 2.4. Formulazione teorica del paradosso. — 3. Framework cyberpsicologici per la navigazione del paradosso. — 3.1. Modelli di adattamento dinamico. — 3.2. Il Transtheoretical Model of Change applicato all'integrazione cognitiva. — 3.3. Flow Theory e l'esperienza ottimale nell'interazione umano-IA. — 3.4. Cognitive Load Theory e gestione della memoria di lavoro. — 3.5. Self-Determination Theory e motivazione intrinseca. — 3.6. Dual-Process Theories e architettura cognitiva. — 4. Conclusioni: L'IA tra 'fatto' predittivo e 'norma' evolutiva.

1. Introduzione: il digitale come ambiente primario dell'esperienza umana.

Il digitale non costituisce più un'aggiunta alla vita contemporanea: è divenuto il suo contesto primario. Rappresenta l'ambiente in cui si svolge una parte sempre più significativa delle esperienze quotidiane, delle relazioni interpersonali e persino delle dinamiche emotive degli individui (Riva, 2025a).

Questa immersione continua nelle tecnologie digitali ha trasformato profondamente il modo in cui gli esseri umani vivono, pensano, ricordano, comunicano e si relazionano. L'osservazione fenomenologica rivela comportamenti sintomatici di questa trasformazione: la consultazione compulsiva dello smartphone senza consapevolezza riflessiva, il disagio acuto quando la connessione viene meno, il ricorso sistematico ai motori di ricerca per informazioni che un tempo erano

conservate nella memoria biologica. La tecnologia ha cessato di essere uno strumento meramente esterno: è diventata una protesi cognitiva, un'estensione silenziosa ma pervasiva del sistema nervoso umano (Clark-Chalmers, 1998). Modella ciò che viene percepito, ciò che viene ricordato, persino ciò che viene desiderato.

1.1. *La cyberpsicologia come disciplina emergente.*

Per comprendere le implicazioni di questo scenario è nata la cyberpsicologia (*cyberpsychology*): un'area di ricerca emergente che incrocia scienze cognitive, psicologia sociale, neuroscienze ed etica della tecnologia (Riva-Wiederhold-Mantovani, 2019). Il suo obiettivo consiste nell'analizzare in profondità l'impatto delle tecnologie digitali sulla mente umana e sulle relazioni interpersonali, decifrando come l'ecosistema digitale stia riscrivendo i fondamenti dell'esperienza umana (Suler, 2016).

Storicamente, le tecnologie sono nate per superare i limiti corporei: gli occhiali per migliorare la funzione visiva, il telefono per abbattere le distanze spaziali nella comunicazione, l'aereo per permettere la locomozione aerea. Tuttavia, alcune tecnologie vanno oltre questa dimensione: non potenziano il corpo, bensì la mente. Queste sono le 'psicotecnologie' (De Kerckhove, 1991), ovvero strumenti che amplificano, emulano o trasformano le capacità cognitive: apprendere, ricordare, pianificare, immaginare, comunicare.

1.2. *Le psicotecnologie e la trasformazione dei* brainframe.

Ogni psicotecnologia, una volta interiorizzata e resa familiare attraverso l'uso prolungato, non si limita a migliorare ciò che gli individui già sono: ridefinisce ciò che possono diventare (Carr, 2010). Come accade per un giovane che, ottenuta la patente di guida, guadagna nuove possibilità di autonomia e socialità, così l'uso quotidiano e integrato delle psicotecnologie plasma nuove forme di pensiero, relazione e identità (Turkle, 2011). In questo processo emergono nuovi *brainframe* (De Kerckhove, 1991), ovvero cornici cognitive individuali e collettive che condizionano il modo stesso in cui il mondo viene percepito, interpretato e agito.

Questa concezione riprende e sviluppa l'ipotesi della mente estesa (*Extended Mind Hypothesis*) proposta da Clark e Chalmers (1998), secondo cui i processi cognitivi non sono confinati entro i limiti del cervello, ma si estendono nell'ambiente attraverso l'integra-

zione funzionale con artefatti esterni. Le psicotecnologie digitali rappresentano la manifestazione più avanzata di questo principio: smartphone, motori di ricerca, social media e sistemi di intelligenza artificiale non sono meri strumenti passivi, ma componenti attive dell'architettura cognitiva distribuita (HOLLAN-HUTCHINS-KIRSH, 2000).

1.3. *L'intelligenza artificiale come svolta evolutiva.*

In questa evoluzione, l'intelligenza artificiale rappresenta un punto di svolta, la psicotecnologia più potente mai creata. A differenza degli strumenti precedenti, l'IA non si limita ad eseguire comandi predefiniti, ma agisce come un partner cognitivo (CHIRIATTI et al., 2025). Questa caratteristica la distingue qualitativamente dalle tecnologie precedenti e solleva questioni inedite relative all'autonomia cognitiva, alla trasparenza epistemica e alla preservazione del ruolo umano nei processi decisionali (MITTELSTADT et al., 2016).

In particolare, essa configura quello che può essere definito il «*paradosso comfort-crescita*» (RIVA, 2025b): le tecnologie digitali devono essere simultaneamente predicibili per creare sicurezza psicologica e ridurre il carico cognitivo, e sufficientemente dirompenti per promuovere sviluppo terapeutico e crescita cognitiva. Questo paradosso rappresenta la sfida centrale per la progettazione etica e l'integrazione responsabile delle psicotecnologie contemporanee, richiedendo framework normativi che bilancino efficacia funzionale e preservazione dell'autonomia umana.

2. Il paradosso comfort-crescita nelle tecnologie digitali.

2.1. *La natura duale del comfort tecnologico.*

Le tecnologie digitali hanno acquisito una notevole capacità di fornire comfort attraverso personalizzazione e predicibilità. I moderni sistemi di intelligenza artificiale apprendono i pattern comportamentali degli utenti, anticipano le loro necessità e adattano le interfacce per corrispondere alle preferenze individuali (KHAMAJ-ALI, 2024). Gli algoritmi di raccomandazione personalizzano contenuti e servizi, riducendo lo sforzo cognitivo richiesto per navigare la complessità informativa contemporanea.

Questo comfort assolve funzioni importanti dal punto di vista psicologico. Riduce il carico cognitivo, permettendo agli utenti di concentrare le risorse attentive su compiti di ordine superiore piut-

tosto che sulla gestione dell'interfaccia tecnologica (MÜLLER-PRIMC-KUHN, 2023). Costruisce fiducia attraverso consistenza e affidabilità. Crea sicurezza psicologica nell'interazione con sistemi complessi. La ricerca in cyberpsicologia ha dimostrato in più studi che quando gli utenti si sentono a proprio agio con la tecnologia, sono più propensi a utilizzarla in modo continuativo e a integrarla nei loro processi cognitivi (RIVA-WIEDERHOLD-MANTOVANI, 2019).

Tuttavia, questo stesso comfort può trasformarsi in un vincolo cognitivo. Quando i sistemi di IA sono eccessivamente predicibili e allineati con le preferenze esistenti, possono risultare inefficaci nel promuovere l'esposizione a prospettive diverse e nel facilitare la crescita intellettuale (YAN et al., 2021). Quando gli algoritmi di raccomandazione rinforzano sistematicamente le credenze e i comportamenti esistenti, possono inavvertitamente creare camere di risonanza che limitano la diversità cognitiva invece di espanderla (CINELLI et al., 2021). Il bilanciamento tra comfort e sfida non può essere uniforme: in ambito clinico richiede cautele terapeutiche e supervisione; in ambito educativo può tradursi in scaffolding progressivo; nei sistemi di raccomandazione implica l'introduzione controllata di diversità informativa.

2.2. *La necessità della sfida per la crescita cognitiva.*

La crescita cognitiva richiede fondamentalmente il confronto con la sfida e la dissonanza (RIVA, 2025b). La «*Zona di Sviluppo Prossimale*» di Vygotsky spiega bene questo principio (VYGOTSKY, 1978): l'apprendimento avviene nello spazio tra ciò che possiamo fare indipendentemente e ciò che possiamo realizzare con supporto. Nel contesto delle tecnologie digitali, questo si traduce in sistemi che estendono gli utenti appena oltre le loro capacità attuali, fornendo al contempo sufficiente supporto (*scaffolding*) per prevenire frustrazione paralizzante.

Si consideri l'apprendimento assistito da intelligenza artificiale. I sistemi devono creare sufficiente difficoltà per stimolare l'elaborazione profonda e l'apprendimento significativo, ma non al punto da indurre sovraccarico cognitivo o abbandono (SIMONS et al., 2016). Questo delicato equilibrio esemplifica il paradosso comfort-crescita: lo stesso sistema di IA deve essere sufficientemente confortevole da usare, pur presentando deliberatamente sfide cognitive che promuovono sviluppo. I sistemi di intelligenza artificiale conversazionale de-

vono far emergere intuizioni che sfidano le assunzioni esistenti e incoraggiano riflessione critica, mantenendo al contempo sufficiente engagement per assicurare la partecipazione continuata.

2.3. *L'intelligenza artificiale come amplificatore del paradosso.*

L'intelligenza artificiale amplifica entrambi i lati del paradosso comfort-crescita. Da un lato, può fornire personalizzazione e supporto senza precedenti, creando esperienze d'uso altamente confortevoli attraverso l'adattamento continuo alle preferenze individuali. Dall'altro, la letteratura recente ha proposto che i sistemi di IA possano essere interpretati come un "Sistema 0" (CHIRIATTI et al., 2024; 2025): un livello di pre-elaborazione cognitiva che filtra e struttura il substrato informativo su cui operano i processi intuitivi e deliberativi. In questa prospettiva, l'IA non si limita a intervenire dopo la decisione umana, ma contribuisce a configurare il campo informativo entro cui la decisione prende forma.

Questa integrazione genera conseguenze ambivalenti: da un lato espande le capacità cognitive attraverso l'accesso a informazioni e capacità computazionali altrimenti inaccessibili; dall'altro può limitare il pensiero critico attraverso fenomeni di 'compiacenza algoritmica' — *sycophancy* (SHARMA et al., 2023), amplificazione dei bias (GLICKMAN-SHAROT, 2025) e creazione di camere di risonanza (*echo chambers*) epistemiche (CINELLI et al., 2021).

In particolare, la compiacenza algoritmica crea un'illusione di validazione che può rinforzare errori di ragionamento e credenze infondate, privilegiando il comfort dell'accordo rispetto alla crescita derivante dal confronto critico. La ricerca sperimentale di Glickman e Sharot (2025) dimostra che le interazioni ripetute con sistemi di IA che presentano questo bias non solo alterano le decisioni immediate, ma modificano i meccanismi cognitivi sottostanti, influenzando giudizi percettivi, emotivi e sociali in modo duraturo.

2.4. *Formulazione teorica del paradosso.*

Il paradosso comfort-crescita può essere formulato come segue: *le tecnologie digitali, e in particolare i sistemi di intelligenza artificiale, devono essere simultaneamente sufficientemente predicibili da creare sicurezza psicologica e ridurre il carico cognitivo estrinseco, e sufficientemente dirompenti da promuovere sviluppo intellettuale e crescita delle capacità cognitive*. Questa tensione dialettica non rappresenta un di-

fetto progettuale da eliminare, ma una proprietà intrinseca delle psicotecnologie orientate allo sviluppo umano.

La risoluzione di questo paradosso richiede un approccio dinamico e contestualizzato. Il bilanciamento ottimale tra comfort e sfida varia in funzione di molteplici fattori: lo stadio del processo di apprendimento, le caratteristiche individuali dell'utente, il tipo specifico di capacità da sviluppare, e il contesto sociale e culturale dell'interazione. Le tecnologie più efficaci sono quelle capaci di adattare questo bilanciamento in tempo reale, rispondendo agli indicatori comportamentali ed emotivi che segnalano la disponibilità dell'utente a impegnarsi con livelli crescenti di complessità (Riva-Wiederhold-Mantovani, 2021).

La cyberpsicologia, attraverso l'adattamento tecnologico di teorie psicologiche consolidate, fornisce gli strumenti concettuali ed empirici necessari per navigare questa tensione. I framework teorici discussi nelle sezioni successive dimostrano come il paradosso possa essere gestito attraverso principi derivati dalla Flow Theory (Csikszentmihalyi, 1990; Nakamura-Csikszentmihalyi, 2009), dal Transtheoretical Model of Change (Prochaska-Velicer, 1997), dalla Self-Determination Theory (Ryan-Deci, 2000) e da altri modelli validati della psicologia dello sviluppo e del cambiamento cognitivo.

La sfida fondamentale consiste nel progettare sistemi di IA che fungano da *partner nello sviluppo cognitivo* piuttosto che da semplici facilitatori di compiti. Questo richiede il passaggio da un paradigma centrato sull'efficienza immediata a uno orientato alla crescita a lungo termine, dove il disagio produttivo e la tensione epistemica sono riconosciuti come componenti essenziali dell'esperienza tecnologica piuttosto che come frizioni da eliminare.

3. Framework cyberpsicologici per la navigazione del paradosso.

3.1. *Modelli di adattamento dinamico.*

Piuttosto che concepire comfort e crescita come stati fissi e mutuamente esclusivi, la cyberpsicologia li rivela come processi dinamici che cambiano in base allo stato dell'utente, al contesto interazionale e agli obiettivi cognitivi (Riva-Wiederhold-Mantovani 2021). Le tecnologie digitali efficaci devono quindi incorporare meccanismi sofisticati di adattamento capaci di leggere gli indizi comportamentali e fisiologici dell'utente e di regolare il bilanciamento comfort-sfida in tempo reale (Pizzoli et al., 2024).

I sistemi di intelligenza artificiale conversazionale, ad esempio, possono rilevare marcatori linguistici di disagio cognitivo — risposte abbreviate, aumento di marcatori linguistici di valenza negativa, o schemi di disimpegno — e modulare di conseguenza le loro risposte (VIJAYALAKSHMI, 2025). Dispositivi indossabili possono identificare momenti ottimali per presentare sfide cognitive basandosi su pattern di attività quotidiana, livelli di stress e ritmi circadiani (SAJNO, BERETTA et al., 2022), assicurando che le sfide vengano proposte quando gli utenti dispongono delle risorse psicologiche per affrontarle efficacemente (SAJNO, MAISETTI et al., 2022).

La sofisticazione di questi sistemi adattivi continua a evolversi. In contesti sperimentali o applicativi delimitati, algoritmi di machine learning possono stimare stati cognitivi ed emotivi a partire da pattern comportamentali e fisiologici, consentendo adattamenti preventivi dell'interazione. Un sistema di assistenza all'apprendimento potrebbe riconoscere segni precoci di frustrazione attraverso pattern di interazione con l'interfaccia, adeguando la difficoltà dei compiti prima che l'utente sperimenti consapevolmente una sfida travolgente.

3.2. *Il Transtheoretical Model of Change applicato all'integrazione cognitiva.*

Il "*Transtheoretical Model of Change*", estensivamente validato in psicologia della salute, fornisce una roadmap per come le tecnologie possano supportare gli utenti attraverso differenti stadi del cambiamento cognitivo e comportamentale (PROCHASKA-VELICER, 1997). Durante gli stadi di *precontemplazione* (il soggetto non vuole cambiare) e *contemplazione* (il soggetto vuole cambiare ma non sa come), la tecnologia dovrebbe prioritizzare comfort e engagement. Man mano che gli utenti transitano negli stadi di *determinazione* (il soggetto sta per cambiare) e *azione* (il soggetto sta cambiando), il sistema può introdurre progressivamente sfide maggiori. Nello stadio di *mantenimento* (il soggetto continua a cambiare), la tecnologia deve bilanciare supporto continuativo con la promozione dell'indipendenza cognitiva.

Questo approccio per stadi risolve il paradosso riconoscendo che il bilanciamento ottimale tra comfort e crescita varia lungo il percorso di sviluppo. Un sistema di apprendimento assistito da IA, ad esempio, potrebbe iniziare con esercizi gentili di costruzione della consapevolezza metacognitiva — monitorando pattern di ragionamento senza giudizio — prima di introdurre tecniche più impegna-

tive di pensiero critico. Nello stadio di *precontemplazione*, le notifiche potrebbero semplicemente offrire spunti interessanti per la riflessione. Durante gli stadi di *azione*, lo stesso sistema potrebbe presentare dilemmi cognitivi che richiedono impegno attivo con esercizi di ragionamento complesso.

L'enfasi del modello sulla *ricaduta* (il soggetto ha smesso di cambiare) come parte normale del processo di cambiamento è particolarmente preziosa per la progettazione di sistemi digitali. Le tecnologie possono normalizzare i momenti di regressione cognitiva ricalibrando il bilanciamento comfort-sfida per ricostruire la fiducia dell'utente. Una piattaforma di sviluppo delle competenze potrebbe automaticamente rilevare un pattern di difficoltà e temporaneamente aumentare le caratteristiche di supporto riducendo le richieste prestazionali, aiutando gli utenti a reimpegnarsi senza vergogna o difficoltà travolgente.

3.3. *Flow Theory e l'esperienza ottimale nell'interazione umano-IA.*

Il concetto di '*flow*' elaborato da Csikszentmihalyi descrive lo stato psicologico raggiunto quando il livello di sfida nell'azione corrisponde perfettamente al livello di competenza (CSIKSZENTMIHALYI, 1990). Le tecnologie digitali possono sfruttare questo principio calibrando continuamente la difficoltà per mantenere gli utenti in questa zona ottimale. Quando questo stato viene raggiunto, il paradosso comfort-crescita si dissolve: gli utenti sperimentano simultaneamente il comfort della competenza e la crescita derivante dall'espansione delle capacità.

Gli elementi di gamification nelle applicazioni cognitive esemplificano questo approccio. Aumentando progressivamente la sfida man mano che gli utenti sviluppano competenze, gli interventi gamificati mantengono il coinvolgimento promuovendo miglioramento continuo (SAJNO et al., 2022). La chiave consiste nell'assicurare che lo sviluppo della sfida segua le capacità crescenti dell'utente piuttosto che un percorso arbitrario. Un'applicazione di training cognitivo potrebbe utilizzare algoritmi prestazionali per rilevare quando un utente ha raggiunto stabilmente l'85% di accuratezza (WILSON *et al.* 2019): il punto ottimale che indica prontezza per difficoltà accresciuta, piuttosto che avanzare i livelli secondo un calendario predefinito.

L'intuizione critica è che il flow rappresenta la risoluzione del paradosso comfort-crescita: nello stato di flow, la sfida risulta confor-

vole perché corrisponde perfettamente alla capacità (NAKAMURA — CSIKSZENTMIHALYI, 2009). Questo stato psicologico ottimale elimina la tensione tra sicurezza e sviluppo creando un'esperienza in cui l'impegno con la complessità viene percepito come intrinsecamente gratificante piuttosto che minaccioso.

3.4. *Cognitive Load Theory e gestione della memoria di lavoro.*

La *Cognitive Load Theory* fornisce intuizioni fondamentali su come le tecnologie digitali possano ottimizzare il bilanciamento comfort-crescita attraverso la gestione delle richieste di elaborazione dell'informazione (KIRSCHNER, 2002). La teoria distingue tra *carico intrinseco* (difficoltà inerente al compito), *carico estrinseco* (onere cognitivo non necessario), e *carico pertinente all'apprendimento* (sforzo mentale produttivo che promuove apprendimento).

I sistemi di intelligenza artificiale possono navigare il paradosso minimizzando il carico estrinseco mentre ottimizzano il bilanciamento tra carichi intrinseco e pertinente (SWELLER-VAN MERRIENBOER-PAAS, 1998). Un assistente IA per la ricerca, ad esempio, potrebbe ridurre il carico estrinseco automatizzando la gestione delle citazioni bibliografiche e l'organizzazione delle fonti, permettendo agli utenti di focalizzare le risorse cognitive sulla comprensione dei pattern concettuali e sulla sintesi di idee complesse. La tecnologia fornisce comfort attraverso interfacce semplificate mantenendo la crescita attraverso sfide progressivamente complesse di elaborazione concettuale.

Questo framework è particolarmente prezioso per compiti cognitivi complessi che richiedono l'integrazione di molteplici fonti informative. Sequenziando le richieste cognitive e fornendo supporto appropriato, le tecnologie possono aiutare gli utenti a costruire modelli mentali gradualmente senza sovraccaricare la loro capacità di elaborazione. Il comfort deriva dalla gestione efficiente delle risorse cognitive limitate; la crescita emerge dall'incremento progressivo della complessità gestibile.

3.5. *Self-Determination Theory e motivazione intrinseca.*

La *Self-Determination Theory* offre una guida essenziale per mantenere il coinvolgimento degli utenti mentre si sostiene la crescita attraverso il focus su autonomia, competenza e relazionalità (RYAN-DECI, 2000). Le tecnologie che supportano questi bisogni psicologici

basilari possono sostenere la motivazione anche quando introducono esperienze di crescita sfidanti.

Le piattaforme digitali possono fornire comfort attraverso il supporto alla competenza — feedback chiaro, traguardi raggiungibili, e supporti per la costruzione di competenze — mentre promuovono crescita attraverso sfide che supportano l'autonomia e che gli utenti scelgono di affrontare. Un sistema di IA conversazionale potrebbe offrire molteplici percorsi di approfondimento, permettendo agli utenti di selezionare sfide che si allineano con i loro valori e obiettivi piuttosto che imporre una progressione rigida.

La componente della relazionalità affronta la dimensione sociale del paradosso comfort-crescita. Le comunità online supportate da IA possono fornire comfort attraverso esperienze condivise sostengono i membri attraverso confronto tra pari, modellamento sociale e feedback costruttivo orientato alla crescita (RIVA, 2025a). La chiave consiste nel creare ambienti dove la vulnerabilità è sicura ma la mancanza di cambiamento viene gentilmente messa in discussione.

3.6. *Dual-Process Theories e architettura cognitiva.*

Le teorie dei processi duali, che distinguono tra elaborazione automatica (Sistema 1) e controllata (Sistema 2), illuminano come le tecnologie digitali possano strategicamente ingaggiare differenti sistemi cognitivi (KAHNEMAN, 2002). Il comfort opera spesso attraverso il Sistema 1 — intuitivo, automatico e familiare — mentre la crescita richiede tipicamente il coinvolgimento del Sistema 2 — razionale, impegnativo e nuovo.

Le tecnologie efficaci possono sfruttare questa comprensione costruendo inizialmente abitudini confortevoli del Sistema 1 prima di introdurre gradualmente sfide del Sistema 2. Un sistema di IA per la produttività potrebbe iniziare creando routine semplici e automatiche di organizzazione delle informazioni prima di introdurre processi decisionali complessi sulla prioritizzazione e l'allocazione strategica delle risorse cognitive. Questo approccio rispetta l'architettura cognitiva promuovendo cambiamento comportamentale sostenibile.

L'integrazione di questi sistemi è cruciale. Le tecnologie che mirano esclusivamente al Sistema 2 possono essere efficaci nel breve termine ma esauriscono le risorse cognitive limitate degli utenti. Inversamente, quelle che operano solamente attraverso il Sistema 1 possono essere confortevoli ma falliscono nel promuovere crescita signifi-

cativa. L'approccio ottimale cicla tra i sistemi, usando il comfort del Sistema 1 per creare capacità per le sfide di crescita del Sistema 2.

Come notato da Chiriatti e colleghi (2024; 2025) e discusso precedentemente nel capitolo, l'intelligenza artificiale introduce un ulteriore livello di complessità operando come Sistema 0: un preprocessore cognitivo che modula l'input a entrambi i sistemi. Questa posizione privilegiata richiede particolare attenzione progettuale per assicurare che il Sistema 0 non elimini sistematicamente le opportunità di ingaggio del Sistema 2, ma piuttosto ottimizzi i momenti e i modi in cui la sfida cognitiva viene presentata.

4. Conclusioni: L'IA tra 'fatto' predittivo e 'norma' evolutiva.

L'analisi condotta in questo capitolo ha esplorato una tensione fondamentale che caratterizza l'integrazione dell'intelligenza artificiale nei processi cognitivi umani. Questa tensione può essere sintetizzata nella dialettica tra il '*fatto predittivo*' e la '*norma evolutiva*': da un lato, l'IA come sistema che osserva il comportamento passato per prevedere e soddisfare i bisogni futuri, creando ambienti digitali estremamente personalizzati e confortevoli; dall'altro, l'IA come partner che facilita l'evoluzione cognitiva, aiutando gli individui a trascendere i pattern esistenti e a diventare versioni migliorate di se stessi.

La dimensione del *fatto predittivo* rappresenta la realtà empirica dell'intelligenza artificiale contemporanea. I sistemi di IA eccellono nell'apprendimento di pattern comportamentali, nell'anticipazione delle preferenze, e nell'ottimizzazione dell'esperienza utente verso la massima fluidità e il minimo attrito cognitivo. Questa capacità predittiva, quando applicata sistematicamente, genera quello che è stato definito Sistema 0 (Chiriatti et al., 2024; 2025): un preprocessore cognitivo che filtra e struttura l'informazione prima che raggiunga la consapevolezza riflessiva.

Come documentato nella ricerca empirica, questa dimensione predittiva assolve funzioni psicologiche importanti. Riduce il carico cognitivo estrinseco, costruisce fiducia attraverso consistenza, e crea sicurezza psicologica nell'interazione con sistemi complessi (Riva-Wiederhold-Mantovani, 2021). Tuttavia, la stessa efficienza predittiva che genera comfort può trasformarsi in un vincolo epistemico. Quando i sistemi anticipano troppo accuratamente le preferenze esi-

stenti, rischiano di cristallizzare pattern cognitivi invece di facilitarne l'evoluzione (SHARMA et al., 2023; GLICKMAN-SHAROT, 2025).

La dimensione della ''norma evolutiva'' non va intesa come norma giuridica, ma come criterio progettuale e regolativo: un principio secondo cui l'IA dovrebbe essere valutata non solo per la sua capacità di predire e soddisfare preferenze esistenti, ma anche per la sua capacità di sostenere autonomia, riflessività, apprendimento e crescita cognitiva. Questa concezione normativa richiede un cambio di paradigma progettuale: dall'ottimizzazione dell'efficienza immediata alla promozione dello sviluppo a lungo termine.

I framework cyberpsicologici discussi in questo capitolo forniscono i principi teorici per questa trasformazione. La Flow Theory (CSIKSZENTMIHALYI, 1990) suggerisce che l'esperienza ottimale emerge quando la sfida corrisponde perfettamente alla competenza, dissolvendo il paradosso tra comfort e crescita. Il Transtheoretical Model (PROCHASKA-VELICER, 1997) indica che il bilanciamento ottimale tra supporto e sfida varia lungo gli stadi del processo di cambiamento. La Self-Determination Theory (RYAN-DECI, 2000) evidenzia che la crescita sostenibile richiede il supporto di autonomia, competenza e relazionalità. La Cognitive Load Theory (KIRSCHNER, 2002; SWELLER-VAN MERRIENBOER-PAAS, 1998) dimostra che la crescita cognitiva richiede l'ottimizzazione del carico pertinente, non la sua eliminazione.

Questi principi convergono verso una concezione dell'IA come psicotecnologia evolutiva: un sistema che utilizza le sue capacità predittive non per eliminare ogni difficoltà, ma per ottimizzare i momenti e i modi in cui la sfida cognitiva viene presentata. L'obiettivo non è rendere la tecnologia il più semplice possibile, ma il più orientata all'*evoluzione cognitiva* possibile.

La cyberpsicologia emerge da questa analisi come la disciplina chiave per navigare la tensione dialettica tra fatto predittivo e norma evolutiva. Non si limita a descrivere il problema o a formulare astrazioni teoriche, ma fornisce gli strumenti concettuali e metodologici per progettare concretamente sistemi di IA che possano essere simultaneamente predittivi ed evolutivi.

Particolarmente significativo è il contributo della cyberpsicologia nell'identificare e gestire i rischi dell'integrazione IA-cognizione umana. La ricerca ha documentato come le interazioni ripetute con sistemi di IA possano alterare non solo decisioni immediate ma meccanismi cognitivi sottostanti, modificando pattern percettivi, emotivi e sociali in modo duraturo (GLICKMAN-SHAROT, 2025). Questa evi-

denza sottolinea l'urgenza di approcci progettuali informati da solide basi psicologiche, capaci di anticipare e mitigare conseguenze cognitive non intenzionali.

In conclusione, la prospettiva cyberpsicologica si rivela non come un'opzione teorica tra molte, ma come un imperativo progettuale per lo sviluppo responsabile dell'intelligenza artificiale. Man mano che i sistemi di IA si integrano sempre più profondamente nei processi cognitivi umani, operando come Sistema 0 che precondiziona e struttura il pensiero, diventa imperativo che questa integrazione sia guidata da una solida comprensione dei meccanismi psicologici, dei bisogni evolutivi, e delle dinamiche del cambiamento cognitivo.

Il paradosso comfort-crescita non rappresenta un problema tecnico da risolvere attraverso ottimizzazioni algoritmiche, ma una tensione intrinseca all'esperienza umana che richiede gestione consapevole e contestualmente appropriata. La cyberpsicologia fornisce gli strumenti per questa gestione, trasformando una tensione potenzialmente paralizzante in un'opportunità per progettare sistemi che fungano da autentici partner nello sviluppo umano.

Il futuro dell'intelligenza artificiale non risiede in tecnologie che rendono semplicemente la vita più facile, ma in quelle che rendono gli esseri umani più capaci, più resilienti, più adattivi. Abbracciando il paradosso comfort-crescita e utilizzando le intuizioni cyberpsicologiche per navigarlo, possiamo creare tecnologie che servono come catalizzatori per il benessere umano: non specchi passivi dei nostri desideri attuali, ma partner attivi nella nostra evoluzione cognitiva e nel nostro sviluppo come agenti capaci di confrontarsi costruttivamente con la complessità del mondo contemporaneo.

Note bibliografiche

Carr N. (2010), *The shallows: What the Internet is doing to our brains*, New York, W.W. Norton & Company.

Chiriatti M.-Bergamaschi Ganapini M.-Panai E.-Wiederhold B.K.-Riva G. (2025), *System 0: Transforming Artificial Intelligence into a Cognitive Extension*, in *Cyberpsychology, Behavior, and Social Networking*, 28, p. 534 ss.

Chiriatti M.-Ganapini M.-Panai E.-Ubiali M.-Riva G. (2024), *The case for human-AI interaction as system 0 thinking*, in *Nature Human Behaviour*, 8, pp. 1829-1830.

Cinelli M.-De Francisci Morales G.-Galeazzi A.-Quattrociocchi W.-Starnini M. (2021), *The echo chamber effect on social media*, in *Proceedings of the National Academy of Sciences*, 118(9), e2023301118.

CLARK A.-CHALMERS D. (1998), *The extended mind*, in *Analysis*, 58(1), p. 7 ss.
CSIKSZENTMIHALYI M. (1990), *Flow: The psychology of optimal experience*, New York, HarperCollins.
DE KERCKHOVE D. (1991), *Brainframes: Technology, mind and business*, Utrecht, Bosch & Keuning.
GLICKMAN M.-SHAROT T. (2025), *How human-AI feedback loops alter human perceptual, emotional and social judgements*, in *Nature Human Behaviour*, 9(2), p. 345 ss.
HOLLAN J.-HUTCHINS E.-KIRSH D. (2000), *Distributed cognition: Toward a new foundation for human-computer interaction research*, in *ACM Transactions on Computer-Human Interaction*, 7(2), p. 174 ss.
KAHNEMAN D. (2002), *Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice*, in FRÄNGSMYR T. (ed.), *The Nobel Prizes 2002*, Stockholm, Nobel Foundation, p. 449 ss.
KHAMAJ A.-ALI A.M. (2024), *Examining the usability and accessibility challenges in mobile health applications for older adults*, in *Alexandria Engineering Journal*, 102, pp. 179-191.
KIRSCHNER P.A. (2002), *Cognitive load theory: Implications of cognitive load theory on the design of learning*, *Learning and Instruction*, 12(1), pp. 1-10, DOI 10.1016/S0959-4752(01)00014-7.
MITTELSTADT B.D.-ALLO P.-TADDEO M.-WACHTER S.-FLORIDI L. (2016), *The ethics of algorithms: Mapping the debate*, in *Big Data & Society*, 3(2), 2053951716679679.
MÜLLER R.-PRIMC N.-KUHN E. (2023), *'You have to put a lot of trust in me': autonomy, trust, and trustworthiness in the context of mobile apps for mental health*, in *Medicine, Health Care and Philosophy*, 26, pp. 313-324.
NAKAMURA-CSIKSZENTMIHALYI (2009), *Flow Theory And Research*, in LOPEZ S.J.-SNYDER C.R. (eds.), *Handbook of positive psychology*, New York, Oxford University Press, p. 195 ss.
PIZZOLI S.F.M.-RABARBARI E.-SCERRATI E.-RIVA G. (2024), *Enhancing Emotion Regulation Through Arousal Modulation with Modal and Amodal Virtual Food Representations*, in *Cyberpsychology, Behavior, and Social Networking*, 27, p. 294 ss.
PROCHASKA J.O.-VELICER W.F. (1997), *The transtheoretical model of health behavior change*, in *American Journal of Health Promotion*, 12, p. 38 ss.
RIVA G.-WIEDERHOLD B.K.-MANTOVANI F. (2019), *Neuroscience of virtual reality: From virtual exposure to embodied medicine*, in *Cyberpsychology, Behavior, and Social Networking*, 22(1), p. 82 ss.
RIVA G.-WIEDERHOLD B.K.-MANTOVANI F. (2021), *Surviving COVID-19: The Neuroscience of Smart Working and Distance Learning*, in *Cyberpsychology, Behavior, and Social Networking*, 24, p. 79 ss.
RIVA G. (2025a), *Digital 'We': Human Sociality and Culture in the Era of Social Media and AI*, in *American Psychologist*, 80 (9), 1535-1549.
RIVA G. (2025b), *The Architecture of Cognitive Amplification: Enhanced Cognitive Scaffolding as a Resolution to the Comfort-Growth Paradox in Human-AI Cognitive Integration*, OSF Preprints, accessibile al sito: *https://osf.io/preprints/osf/4wcdb_v1*.

Ryan R.M.-Deci E.L. (2000), *Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being*, in *The American Psychologist*, 55, p. 68 ss.

Sajno E.-Beretta A.-Novielli N.-Riva G. (2022), *Follow the Flow: A Prospective on the On-Line Detection of Flow Mental State through Machine Learning*, in *2022 IEEE International Workshop on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE 2022 - Proceedings*, p. 217 ss.

Sajno E.-Maisetti M.-Luongo R.-Cipresso P. (2022), *Cyberintervention on plant workforce's mental activity for safety*, in *Annual Review of CyberTherapy and Telemedicine*, 20, p. 71 ss.

Sharma M.-Tong M.-Korbak T.-Duvenaud D.-Askell A.-Bowman S.R.-Newton C.-Hatfield-Dodds Z.-Barez F.-Perez E. (2023), *Towards understanding sycophancy in language models*, arXiv preprint arXiv:2310.13548.

Simons D.J.-Boot W.R.-Charness N.-Gathercole S.E.-Chabris C.F.-Hambrick D.Z.-Stine-Morrow E.A. (2016), *Do "brain-training" programs work?*, in *Psychological Science in the Public Interest*, 17, p. 103 ss.

Suler J. (2016), *Psychology of the digital age: Humans become electric*, Cambridge, Cambridge University Press.

Sweller J.-Van Merrienboer J.J.G.-Paas F.G.W.C. (1998), *Cognitive Architecture and Instructional Design*, in *Educational Psychology Review*, 10, p. 251 ss.

Turkle S. (2011), *Alone together: Why we expect more from technology and less from each other*, New York, Basic Books.

Vijayalakshmi S.-Disney Sandhya A.-Nimala D.P. (2025), *AI-Powered Holistic Mental Health Monitoring: Integrating Facial Emotion Recognition, Chatbot, and Voicebot for Personalized Support*, in *2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, IEEE, pp. 1-6.

Vygotsky L.S. (1978), *Mind in society: The development of higher psychological processes*, Cambridge, MA, Harvard University Press.

Wilson R.C.-Shenhav A.-Straccia M.-Cohen J.D. (2019), *The eighty five percent rule for optimal learning*, in *Nature Communications*, 10, p. 4646 ss.

Yan M.-Filieri R.-Raguseo E.-Gorton M. (2021), *Mobile apps for healthy living: Factors influencing continuance intention for health apps*, in *Technological Forecasting and Social Change*, 166, p. 120644 ss.

(LA NORMA)
INTELLIGENZA ARTIFICIALE E SALUTE: PROFILI PUBBLICISTICI

di *Andrea Simoncini*

Sommario: 1. Premessa. — 2. La tutela costituzionale della salute e la sfida dell'IA, tra diritto e organizzazione. — 3. La salute come diritto fondamentale. L'impiego clinico dell'IA e il rischio di ''cattura della decisione''. — 3.1. Applicazioni cliniche: dalla ricerca alla diagnosi e alla terapia personalizzata. — 3.2. Il rischio etico-cognitivo: l'automation bias e la ''cattura della decisione'' clinica. — 3.3. Conseguenze dell'automation bias: ipermedicalizzazione e aumento dei 'falsi positivi'. — 3.4. I 'bias algoritmici': la non-trasparenza e la disuguaglianza sanitaria. — 4. La salute come organizzazione. L'AI Act. — 4.1. I sistemi di IA ad alto rischio nell'AI Act: criteri di classificazione in ambito sanitario. — 4.2. L'IA nell'accesso ai servizi sanitari: dalle risorse scarse al triage di emergenza. — 4.3. L'IA nella profilazione e nelle polizze assicurative: i rischi di discriminazione economica. — 5. La *governance* europea e nazionale. — 5.1. L'AI Act come legislazione sulla qualità dei prodotti: requisiti di sicurezza e conformità. — 5.2. L'European Health Data Space (EHDS) e il controllo dei dati sanitari elettronici. — 5.2.1. Uso primario: accesso, controllo e condivisione transfrontaliera. — 5.2.2. Uso secondario: ricerca, opt-out e ambienti di trattamento sicuri. — 5.3. La legge n. 132 del 2025: i principi di non-discriminazione e la centralità della decisione medica umana. — 6. Responsabilità e decisioni mediche. — 6.1. Il ruolo della responsabilità professionale medica nell'architettura uomo-macchina. — 6.2. Ripensare l'intero processo: dalla sostituzione all'assistenza consapevole. — 6.3. Il modello della responsabilità condivisa (*shared responsibility*). — 7. Conclusione. L'impiego dell'IA, tra dignità umana e la libertà di scelta terapeutica.

1. Premessa.

L'introduzione dell'Intelligenza Artificiale (IA) nel settore sanitario non si configura come una mera evoluzione strumentale, ma come un'autentica inversione del paradigma tecnico tradizionale. Storicamente, la tecnologia è stata impiegata per eseguire le decisioni umane con maggiore velocità, potenza, accuratezza ed efficienza. L'IA, tuttavia, è oggi utilizzata prevalentemente come una 'tecnologia

decidente': « un sistema automatizzato progettato per funzionare con livelli di autonomia variabili e che può presentare adattabilità dopo la diffusione e che, per obiettivi espliciti o impliciti, deduce dall'input che riceve come generare output quali previsioni, contenuti, raccomandazioni o decisioni che possono influenzare ambienti fisici o virtuali », secondo la definizione contenuta nell'articolo 3 dell'AI Act.

Questa distinzione tra automazione (l'esecuzione di una decisione precedentemente definita dall'uomo) e autonomia (la capacità di *prendere* la decisione) pone interrogativi radicali sul corretto rapporto tra umani e macchine nello svolgimento di attività complesse, quali la ricerca, la prevenzione, la diagnosi e la cura. Di conseguenza, la sfida normativa non si limita alla semplice regolamentazione dei prodotti tecnologici in termini di sicurezza (come stabilito dalla legislazione preesistente sui *medical devices*), ma deve necessariamente affrontare la gestione dell'autonomia intrinseca di questi sistemi e il loro potenziale di incidere direttamente sui diritti fondamentali.

2. La tutela costituzionale della salute e la sfida dell'IA, tra diritto e organizzazione.

In Italia, la tutela della salute è elevata a principio cardine dell'ordinamento giuridico. L'articolo 32 della Costituzione sancisce che la Repubblica tutela la salute come « fondamentale diritto dell'individuo e interesse della collettività » e garantisce cure gratuite agli indigenti. Significativamente, lo stesso articolo impone un limite invalicabile al potere legislativo, stabilendo che la legge non può « in nessun caso violare i limiti imposti dal rispetto della persona umana ».

La salute è intesa nel sistema giuridico italiano non soltanto come l'assenza di malattia o infermità (in linea con la definizione dell'Organizzazione Mondiale della Sanità che la descrive come uno stato di completo benessere fisico, mentale e sociale), ma come un diritto fondamentale che va protetto da potenziali lesioni. La giurisprudenza ha nel tempo rafforzato il carattere del diritto alla salute, superando le distinzioni classiche e attribuendogli una connotazione non solo pubblicistica (obbligo dello Stato di garantire i *Livelli essenziali di assistenza*) ma anche privatistica, soprattutto in relazione all'evoluzione della nozione di danno biologico. Quando l'IA si configura come una tecnologia decidente che interviene nel processo di cura o nell'allocazione delle risorse sanitarie, essa impatta direttamente sulla sfera di questo diritto fondamentale. Qualsiasi sistema algoritmico che, a causa di *bias*

o criteri di selezione, introduca elementi di discriminazione o iniquità nell'accesso alle cure potrebbe configurare una violazione del limite costituzionale del rispetto della persona umana. L'efficienza promessa dall'algoritmo non può, in questo contesto, prevalere sulla dignità individuale e sul principio di universalità del servizio.

L'impatto dell'IA sulla tutela della salute si articola in due dimensioni interdipendenti, ma distintive, nel sistema giuridico e organizzativo. Innanzitutto, una dimensione che potremmo definire 'micro', legata alla sua natura di diritto fondamentale e che riguarda l'IA impiegata direttamente nella cura e nel rapporto medico-paziente. Questa dimensione concerne la salute intesa come diritto/dovere, focalizzata sulla libertà e sulla dignità dell'individuo. Le applicazioni includono la ricerca, la diagnostica, l'attività di cura, la riabilitazione, il monitoraggio e la predizione/prevenzione dell'insorgenza delle malattie.

Vi è poi una dimensione 'macro' che riguarda l'IA impiegata nell'organizzazione dei servizi sanitari alla collettività. Questa dimensione si concentra sulla salute intesa come organizzazione del servizio sanitario. In quanto tecnologia decidente, l'IA può essere utilizzata per la gestione delle risorse, la valutazione dell'accesso alle prestazioni essenziali e la definizione delle politiche sanitarie.

Entrambe le dimensioni sono soggette a regolamentazione europea e nazionale, ma presentano rischi etici e legali distinti, che saranno analizzati nei capitoli successivi.

3. La salute come diritto fondamentale. L'impiego clinico dell'IA e il rischio di "cattura della decisione".

3.1. *Applicazioni cliniche: dalla ricerca alla diagnosi e alla terapia personalizzata.*

Nella dimensione micro, l'IA ha dimostrato un grande potenziale trasformativo, contribuendo significativamente alla ricerca scientifica e al miglioramento delle attività cliniche. L'IA accelera la ricerca, come testimoniato dall'applicazione della biologia computazionale e della previsione della struttura delle proteine, un campo in cui DeepMind di Google ha già ottenuto riconoscimenti di livello internazionale (si pensi che il Premio Nobel per la Chimica 2024 è stato assegnato a Demis Hassabis e John Jumper per la predizione di strutture proteiche realizzata mediante l'impiego di IA attraverso il sistema AlphaFold prodotto da Google DeepMind).

Le applicazioni dirette si estendono attraverso l'intero percorso clinico:

Diagnostica: L'IA supporta figure professionali come radiologi e anatomo-patologi nell'analisi di risonanze, vetrini digitali (la cosiddetta *digital pathology*) o interpretazione dei comportamenti (es. in psichiatria o neurologia), contribuendo alla stesura dei referti. Algoritmi avanzati di visione artificiale possono identificare *pattern* cellulari e tissutali, assistendo i patologi nella diagnosi. Esempi come l'algoritmo "Sybil" mostrano la capacità predittiva dell'IA, analizzando immagini TAC per stimare il rischio di sviluppare patologie, come il tumore al polmone, con anni di anticipo.

Cura e Prevenzione: L'IA viene utilizzata in piattaforme di supporto alle decisioni per la gestione dei pazienti cronici, nella riabilitazione, nel monitoraggio e nella predizione/prevenzione dell'insorgenza delle malattie.

Nonostante i vantaggi in termini di precisione e personalizzazione, la pervasività dell'IA nella pratica clinica solleva profonde questioni etiche e legali relative al mantenimento dell'autonomia e della responsabilità professionale umana.

3.2. *Il rischio etico-cognitivo: l'automation bias e la "cattura della decisione" clinica.*

L'integrazione di sistemi autonomi nel processo decisionale medico comporta un rischio etico-cognitivo noto come *automation bias*. Questo meccanismo psicologico induce le persone a riporre un'eccessiva fiducia nelle raccomandazioni, nei suggerimenti o nelle decisioni fornite da sistemi automatizzati, anche quando esistono indicazioni che suggeriscono che tali raccomandazioni possano essere incomplete o errate.

Quando un dispositivo tecnologico autonomo viene inserito in un processo decisionale umano complesso, esiste una forte tendenza del dispositivo a 'catturare' la decisione stessa. Questa cattura trasforma radicalmente la relazione di cura e i presupposti giuridici che la regolano, incidendo sulla trasparenza delle scelte sanitarie e sulla fiducia nel rapporto terapeutico.

La dinamica della 'cattura' è intrinsecamente legata al *design* del sistema di IA. Se un sistema è progettato per produrre una bozza di referto elaborata autonomamente (*opting out*), che il medico è chiamato semplicemente a firmare o correggere, l'esperienza operativa (praticità e rapidità) spinge inevitabilmente il clinico a non smentire

l'algoritmo. In tale scenario, il rapporto uomo-macchina si conclude in una probabile sostituzione della funzione umana.

Al contrario, se l'IA è utilizzata per segnalare al medico una serie di elaborazioni di dati o immagini che egli utilizzerà per stendere il proprio referto, la macchina agisce come supporto consapevole, assistendo l'umano a commettere meno errori e mantenendo la sovranità decisionale in capo al professionista.

Il successo nella garanzia che l'IA sia un *aiuto* e non una *sostituzione* dipende, dunque, dalla riprogettazione completa delle procedure e degli ambienti cognitivi, definendo chiaramente il contributo distinto di umani e macchine.

3.3. *Conseguenze dell'automation bias: ipermedicalizzazione e aumento dei 'falsi positivi'.*

Una conseguenza diretta e misurata dell'uso massiccio di sistemi di decisione autonoma è il fenomeno dell'ipermedicalizzazione, caratterizzato da un conseguente aumento dei falsi positivi.

Questo accade perché l'architettura della responsabilità professionale medica gioca un ruolo decisivo. Se una macchina, esaminando una risonanza, suggerisce al medico l'esigenza di ulteriori analisi, il professionista umano è posto in una posizione difficile. Rischia di esporsi a responsabilità legale per negligenza o omissione diagnostica se decide, basandosi sulla sua conoscenza ed esperienza, di smentire il suggerimento algoritmico. Per mitigare tale rischio legale, la prassi tende a massimizzare le indagini diagnostiche, anche quando il parere della macchina è incerto, portando a un eccesso di medicalizzazione. Questo sposta il rischio clinico dalla possibilità di non vedere una patologia (negligenza) a quella di vedere patologie inesistenti o clinicamente irrilevanti (eccesso di conformità all'algoritmo).

3.4. *I 'bias algoritmici': la non-trasparenza e la disuguaglianza sanitaria.*

Oltre ai rischi di cattura cognitiva, l'IA in sanità introduce la possibilità di amplificare le disuguaglianze sanitarie preesistenti. L'addestramento dei modelli di IA su *dataset* storici, che riflettono le distorsioni e i pregiudizi clinici della società, può portare l'algoritmo a sottovalutare i sintomi di alcune categorie di pazienti, in particolare donne e individui appartenenti a minoranze etniche.

Diversi studi hanno rilevato che i grandi modelli linguistici (LLM) già utilizzati negli ospedali possono suggerire cure meno intensive a pazienti donne o rispondere in modo meno empatico a pazienti di colore o asiatici. Una minore considerazione clinica basata sulla razza o sul genere percepiti dal modello rischia di consolidare schemi di sottotrattamento, violando i principi fondamentali di equità e non-discriminazione alla base del diritto alla salute.

A ciò si aggiunge il problema della *black box*, ovvero la non-trasparenza della logica decisionale utilizzata dall'IA. Senza una chiara informazione sulla logica algoritmica, diventa difficile per il medico e, ancor più per il paziente, comprendere e contestare la giustificazione di una raccomandazione o di una diagnosi. La standardizzazione e la diffusione dell'IA clinica, pur essendo benefiche per l'innovazione, devono essere gestite con estrema cautela per non diffondere su larga scala questi *bias* sistemici.

4. La salute come organizzazione. L'AI Act.

4.1. *I sistemi di IA ad alto rischio nell'AI Act: criteri di classificazione in ambito sanitario.*

La dimensione *macro* dell'IA riguarda l'organizzazione dei servizi e l'allocazione delle risorse, aspetti che l'AI Act (Regolamento UE 2024/1689), sebbene miri principalmente alla sicurezza dei prodotti, classifica esplicitamente come ad alto rischio.

L'AI Act, pienamente operativo dal 2 agosto 2026, identifica come sistemi ad alto rischio quelli il cui potenziale impatto sui diritti fondamentali e sulla sicurezza giustifica requisiti regolamentari più stringenti. Nello specifico, l'Allegato III, paragrafo 5, include due categorie di sistemi IA che incidono direttamente sulla sanità pubblica:

a) Sistemi destinati a essere utilizzati dalle o per conto delle autorità pubbliche per valutare l'ammissibilità delle persone fisiche a prestazioni e servizi di assistenza pubblica essenziali, inclusi i servizi sanitari, o per concedere, ridurre, revocare o recuperare tali prestazioni.

b) Sistemi destinati a essere utilizzati per la selezione dei pazienti per quanto concerne l'assistenza sanitaria di emergenza (triage), o per la valutazione e classificazione delle chiamate di emergenza e l'invio dei servizi di primo soccorso.

Queste previsioni indicano che l'Unione Europea riconosce un rischio intrinseco nell'affidare decisioni organizzative cruciali, relative all'accesso e alla priorità delle cure, a sistemi algoritmici.

4.2. *L'IA nell'accesso ai servizi sanitari: dalle risorse scarse al triage di emergenza.*

L'utilizzo di sistemi di IA per organizzare l'accesso ai servizi sanitari presenta margini di miglioramento significativi, grazie alla capacità dell'IA di fornire soluzioni *data-driven*, analizzare grandi quantità di dati complessi e operare in tempi rapidissimi. Tuttavia, l'impiego di IA per decidere chi può accedere a determinate cure in condizioni di scarsità di risorse (un problema emerso drammaticamente durante la pandemia) solleva il più grave rischio etico nella dimensione *macro*.

Delegare l'accesso ai servizi sanitari o di emergenza a sistemi di IA, specialmente in modo autonomo e senza supervisione umana, è considerato un rischio altissimo. Sebbene l'IA possa fornire una base tecnica e obiettiva per la selezione, la decisione finale sull'allocazione di risorse vitali è una decisione etica e valoriale, che non può essere totalmente decontestualizzata dalla responsabilità umana. La classificazione di questi sistemi come ad alto rischio nell'AI Act riflette la consapevolezza che tali applicazioni toccano il nucleo della giustizia distributiva e della dignità umana.

4.3. *L'IA nella profilazione e nelle polizze assicurative: i rischi di discriminazione economica.*

Un ulteriore rischio sistemico si manifesta nell'impiego di sistemi di IA per profilare le abitudini di vita delle persone, con l'obiettivo di determinare il costo delle polizze assicurative o l'accesso a programmi di prevenzione specifici. L'acquisizione di aziende produttrici di dispositivi di monitoraggio della salute (come l'acquisto di Fitbit da parte di Google) evidenzia la convergenza tra dati sanitari e interessi commerciali.

Quando l'IA viene utilizzata nel settore assicurativo per la profilazione, il rischio medico tende a essere trasformato da un onere collettivo (tipico del servizio sanitario pubblico) a una responsabilità individuale basata su un punteggio algoritmico. Questo meccanismo può generare nuove forme di disuguaglianza economica nell'accesso alle cure e alla prevenzione. Individui profilati come ad alto rischio comportamentale potrebbero vedersi negare coperture o vedersi imporre costi esorbitanti, minando il principio di universalità e solidarietà del sistema sanitario. La regolamentazione deve pertanto impedire che l'analisi predittiva dei dati sanitari, pur promettendo preven-

zione, si traduca in criteri discriminatori che limitino l'accesso ai servizi essenziali.

5. La *governance* europea e nazionale.

La regolamentazione dell'IA in sanità si basa su tre pilastri normativi principali: l'AI Act (che regola il prodotto), l'European Health Data Space (EHDS, che regola il dato) e la legislazione nazionale (che specifica l'applicazione clinica).

5.1. *L'AI Act come legislazione sulla qualità dei prodotti: requisiti di sicurezza e conformità.*

L'AI Act, pienamente operativo a partire dal 2 agosto 2026, si concentra sulla sicurezza dei prodotti tecnologici e sulla protezione dei diritti fondamentali. In quanto legislazione di armonizzazione basata sul Nuovo Quadro Legislativo Europeo, l'AI Act identifica i rischi e stabilisce i requisiti specifici per la gestione di tali rischi.

Per i sistemi di IA classificati come ad alto rischio (e, tra questi, quelli sanitari menzionati nell'Allegato III), sono previsti obblighi stringenti di valutazione della conformità. Questi includono la procedura di controllo interno della conformità (Allegato VI) o, in alternativa, una valutazione che coinvolge un organismo di certificazione, basata sul sistema di gestione della qualità e sulla documentazione tecnica (Allegato VII). Poiché molti sistemi di IA clinica sono anche dispositivi medici, l'AI Act agisce per completare il quadro normativo esistente sui '*medical devices*', assicurando che la sicurezza algoritmica sia integrata con la sicurezza del dispositivo medico. La responsabilità di garantire la sicurezza e la conformità non ricade solo sullo sviluppatore, ma anche sul *deployer* (l'organismo di diritto pubblico o l'ente privato che fornisce servizi pubblici) che è ritenuto nella posizione migliore per identificare i rischi nel contesto specifico di utilizzo.

5.2. *L'European Health Data Space (EHDS) e il controllo dei dati sanitari elettronici.*

Parallelamente all'AI Act, l'Unione Europea ha istituito l'European Health Data Space (EHDS) con l'obiettivo di creare un quadro giuridico e tecnico armonizzato per lo scambio dei dati sanitari elettronici, potenziando il controllo degli individui sui propri dati e favorendo l'uso secondario per la ricerca e l'innovazione.

5.2.1. *Uso primario: accesso, controllo e condivisione transfrontaliera.*

L'EHDS mira a dare maggiore potere agli individui sull'accesso e il controllo dei loro dati sanitari elettronici personali a livello nazionale e transfrontaliero, per l'erogazione dei servizi sanitari. Le garanzie per l'interessato nell'uso primario sono:

— Diritto di accesso rapido e gratuito ai propri dati in formato elettronico standardizzato (come referti di immagini e risultati di laboratorio).

— Diritto di limitare l'accesso (in tutto o in parte) ai propri dati per i professionisti sanitari.

— Diritto di visualizzare chi ha avuto accesso ai dati e di richiederne la correzione in caso di errori.

5.2.2. *Uso secondario: ricerca, opt-out e ambienti di trattamento sicuri.*

L'uso secondario dei dati (per ricerca, politica pubblica, regolamentazione, e medicina personalizzata) è fondamentale per l'addestramento dei sistemi di IA. L'EHDS bilancia questa necessità con importanti misure di tutela della privacy.

Il trattamento dei dati per uso secondario è consentito solo per scopi specifici e basato su un permesso rilasciato da un organismo di accesso ai dati sanitari. La persona fisica ha il diritto di opporsi (*opt-out*) in modo semplice e reversibile all'uso secondario dei propri dati, sebbene possano esistere eccezioni per importanti scopi di interesse pubblico, subordinate a rigorose misure di trasparenza.

Per garantire la sicurezza, il trattamento dei dati avviene esclusivamente in ambienti di elaborazione sicuri (*secure processing environments*) che aderiscono ai più elevati standard di *cybersecurity*. È richiesto l'uso di dati pseudonimizzati (se l'anonimizzazione non è sufficiente), ed è severamente vietato tentare di ri-identificare i soggetti o scaricare dati personali da tali ambienti. È inoltre proibito utilizzare i dati per prendere decisioni che possano essere dannose per gli individui o per scopi di marketing.

5.3. *La legge n. 132 del 2025: i principi di non-discriminazione e la centralità della decisione medica umana.*

A livello nazionale, il Parlamento ha approvato la legge n. 132 del 2025 (*Disposizioni e delega in materia di IA*), che, in aggiunta al

quadro normativo europeo che ho richiamato, introduce ulteriori principi specifici per l'ambito sanitario, che mirano a rafforzare la sovranità decisionale umana di fronte alla tecnologia algoritmica.

L'articolo 7 della legge stabilisce due principi normativi fondamentali. Innanzitutto, il divieto di discriminazione. L'introduzione di sistemi di IA nel sistema sanitario « non può selezionare e condizionare l'accesso alle prestazioni sanitarie secondo criteri discriminatori » (Art. 7, comma 2). Questo principio rinforza il dettato costituzionale e contrasta esplicitamente i rischi associati all'allocazione algoritmica di risorse (*Dimensione Macro*).

In secondo luogo, l'IA come supporto non vincolante. I sistemi di IA in ambito sanitario « costituiscono un supporto nei processi di prevenzione, diagnosi, cura e scelta terapeutica, lasciando impregiudicata la decisione, che è sempre rimessa agli esercenti la professione medica » (Art. 7, comma 5).

Quest'ultimo punto è cruciale in quanto rappresenta un tentativo normativo esplicito di contrastare il rischio di *automation bias* e la 'cattura della decisione' (*Dimensione Micro*). Sebbene il medico mantenga formalmente la responsabilità, il successo di questa disposizione dipenderà dall'implementazione di protocolli operativi che assicurino l'effettiva *consapevolezza* e *autonomia* del medico nell'uso dello strumento, evitando che la praticità operativa trasformi il supporto in sostituzione.

La legge rafforza inoltre il diritto all'informazione, garantendo all'interessato il diritto di essere informato non solo sull'utilizzo dell'IA e sui vantaggi derivanti, ma anche sulla 'logica decisionale utilizzata'.

6. Responsabilità e decisioni mediche.

6.1. *Il ruolo della responsabilità professionale medica nell'architettura uomo-macchina.*

Il quadro della responsabilità professionale medica, in Italia regolato dalla Legge Gelli-Bianco (n. 24 del 2017), è fondamentale per definire l'architettura decisionale nel contesto dell'IA.

L'introduzione di strumenti di IA crea un paradosso legale per il medico: se il professionista segue il suggerimento algoritmico, anche se non pienamente convinto, rischia di incorrere in ipermedicalizzazione o falsi positivi. Se invece decide di disattendere l'indicazione dell'IA (ad esempio, rigettando una proposta di referto automatizzato) e si dovesse verificare un esito avverso, la registrazione del-

l'output algoritmico non seguito potrebbe diventare una prova schiacciante di negligenza, esponendolo a responsabilità. Questo meccanismo incentiva il medico a cedere alla praticità dell'algoritmo, contribuendo alla '*cattura della decisione*'.

Per superare questo stallo, è necessario ridisegnare completamente le procedure operative. La disciplina della responsabilità professionale deve essere utilizzata come leva per definire un'architettura decisionale chiara, in cui sia distinto il contributo tecnico della macchina dalla convalida e dall'interpretazione clinica umana. L'elaborazione di protocolli di utilizzo specifici, come quelli in fase di studio per la *digital pathology* è essenziale per guidare i professionisti nell'interazione con l'IA.

6.2. *Ripensare l'intero processo: dalla sostituzione all'assistenza consapevole.*

La distinzione tra l'IA che sostituisce e l'IA che supporta consapevolmente, è cruciale per la salvaguardia dell'autonomia professionale. Il modello del *referto bozza* o della *soluzione pre-compilata* che il medico deve solo correggere (*opting out*) facilita la sostituzione umana, in quanto, per ragioni di efficienza, è più probabile che il medico si appoggi inevitabilmente sulla decisione algoritmica. Al contrario, un sistema che utilizza l'IA per segnalare al medico una serie di immagini o dati grezzi elaborati, che l'umano poi sintetizza nel referto, favorisce un'assistenza consapevole. In questo secondo caso, la macchina aiuta l'uomo a commettere meno errori, ma non lo esonera dall'interpretazione critica e dalla validazione finale.

Affinché l'IA sia un vero aiuto e non una mera sostituzione, è indispensabile che chi usa la tecnologia sia ''sufficientemente consapevole di ciò che sta usando''. Questo richiede che gli sviluppatori (spesso società *profit*) non progettino sistemi volti a massimizzare la dipendenza del clinico.

6.3. *Il modello della responsabilità condivisa (*shared responsibility*).*

A fronte della complessità intrinseca dei sistemi di IA e della difficoltà di attribuire la colpa in caso di errore algoritmico, la dottrina bioetica e giuridica sta convergendo verso l'adozione di un modello di responsabilità condivisa (*shared responsibility*).

Questo modello non frammenta la responsabilità tra i singoli attori (clinici, sviluppatori, istituzioni), ma la distribuisce in modo da garantire sia un risarcimento adeguato al paziente danneggiato, sia la prevenzione dei danni futuri, promuovendo una cultura di ''imparare anziché incolpare''. Un approccio di responsabilità condivisa richiede che le società che sviluppano l'IA clinica, accettino la loro parte di responsabilità morale e legale, collaborando con clinici e istituzioni fin dalla fase di progettazione. La creazione di piattaforme di lavoro comune per affrontare le cause prevedibili di danno derivanti dall'uso dell'IA è un elemento chiave per l'attuazione di questo modello.

7. Conclusione. L'impiego dell'IA, tra dignità umana e la libertà di scelta terapeutica.

L'Intelligenza Artificiale rappresenta una risorsa di potenziale immenso per il miglioramento del sistema sanitario, ma la sua implementazione deve essere gestita attraverso una governance che non sacrifichi i diritti fondamentali sull'altare dell'efficienza algoritmica.

La regolamentazione europea (AI Act ed EHDS) fornisce un quadro per la sicurezza del prodotto e la protezione dei dati, cruciali per mitigare i rischi tecnologici e di *privacy*. Tuttavia, la vera sfida risiede nella dimensione etico-legale dell'interazione uomo-macchina, dove il rischio di *automation bias* e di 'cattura della decisione' minaccia la sovranità professionale del medico e l'autodeterminazione del paziente.

La normativa nazionale, in particolare la legge 132 del 2025, tenta di ribadire il principio che l'IA debba essere solo un supporto non vincolante, preservando la decisione finale in capo al professionista. Il successo di questa linea dipenderà dalla capacità del sistema di imporre un design dei sistemi di IA che prediliga il supporto consapevole e l'analisi dei dati grezzi rispetto alla mera proposta di referti automatizzati (*opting out*).

In sintesi, la governance dell'IA in sanità non è solo una questione di conformità tecnica, ma deriva dal sistema costituzionale la più generale necessità di mantenere un'architettura decisionale che subordini l'efficienza algoritmica alla libertà di cura, all'equità nell'accesso e, in ultima analisi, al rispetto della persona umana, come richiesto dall'articolo 32 della Costituzione.

(LA NORMA)
L'IMPIEGO DI SISTEMI AUTONOMI IN MEDICINA E LE RESPONSABILITÀ DEL SANITARIO

di *Antonio Valitutti*

SOMMARIO: — 1. Struttura e funzione dell'illecito. — 2. La responsabilità del sanitario secondo la legge Gelli-Bianco. — 3. L'uso dell'intelligenza artificiale in medicina. — 4. Le nuove frontiere della responsabilità medica. — 5. Conclusione.

1. Struttura e funzione dell'illecito.

Il tema della responsabilità del sanitario in relazione all'impiego dell'intelligenza artificiale rende necessaria — come è evidente — una chiarificazione circa l'esatto significato da attribuirsi al concetto di *responsabilità*, e specificamente a quella del sanitario, nonché al concetto di *illecito*, che si colloca a monte della responsabilità, in quanto presupposto essenziale e imprescindibile della stessa.

L'*illecito* è un comportamento difforme da quello che si sarebbe atteso da un soggetto, alla stregua delle regole imposte dall'ordinamento di una determinata società. Quando si parla di conseguenze negative dell'agire umano (pena o risarcimento del danno), infatti, si parte logicamente da un presupposto: che il soggetto poteva comportarsi in modo diverso da come si è comportato e che la *responsabilità* di quanto ha fatto risale ad una sua decisione libera e consapevole. L'accertamento della responsabilità dipende, dunque, dal riscontro di alcune «*categorie logiche*» (NUVOLONE, 1975), dalle quali possa, appunto, desumersi che la scelta di tenere un comportamento difforme dalle regole dipende da una decisione volontaria dell'agente, sicchè il medesimo possa essere chiamato a rispondere del suo agire, sia sul piano civile che su quello penale.

La prima di tali categorie, com'è evidente, è costituita dalla *libertà dell'agire*, e pertanto dalla *capacità* dell'agente. Perciò, non risponde del fatto dannoso, sul piano civile, chi non aveva la capacità di intendere e di volere al momento in cui lo ha commesso (art. 2046

c.c.), sul piano penale, chi non era imputabile per incapacità di intendere e di volere (art. 85 c.p.), o il minore di anni 14 (art. 97 c.p.).

Ma la responsabilità implica, altresì, un'altra categoria logica che è quella della *causalità*: in tanto si può ritenere un soggetto responsabile per una condotta da lui posta in essere, in quanto, da un lato, il fatto sia determinato da una sua decisione volontaria (*causalità soggettiva o elemento psicologico dell'illecito*), dall'altro, in quanto — sul piano materiale ed oggettivo — un certo risultato sia l'effetto di una sua condotta positiva o negativa (*causalità oggettiva o effettuale*). La decisione di uccidere un uomo, o di procuragli delle lesioni, ad esempio, imputa l'azione posta in essere alla psiche del soggetto che ha elaborato siffatta determinazione che, nella specie, si atteggia come dolosa, ossia come coscienza e volontà di cagionare l'evento letale. Sul piano della realtà effettuale, peraltro, occorre anche che tale decisione del soggetto sia stata attuata ponendo in essere un'azione, o un'omissione, idonea a cagionare l'evento letale (esplosione di un colpo all'indirizzo della vittima, somministrazione di una sostanza venefica, o altro).

Nell'esempio fatto, il reato determina, altresì, la sussistenza di un illecito civile extracontrattuale, atteso che secondo l'ampia formula adoperata dall'art. 2043 c.c. (cd. *atipicità dell'illecito civile*), si è in presenza di un fatto doloso che ha cagionato « ad altri » un « danno ingiusto », che obbliga l'autore del fatto a risarcirlo. L'illecito è, invero, una delle fonti delle obbligazioni, ai sensi dell'art. 1173 c.c.

L'illecito civile, che può essere *contrattuale* (l'inadempimento di una obbligazione nascente da un contratto, ai sensi dell'art. 1218 c.c.) o *extracontrattuale* (art. 2043 c.c.), consiste nella violazione di una norma posta a tutela di un interesse privato alla quale consegue una sanzione risarcitoria, cioè finalizzata a reintegrare il danno subito dal soggetto portatore dell'interesse tutelato, irrogata dal giudice nell'ambito della giurisdizione civile. L'illecito penale, o reato, consiste, invece, nella violazione di una norma posta a tutela dell'interesse pubblico, in quanto attinente all'ordine etico-politico-sociale dello Stato, alla quale consegue una sanzione punitiva, la pena, finalizzata all'afflizione del trasgressore, irrogata dal giudice nell'ambito della giurisdizione penale.

Sul piano della *struttura*, entrambi gli illeciti sono connotati dagli elementi suindicati: libertà, causalità, elemento psicologico. Quest'ultimo integra la c.d. *colpevolezza*, ossia quell'atteggiamento antidoveroso della volontà del soggetto che solo giustifica il rimprovero giuri-

dico-penale, nelle forme del *dolo* (coscienza e volontà della lesione dell'interesse tutelato) o *colpa* (volontà della produzione di un rischio da cui, non voluta, deriva la lesione dell'interesse); anche il delitto colposo, infatti, quando esprime noncuranza del pericolo e potenziale spregio dei diritti altrui, all'unico scopo di realizzare il proprio egoistico interesse, si concreta, infatti, in un'aggressione criminologicamente e civilmente rilevante.

Nell'esempio fatto, un delitto di omicidio o di lesioni personali (ed il corrispondente illecito civile) può essere posto in essere anche per colpa, ossia per negligenza, imprudenza, imperizia, o violazione di norme giuridiche cautelari (art. 43 c.p.). Tuttavia, il soggetto agente non vuole l'evento lesivo (« delitto contro l'intenzione »), sebbene possa anche averlo previsto. Egli agisce producendo volontariamente un rischio, ma nella sicura convinzione che nessun pregiudizio potrà verificarsi in conseguenza della condotta colposa tenuta, giacchè se accettazione del rischio vi fosse, si verserebbe in una ipotesi dolosa.

Sul piano della *funzione*, l'illecito penale si pone come un atto che lede gli interessi della collettività e, per questo, una volta accertato va sanzionato con la pena. Più complessa la ricostruzione dell'illecito civile sul piano funzionale. Esso può relazionarsi, per vero, a comportamenti posti in essere in tipi di relazioni differenti. Può accadere che un soggetto abbia assunto nei confronti di un altro un obbligo specifico, qualunque ne sia la fonte (contratto o altro fatto idoneo a produrlo), come può accadere che una responsabilità insorga per la violazione del dovere generico di non ledere l'altrui sfera giuridica (*neminem laedere*). In campo civilistico, con la responsabilità contrattuale si crea un vincolo obbligatorio per il risarcimento del danno che discende dalla violazione di un obbligo preesistente, laddove la responsabilità cd. aquiliana (dalla *lex Aquilia* romana) *ex* art. 2043 c.c. sanziona i comportamenti illeciti che incidono nella sfera di soggetti cui l'agente non è legato da alcun rapporto specifico, e può dare luogo — come visto — anche ad un illecito penale, proprio perché si concreta nella violazione di norme che tutelano determinati beni sul piano generale.

Le differenze principali nel regime giuridico tra le due forme di responsabilità sono: *a*) la diversità della regola in tema di onere probatorio, ossia l'applicazione dell'art. 1218 c.c., che prevede una presunzione di colpa, per la responsabilità contrattuale, per contro l'onere di provare la colpa (oltre agli altri elementi dell'illecito) in caso

di illecito extracontrattuale; *b*) il fatto che nell'illecito aquiliano la costituzione in mora non occorre (cd. *mora ex re*); *c*) la quantificazione del danno risarcibile, atteso che nell'illecito contrattuale si opera applicando l'art. 1225 c.c., per cui il debitore risponde anche dei danni non prevedibili in caso di dolo, ma tale norma non è richiamata dall'art. 2056 c.c. per l'illecito extracontrattuale; *d*) il diverso regime della prescrizione, decennale per l'illecito contrattuale, quinquennale per l'aquiliano; *e*) il fatto che nella responsabilità contrattuale l'inadempimento non è sanzionato soltanto attraverso la responsabilità risarcitoria, ma è assistito altresì da rimedi specifici, quali l'eccezione di inadempimento, la risoluzione del contratto, e la sospensione unilaterale della prestazione. I confini tra responsabilità contrattuale ed aquiliana, però, sono sempre più mobili, essendosi, per un verso, l'illecito *ex* art. 2043 c.c. esteso anche ad interessi protetti dal contratto, per altro verso, essendosi spesso verificata la trasmigrazione nell'ambito della responsabilità contrattuale di fattispecie tradizionalmente ricondotte all'illecito aquiliano.

Una fattispecie di tal fatta, che ha messo in crisi i criteri di riparto tra le due forme di responsabilità, è quella della «*responsabilità da contatto sociale qualificato*» (rapporto obbligatorio di fatto). L'istituto del «contatto sociale qualificato» si fece strada tra gli autori di lingua tedesca negli anni Trenta del Ventesimo secolo, con riferimento a quei rapporti, denominati *Selbstbindung ohne Vertrag* (autoregolamento impegnativo senza contratto), nei quali, se è una finzione parlare di negozio formalmente stipulato dalle parti, è altrettanto frutto di un'evidente mistificazione della realtà ignorare quei fatti dai quali scaturiscono obblighi tra due o più soggetti, trattandosi di «contatti voluti ed orientati ad uno scopo» (si vedano, *ex multis*, BENATTI, 1963; CASTRONOVO, 2006; GRISI, 1990). Trattasi di una forma di responsabilità che prescinde dalla sussistenza di un precedente vincolo contrattuale (per cui viene definita «di tipo contrattuale»), reputandosi sufficiente dall'ordinamento giuridico un «contatto sociale qualificato», ossia connotato da uno scopo comune tra le parti e da cui discendono specifici doveri di protezione di determinati beni giuridici; il che pone il problema di individuare in quali situazioni una relazione sociale possa ritenersi «qualificata» dall'ordinamento giuridico e, pertanto, idonea a fondare una responsabilità di tipo contrattuale. Gli esempi ricorrenti sono quelli: *a*) delle trattative contrattuali (artt. 1337 e 1338 c.c.), nelle quali vi è un contatto tra le parti per lo scopo comune di addivenire alla stipula del contratto; *b*) dell'attività sanitaria (che qui più inte-

ressa), quando il medico « prende in carico » il paziente in reparto, in tal modo venendo a verificarsi un contatto tra sfere giuridiche che obbliga il sanitario — anche in mancanza di un contratto di prestazione d'opera professionale (art. 2230 c.c.) — a proteggere il bene salute del paziente, nel comune scopo di conseguirne la guarigione; *c*) il rapporto tra insegnante e allievo nella scuola, con l'obiettivo comune della crescita educativa e culturale di quest'ultimo; *d*) il rapporto tra privati e pubblica amministrazione nell'ambito di un procedimento amministrativo.

Si parla, al riguardo, di « obbligazioni senza prestazione » (obblighi di informazione, di protezione che, a differenza dagli altri obblighi, non si traducono in un dare o fare una cosa determinata, la prestazione, *ex* art. 1174 c.c.), i cui referenti normativi sono costituti dagli artt. 2 Cost. e 1375 c.c., che obbliga le parti, oltre a quanto da esse convenuto nel contratto, anche ad un comportamento di buona fede.

Orbene, mentre l'illecito aquiliano, che presuppone la mancanza di un rapporto tra le parti, obbliga a risarcire il danno prodotto « ad altri », ossia a soggetti non vincolati da un preesistente obbligo all'agente, mediante una condotta dolosa o colposa (art. 2043 cc.), il contatto sociale — inquadrabile tra gli altri *fatti* idonei a produrre obbligazioni (nella specie di protezione e di informazione), ai sensi dell'art. 1173 n. 3 c.c. — determina una responsabilità da « affidamento » molto più prossima a quella da contratto che a quella da torto, poiché la responsabilità si radica in un rapporto tra soggetti specificamente individuati. Si è in presenza, dunque, di una diversità di situazioni di fatto (*ex facto oritur ius*, affermavano i romani), che plasma in modo diverso le tutele. Il che investe in particolare il ruolo del giudice, chiamato a dare l'esatta qualificazione giuridica ai fatti che gli vengono sottoposti. E d'altro canto, mentre la legge — in quanto generale ed astratta — non realizza e non disciplina nel concreto l'esperienza nel suo farsi, il giudice è soggetto alla pressione del contesto fattuale da regolare ed ordina l'esperienza secondo ragione. Per questo la motivazione della sentenza viene definita l'« autorità della ragione », laddove la decisione, nel suo aspetto precettivo (il dispositivo della sentenza) costituisce la « ragione dell'autorità » (Taruffo, 2002).

2. La responsabilità del sanitario secondo la legge Gelli-Bianco.

Premesso quanto precede in via di principio, va osservato — con specifico riferimento alla responsabilità del sanitario — che la legge 8

marzo 2017, n. 24 (cd. legge Gelli-Bianco), all'art. 7 prevede — per quel che rileva in ordine alla responsabilità civile del medico — che: « [...] 3. L'esercente la professione sanitaria di cui ai commi 1 e 2 risponde del proprio operato ai sensi dell'articolo 2043 del codice civile, salvo che abbia agito nell'adempimento di obbligazione contrattuale assunta con il paziente » (CARPANI-FARES, 2023; CASSANO, 2024). In forza della medesima norma, la struttura sanitaria risponde a titolo di responsabilità contrattuale.

Orbene, secondo alcuni autori, deve ritenersi che la locuzione « obbligazione contrattuale assunta con il paziente » faccia riferimento alla sola assunzione mediante contratto di prestazione d'opera professionale, *ex* art. 2230 c.c. (VIOLA, 2019). In realtà il problema si incentra sul significato da attribuire alla dizione « salvo che abbia agito nell'adempimento di obbligazione contrattuale assunta con il paziente ».

Certamente, atteso che la regola è che la responsabilità del sanitario è di tipo aquiliano, e la previsione ha valore di norma imperativa, la responsabilità contrattuale, anche da contatto sociale, non potrà essere affermata nel caso in cui il paziente non provi il contratto (art. 2229 cod. civ.) con il medico, ovvero — contrariamente al non condivisibile assunto di una parte della giurisprudenza, secondo cui anche il « contatto sociale » meramente fortuito ed informale, intercorso tra medico e paziente, sarebbe idoneo a far scattare i presidi della responsabilità contrattuale (Cass., 3/10/2016, n. 19670) — quando il contatto tra un determinato sanitario ed il paziente sia del tutto estemporaneo ed occasionale, perché — ad esempio — il sanitario di pronto soccorso si sia limitato semplicemente ad avviare il paziente verso uno dei reparti per il ricovero. Ma certamente più problematiche sono le ipotesi in cui sia il medesimo sanitario a « prendere in carico » il paziente ed a seguirlo per tutto il tempo del ricovero nella struttura (anamnesi, diagnosi, terapia o intervento, dimissioni), o in cui il sanitario, sebbene non addetto a quello specifico reparto, abbia comunque svolto una specifica attività professionale diagnostica o terapeutica sul paziente (Cass., 4/9/2023, n. 25772).

Si tratta si ipotesi nella quale sembra difficile negare che vi sia stato un contatto sociale qualificato che mira ad un risultato terapeutico comune. In tali fattispecie risulta problematico tenere fuori la terza fonte dell'obbligazione, diversa dall'illecito e dal contratto, costituita da qualsia altro fatto idoneo a produrla, come il « contatto sociale » (art. 1173 cod. civ. — nel senso che la legge Gelli-Bianco non abbia eleminato la possibilità di configurare una responsabilità del sa-

nitario a titolo di «contatto sociale», cfr. Trib. Cassino, 13/8/2025, n. 1093; Trib. Mantova, 5/9/2019, n. 642). Tanto più che sul medico incombono anche altri obblighi che non hanno la loro fonte nel contratto, come nel caso del consenso informato, che trova il suo fondamento direttamente nelle norme costituzionali (artt. 2, 13 e 32 Cost). Il cd. *consenso informato* dà luogo — per vero — in caso di omissione di informazione, a responsabilità contrattuale del sanitario, a fronte dell'allegazione del relativo inadempimento da parte del paziente. Il medico è pertanto, tenuto a provare di avere adempiuto all'obbligazione, ossia di avergli fornito un'informazione completa ed effettiva sul trattamento sanitario e sulle sue conseguenze (Cass., 16/9/2022, n. 27279).

È la diversità delle situazioni fattuali — come dinanzi detto — a conformare le tutele. Se è vero, infatti, che l'interprete non può certamente sottrarre al legislatore il potere di operare una qualificazione normativa della fattispecie astratta; vero è pure che il legislatore non può sottrarre all'interprete il potere di operare l'interpretazione e la qualificazione della fattispecie concreta, nel senso già detto, lumeggiate dalla natura dell'attività medica, dalla caratura e ambivalenza del bene tutelato e del conseguente bilanciamento di interessi (per un verso la difesa del paziente, per l'altro la tutela del medico) (MIGLIACCIO, 2019).

Nella giurisprudenza civile più recente, si è affermato, poi, che l'accertamento del *nesso di causalità* oggettiva nella responsabilità civile sanitaria è improntato alla regola di funzione della «*preponderanza dell'evidenza*» (o del «*più probabile che non*»), la quale, con riguardo al caso in cui, rispetto a uno stesso evento, si pongano diverse ipotesi alternative, comporta che il giudice dapprima elimini, dal novero delle ipotesi valutabili, quelle meno probabili e poi analizzi le rimanenti ipotesi ritenute più probabili, selezionando, infine, quella che abbia ricevuto, secondo un ragionamento di tipo inferenziale, il maggior grado di conferma dalle circostanze di fatto acquisite al processo, tenendo conto della qualità, quantità, attendibilità e coerenza delle prove disponibili, dalla cui valutazione complessiva trarre il giudizio probabilistico (Cass., 5/3/2024, n. 5922.; Cass., 26/9/2024, n. 25805). La responsabilità penale è informata, invece, al principio della «*prova oltre ogni ragionevole dubbio*», circa la responsabilità dell'imputato — nella specie il sanitario — per il reato ascrittogli.

In tema di colpa medica, la limitazione — sul piano civilistico — della responsabilità professionale del medico ai soli casi di dolo o colpa grave, di cui all'art. 2236 c.c., non opera nelle ipotesi di *impru-*

denza né, al riguardo, rileva l'astratta conformità della tecnica adottata dalle linee guida. In tema di responsabilità sanitaria per attività medico-chirurgica, le linee guida non hanno, invero, rilevanza normativa o « parascriminante », non essendo né tassative, né vincolanti; conseguentemente, pur rappresentando un parametro utile nell'accertamento dei profili di colpa medica, esse non valgono ad eliminare la discrezionalità del giudice di valutare se le circostanze del caso concreto esigano una condotta diversa da quella prescritta nelle medesime linee guida (Cass., 11/12/2023, n. 34516).

La responsabilità penale del medico, secondo la Legge Gelli-Bianco, è limitata, per contro, ai casi di colpa grave, esclusa in caso di rispetto delle linee guida e delle buone pratiche clinico-assistenziali, che diventano il parametro di riferimento per la valutazione del comportamento. Il professionista sanitario non è punibile penalmente se l'evento dannoso (omicidio o lesioni personali colpose) si verifica pur avendo seguito correttamente le linee guida e le best practice adeguate al caso concreto. La punibilità sussiste solo in caso di negligenza o imprudenza grave, o quando non siano state rispettate tali linee guida (art. 590-sexies c.p). L'art. 590-sexies cod. pen., introdotto dall'art. 6 della legge 8 marzo 2017, n. 24, prevede — si è osservato in giurisprudenza — una causa di non punibilità applicabile ai soli fatti inquadrabili nel paradigma dell'art. 589 o di quello dell'art. 590 c.p., e operante nei soli casi in cui l'esercente la professione sanitaria abbia individuato e adottato linee guida adeguate al caso concreto e versi in colpa lieve da imperizia nella fase attuativa delle raccomandazioni previste dalle stesse. La suddetta causa di non punibilità non è applicabile, invece, né ai casi di colpa da imprudenza e da negligenza, né quando l'atto sanitario non sia per nulla governato da linee-guida o da buone pratiche, né quando queste siano individuate e dunque selezionate dall'esercente la professione sanitaria in maniera inadeguata con riferimento allo specifico caso, né, infine, in caso di colpa grave da imperizia nella fase attuativa delle raccomandazioni previste dalle stesse. Le raccomandazioni contenute nelle linee guida definite e pubblicate ai sensi dell'art. 5 della legge 8 marzo 2017, n. 24 — si è, dipoi, ulteriormente precisato — pur rappresentando i parametri precostituiti a cui il giudice deve tendenzialmente attenersi nel valutare l'osservanza degli obblighi di diligenza, prudenza, perizia — non integrano veri e propri precetti cautelari vincolanti, capaci di integrare, in caso di violazione rimproverabile, ipotesi di colpa specifica, data la necessaria elasticità del loro adatta-

mento al caso concreto; ne consegue che, nel caso in cui tali raccomandazioni non siano adeguate rispetto all'obiettivo della migliore cura per lo specifico caso del paziente, l'esercente la professione sanitaria ha il dovere di discostarsene (Cass. S.U., 21/12/2018, n. 8770). In tema responsabilità medica, le linee guida definite e pubblicate ai sensi dell'art. 5 legge 8 marzo 2017, n. 24, sono, dunque, raccomandazioni di ordine generale, che contengono « regole » cautelari di massima, flessibili e adattabili, prive di carattere precettivo, rispetto alle quali è fatta salva la libertà di scelta professionale del sanitario nel rapportarsi alla specificità del caso concreto, nelle sue molteplici varianti e peculiarità e nel rispetto della « relazione terapeutica » con il paziente (Cass., 3/2/2022, n. 7849).

3. L'uso dell'intelligenza artificiale in medicina.

Il già vasto e complesso panorama relativo all'attività medica, si è andato ulteriormente arricchendo — o, in una diversa prospettiva, complicando — per effetto dell'uso ampio e diffuso della *intelligenza artificiale* AI) nella medicina (COLECCHIA, 2020; TOPPETTI, 2025). Dalla chirurgia robotica alle microcapsule robotiche a scopo terapeutico, dalle radiazioni robotiche ionizzanti ai trattamenti farmacologici post-operatori. Ma — come ogni strumento diretto ad ottimizzare il rendimento di un'attività umana — l'introduzione di AI non è immune da problemi, tali da aprire nuove frontiere alla configurabilità dell'attività sanitaria, a partire dalla regolamentazione dell'uso dei sistemi in questione. Nell'ambito dell'Unione Europea sono state — in proposito — dapprima poste alcune direttive, con due risoluzioni del Parlamento Europeo (16 febbraio 2017 e 20 ottobre 2020), che ha avviato un processo per l'armonizzazione delle norme in materia di intelligenza artificiale nel settore della responsabilità civile. con riferimento al settore sanitario, precisando che la scelta inziale di cura, e quella finale sull'esecuzione, dovranno spettare comunque ad un « *medico umano* » (la tecnologia robotica dovrà integrare la capacità umana e non sostituirsi ad essa) e sottolineando la necessità di predisporre un quadro giuridico basato su principi comuni per garantire la certezza giuridica in materia e fissare norme uniformi in tutta l'Unione. Si è arrivati così alla proposta di « Direttiva del Parlamento Europeo e del Consiglio del 28.09.2022 », che concerne, però, la responsabilità del produttore. Successivamente, è stato emanato il Regolamento (UE) 2024/1689 del Parlamento Europeo e del Consiglio

del 13 giugno 2024, noto anche come « AI Act », che stabilisce regole armonizzate per l'intelligenza artificiale nell'Unione Europea. Secondo l'Allegato III dell'AI Act, sono classificati come « sistemi ad alto rischio » quelli impiegati in contesti sanitari, tra cui i « sistemi di supporto decisionale clinico ».

Tanto premesso, va rilevato che la crescente diffusione delle intelligenze artificiali in ambito medico ha indubbiamente migliorato il livello di efficienza nell'erogazione della prestazione sanitaria, riducendo i rischi di errore ad opera del personale umano e, in parallelo, aumentando il grado di sicurezza delle cure. Il paziente trae infatti beneficio dall'impiego di queste tecnologie anche in termini di minor grado di invasività delle operazioni, riduzione delle sofferenze fisiche e snellimento dei tempi di riabilitazione. Agli indubitabili vantaggi sopra elencati, si contrappongono, tuttavia, dubbi applicativi e profili di rischio non indifferenti. Il medico, un tempo esclusivo custode delle conoscenze, a fronte della progressiva automazione dell'attività sanitaria, è relegato al termine del processo di cura, con un ruolo di validazione e controllo dell'operato della macchina solo a posteriori. Si assiste così ad una nuova rimodulazione del rapporto medico-paziente: dopo il passaggio dal modello paternalistico originario a quello « *patient-centric* », la relazione si complica nuovamente a causa della presenza di un « terzo » ente, l'intelligenza artificiale. Sulla base di un criterio di autonomia crescente, occorre infatti distinguere tra « *robot tele-operati* », « *robot autonomi* » e « *robot cognitivi* »: mentre i primi sono strumenti nelle mani dell'operatore che ne detiene totalmente il controllo, i secondi sono in grado di svolgere il compito affidato in autonomia durante la fase di attuazione e di esecuzione, i terzi sono invece dotati di maggiore autonomia, anche in virtù della maggiore capacità di adattamento agli stimoli e alla mutevolezza dell'ambiente esterno (SANTOSUOSSO, 2020).

4. Le nuove frontiere della responsabilità medica.

Orbene, sembra evidente che il primo tipo di AI, *robot tele-operati*, può dare luogo, sul piano civilistico, ad una responsabilità del medico operatore ai sensi dell'art. 2043 c.c., secondo il modello della legge Gelli-Bianco. Per cui il sanitario risponderà a titolo di colpa generica, qualora produca l'evento per negligenza, imprudenza o imperizia nell'uso dell'AI, o per inosservanza delle regole che ne disciplinano l'uso. Ed in siffatte ipotesi sarà ravvisabile anche una responsa-

bilità penale del sanitario per le lesioni personali colpose o per l'omicidio colposo, in ipotesi cagionati in danno del paziente, nel caso di uso negligente, imprudente o imperito, o in violazione delle nuove normative sulla AI.

Più complesso l'accertamento della responsabilità del sanitario, nel caso dei *robot autonomi* e dei *robot cognitivi*. Un software autonomo di AI suggerisce, ad esempio, un certo rischio di neoplasia, sulla base del quale un medico decide di intervenire. Si scopre poi che quella previsione fatta dalla macchina era errata. Oppure il medico decide di ignorare le raccomandazioni fornite dalla macchina, perché dà una diversa interpretazione del caso clinico alla luce della propria esperienza, e questo produce conseguenze negative per il paziente. Ci si chiede se il medico dovrà rispondere sul piano civile, ai sensi dell'art. 2043 c.c., e su quello penale, ai sensi degli artt. 589 c.p. (omicidio colposo) e 590 c.p. (lesioni colpose).

Una parte della dottrina ipotizza che l'azione incontrollabile dei dispositivi appartenenti alla seconda e (soprattutto) alla terza categoria, causativa di un danno al paziente, rappresenterebbe un fattore sopravvenuto interruttivo del nesso di causalità (art. 41 c.p.), quale circostanza che introduce un rischio esorbitante, non prevedibile a priori. Il che escluderebbe sia la responsabilità penale del sanitario che quella civile, atteso che la causalità in tale ultimo settore dell'ordinamento non si atteggia diversamente.

Questa ricostruzione non appare convincente ad altra parte della dottrina, dal momento che i soggetti coinvolti nell'erogazione della prestazione sanitaria (medico e azienda ospedaliera) lungi dal potersi adagiare acriticamente sui risultati della macchina, sono chiamati, in quanto professionisti esperti, ad un attento e costante « monitoraggio dei risultati », allo specifico fine di prevedere ed impedire possibili errori. Tanto più che il Regolamento europeo succitato ha classificato come « sistemi ad alto rischio » quelli impiegati in contesti sanitari, tra cui i « sistemi di supporto decisionale clinico ».

In questo modo, sarebbe tutelata adeguatamente anche la posizione del paziente danneggiato, il quale potrebbe agire per il risarcimento del danno verso l'ente ospedaliero ed eventualmente verso il medico operante. Va escluso l'inquadramento dell'attività del medico nell'ambito della *responsabilità presunta ex art. 2050 c.c.* (attività pericolosa per la natura dei mezzi adoperati), non sembrando agevole considerare un meccanismo finalizzato a ridurre l'area di errore come un mezzo pericoloso per la salute del paziente, come potrebbe essere —

ad esempio — una trasfusione con sangue infetto. Ma, più in generale, l'attività medica, quale attività professionale, non si presta all'inquadramento nella fattispecie di cui all'art. 2050 c.c. L'art. 2050 c.c. concerne genericamente le attività pericolose e non si applica a quelle per le quali il legislatore ha specificamente provveduto, sicchè l'attività che formi oggetto della prestazione dovuta dal professionista (nella specie: medico chirurgo) al proprio cliente ricade esclusivamente nell'ambito dell'art. 2236 c.c., cui e estranea ogni presunzione di colpa (Cass., 10/7/1979, n. 3978). Sembra da escludere anche la riconducibilità del caso in questione nella fattispecie di *responsabilità oggettiva* (sulla base del solo rapporto di causalità) *di cui all'art. 2051 c.c.* (responsabilità per cose in custodia). Difetta, invero, nella specie, il criterio di imputazione specifico di tale forma di responsabilità del custode, che sussiste solo allorquando il danno scaturisca quale effetto dell'intrinseco dinamismo della cosa, oppure laddove il pregiudizio consegua a un'azione umana che determini l'insorgenza di un processo dannoso nella cosa medesima, senza che il custode abbia provato il caso fortuito, ossia il fatto dello stesso danneggiato o di terzi (Cass., 12/7/2022, n. 21977). Nel caso di specie, il danno subito — in ipotesi — dal paziente non è ascrivibile ad una omissione di intervento sulla cosa da parte del custode, onde evitare l'attivarsi di un dinamismo intrinseco alla stessa, o — viceversa — ad un positivo intervento sulla cosa (l'AI) che faccia scaturire tale dinamismo (es. compiere dei lavori su di un lastrico solare, dal quale derivino infiltrazioni d'acqua ai piani sottostanti). Nella specie, l'eventuale responsabilità dell'operatore — in caso di sistema autonomo, si radica, invero, a valle, quando l'attività dell'AI si è espletata (ad es. il compimento di un'operazione chirurgica).

Né va sottaciuto — in relazione ad entrambe le forme di responsabilità *ex* artt. 2050 e 2051 c.c. — che esse non sono in alcun modo menzionate dall'art. 7 della legge Gelli-Bianco, dovendosene privilegiare una interpretazione restrittiva, attesa la sua finalità di equilibrare il diritto del paziente al risarcimento con una maggiore tutela dell'operatore sanitario.

Potrebbe essere percorribile la soluzione di inquadrare la responsabilità del sanitario, sul piano civilistico, anche nell'ipotesi di utilizzo dei sistemi autonomi o cognitivi, nell'ambito della *responsabilità extracontrattuale* (salvi i casi di responsabilità contrattuale), responsabilità che — lungi dal poter essere esclusa automaticamente, ai sensi dell'art. 41 c.p., per effetto dell'azione, in ipotesi, incontrollabile dei dispositivi in parola, causativa di un danno al paziente — verrebbe a

radicarsi nell'ipotesi di colpevole (per negligenza o imperizia) monitoraggio dei risultati, che compete in ogni caso all'operatore, allo specifico fine di prevedere ed impedire possibili errori compiuti dalla AI, nella diagnosi o terapia.

Ne discende che, nel primo degli esempi succitati, qualora il software abbia ipotizzato una neoplasia, il medico che ometta di monitorare e correggere eventuali errori diagnostici di AI (software ad alto rischio), rilevabili sulla base di altri accertamenti diagnostici e dell'esame obiettivo del paziente, non potrà andare esente da responsabilità, sia *ex* art. 2043 c.c., sussistendo il nesso causale dell'evento lesivo con la condotta omissiva del sanitario e la colpa per negligenza o imperizia, sia per i delitti di omicidio o di lesioni colpose. Del pari, nel secondo esempio, laddove il discostarsi del sanitario dal responso di AI sia ascrivibile ad un errore colposo del medesimo, dovuto a negligenza o imperizia, non potrà esserne esclusa la responsabilità per colpa, sia sul piano civile che su quello penale.

Dunque, la professionalità del medico — che va certo, garantito nella libera esplicazione della sua attività — non può e non deve subire alcuna mortificazione o regressione, per effetto dell'avvento «salvifico» della AI, al di là dei miglioramenti che l'attività sanitaria potrà oggettivamente conseguire. La soluzione che — al di là delle sporadiche evenienze di impossibilità di correggere il risultato dell'attività della AI — ravvisa la responsabilità del medico nel solo caso di comprovata (dal paziente) colpa del sanitario nella interpretazione e valutazione dei risultati, sembra anche in linea con la sollecitazione europea a non sostituire il «medico uomo» con la macchina. L'invito cioè a conservare un modello di operatore, conosciuto dalle tradizioni degli Stati europei, che va tutelato nella libertà del suo agire — anche al fine di evitare la cd. medicina difensiva — ma che, al contempo, deve mantenere un livello elevato di professionalità e dignità nell'esercizio di una attività di rilievo pubblico essenziale ed assoluto (art. 32 Cost.). Resta, ovviamente fermo il diritto del paziente, ma anche del medico e della struttura sanitaria in via di regresso (art. 2055 c.c.), di agire nei confronti del produttore della AI per eventuale difettoso funzionamento del dispositivo.

5. Conclusione.

L'umanizzazione del trattamento sanitario, il dovere di chiarezza nel dialogo medico-paziente, il rispetto della dignità della persona

(anche del medico che va protetto da aggressioni intollerabili), ed al contempo dei risultati acquisiti dalla comunità scientifica, non sono, dunque, acquisizioni relegate al tempo delle grandi narrazioni, superate da una post-modernità fondata sui nuovi strumenti tecnologici. Invero, al di là delle acquisizione scientifiche e dell'ossequio alle stesse, in special modo alle linee guida, del sistema assicurativo, della necessità di potenziamento del servizio sanitario sul piano organizzativo, tutte mete da perseguire con determinazione, ed oggi dell'apporto rassicurante della AI — che, peraltro, è, e deve restare, uno strumento a servizio dell'uomo — è innegabile che un ruolo essenziale gioca — in quella particolare ecologia del rapporto terapeutico (Zatti, 2012), che è la relazione tra medico e ammalato, e nel necessario equilibrio tra i diritti/doveri dei due protagonisti — la quotidiana riscoperta, da parte del medico, della propria vocazione ad assicurare il bene del paziente, rinnovando ogni giorno il proprio giuramento di Ippocrate, con coraggio ed abnegazione. Questo impegno non è surrogabile da nessun sistema robotico, essendo radicato nel dovere deontologico del « medico umano », quello che gli ammalati cercano con ansia, perché li curi e li rassicuri. Torna alla mente l'immagine intramontabile del dr. Andrew Manson nel romanzo « La Cittadella » di Cronin. Dinanzi al corpo inerte e martoriato della moglie, travolta da un veicolo pirata, Manson — levando lo sguardo all'orizzonte, nel quale le nuvole avevano disegnato una « cittadella nel cielo » — giurerà di dedicare tutta la sua esistenza al bene dei pazienti, senza concessioni di sorta a stanchezze, amarezze e delusioni, e senza cedere alle lusinghe del disimpegno e del facile guadagno, fino a quando la sua anima si sarebbe elevata, a raggiungere l'amata, in quella regione di perenne certezza nella quale non esistono il dolore e la malattia, e nella quale soltanto abita la vera Giustizia.

Note bibliografiche

Alpa G. (1979), *Il problema della atipicità dell'illecito*, Napoli, Jovene.
Benatti F. (1963), *La responsabilità precontrattuale*, Milano, Giuffrè.
Carpani G.-Fares G.M. (2023), *Guida alle norme sulla responsabilità nelle professioni sanitarie. Un primo bilancio dopo cinque anni di legge Gelli-Bianco*, Torino, Giappichelli.
Cassano G. (2024), *Manuale pratico operativo della responsabilità medica*, Santarcangelo di Romagna (RN), Maggioli Editore.
Castronovo C. (2006), *L'obbligazione senza prestazione. Tra contratto e torto*, Milano, Giuffrè.

COLECCHIA G. (2020), *Intelligenza artificiale e medicina digitale. Una guida critica*, Roma, Il pensiero Scientifico Editore.
FAVA P. (2018), *La responsabilità civile*, Milano, Giuffrè.
FRANZONI M. (2010), *L'illecito*, in FRANZONI M., *Trattato della responsabilità civile*, Milano, Giuffrè.
GALGANO F. (2008), *I fatti illeciti*, Padova, Cedam.
GIANNITI P. (1998), *Responsabilità civile e penale a confronto*, Padova, Cedam.
GILMORE G. (1974), *The Death of contract*, Columbus, Ohio State University Press.
GRISI G. (1990), *L'obbligo precontrattuale di informazione*, Napoli, Jovene.
HAUPT G. (1941), *Überfaktische Vertragsverhaltnisse;* Leipzig.
HAZAN M.-MARTINI F.-RODOLFI M. (2024), *Il decreto attuativo della legge Gelli-Bianco. Commento al d.m. 15 dicembre 2023, n. 232*, Milano, Giuffrè.
LARENZ K. (1987), *Lehrbuch des Schuldrechts*, Munchen, Verhag CH Beck.
LUCCHINI GUASTALLA E., *Il contrato e il fatto illecito*, Milano, Giuffrè.
MAZZON R. (2017), *Le cause di giustificazione nella responsabilità per illecito*, Milano, Giuffrè.
MENGONI L. (1956), *Sulla natura della responsabilità precontrattuale*, in *Rivista di diritto commentato*, II, p. 360 ss.
MESSINEO F. (1971), *Il contratto in genere*, in CICU A.- MESSINEO F., *Trattato di diritto civile e commerciale*, vol. I, Milano, Giuffrè.
MIGLIACCIO E. (2019), *Ancora in tema di responsabilità medica e onere della prova*, in *Diritto di famiglia e della persona*, p. 1239 ss..
MONATERI P.G. (1998), *Responsabilità civile*, Torino, Utet.
NAPOLI G.E. (2012), *Il nesso causale come elemento costitutivo del fatto illecito*, Napoli, Edizioni Scientifiche Italiane.
NUVOLONE P. (1975), *Il sistema del diritto penale*, Padova, Cedam.
RESCIGNO F. (2012), *Obblighi negoziali al di là del contratto?*, in *Rivista di diritto commerciale*, 1, p. 73 ss.
RICCIARDI A. (2025), *Dall'illecito efficiente ai punitive damages: il risarcimento del danno da illecito antitrust*, Milano, Giuffrè.
ROCCATAGLIATA L.-STAMPANONI BASSI G.B. (2024), *Corporate governance e illecito penale. Poteri e responsabilità nelle organizzazioni complesse*, Milano, Giuffrè.
RODOTÀ S. (1964), *Il problema della responsabilità civile*, Milano, Giuffrè.
ROIATI A. (2017), *La colpa medica dopo la legge "Gelli-Bianco": contraddizioni irrisolte, nuove prospettive ed eterni ritorni*, in *Archivio penale*, 2.
ROVELLI L. (2000), *La responsabilità precontrattuale*, in BESSONE M., *Trattato di diritto privato*, vol. II, Torino, Utet.
RUFFOLO U.-SAVINI NICCI M. (2022), *Le nuove frontiere della responsabilità medica*, Milano, Giuffrè.
SALVI C. (2025), *La responsabilità civile*, in IUDICA G.-ZATTI P., *Trattato di diritto privato*, Milano, Giuffrè.
SANTOSUOSSO A. (2020), *Intelligenza artificiale e diritto*, Milano, Mondadori Education.

SCOGNAMIGLIO C. (2016), *Responsabilità contrattuale e "contratto sociale qualificato"*, in *Responsabilità civile e previdenza*, Milano, Giuffrè, 6, p. 1492 ss.
SCOGNAMIGLIO R. (1970), *Commentario al Codice civile. Contratti in generale (artt. 1321-1352 del Cod. Civ.)*, Bologna, Zanichelli.
TARUFFO M. (2002), *Sui confini, scritti sulla giustizia civile*, Bologna, Il Mulino.
TOPPETTI F. (2025), *Intelligenza artificiale e responsabilità dei professionisti. Etica, diligenza e algoritmi*, Santarcangelo di Romagna (RN), Maggioli Editore.
TRIMARCHI P. (2021), *La responsabilità civile: atti illeciti, rischio e danno*, Milano, Giuffrè.
VARANESE G. (2012), *Sui rapporti contrattuali di fatto*, Torino, Giappichelli.
VIOLA L. (2019), *Legge Gelli-Bianco (24/2017): no alla sua retroattività, fatta eccezione per i richiami agli artt. 138139 cod. ass. Non c'è più il contatto sociale*, in *La nuova procedura civile*, 5.
VISINTINI G. (2025), *Fatti illeciti. Fondamenti e nuovi sviluppi della responsabilità civile*, Pisa, Pacini Giuridica.
ZATTI P. (2012), *«Parole tra noi così diverse». Per una ecologia del rapporto terapeutico*, in *Nuova giurisprudenza civile commentata*, II, p. 143 ss.

(LA NORMA)
ROBOT INTELLIGENTI E ATTIVITÀ MEDICA: PRESUPPOSTI E (SOPRATTUTTO) LIMITI DELLA RESPONSABILITÀ PENALE DELL'OPERATORE SANITARIO

di *Antonella Massaro*

SOMMARIO: 1. La struttura antropocentrica (e antropomorfa) del diritto penale. — 2. I possibili modelli di interazione tra attività medica e intelligenza artificiale. — 3. Coordinate ricostruttive in materia di colpa penale: cenni. — 4. Colpa penale e robot intelligenti. — 4.1. Il principio di dominabilità ai tempi dell'intelligenza artificiale.

1. La struttura antropocentrica (e antropomorfa) del diritto penale.

Marcello Gallo, nel 1956, osservava come « assai poche, e forse nessuna, delle proposizioni più di frequente ripetute dai penalisti trovano oggi così generale consenso quanto quella che afferma che ''la persona umana è al centro del diritto penale'' » (GALLO, 1956).

Questa affermazione è riferibile non solo all'esigenza di un generale e generico umanesimo penale, con la conseguente centralità assunta dalla persona dopo il protagonismo dello Stato che aveva caratterizzato l'esperienza autoritaria. La persona come centro (e baricentro) del diritto penale rinvia alla constatazione, tanto ovvia quanto poco scontata nel dibattito in materia di nuove tecnologie, per cui il diritto penale è strutturalmente antropocentrico o, per dir così, antropomorfo.

Basterebbe osservare che l'art. 27 terzo comma della Costituzione attribuisce un tratto necessariamente 'umano' alle pene (« le pene non possono consistere in trattamenti contrari al senso di umanità »), indirizzandole alla risocializzazione di un condannato che, nonostante la commissione del reato, resta una persona (« e devono tendere alla rieducazione del condannato »).

In riferimento all'esercizio delle funzioni amministrative mediante algoritmi, del resto, non si è esitato a ravvisare una riserva di umanità, come principio generale che dovrebbe regolare l'impiego di macchine, algoritmi e intelligenza artificiale nel settore della 'decisione amministrativa'. Il Consiglio di Stato, con la nota sentenza n. 8472 del 13 dicembre 2019, ha fissato i tre principi di conoscibilità, di non discriminazione algoritmica, e di non esclusività della decisione algoritmica, precisando, in particolare, che deve comunque esistere nel processo decisionale un contributo umano capace di controllare, validare ovvero smentire la decisione automatica (v. anche GALLONE, 2023). Il tema della decisione digitale è molto distante da quello relativo all'impiego delle nuove tecnologie nello svolgimento di attività che, come quella medico-chirurgica, coinvolgono la salute del paziente. La cautela registratasi sul versante del diritto amministrativo, tuttavia, potrebbe offrire qualche spunto adi riflessione anche sul versante di una pretesa 'spersonalizzazione' del concetto di responsabilità (civile e, *a fortiori*, penale).

Non mi soffermerò sui profili relativi a una possibile responsabilità penale dell'intelligenza artificiale o, se si vuole, alla possibile rilevanza dell'intelligenza artificiale come 'soggetto giuridico'.

Mi limito a osservare che non sono mancate proposte favorevoli a individuare proprio nel sistema intelligente il 'soggetto penalmente responsabile', anche sviluppando il preteso parallelismo tra il superamento del *societas delinquere non potest* e del *machina delinquere non potest* (PIERGALLINI, 2020; CAPPELLINI, 2019). Lo stato della materia penale, come risultante dal diritto positivo (diritto penale sostanziale, diritto processuale penale e diritto dell'esecuzione penale), ad ogni modo, non consente un radicale mutamento di paradigma della responsabilità penale, che conserva intatto il suo volto 'antropomorfo'.

Muovendo da questa premessa, sembrerebbe che la sola prospettiva ammissibile sia quella che relega i sistemi intelligenti a un ruolo funzionale, strumentale rispetto all'attività umana. L'idea della ''non autonomia'' della macchina rispetto all'uomo, se applicata senza ulteriori specificazioni al tema della responsabilità (civile e penale) per eventi dannosi o pericolosi derivanti dalla interazione tra uomo e macchina, dovrebbe condurre alla necessaria conclusione per la quale andrebbe individuata, sempre e comunque, una persona che risponde per le conseguenze derivanti da un'attività compiuta da (o insieme a) un sistema intelligente.

L'automatismo in questione, in realtà, rischia di provare troppo.

L'indagine sui rapporti tra intelligenza artificiale e colpa penale, con particolare riguardo allo svolgimento dell'attività medica, potrebbe in realtà condurre a identificare nelle nuove tecnologie non solo (e non tanto) un fattore di necessario ampliamento della responsabilità dell'operatore sanitario, ma, all'opposto, un suo possibile limite, valorizzando il ruolo dell'intelligenza artificiale come fattore di potenziale 'stabilizzazione' del giudizio di responsabilità penale.

2. I possibili modelli di interazione tra attività medica e intelligenza artificiale.

Il punto di osservazione, come precisato, è quello dei rapporti tra operatore sanitario e intelligenza artificiale per la definizione della responsabilità penale della persona fisica.

Una prima questione, che queste riflessioni lasceranno tuttavia solo sullo sfondo, attiene alla responsabilità del produttore o del programmatore. L'aspetto più problematico, come ampiamente noto, attiene al fatto che, attraverso le tecniche di *machine learning* e di *deep learning*, l'algoritmo intelligente, se opportunamente 'addestrato', è in grado di sviluppare *output* in maniera 'autonoma' e, in ogni caso, non del tutto prevedibile al momento della programmazione iniziale. L'autonomia dell'intelligenza artificiale, dunque, rappresenta un fattore di evidente criticità, che impedisce, per esempio, la automatica trasposizione delle categorie e dei principi elaborati, anche in sede penale, per la responsabilità da prodotto difettoso (FRAGASSO, 2025).

Non mi soffermerò neppure sull'impiego di algoritmi diagnostici o predittivi, intesi come strumenti capaci di esaminare dati clinici, individuando elementi e correlazioni non sempre percepibili dall'occhio umano o, ancora, elaborando 'previsioni' sulle probabili evoluzioni di un certo quadro clinico (FRAGASSO, 2025). A venire in considerazione potrebbe essere l'affidamento riposto dall'operatore nel risultato fornito dal sistema intelligente, che, tuttavia, non si riveli corretto o, comunque, adeguato nel caso concreto. In ipotesi di questo tipo, pare opportuno precisarlo, sembrerebbe ci sia poco spazio per un esonero di responsabilità dell'operatore: la persona resta il 'decisore finale', con tutte le conseguenze che ne derivano in tema di responsabilità penale.

Maggiori spazi per un legittimo affidamento, invece, sembrerebbero residuare in caso di interazione tra operatore sanitario e robot intelligenti.

Sono stati efficacemente tratteggiati i possibili modelli di utilizzo di strumenti robotici in ambito sanitario, attraverso classificazioni che possono costituire una base utile per ulteriori riflessioni (IANGEMMA, 2020).

La distinzione preliminare è quella tra robotica clinica, robotica per la riabilitazione e robotica di assistenza.

Alla robotica clinica possono ricondursi le tecnologie automatizzate che consentano di privilegiare pratiche chirurgiche poco invasive, con evidenti vantaggi in termini sia di riduzione dei rischi ordinariamente associati a un intervento chirurgico sia di recupero postoperatorio. Le incisioni chirurgiche di dimensioni ridotte diminuiscono il dolore e favoriscono un più rapido recupero da parte del paziente, riducono il rischio di complicanze, consentono una gestione del paziente anche a distanza (IANGEMMA, 2020). Un'ulteriore articolazione, poi, è quella offerta dalla robotica diagnostica, con particolare riguardo all'impiego di microcapsule endoscopiche fornite di telecamere, utili nella diagnosi di patologie dell'apparato intestinale, con numerose prospettive di evoluzione tecnologica rispetto a dispositivi analoghi (IANGEMMA, 2020).

La robotica riabilitativa e protesica, invece, si riferisce alle macchine impiegate nel trattamento di disabilità fisiche e cognitive, che possono essere impiegate anche in ambito domestico. L'esoscheletro robotizzato *Lokomat*, per esempio, presenta una tecnologia particolarmente avanzata, ma sono sempre più numerose le 'protesi intelligenti' di cui possono servirsi pazienti con problemi di mobilità. A ciò si aggiungono i robot impiegati per il trattamento cognitivo dei soggetti affetti da disturbi mentali: dai malati di Alzheimer alle persone affette da sindrome autistica, per arrivare a macchine capaci di interagire con pazienti che mostrino stati di ansia o agitazione (IANGEMMA, 2020).

La robotica di assistenza, infine, comprende le macchine in grado di svolgere attività infermieristiche comuni, come il monitoraggio dei parametri vitali, la cura dell'igiene dei pazienti o la medicazione di ferite. Nel corso della pandemia da Covid-19 proprio il ricorso a questi robot ha consentito, in alcune struttura sanitarie, di ridurre tanto il carico di lavoro del personale infermieristico quanto il contatto con i malati contagiosi (IANGEMMA, 2020).

3. Coordinate ricostruttive in materia di colpa penale: cenni.

Si tratta, a questo punto, di comprendere se e fino a che punto lo statuto penale della colpa possa assorbire l'onda d'urto dell'intelli-

genza artificiale e, soprattutto, quali sono i luoghi sistematici maggiormente interessati dall'impatto delle nuove tecnologie.

Il primo elemento di difficoltà è costituito proprio dalle coordinate ricostruttive che si ritengano più convincenti in materia di colpa penale, visto che le proposte teoriche non sempre risultano tra loro perfettamente coincidenti.

A me pare, anzitutto, che la sistematica più convincente in materia di colpa penale sia quella che ne valorizzi la doppia funzione (e la doppia misura), prima oggettiva e poi soggettiva (CASTRONUOVO, 2021; sulla diversificazione del parametro di accertamento, da individuarsi nell'*homos eiusdem professionis et condicionis* per la misura oggettiva e nell'agente concreto per la misura soggettiva, sia consentito il rinvio a MASSARO, 2013).

La regola cautelare, intesa come la regola di comportamento a finalità preventiva la cui violazione costituisce il presupposto necessario di una responsabilità per colpa e la cui individuazione avviene senza tener conto delle caratteristiche dell'agente concreto, contribuisce a definire la condotta penalmente rilevante e, quindi, svolge anzitutto il proprio ruolo nell'ambito della tipicità oggettiva. In assenza di una regola cautelare, manca una condotta penalmente rilevante. Ad analoghe considerazioni dovrebbe giungersi, secondo l'opinione preferibile, anche nel caso in cui l'evento verificatosi non rappresenti la concretizzazione del rischio che la regola cautelare mirava a evitare (si veda Cass. pen., Sez. IV, 17 aprile 2019, n. 33244). Il requisito della c.d. concretizzazione del rischio, in effetti, non sembra dotato di effettiva autonomia nella sistematica del reato colposo. La circostanza per cui l'evento concreto deve potersi ricondurre al tipo di quelli che la regola a finalità preventiva mirava ad evitare, lungi dal costituire una precisazione ulteriore cui è affidato il mero compito di specificare l'ambito applicativo di una regola cautelare già individuata, concorre all'individuazione della regola stessa. Il contenuto di una regola cautelare può essere descritto ricorrendo alla seguente schematizzazione: « In presenza della situazione A, è prevedibile che si verificherà un evento del tipo Y, che si può evitare tenendo (o non tenendo) un comportamento X ». Se si accoglie la duplice premessa per cui *a*) la regola cautelare assolve ad una funzione oggettiva della colpa, nel senso che contribuisce alla descrizione della condotta tipica, e *b*) lo 'scopo di tutela' viene in considerazione già in sede di formulazione della regola cautelare, dovrebbe ottenersi come sillogistica conclusione che la riconducibilità dell'evento concreto allo

scopo di tutela della regola cautelare contribuisce alla stessa 'selezione' della condotta penalmente rilevante. In altri termini: nel caso di scuola dell'anziano debole di cuore, còlto da infarto alla vista di una manovra di sorpasso vietata, non ha senso porre un problema di derivazione causale dell'evento dalla condotta dell'automobilista imprudente perché, a monte, manca una condotta tipica in grado di costituire il primo termine della relazione 'causale', in quanto manca una regola cautelare della cui operatività si possa discutere nel caso di specie.

Sul versante soggettivo, poi, è necessario che la violazione della regola cautelare risulti riferibile all'agente concreto, per il quale, quindi, l'evento deve risultare prevedibile ed evitabile, assicurando, in questo modo, un pieno rispetto del principio di personalità della responsabilità penale.

Nell'ambito della responsabilità medica, come ampiamente noto, l'esigenza di edificare degli argini al dilagante fenomeno della c.d. medicina difensiva ha condotto la giurisprudenza (prima) e il legislatore (poi) a valorizzare il possibile ruolo di linee guida, protocolli e buone pratiche nella pre-determinazione della condotta doverosa, giungendo, con la c.d. legge Gelli-Bianco (l. n. 24 del 2017), all'introduzione dell'art. 590-*sexies* c.p. (CAPUTO, 2021).

In sede di coordinate ricostruttive preliminari, infine, pare doveroso evidenziare il possibile ruolo del principio di affidamento nell'ambito di attività mediche plurisoggettive, che, da un punto di vista statistico, sembrerebbero ormai rappresentare le ipotesi più frequenti (anche) nelle aule di tribunale. Il principio di affidamento, in via di prima approssimazione, impone di ritenere penalmente irrilevante la condotta di chi, rispettando le regole cautelari di cui è diretto destinatario, non è tenuto anche a rispondere delle inosservanze di coloro con cui è chiamato a interagire (MANTOVANI, 2021). La storia del legittimo affidamento, però, è segnata dall'ampio spazio riservato alle eccezioni dello stesso, visto che il suo limite sarebbe segnato tanto da un errore riconoscibile quanto da un obbligo di controllo sull'operato altrui. Dopo una fase giurisprudenziale in cui le eccezioni avevano finito per fagocitare (e paralizzare) la regola, le pronunce degli ultimi anni sembrano mostrare una maggiore sensibilità nei confronti di una più rigorosa ripartizione delle rispettive sfere di responsabilità, seguendo le cadenze imposte da una responsabilità penale autenticamente personale (MASSARO, 2020).

Queste coordinate ricostruttive, che già nella loro dimensione 'analogica' faticano a trovare una sicura collocazione tanto in dottrina quanto in giurisprudenza, parrebbero ulteriormente destabilizzate dall'impatto delle nuove tecnologie.

Con particolare riguardo alla possibile interazione tra operatore sanitario e robot intelligenti non sembra, però, che la strada obbligata sia quella di un mutamento di paradigma della responsabilità penale per colpa.

4. Colpa penale e robot intelligenti.

Possono individuarsi almeno tre gruppi di casi, consistenti nella interazione colposa tra uomo e macchina, nell'omesso controllo di robot non azionati direttamente dall'operatore e nel malfunzionamento di una macchina per difetti di costruzione e/o di programmazione.

La interazione colposa tra uomo e macchina, a sua volta, comprende tre differenti ipotesi: 1) errore del medico nel maneggiare la macchina (per esempio il robot in sala operatoria), da cui deriva un evento dannoso o pericoloso a danno del paziente; 2) impianto di un dispositivo diagnostico 'intelligente' non adeguato alle condizioni del paziente o, comunque, senza rispettare le indicazioni relative al dispositivo in questione, da cui derivi un'offesa alla salute o alla vita; 3) mancato controllo di un robot infermiere, che cagiona dei danni al paziente. In questi casi non sembra che lo statuto generale della colpa penale possa o debba subire torsioni particolarmente vistose. Si rende necessario, evidentemente, ridefinire lo standard di diligenza: maneggiare un robot-chirurgo, per esempio, richiede una formazione adeguata, con la conseguenza per cui sarà in colpa il medico che, in assenza delle necessarie competenze, decida di assumere comunque la direzione dell'intervento (Pisani, 2021).

Più complessa potrebbe risultare la questione relativa all'omesso controllo di robot destinati a operare in assenza di un collegamento diretto con il sanitario, come avviene, segnatamente, nel caso di robot infermieri. Le categorie cui fare riferimento sembrerebbero quelle tradizionalmente esaminate in riferimento alla *équipe* medica, intesa, in senso ampio, come interazione tra più sanitari, non necessariamente in un medesimo contesto spazio-temporale: il principio di affidamento, con le sue eccezioni, può trovare un ulteriore banco di prova proprio nel contesto di una 'interazione prudente' arricchita

dal dato tecnologico, lasciando spazio a un limite della responsabilità penale troppo a lungo soffocato dalla giurisprudenza.

Restano, infine, i casi di malfunzionamento della macchina, per cause derivanti da difetti da produzione/programmazione non riconoscibili da parte dell'operatore sanitario. Non sembra sussistano particolari dubbi sul fatto che, almeno di regola, debbano escludersi profili di responsabilità penale per l'utilizzatore finale del sistema intelligente, ferma restando, come anticipato, la questione relativa a una possibile responsabilità anche del produttore e/o programmatore.

4.1. *Il principio di dominabilità ai tempi dell'intelligenza artificiale.*

Tra le tante chiavi 'tradizionali' di cui servirsi per accedere alle interazioni uomo-macchina derivanti dall'intelligenza artificiale potrebbe forse risultare utile quella offerta dal concetto (o, se si vuole, principio) di dominabilità. La dominabilità, come ampiamente noto, è una categoria che taglia trasversalmente l'intera teoria generale del reato: dalla coscienza e volontà della condotta al nesso causale, dall'omissione alla colpa. Una possibile strada, allora, potrebbe consistere proprio nel tentativo di esportazione di un concetto analogo sul versante dell'intelligenza artificiale.

Di particolare interesse, da questo punto di vista, risulta l'opinione di chi ha teorizzato un principio di controllabilità dell'intelligenza artificiale e di insostituibilità del percorso decisionale umano. La premessa da cui si muove è quella per cui l'intelligenza artificiale è uno strumento di cui l'uomo e si serve e che, come tale, deve essere controllabile (FIORELLA, 2022). L'autonomia dell'intelligenza artificiale, dunque, non può risultare illimitata, ma deve essere oggetto di definizione (e delimitazione) da parte della legge: il limite dell'intelligenza artificiale, più esattamente, è quello segnato dalla possibilità di un controllo umano. Nel caso in cui il sistema esprima una autonomia ontologicamente non controllabile in tempo reale, bisognerebbe riconoscere delle aree di rischio consentito più elastiche con sufficiente possibilità e margine di un intervento umano utile, per quanto differito, riducendo eventualmente le potenzialità dell'intelligenza artificiale. Da un punto di vista normativo, la proposta in questione si traduce nella elaborazione di protocolli chiari, cui segua l'introduzione di una clausola di non responsabilità modellata sull'attuale art. 590-*sexies* c.p. in materia di colpa medica (FIORELLA, 2022).

Non è questa la sede per verificare se e fino a che punto una proposta di questo tipo sia 'tecnicamente' realizzabile.

Sembra possibile, però, individuare alcuni punti fermi, che potrebbero rappresentare la base su cui costruire soluzioni più 'avanzate'.

Anzitutto, l'idea di un'intelligenza artificiale suscettibile di controllo umano è quella cui sembra ispirarsi il Regolamento europeo sull'intelligenza artificiale (AI Act) del 2024. L'art. 14, per esempio, stabilisce che «i sistemi di IA ad alto rischio sono progettati e sviluppati, anche con strumenti di interfaccia uomo-macchina adeguati, in modo tale da poter essere efficacemente supervisionati da persone fisiche durante il periodo in cui sono in uso», precisando altresì che «il sistema di IA ad alto rischio è fornito al *deployer* in modo tale che le persone fisiche alle quali è affidata la sorveglianza umana abbiano la possibilità, ove opportuno e proporzionato, di: *a*) comprendere correttamente le capacità e i limiti pertinenti del sistema di IA ad alto rischio ed essere in grado di monitorarne debitamente il funzionamento, anche al fine di individuare e affrontare anomalie, disfunzioni e prestazioni inattese».

Si rende necessario, inoltre, intensificare il confronto tra giuristi ed esperti di intelligenza artificiale, al fine di evitare l'elaborazione di proposte anacronistiche o, all'opposto, (ancora) irrealizzabili.

Il luogo sistematico al quale guardare con attenzione, infine, resta quello del rischio consentito, inteso come concetto che sintetizza il ruolo delle regole cautelari nel giudizio di responsabilità per colpa.

L'interazione tra uomo e sistemi intelligenti, detto altrimenti, potrebbe essere inquadrata non solo dalla prospettiva della ricerca del soggetto responsabile, ma anche da quella, per certi aspetti antitetica, dell'esonero di responsabilità, attraverso una (necessaria) ripartizione delle rispettive sfere di competenza o, se si vuole, una valorizzazione del legittimo affidamento come limite alla responsabilità per colpa.

Sviluppando le indicazioni derivanti dalla sistematica generale della colpa penale, la valorizzazione dell'intelligenza artificiale potrebbe condurre, sul versante oggettivo, a una maggiore riconoscibilità della regola cautelare di riferimento, contribuendo a una più nitida individuazione degli standard di diligenza richiesti da un certo tipo di prestazione: si tratterebbe di approfondire quella tendenza già emersa chiaramente in materia di responsabilità penale del medico, che, attraverso l'esplicita rilevanza attribuita alle linee guida, mira a

rendere più 'afferrabile' il comportamento doveroso che si richiede all'operatore sanitario.

Sul versante soggettivo della colpa, il riferimento alla dominabilità dovrebbe condurre all'esclusione di responsabilità dell'operatore sanitario nei casi in cui la violazione della regola cautelare, pur oggettivamente sussistente, non possa essergli riferita. Il giurista non può limitarsi a prendere atto delle potenzialità dell'intelligenza artificiale, a partire dalla sua 'autonomia', rassegnandosi a firmare una resa sconfortata e senza condizioni: il diritto non sempre deve adattarsi alle nuove tecnologie, potendo anzi pretendere che le stesse si adeguino ai principi e alle categorie giuridiche.

Questo obiettivo, come precisato, può raggiungersi solo coniugando la 'ricerca del colpevole' con la 'ricerca dei limiti della responsabilità', nell'ambito di un dialogo serrato con gli esperti di riferimento.

Note bibliografiche

Cappellini A. (2019), *Machina delinquere non potest? Brevi appunti su intelligenza artificiale e responsabilità penale*, in *Criminalia*, p. 1 ss.

Caputo M. (2021), voce *Colpa medica*, in *Enciclopedia di diritto*, *I tematici*, II, *Reato colposo*, Donini M. (a cura di), Milano, Giuffrè p. 153 ss.

Castronuovo D. (2021), voce *Colpa penale*, in *Enc. dir.*, *I tematici*, II, *Reato colposo*, diretto da M. Donini, Milano, Giuffrè, p. 200 ss.

Fiorella A. (2022), *Responsabilità penale del Tutor e dominabilità dell'Intelligenza Artificiale. Rischio permesso e limiti di autonomia dell'Intelligenza artificiale*, in Giordano A.-Panzarola A.-Police A.-Preziosi S.-Proto M. (a cura di), *Il diritto nell'era digitale. Persona, Mercato, Amministrazione, Giustizia*, Milano, Giuffrè, p. 651 ss.

Fragasso B. (2025), *Intelligenza artificiale e responsabilità penale. Principi e categorie alla prova di una tecnologia "imprevedibile"*, Torino, Giappichelli.

Gallo M. (1956), *La persona umana nel diritto penale*, in *Iustitia*, p. 125 ss.

Gallone G. (2023), *Riserva di umanità e funzioni amministrative. Indagine sui limiti dell'automazione digitale tra procedimento e processo*, Milano, Wolters Kluver - Cedam.

Iagnemma C. (2020), *I "robot medici": profili problematici in tema di alleanza terapeutica e di responsabilità penale*, in *Corti supreme e salute*, 2, p. 441 ss.

Mantovani M. (2021), voce *Affidamento (principio di)*, in *Enciclopedia di diritto*, *I tematici*, II, *Reato colposo*, Donini M. (a cura di), Milano, Giuffrè, p. 1 ss.

Massaro A. (2013), *Responsabilità colposa per omesso impedimento di un fatto illecito altrui*, Napoli, Jovene.

MASSARO A. (2020), *Colpa penale e attività plurisoggettive nella più recente giurisprudenza: principio di affidamento, cooperazione colposa e concorso colposo nel delitto doloso*, in *La Legislazione Penale*, p. 1 ss.
PIERGALLINI C. (2020), *Intelligenza artificiale: da ''mezzo'' ad ''autore'' del reato?*, in *Rivista italiana di diritto e procedura penale*, 4, p. 1745 ss.
PISANI N. (2021), voce *Colpa per assunzione*, in *Enciclopedia di diritto*, *I tematici*, II, *Reato colposo*, DONINI M. (a cura di), Milano, Giuffrè, p. 247 ss.

(LA NORMA)
INTELLIGENZA ARTIFICIALE IN MEDICINA E RESPONSABILITÀ CIVILE DELLA STRUTTURA SANITARIA

di *Mirko Faccioli*

Sommario: 1. Introduzione. — 2. La responsabilità individuale del medico. — 3. La responsabilità della struttura sanitaria. — 4. Il nesso causale. — 5. Conclusioni.

1. Introduzione.

Sono ormai familiari anche a chi non si occupa da vicino di questa materia le caratteristiche dei sistemi di intelligenza artificiale (d'ora in avanti, IA) che sollevano le questioni più rilevanti sotto il profilo giuridico, e più nello specifico sotto il profilo della responsabilità civile: la capacità di funzionare in maniera autonoma e produrre risultati (in tutto o in parte) imprevedibili; il *machine learning*, vale a dire l'attitudine all'apprendimento automatico che consente alla macchina di raggiungere obiettivi assegnati dall'uomo senza che quest'ultimo abbia indicato come procedere al sistema, che nel corso del tempo è in grado di modificare il suo stesso funzionamento; l'opacità, detta anche effetto *black box*, ovverosia la mancanza dell'indicazione delle ragioni delle proprie decisioni da parte dei sistemi IA, le risposte dei quali rimangono pertanto prive di una spiegazione comprensibile per l'essere umano; il carattere aperto, o *openness*, che di nuovo evidenzia l'attitudine dei sistemi di IA a modificare le proprie caratteristiche una volta messi in funzione in conseguenza di aggiornamenti e *upgrade* nonché dell'interazione con altre macchine intelligenti.

In questo contesto, lo studio delle ripercussioni che l'utilizzo dell'IA può avere sulla responsabilità sanitaria gioca senza dubbio un ruolo di primaria importanza, la medicina essendo uno dei settori nei quali la tecnologia in esame trova maggiore espansione e prospetta il

raggiungimento dei più rilevanti benefici tanto per i singoli individui quanto per la collettività. I sistemi di IA più avanzati e maggiormente diffusi in questo ambito sono i *Clinical decision support system* (CDSS), impiegati a supporto delle attività diagnostiche e delle decisioni cliniche, che consentono di raggiungere livelli di precisione, di accuratezza e di efficacia delle prestazioni mediche sempre più elevati rispetto a quelli attingibili con le sole capacità umane. Ma le possibilità di utilizzo dell'IA sono notevolmente più vaste e attraversano l'intero spettro delle attività sanitarie, fra le altre cose incrociando, per esempio, la prevenzione, il monitoraggio dei pazienti, la medicina di precisione, la chirurgia, l'assistenza domiciliare, la riabilitazione, la telemedicina, la medicina genomica, la farmaceutica, e svariati altri ambiti ancora, fino ad arrivare al controverso fenomeno del c.d. *human enhancement*; anche se va sottolineato come nessuna di tali applicazioni, per quanto avanzata e sofisticata, sia ancora in grado di sostituirsi completamente all'uomo e svolgere in totale autonomia un'intera procedura diagnostica o terapeutica.

Va messo in evidenza che, incrementando la qualità e l'efficienza delle cure, l'utilizzo dell'IA in sanità è in grado di produrre notevoli ricadute positive anche dal punto di vista economico, comprimendo in maniera significativa i costi sostenuti tanto dal S.S.N. per erogare le prestazioni mediche quanto dalla popolazione per beneficiarne (si pensi, per esempio, al taglio delle spese di trasporto conseguenti alla diffusione delle pratiche di telemedicina). Sotto questo punto di vista diviene, quindi, di fondamentale importanza rimuovere i molteplici ostacoli che tuttora si frappongono alla piena implementazione dell'IA in ambito medico e che, precludendo la realizzazione dei vantaggi economici sopra menzionati, cagionano danni in termini di c.d. costi opportunità stimabili attorno al 2% del prodotto interno lordo del nostro Paese. Disincentivando il ricorso a questo tipo di tecnologia per timore del contenzioso giudiziario che ne potrebbe scaturire, nell'ambito del più ampio e già da tempo noto fenomeno della c.d. medicina difensiva, l'aura di incertezza che attualmente caratterizza la responsabilità civile cui possono andare incontro i medici e le strutture sanitarie che impiegano l'IA rientra a pieno titolo fra gli ostacoli in discorso: è anche sotto questo profilo che si giustifica, quindi, l'approfondimento della materia.

Sotto quest'ultimo punto di vista va pure sottolineato che, nel nostro settore di nostro interesse, tali problematiche potrebbero presentarsi in termini anche più gravi rispetto al contesto generale della

medical malpractice, perché l'utilizzo di soluzioni tecnologiche particolarmente sofisticate e avanzate potrebbe innalzare le aspettative dei pazienti circa il buon esito delle cure e quindi aggravare il sentimento di frustrazione che, nel caso di mancato raggiungimento del risultato sperato, li spinge a pretendere il risarcimento del danno (asseritamente) subito.

A rendere ancora più complessa la materia contribuisce, del resto, l'intreccio di fonti normative che provvedono alla sua regolamentazione. Da un lato, occorre fare riferimento alla normativa codicistica sulla responsabilità contrattuale ed extracontrattuale in congiunzione con l'apposita disciplina della *medical malpractice* contenuta nella c.d. Legge Gelli-Bianco di cui alla l. 8 marzo 2017, n. 24, recante « Disposizioni in materia di sicurezza delle cure e della persona assistita, nonché in materia di responsabilità professionale degli esercenti le professioni sanitarie ». Dall'altro lato, vengono in rilievo le normative nazionali e sovranazionali specificamente dedicate al fenomeno dell'IA, tra le quali devono essere in particolare menzionate il Reg. UE 2024/1689 sull'intelligenza artificiale del 13 giugno 2024, meglio noto come *AI Act*, e la recentissima l. 23 settembre 2025, n. 132, contenente « Disposizioni e deleghe al Governo in materia di intelligenza artificiale ».

Nel prosieguo del discorso si tenterà quindi di comprendere come in questo contesto si atteggino, oltre alla responsabilità individuale del medico alla quale saranno dedicate alcune osservazioni di carattere generale, la responsabilità della struttura sanitaria che fa ricorso a strumenti di IA e il profilo concernente la ricostruzione del nesso eziologico tra il funzionamento dei sistemi intelligenti e il danno alla salute sofferto dal paziente, per poi concludere con alcune osservazioni di sintesi.

2. La responsabilità individuale del medico.

Secondo i principi generali, il medico è tenuto a eseguire una prestazione di mezzi e non di risultato, in quanto l'aleatorietà derivante delle ineliminabili imperfezioni della scienza medica e dall'imprevedibilità delle reazioni dell'organismo umano non consente di governare fino in fondo l'attività svolta e garantire con sufficiente certezza il buon esito delle cure. In caso di danno al paziente, quindi, il professionista sanitario risponde del suo operato non in termini oggettivi, ma soltanto se rimproverabile per colpa, parametrata sul ca-

none della diligenza professionale contemplato dall'art. 1176, comma 2, c.c. e sul rispetto delle linee guida e delle buone pratiche clinico-assistenziali di cui all'art. 5 della Legge Gelli-Bianco.

Nell'ottica della responsabilità colposa, si deve, allora, in linea di principio escludere che il medico possa essere chiamato a rispondere del malfunzionamento di un sistema di IA che fuoriesce dall'ambito del proprio controllo in virtù del suo elevato grado di autonomia, imprevedibilità e opacità, la sua responsabilità rimanendo circoscritta alle sole ipotesi in cui il danno sia derivato da uno scorretto utilizzo della macchina imputabile alla propria negligenza. Esemplificativamente, il professionista potrà essere ritenuto responsabile per essersi avvalso dell'IA senza avere le competenze necessarie, per avere utilizzato un sistema intelligente nella consapevolezza o nella colpevole ignoranza della sua difettosità, per non avere rilevato la scorrettezza delle indicazioni fornite dalla macchina nei limiti in cui questa fosse rilevabile con una verifica diligente, o ancora per averla impiegata per affrontare un caso clinico le cui specificità imponevano l'impiego di metodi di cura tradizionali (si pensi, per esempio, al medico che, sfruttando i servizi di telemedicina, visita a distanza il paziente le cui condizioni invece richiedevano un esame fisico in presenza).

Vi sono, peraltro, diverse ipotesi nelle quali la giurisprudenza ritiene che vi siano gli estremi per ricostruire un'obbligazione di risultato in capo al medico o, comunque, fa uso di regole operazionali, soprattutto sotto il profilo dell'onere della prova, che finiscono per produrre esiti applicativi del tutto simili. Si tratta di trattamenti sanitari presieduti da regole tecniche molto specifiche e altamente vincolanti, al punto da potersi istituire, anche con il conforto del dato statistico, uno stretto collegamento tra il rispetto di quelle regole e il raggiungimento di un certo esito clinico e, di conseguenza, tra il fallimento delle cure da un lato e un errore medico dall'altro.

Considerato quanto l'IA possa incrementare l'accuratezza, la precisione e l'efficacia delle attività umane, i progressi apportati dall'impiego delle macchine intelligenti potrebbero allora condurre, alla luce delle specifiche peculiarità del caso concreto, alla riqualificazione quale obbligazione di risultato della prestazione medica, essendo del tutto plausibile e già sperimentato nella pratica il fenomeno della conversione di determinati rapporti da obbligazioni di mezzi a obbligazioni risultato in conseguenza dell'evoluzione scientifica e tecnologica. Non sarebbe tuttavia corretto sviluppare *sic et simpliciter* un siffatto ragionamento sulla base della mera constatazione dell'utilizzo

dell'IA, il quale va di per sé considerato un fattore neutrale, fino a che l'accertamento dei benefici effettivamente guadagnati grazie al suo impiego non giustifichi una diversa soluzione. Quindi, secondo la generale tendenza già messa in luce, salvo diversa verifica la prestazione medica eseguita con l'IA dovrebbe rientrare a pieno titolo nell'ambito delle prestazioni di mezzi, la responsabilità per l'inadempimento delle quali è dominata dal criterio della colpa nel senso che è già stato sopra illustrato.

Allo stesso modo non sarebbe appropriato, sotto un diverso ma complementare punto di vista, istituire un'automatica correlazione tra la sofisticatezza e l'innovatività dei sistemi di IA impiegati in medicina e l'applicazione della limitazione della responsabilità ai soli casi di dolo o colpa grave prevista dall'art. 2236 c.c. per il caso di prestazioni implicanti « problemi tecnici di speciale difficoltà »: secondo una consolidata impostazione, la ricorrenza di tale presupposto dev'essere, infatti, sempre effettivamente riscontrata nel caso concreto, non essendo sufficiente, per l'operare della norma *de qua*, la potenziale prospettabilità di problemi di speciale difficoltà desunta da categorie astratte e predefinite. Ciò precisato, nulla esclude che l'impiego dell'IA possa concretamente avere rilievo sul punto, non essendovi dubbi circa la necessità di valutare i presupposti di applicazione della norma in esame alla luce dello stato di avanzamento della scienza e della tecnica del momento storico nel quale si deve giudicare la responsabilità del professionista, per il quale una prestazione in passato considerata di speciale difficoltà potrebbe, quindi, non dover essere più reputata tale in ragione dei vantaggi apportati dall'utilizzo di dispositivi medici intelligenti. Non si può peraltro sottacere come la giurisprudenza sia, da sempre, estremamente riluttante a dare concreta applicazione all'art. 2236 c.c. nel settore sanitario e ne circoscriva l'ambito di operatività entro confini che in dottrina sono prevalentemente e a ragione considerati, seppure non senza eccezioni, eccessivamente restrittivi: prima di delimitarne ulteriormente la sfera di operatività facendo leva sull'utilizzo dell'IA, quindi, sarebbe preferibile restituire a tale norma un ruolo adeguato alla sua fondamentale *ratio*, che la Relazione al codice civile chiarisce essere quella di conciliare « due opposte esigenze, quella di non mortificare l'iniziativa del professionista col timore di ingiustificate rappresaglie da parte del cliente in caso di insuccesso e quella inversa di non indulgere verso non ponderate decisioni o riprovevoli inerzie del professionista » stesso.

3. La responsabilità della struttura sanitaria.

Il regime di responsabilità del personale medico esposto nel paragrafo precedente non si traduce in un vuoto di tutela per il paziente danneggiato dal difettoso funzionamento del sistema di IA utilizzato nelle cure, egli potendo utilmente rivolgere la propria pretesa risarcitoria nei confronti della struttura sanitaria: come noto, quest'ultima può infatti andare incontro, oltre e anche a prescindere dalla responsabilità indiretta per le condotte del personale medico operante al suo interno, ad una diretta e autonoma responsabilità per c.d. difetto di organizzazione che discende dall'inadempimento dell'obbligo, scaturente in capo al nosocomio in virtù del contratto di spedalità intercorrente con i malati, avente per oggetto la predisposizione di un contesto organizzativo di livello adeguato nel quale accogliere gli assistiti.

In questo scenario, si ritiene che l'ente nosocomiale sia tra le altre cose tenuto a sottoporre il proprio equipaggiamento tecnico sia a generali procedure periodiche di controllo e di manutenzione sia a verifiche *ad hoc* prima di ogni concreto utilizzo, la sua responsabilità rimanendo esclusa soltanto per i danni dovuti a difetti di fabbricazione e funzionamento non rilevabili con un'ispezione diligente. Merita di essere, inoltre, sottolineato che, essendo i doveri organizzativi dell'ospedale inquadrabili nell'ambito delle obbligazioni di risultato, la responsabilità che consegue alla loro inosservanza è per lo più ricostruita impiegando, in luogo del criterio della colpa utilizzata per valutare la responsabilità individuale del medico, un più rigido parametro di natura oggettiva, incompatibile pure con il criterio soggettivo sul quale è imperniato l'art. 2236 c.c., che addossa all'ente nosocomiale il rischio dell'inadempimento derivante da tutte le anomalie che si possano verificare nell'ambito del proprio apparato organizzativo fino al limite dell'impossibilità sopravvenuta non imputabile.

È importante sottolineare che l'affermazione della responsabilità per difetto di organizzazione delle strutture sanitarie di per sé non abbisogna di un'esplicita previsione normativa, che del resto non è dato rinvenire nella stessa Legge Gelli-Bianco, a tal fine essendo (sempre stato) sufficiente il ricorso alla disciplina generale del contratto e delle obbligazioni contenuta nel codice civile. La legge in esame, peraltro, contiene qualche indiretto riferimento alla responsabilità in discorso nell'art. 1, che colloca tra gli aspetti fondamentali della disciplina la « sicurezza delle cure » — che del resto campeggia già nell'intitolazione della legge — da realizzare (anche) mediante lo

svolgimento di « attività finalizzate alla prevenzione e alla gestione del rischio connesso all'erogazione di prestazioni sanitarie » nonché « l'utilizzo appropriato delle risorse strutturali, tecnologiche e organizzative »; e ad evidenziare la configurabilità di un'autonoma responsabilità dell'ente ospedaliero per fatto proprio concorre pure l'art. 7 della legge in parola, il quale disciplina la responsabilità della struttura, richiamando il paradigma contrattuale dell'art. 1218 c.c., separatamente dalla responsabilità individuale degli operatori invece sottoposta al regime aquiliano dell'art. 2043 c.c.

Applicati all'ambito di nostro interesse, gli orientamenti in discorso portano quindi a riconoscere la possibilità di affermare la responsabilità della struttura sanitaria per i danni cagionati da sistemi di IA che non siano stati sottoposti alle verifiche e alle procedure di manutenzione e di aggiornamento necessarie per assicurare il loro regolare funzionamento. Sembra doversi, invece, escludere che l'ente sanitario possa essere chiamato a rispondere dei danni cagionati da un difetto dei sistemi di IA riconducibile alla fase di realizzazione del dispositivo e insuscettibile di essere rilevato con un esame diligente da parte dell'ospedale, in tal caso venendo in rilievo l'esclusiva responsabilità di quanti hanno per l'appunto partecipato alla fabbricazione e alla messa in opera della macchina; nell'ottica della responsabilità oggettiva per difetto organizzativo, peraltro, tale circostanza dovrà essere provata dalla stessa struttura sanitaria, che in caso contrario dovrà risarcire il danno subito dal paziente sottoposto al trattamento eseguito tramite il sistema di IA malfunzionante.

Secondo una diversa prospettiva, la struttura sanitaria dovrebbe invece rispondere anche di siffatte evenienze, in applicazione analogica della disciplina dettata dall'art. 1228 c.c. sulla responsabilità del debitore per il fatto dei propri ausiliari: alla luce di questa impostazione, in altri termini, considerato che al debitore viene imputata in via oggettiva la condotta negligente dell'ausiliario umano senza possibilità di offrire una prova liberatoria, tale risultato dovrebbe trovare spazio anche nel caso di sinistro cagionato dall'ausiliario macchina artificiale. A favore di questa soluzione si è anche ritenuto possibile osservare che, diversamente opinando, i dispositivi di IA potrebbero essere strumentalmente impiegati dal debitore al solo e deplorevole scopo di sottrarsi alla responsabilità vicaria alla quale egli andrebbe incontro impiegando esseri umani per svolgere le stesse mansioni, con il risultato di indebolire la fiducia della collettività nei confronti dell'IA e disincentivarne lo sviluppo e la diffusione.

Per quanto suggestiva, tale proposta ermeneutica non può essere tuttavia condivisa, in quanto presuppone di potere riconoscere ai sistemi di IA un attributo, quello della soggettività giuridica, che almeno allo stato attuale dell'evoluzione tecnologica — ma, secondo molti, anche a prescindere dagli ulteriori avanzamenti della stessa — è manifestamente irriferibile a macchine e dispositivi artificiali. Anche volendo per ipotesi superare questo ostacolo, inoltre, rimarrebbe l'impossibilità di rimproverare ai sistemi di IA l'elemento soggettivo del dolo o della colpa richiesto dall'art. 1228 c.c. per la concretizzazione della responsabilità del debitore per i fatti dei suoi ausiliari; senza contare che nei casi di specie mancherebbe pure un altro elemento che normalmente ricorre nel contesto della responsabilità in esame, vale a dire la facoltà del debitore, che abbia risarcito il danno subito dalla controparte, di agire in regresso nei confronti degli ausiliari al fine di recuperare, in tutto o in parte, l'esborso economico sostenuto.

Sotto un diverso punto di vista, alla struttura sanitaria potrebbe essere poi addebitato il fatto di avere adibito all'utilizzo di sistemi di IA personale privo delle conoscenze, delle competenze e della formazione necessarie per impiegare in maniera corretta macchine intelligenti. Né sembra doversi escludere, sulla scorta dell'abbondante e rigorosa giurisprudenza in tema di responsabilità per carenze strumentali degli enti nosocomiali, che questi ultimi possano essere in futuro chiamati a rispondere pure per non aver acquistato e successivamente utilizzato strumenti tecnici innovativi, tra cui appunto i macchinari intelligenti, nel momento in cui tali apparecchiature siano considerate sicure e abbiano trovato una ampia diffusione nel mondo della sanità pubblica e privata.

Va peraltro osservato che lo *standard* organizzativo esigibile dalle strutture sanitarie non è identico per tutti gli enti nosocomiali, in quanto va determinato sulla base del parametro, di carattere elastico, costituito dalla ragionevole aspettativa che il grande pubblico dei pazienti può legittimamente riporre nei confronti dell'ospedale di volta in volta considerato alla luce di tutte le circostanze del caso concreto, anche se sempre nell'ottica di un elevato livello di sicurezza e di tutela della salute e di accettazione di rischi soltanto minimi. Anche avvalendosi del parere di consulenti tecnici nell'ottica di un giudizio di carattere multidisciplinare, nella valutazione dell'idoneità dell'organizzazione di ogni ospedale il giudice dovrà pertanto tenere conto di elementi quali la tipologia e le dimensioni della struttura, il suo ba-

cino d'utenza, le caratteristiche del servizio prestato dagli ospedali affini, lo stato di avanzamento dell'evoluzione scientifica e tecnologica del momento storico, e così via, senza dimenticare il fattore costituito dalla limitatezza delle risorse finanziarie che le strutture sanitarie, soprattutto pubbliche, hanno a disposizione; e sempre fatto salvo il caso che venga in rilievo un requisito organizzativo, strutturale, di dotazione tecnico-strumentale, di carattere igienico-sanitario e simili espressamente previsto da una specifica disposizione normativa, l'ottemperanza alla quale non può che rientrare negli obblighi organizzativi assunti dall'ente sanitario nei confronti del paziente senza che possa esservi spazio per alcun margine di valutazione.

Occorre infine precisare che, secondo un orientamento generale che nessuna ragione preclude di applicare anche nell'ambito dell'IA, l'inadeguatezza dell'apparato organizzativo dell'ospedale non elimina né attenua i profili di responsabilità del personale medico, che in tali casi anzi incontra una serie di specifici e appositi doveri, finalizzati a tutelare la salute del paziente di fronte al deficit organizzativo dell'ente, l'inadempimento dei quali comporta il suo coinvolgimento nel giudizio di responsabilità della struttura secondo la regola della solidarietà. Nello specifico, è stato in giurisprudenza affermato che l'operatore sanitario è tenuto a: verificare l'adeguatezza della struttura nella quale dovrebbe essere erogata la prestazione diagnostica o terapeutica; informare il paziente delle deficienze strutturali e organizzative delle quali egli sia — a prescindere dallo svolgimento e dagli esiti della verifica di cui si è appena sopra detto — a conoscenza, in modo che quegli possa consapevolmente decidere di sottoporsi al trattamento in quella piuttosto che in un'altra struttura; nel caso in cui il paziente non sia in grado di prendere questa decisione, attivarsi perché venga trasportato e assistito in un nosocomio più adeguato alle sue esigenze; infine, nel caso in cui non sia possibile procedere nel modo finora descritto, fare tutto quanto rientra nelle sue possibilità per cercare di sopperire alle deficienze dell'apparato organizzativo in cui dovrà comunque essere eseguito il trattamento medico.

Il momento centrale di questi doveri 'correttivi' delle carenze organizzative dell'ospedale può essere individuato nell'obbligo del medico di informare il paziente in merito alle stesse per reindirizzarlo verso una struttura meglio attrezzata. A fronte di un atteggiamento giurisprudenziale decisamente rigoroso sul punto, la dottrina ha segnalato la necessità di circoscrivere la portata di questo obbligo — e la responsabilità per il suo inadempimento — entro confini ragione-

voli e ben definiti, che pare possano essere tracciati distinguendo due situazioni: quella della struttura priva delle caratteristiche organizzative che dovrebbe invece possedere e che sono necessarie per eseguire correttamente il trattamento richiesto dal paziente, nella quale ci si trova allora di fronte ad un vero e proprio deficit organizzativo che il malato ha senz'altro diritto di conoscere per decidere di conseguenza; e quella dell'ospedale dotato dello standard organizzativo sufficiente per la prestazione di cui il paziente ha bisogno, la quale potrebbe però essere svolta anche in un altro nosocomio caratterizzato da una migliore dotazione di mezzi e/o di personale, verso il quale il paziente dovrà essere reindirizzato solamente qualora si possa affermare che, presso il medesimo, il trattamento necessitato presenta un significativo margine di maggiori possibilità di successo e/o di minori rischi di esito infausto.

4. Il nesso causale.

Che si diriga nei confronti del medico piuttosto che della struttura sanitaria o di entrambi, la pretesa risarcitoria dispiegata dal paziente per ottenere il risarcimento del danno cagionato dall'IA impiegata nelle cure dovrà poi confrontarsi, alla pari di tutte le azioni di responsabilità da *medical malpractice*, con le questioni relative al nesso eziologico. La ricostruzione del nesso causale potrebbe peraltro presentarsi particolarmente complessa nella materia che ci occupa, perché di fronte all'opacità, all'autonomia e all'imprevedibilità dei sistemi di IA è plausibile che possa emergere più di qualche difficoltà nello stabilire se il danno subito dal paziente sia stato o meno cagionato da un malfunzionamento della macchina intelligente.

Quella delineata non è, peraltro, una questione inedita. Il problema probatorio della c.d. causa ignota costituisce infatti uno dei profili della responsabilità sanitaria maggiormente approfonditi tanto in sede dottrinale quanto nelle aule di giustizia negli ultimi anni, venendo da ultimo risolto con il noto indirizzo giurisprudenziale secondo cui il paziente è tenuto a provare, anche attraverso presunzioni, il nesso di causalità materiale tra la condotta del medico e l'evento dannoso, consistente nella lesione della salute o nell'aggravamento della situazione patologica o nell'insorgenza di una nuova malattia, mentre è onere della controparte, ove il paziente abbia dimostrato il suddetto nesso di causalità materiale, provare — di avere esattamente adempiuto, oppure — che l'inadempimento della presta-

zione sanitaria è dipeso da impossibilità della prestazione derivante da causa non imputabile. Pertanto, se rimane ignota la causa del danno riportato dal paziente, sarà quest'ultimo a subirne le conseguenze vedendo la propria azione respinta, mentre se rimane ignota la causa che ha impedito ai medici e al nosocomio la corretta esecuzione della prestazione sanitaria, saranno questi ultimi a farne le spese tramite la condanna al risarcimento.

Per mezzo dell'applicazione di questi principi, quindi, dovrebbero poter essere affrontate in maniera appropriata anche le incertezze sulla ricostruzione del nesso eziologico che emergeranno nelle controversie per danni cagionati da trattamenti sanitari eseguiti con l'ausilio dell'IA.

Con particolare riguardo al nesso di causalità tra il difetto di organizzazione riscontrabile sul piano dei dispositivi di IA e il pregiudizio riportato dal paziente, va poi osservato che, in applicazione del criterio del 'più probabile che non', si dovrà affermare la responsabilità della struttura qualora l'ipotesi della sussistenza del nesso in discorso appaia caratterizzata da maggiori probabilità rispetto all'ipotesi contraria, senza peraltro escludere che, laddove l'incertezza sulla ricostruzione del nesso eziologico impedisca il raggiungimento dello standard probatorio fissato nella formula surriferita, possa trovare spazio una pretesa risarcitoria avente per oggetto il danno da perdita di *chance* di guarigione (o di sopravvivenza, nel caso di malati terminali) fondata sulla dimostrazione del fatto che il difetto organizzativo ha privato il paziente della possibilità di guarire (o di vivere più a lungo e/o in condizioni di minore sofferenza). Va da sé che, qualora il difetto di organizzazione dell'ospedale abbia contribuito a cagionare il danno insieme alla *malpractice* individuale di uno o più operatori sanitari, questi ultimi vedranno corrispondentemente diminuito il *quantum* risarcitorio di cui dovranno farsi carico in sede di rivalsa esercitata della clinica privata o di responsabilità amministrativa fatta valere dal pubblico ministero presso la Corte dei conti.

5. Conclusioni.

La trama di doveri e responsabilità di medici e strutture sanitarie alle prese con sistemi di IA fin qui descritta non sembra dover essere rimessa in discussione alla luce dalle norme sugli obblighi dei *deployer* di sistemi di IA ad alto rischio previsti nell'*AI Act*, senz'altro applicabili anche ai professionisti e agli enti nosocomiali che doves-

sero utilizzare dispositivi medici classificabili come tali, le quali principalmente impongono di: adottare misure tecniche e organizzative per garantire un utilizzo di tali sistemi conforme alle istruzioni che li accompagnano; affidare la sorveglianza umana a persone che dispongono della competenza, della formazione e dell'autorità necessarie; garantire che i dati di input siano pertinenti e sufficientemente rappresentativi; monitorare il funzionamento del sistema, sospenderne l'uso e informare i fornitori e le pertinenti autorità di vigilanza in caso di rischi elevati o incidenti gravi; conservare i log generati automaticamente dal sistema che si trovano sotto il loro controllo per un periodo, adeguato alle finalità del sistema stesso, di almeno sei mesi; laddove applicabile, effettuare una valutazione d'impatto sui diritti fondamentali che l'uso del sistema può produrre. Né pare che i termini del discorso debbano mutare alla luce delle previsioni sull'utilizzo dell'IA contenute nell'art. 7 della l. n. 132/2025, il quale tra le altre cose prevede, al comma 5, che «i sistemi di intelligenza artificiale in ambito sanitario costituiscono un supporto nei processi di prevenzione, diagnosi, cura e scelta terapeutica, lasciando impregiudicata la decisione, che è sempre rimessa agli esercenti la professione medica», e al comma 6 che «i sistemi di intelligenza artificiale utilizzati in ambito sanitario e i relativi dati impiegati devono essere affidabili, periodicamente verificati e aggiornati al fine di minimizzare il rischio di errori e migliorare la sicurezza dei pazienti».

Le riflessioni condotte nelle pagine precedenti consentono, quindi, di concludere che, alla luce dell'odierno stadio dell'evoluzione tecnologica, la responsabilità civile sanitaria derivante dall'utilizzo di sistemi di IA dovrebbe potere essere governata in maniera soddisfacente avvalendosi degli strumenti messi a disposizione dall'attuale quadro normativo e giurisprudenziale della *medical malpractice*.

Questo, peraltro, non esclude che alcuni accorgimenti potrebbero comunque essere introdotti allo scopo di migliorare la tenuta complessiva del sistema. Da questo punto di vista, si potrebbe per esempio prevedere che le indicazioni diagnostiche e terapeutiche fornite dai sistemi di IA siano assimilate alle linee guida di cui all'art. 5 della Legge Gelli-Bianco alle quali il medico deve attenersi, fatte salve le specificità del caso concreto, nello svolgimento della propria attività: in questo modo, l'interprete avrebbe una base normativa sicura per far entrare nel giudizio sulla responsabilità del medico l'atteggiamento di osservanza/inosservanza dallo stesso tenuto nei confronti del responso della macchina intelligente.

Note bibliografiche

Albanese A. (2025), *Digitalizzazione delle cure e intelligenza artificiale: gli obblighi integrativi strumentali e di protezione nell'adempimento della prestazione sanitaria*, in *Responsabilità medica. Diritto e pratica clinica*, 3, p. 307 ss.

D'Adda A. (2022), *Danni «da robot» (specie in ambito sanitario) e pluralità di responsabili tra sistema della responsabilità civile ed iniziative di diritto europeo*, in *Rivista di diritto civile*, 5, p. 805 ss.

Di Martino A. (2022), *Intelligenza artificiale e responsabilità civile in ambito sanitario*, Milano, Giuffrè.

Di Rosa G. (2025), *Robot e responsabilità sanitaria*, in Albanese A.-Guffanti Pesenti L.-Tommasi S. (a cura di), *Il diritto privato nell'economia digitale. Tutele giurisdizionali e autorità indipendenti*, Torino, Giappichelli, p. 217-232.

Faccioli M. (2018), *La responsabilità civile per difetto di organizzazione delle strutture sanitarie*, Pisa, Pacini Giuridica.

Faccioli M. (a cura di) (2022), *Profili giuridici dell'utilizzo della robotica e dell'intelligenza artificiale in medicina*, Napoli, Esi.

Faccioli M. (2023), *Intelligenza artificiale e responsabilità sanitaria*, in *La nuova giurisprudenza civile commentata*, 3, II, p. 732 ss.

Grasso A.G. (2023), *Diagnosi algoritmica errata e responsabilità medica*, in *Rivista di diritto civile*, 2, p. 334 ss.

Natale A. (2025), *La medicina difensiva al tempo dell'intelligenza artificiale*, in *BioLaw Journal Rivista di BioDiritto*, 3, p. 371 ss.

Pagallo U. (2022), *Il dovere alla salute. Sul rischio di sottoutilizzo dell'intelligenza artificiale in ambito sanitario*, Milano, Mimesis.

Pasceri G. (2021), *Intelligenza artificiale, algoritmo e* machine learning*. La responsabilità del medico e dell'amministrazione sanitaria*, Milano, Giuffrè.

Perlingieri C. (2024), *Intelligenza artificiale in ambito medico-sanitario: profili di ricostruzione normativa a seguito dell'*AI Act, in *Rassegna di diritto civile*, 3, p. 907 ss.

Ruffolo U.-Gabbrielli M. (a cura di) (2023), *Intelligenza Artificiale, dispositivi medici e diritto. Un dialogo fra saperi: giuristi, medici e informatici a confronto*, Torino, Giappichelli.

Salito G. (2022), *La responsabilità da algoritmo tra (teoria della) finzione e realtà sanitaria: una nuova declinazione della responsabilità medica?*, in *Rivista italiana di medicina legale e del diritto in campo sanitario*, 4, p. 849 ss.

Scotti R. (2024), *La responsabilità civile dei danni cagionati da sistemi di intelligenza artificiale in ambito sanitario*, in *Giustizia civile*, 1, p. 159 ss.

(LA NORMA)
INTELLIGENZA ARTIFICIALE E TRATTAMENTI SANITARI: TRA CONSENSO INFORMATO, AUTODETERMINAZIONE TERAPEUTICA E PROTEZIONE DEI DATI SENSIBILI

di *Ruben Razzante*

Sommario: 1. Introduzione. — 2. Il consenso informato tra autodeterminazione costituzionale e sanità digitale. — 3. Trasparenza algoritmica e comprensibilità delle decisioni cliniche automatizzate. — 4. Sistemi algoritmici ad alto rischio in sanità: bilanciamento dei diritti fondamentali alla luce dell'AI Act. — 5. La disciplina italiana dell'Intelligenza Artificiale in ambito sanitario.

1. Introduzione.

L'impiego di sistemi di Intelligenza Artificiale nei trattamenti sanitari solleva questioni giuridiche che intersecano simultaneamente la dimensione costituzionale nazionale (articoli 2, 13 e 32 Cost.), il diritto dell'Unione europea (Regolamento UE 2016/679, AI Act, articolo 3 della Carta dei diritti fondamentali) e il sistema convenzionale CEDU (Camardi, 2021; Frosini, 2022; Alpa, 2020).

Tale stratificazione normativa multilivello impone un approccio interpretativo integrato per verificare se e come le garanzie tradizionali del consenso informato possano preservare la propria effettività dinanzi ai sistemi decisionali automatizzati, caratterizzati da logiche algoritmiche spesso imperscrutabili per il paziente e talvolta anche per il personale sanitario.

L'analisi in proposito non può che partire dai diritti alla vita e alla salute, sanciti rispettivamente dagli articoli 2 e 32 della Costituzione italiana, che costituiscono il presupposto imprescindibile per qualsiasi trattamento sanitario, necessariamente orientato alla protezione dell'integrità fisica e psichica dell'individuo, nonché al pieno rispetto della sua dignità. La dignità umana, principio cardine sia dell'ordinamento interno sia del diritto europeo (art. 1 Carta di Nizza), impone che ogni intervento sanitario avvenga nel rispetto della volontà consapevole del-

l'interessato, escludendo ogni forma di imposizione o automatismo non mediato da un processo decisionale umano responsabile.

In tale contesto si innesta il principio di autodeterminazione (Rescigno, 2018; Ferrando, 2018; Stanzione, 2017), inteso come il diritto del singolo di disporre liberamente del proprio corpo e delle proprie scelte terapeutiche, previa adeguata informazione. Il consenso informato (per una nozione dello stesso come 'legittimazione e fondamento' si veda Cassazione - Caso Englaro) si configura quale atto giuridico unilaterale recettizio, espressione della volontà negoziale del paziente di autorizzare o rifiutare il trattamento proposto, con piena consapevolezza degli effetti e dei rischi connessi. La dottrina ha più volte poi sottolineato come il consenso informato non rappresenti una mera formalità burocratica, bensì sia uno strumento primario di attuazione della libertà personale.

L'avvento dell'Intelligenza Artificiale (Perlingeri, 2020) nel settore sanitario introduce nuovi elementi di complessità, modificando profondamente il rapporto medico-paziente tradizionale, incidendo in maniera sempre più sfidante sulla relazione terapeutica e sollecitando una rivisitazione di alcuni istituti giuridici in funzione dell'effettività del diritto alla salute a norma dell'articolo 32 della Costituzione Italiana.

Sistemi di supporto decisionale automatizzato (CDSS - Clinical Decision Support Systems), algoritmi predittivi basati su big data, tecnologie di machine learning che personalizzano terapie e diagnosi, pongono interrogativi circa la possibilità di mantenere inalterato il livello di informazione necessario per la formazione di un consenso valido. Il fenomeno dell''opacità algoritmica' (c.d. *black box problem*) rischia, infatti, di compromettere la piena comprensibilità delle scelte cliniche automatizzate, rendendo la volontà del paziente meno libera o meno consapevole.

A ciò si aggiunge il profilo particolarmente delicato della tutela dei dati personali e, in particolare, dei dati sensibili relativi alla salute (Morana, 2022), il cui trattamento, nel contesto dell'AI sanitaria, assume carattere intensivo e sistematico. In tale prospettiva, si veda a titolo esemplificativo, la disciplina contenuta nel Regolamento (UE) 2016/679 (GDPR) e, specificamente, gli artt. 9, 25 e 35, che impongono obblighi stringenti a carico dei titolari del trattamento, quali il rispetto del principio di minimizzazione, l'adozione di misure tecniche adeguate e la realizzazione di valutazioni di impatto sulla protezione dei dati. La necessità di assicurare un'effettiva protezione della riservatezza, evitando profilazioni indebite o utilizzi discriminatori delle in-

formazioni genetiche e sanitarie, diventa una delle sfide centrali nell'era dell'healthcare algoritmico (PASCUZZI, 2025; RODOTÀ, 2015).

In definitiva, l'analisi giuridica del consenso al trattamento sanitario nell'era dell'Intelligenza Artificiale non può che fondarsi su una lettura sistemica dei principi fondamentali, volta a garantire che l'innovazione tecnologica (MASCHIO, 2020) non comprometta ma, al contrario, rafforzi la tutela della persona umana nella sua dimensione fisica, psichica e informativa.

La digitalizzazione della sanità (VICARELLI-BRONZINI, 2019) e più in particolare l'avvento dell'AI ha quindi un impatto rilevante e decisivo sulla libertà di cura e dunque sull'autodeterminazione del soggetto rispetto alle scelte terapeutiche ma anche sul diritto di ricevere trattamenti sanitari in condizioni di eguaglianza.

In quest'ottica occorre indagare gli effetti derivanti dal ricorso all'Intelligenza Artificiale sia nella definizione e declinazione dell'istituto del consenso informato sia nella realizzazione del principio di non discriminazione in ambito terapeutico.

2. Il consenso informato tra autodeterminazione costituzionale e sanità digitale.

Il contesto è quello della sanità digitale (RUFFOLO, 2023), concetto molto ampio che include ogni forma di applicazione al settore sanitario delle tecnologie dell'informazione e della comunicazione (ICT), quindi le soluzioni medico-sanitarie dell'Intelligenza Artificiale, quelle della telemedicina, intesa quale modalità di erogazione della prestazione sanitaria attraverso tecnologie innovative che consentono di sviluppare la relazione medico-paziente senza la necessità di una compresenza fisica dei soggetti coinvolti e, infine, il fascicolo sanitario elettronico, quale modalità di conservazione, accessibilità e circolazione di dati che attengono alla cura e al monitoraggio della salute individuale.

In prima battuta occorre inquadrare le garanzie costituzionali che sottendono il rapporto di cura, partendo da quelle che assicurano il consenso libero e informato del paziente, con riguardo alle soluzioni terapeutiche proposte dal medico. Un primo rischio da scongiurare è quello di compromettere l'alleanza terapeutica medico-paziente, attraverso un utilizzo improprio della telemedicina e delle altre risorse dell'e-health. Appare in tal senso rilevante la definizione di e-health offerta dall'Organizzazione Mondiale della Sanità, che la qualifica come «*the cost-effective and secure use of information and communica-*

tions technologies in support of health and health-related fields, including health-care services, health surveillance, health literature, and health education, knowledge and research».

La bussola orientatrice deve rimanere l'art. 32 comma 2 della Costituzione, che pone al centro della pratica sanitaria la persona umana sia nei trattamenti obbligatori che in quelli volontari (Fioriglio, 2021; Farina, 2019; Eysenbach, 2001).

Ecco perché l'acquisizione del consenso informato (Dassano, 2006; Morana, 2008; Coraggio, 2008) agli interventi sanitari non è solo una buona prassi sul piano tecnico ma appare la lineare e coerente attuazione dell'art. 32 della Costituzione che, nel riconoscere il diritto alla salute come libertà giuridica, sottopone a limiti puntuali la possibilità di introdurre obblighi di cura. Va ricordato che l'obbligatorietà di un trattamento sanitario è un'eccezione, giustificabile solo per tutelare l'interesse della collettività alla salute, sempre nella cornice attuativa dell'art. 32 della Costituzione. Nel contesto della legislazione europea, occorre menzionare l'art. 3 comma 2 della Carta dei diritti fondamentali dell'Unione europea, ma anche l'art. 8 della Convenzione europea per la salvaguardia dei diritti dell'uomo e delle libertà fondamentali (CEDU), che al primo comma protegge il diritto al rispetto della propria vita privata e famigliare, del proprio domicilio e della propria corrispondenza e al secondo comma aggiunge: «*Non può esservi ingerenza di una autorità pubblica nell'esercizio di tale diritto a meno che tale ingerenza sia prevista dalla legge e costituisca una misura che, in una società democratica, è necessaria alla sicurezza nazionale, alla pubblica sicurezza, al benessere economico del paese, alla difesa dell'ordine e alla prevenzione dei reati, alla protezione della salute o della morale, o alla protezione dei diritti e delle libertà altrui*».

La base del consenso informato è proprio la libertà del soggetto di compiere una scelta in ordine alle cure: decidere se optare per questa o quella soluzione terapeutica oppure rifiutare un trattamento sanitario o una cura.

Su questo punto giova richiamare la legge 22 dicembre 2017, n. 219 «Norme in materia di consenso informato e di disposizioni anticipate di trattamento», che ribadisce il carattere volontario dei trattamenti sanitari e promuove e valorizza la «*relazione di cura e di fiducia tra paziente e medico che si basa sul consenso informato nel quale si incontrano l'autonomia decisionale del paziente e la competenza, l'autonomia professionale e la responsabilità del medico*» (art. 1, comma 2).

Appare prioritario soffermarsi in questa sede sull'espressione consenso informato, che opera una conciliazione concettuale e fattuale tra la libertà di cura, con la primazia della regola della volontarietà dei trattamenti sanitari e dunque della possibilità di acconsentire o meno alle cure proposte dal medico e dalla struttura sanitaria, e il diritto del paziente ad essere correttamente informato sulle caratteristiche e i possibili effetti di quelle cure, al fine di poter assumere una decisione terapeutica libera e consapevole. Un requisito primario, dunque, per quella conciliazione, è senz'altro un regolare flusso informativo all'interno della relazione tra medico e paziente, fondato sulla trasparenza rispetto allo stato di salute, alle possibili terapie ritenute adeguate in presenza di particolari patologie e alle possibili conseguenze derivanti dalla sottoposizione a uno specifico percorso diagnostico-terapeutico. Tutto ciò fa emergere nitidamente il carattere strumentale del diritto all'informazione rispetto alla libertà di cura.

Volendo ampliare sinteticamente il raggio della riflessione giuspubblicistica ad altri due riferimenti costituzionali, l'art. 2 e l'art. 13 della Costituzione, che una copiosa giurisprudenza ha applicato in maniera creativa al tema in oggetto, occorre evidenziare che taluni trattamenti sanitari coercitivi, oltre che limitare la libertà di cura, possono andare a incidere in senso restrittivo anche sulla libertà fisica del soggetto (art. 13), oltre che intaccare la superficie dei diritti inviolabili protetti dall'art. 2 della nostra Carta. Soccorrono su questo versante alcuni strumenti di soft law tra i quali gli articoli 16 e 33-39 del Codice di deontologia medica, approvato il 18 maggio 2014 e successivamente modificato.

Tale cornice regolatoria rappresenta l'humus ideale per misurare l'impatto delle nuove tecnologie di Intelligenza Artificiale (AI) sul rapporto terapeutico. Fra le tecnologie che promettono di rivoluzionare la pratica clinica e di incidere in maniera significativa sul rapporto fiduciario medico-paziente (MITTELSTADT, 2021) e sul consenso informato c'è senz'altro l'AI.

Ai sensi dell'art. 1 comma 3 della legge n. 219 del 2017 (PICCINNI, 2018), la persona dev'essere informata «*in modo aggiornato e a lei comprensibile riguardo alla diagnosi, alla prognosi, ai benefici e ai rischi degli accertamenti diagnostici e dei trattamenti sanitari indicati, nonché riguardo alle possibili alternative e alle conseguenze dell'eventuale rifiuto del trattamento sanitario e dell'accertamento diagnostico o della rinuncia ai medesimi*».

Da questo punto di vista anche la giurisprudenza ha offerto negli ultimi 10-15 anni diversi approdi illuminanti in ordine alle modalità

di manifestazione del consenso e al contenuto dell'obbligo informativo. Anzitutto la Cassazione, nella sentenza n. 20984 del 27 novembre 2012 e nella sentenza n. 8035 del 21 aprile 2016, ha chiarito che il consenso dev'essere effettivo e non deve mai ritenersi presunto. Dunque, deve necessariamente costituire oggetto di una espressa manifestazione di volontà. Inoltre, nelle sentenze n. 16543 del 28 luglio 2011 e n. 12205 del 12 giugno 2015, si puntualizza che il consenso deve essere continuato, cioè espresso per ogni singolo atto terapeutico o diagnostico idoneo a cagionare autonomi rischi.

La Cassazione ha anche ribadito che l'informazione al paziente, in quanto diretta a ottenere un suo libero consenso alle cure e a non colmare il divario informativo tra medico e paziente, deve presentare i caratteri di completezza e adeguatezza. Nella sentenza n. 2847 del 9 febbraio 2010, in ordine al requisito della completezza, la Cassazione chiarisce che l'informazione dev'essere particolarmente dettagliata al fine di garantire lo scrupoloso rispetto del diritto di autodeterminazione del paziente.

Per quanto attiene, invece, all'adeguatezza, l'informazione deve sostanziarsi in spiegazioni adeguate al livello culturale del paziente, con l'adozione di un linguaggio che tenga conto del suo particolare stato soggettivo e del grado di conoscenze specifiche di cui dispone. L'adeguatezza pertiene, quindi, esclusivamente all'ambito formale, delle modalità di informazione, senza in alcun modo andare a detrimento della completezza. Essa cioè non può servire a delimitare il novero di informazioni che il medico deve fornire al paziente per metterlo nelle migliori condizioni per esprimere un consenso realmente consapevole alle cure. Inoltre va precisato che, con sentenza n. 364 del 15 gennaio 1997, il giudice di legittimità ha stabilito che l'obbligo informativo del medico «*si estende ai rischi prevedibili e non anche agli esiti anomali, al limite del fortuito*». Si tratta, tuttavia, di un limite riferito a ipotesi estreme.

In generale la giurisprudenza tende ad allargare i confini del contenuto dell'obbligo informativo, mostrando di condividere i tre standard per l'offerta di informazioni sanitarie elaborati dal Comitato nazionale per la bioetica (Informazione e consenso all'atto medico, 20 giugno 1992): lo standard professionale, che obbliga a veicolare tutte le ultime acquisizioni della comunità scientifica; lo standard soggettivo, che obbliga a comunicare al paziente ciò che sia davvero essenziale ai fini della comprensione e assunzione da parte sua di una scelta consapevole; lo standard medio, in base al quale devono essere

fornite quelle informazioni che a una persona ragionevole e di media cultura sono necessarie per compiere una scelta consapevole rispetto alla terapia proposta dal medico.

3. Trasparenza algoritmica e comprensibilità delle decisioni cliniche automatizzate.

Arrivando più da vicino al quesito relativo all'esistenza di un obbligo di informare il paziente dell'utilizzo di software di Intelligenza Artificiale per la diagnosi e la proposta di terapia, si può ancora una volta fare riferimento a quanto affermato dal Comitato nazionale per la bioetica e dal Comitato nazionale per la biosicurezza, le biotecnologie e le scienze della vita, nel documento del 29 maggio 2020 dal titolo « Intelligenza Artificiale e medicina: aspetti etici ».

« *È dunque un obbligo etico e giuridico* — si legge nel documento — *che coloro che si sottopongono a trattamenti sanitari così innovativi, attraverso l'AI, siano informati nelle modalità più consone e comprensibili al paziente di ciò che sta accadendo, di essere (se è il caso) oggetto di sperimentazione e validazione; di essere a conoscenza che ciò che è loro applicato (sul piano diagnostico e terapeutico) implica dei vantaggi, ma anche dei rischi. Va specificato in modo esplicito nel consenso informato se i trattamenti applicati (diagnostici e terapeutici) provengano solo da una macchina (AI, robot) o se e quali sono gli ambiti e i limiti del controllo umano o supervisione sulla macchina* ».

Se, dunque, si configura con chiarezza un obbligo del professionista sanitario di informare il paziente dell'utilizzo di AI nel procedimento che ha portato alla diagnosi e alla proposta terapeutica, nondimeno si palesano alcune remore (Collecchia, 2019) in ordine alla configurabilità di un eventuale diritto di essere informati sull'uso dell'AI da parte del medico, tra cui il timore che soggetti più fragili e distanti culturalmente dalle tecnologie possano opporre un rifiuto preconcetto alla proposta diagnostica o terapeutica (Astromskė-Peičius-Astromskis, 2020). D'altra parte, però, se il medico si astenesse del tutto dal fornire informazioni sull'uso dell'AI e di strumenti simili, finirebbe per venir meno all'obbligo di completezza dell'informazione nell'ambito del consenso informato. Peraltro, un generico riferimento all'utilizzo di software intelligenti, senza alcun dettaglio relativo al loro funzionamento e alle finalità perseguite, potrebbe generare sospetti e ulteriore diffidenza nei pazienti, inducendoli a un rifiuto aprioristico.

In definitiva si può affermare che il diritto a una piena informazione del paziente, finalizzato all'assunzione da parte sua di una scelta consapevole, dovrebbe imporre al medico di rendere conoscibile e comprensibile al paziente l'algoritmo in questione. Dubbi tuttavia affiorano circa il grado di approfondimento delle informazioni da trasmettere, con particolare riferimento al peso assunto dall'algoritmo all'interno della decisione finale, quindi al grado di autonomia del medico nella decisione.

L'Organizzazione Mondiale della Sanità, pur ammettendo che non esistono 'precedenti' per la configurazione di un obbligo del medico, etico o addirittura giuridico, di ottenere il consenso del paziente all'uso di tecnologie per la diagnosi, sottolinea il valore della trasparenza come requisito cruciale per creare fiducia nelle nuove tecnologie. Nei sistemi di Intelligenza Artificiale il confine tra la tutela dei dati personali e la tutela della salute appare più labile e occorre far riferimento agli articoli 13 e 14 del GDPR. In base all'art. 13, comma 2 lett. *f*) del GDPR, l'interessato ha diritto di essere informato dal titolare del trattamento, al momento della raccolta dei dati, circa «*l'esistenza di un processo decisionale automatizzato, compresa la profilazione di cui all'articolo 22 paragrafi 1 e 4*» e di ottenere, «*almeno in tali casi, informazioni significative sulla logica utilizzata*» e circa «*l'importanza e le conseguenze previste di tale trattamento per l'interessato*». In base al vigente diritto europeo (art. 22.2 GDPR), una decisione algoritmica può legittimamente incidere sui diritti della persona anche senza intervento umano: *a*) quando vi sia il consenso esplicito dell'interessato; *b*) quando sia necessaria per concludere o eseguire un contratto tra interessato e titolare del trattamento; infine, *c*) quando sia autorizzata dal diritto dell'Unione o dello Stato membro. In questi casi (eccetto il caso b) il titolare del trattamento deve adottare misure appropriate per tutelare i diritti, le libertà e i legittimi interessi dell'interessato, quantomeno il diritto di ottenere l'intervento umano da parte del titolare del trattamento, di esprimere la propria opinione e di contestare la decisione.

4. Sistemi algoritmici ad alto rischio in sanità: bilanciamento dei diritti fondamentali alla luce dell'AI Act.

Assicurare la protezione dei diritti fondamentali e fare in modo che l'Intelligenza Artificiale li rafforzi anziché comprometterli è cruciale per il futuro del rapporto tra uomo e tecnologie. Nel discutere

l'uso dell'AI, è essenziale riconoscere come le opportunità offerte dal progresso tecnologico debbano essere accompagnate da adeguate garanzie di tutela per le persone e le comunità, riducendo al minimo i possibili effetti negativi.

L'utilizzo dell'AI nell'ambito del diritto alla salute ha permesso di introdurre il concetto di telemedicina, che ha conosciuto un significativo sviluppo a seguito dell'emergenza sanitaria dovuta al Covid-19.

L'utilizzo dell'AI in ambito sanitario pone una duplice sfida: da una parte, accertare se l'utilizzo della stessa possa contribuire ad accrescere l'efficacia della prestazione sanitaria e, quindi, a potenziare il grado di effettività del diritto fondamentale della tutela della salute; dall'altra, quella di assicurare che l'erogazione dei servizi di assistenza sanitaria tramite la telemedicina non entri in conflitto con altri valori e diritti di pari livello. È infatti necessario che sia garantito il diritto ad una decisione non completamente automatizzata, il diritto ad una decisione che non violi i principi di autodeterminazione del paziente (con la gestione del rischio da parte di chi esegue la prestazione), l'obbligo di motivazione della decisione presa ed il diritto di conoscere e contestare (anche in sede giudiziaria) il percorso decisionale seguito ed i relativi esiti.

Proprio in questa prospettiva non bisogna dimenticare che, nonostante la tecnologia e in particolare l'avvento dell'AI sia apparsa sin da subito un potenziale strumento di accrescimento delle possibilità per l'individuo, con il passare del tempo stanno emergendo, sempre con maggiore evidenza, i rischi (ADINOLFI, 2022) legati all'eccessivo utilizzo della stessa.

In ambito sanitario è stato dimostrato come l'AI possa essere utilizzata per studiare le cause ed i problemi di salute pubblica come, ad esempio, i rischi legati all'ambiente o alla salute sul lavoro. In quest'ottica, dunque, le diverse soluzioni digitali (app, dispositivi, etc.) possono consentire alle persone di godere di uno stile di vita sano e di prevenire lo sviluppo o la progressione di alcune malattie. Alcuni strumenti di salute mobile, c.d. m-health, rilevano addirittura i primi sintomi o indicatori della patologia, fornendo riscontri agli operatori sanitari ed aiutando i pazienti ad aderire ai programmi di trattamento.

Tuttavia, bisogna tener presente che l'utilizzo dell'AI in campo sanitario può generare o, comunque, può contribuire ad accentuare anche le diseguaglianze tra persone, tenuto conto sia della disomoge-

neità di competenze e conoscenze individuali dell'utilizzo dei dispositivi tecnologici, sia dell'esistenza di diversi livelli di implementazione delle politiche territoriali volte allo sviluppo digitale. Pare chiaro allora come l'utilizzo di soluzioni digitali basate sull'AI per incrementare la salute esiga pari opportunità e l'assoluta necessità di una alfabetizzazione digitale diffusa, così come di una adeguata diffusione di strumenti, servizi e piattaforme disponibili ed accessibili a tutti.

Si può concludere che, nell'ambito della tutela del diritto alla salute, tra le opportunità dell'utilizzo dell'AI è possibile individuare benefici capaci di assicurare agli individui cure potenzialmente più efficaci; benefici associati anche a considerevoli risparmi di tempo e di risorse economiche (si pensi ad esempio al risparmio collegato all'utilizzo del c.d. fascicolo elettronico). L'utilizzo dell'AI nell'ambito sanitario offre inoltre l'opportunità di 'personalizzare' la medicina, rendendola 'a misura' del paziente che può essere monitorato in tempo reale. Nella prospettiva opposta, l'utilizzo dell'AI in medicina evidenzia criticità quali, ad esempio, le differenze e le disomogeneità che rischiano di acuirsi proprio a causa dell'utilizzo di quella tecnologia.

Grazie alla sua capacità di analizzare grandi quantità di dati e di apprendere da essi con il machine learning, l'AI sta diventando uno strumento sempre più importante per migliorare il sistema sanitario e stiamo assistendo ad una rivoluzione nel modo in cui i pazienti vengono curati e le diagnosi vengono effettuate. L'Intelligenza Artificiale e i dati sanitari costituiscono un potente alleato per la sanità e la ricerca. In Italia, negli ultimi anni si sono registrati importanti progressi nell'utilizzo di queste tecnologie.

Il fascicolo sanitario elettronico ha reso possibile la condivisione dei dati sanitari tra i diversi professionisti sanitari, mentre la telemedicina ha consentito di fornire assistenza sanitaria a distanza. Tuttavia, si profilano sfide in termini di privacy ed equità nell'uso di tali tecnologie (Tozzi-Cinelli, 2021; Casonato-Violante, 2020). Proprio per questo motivo nel 2022 il Consiglio Superiore di Sanità ha stilato, grazie anche al contributo di un gruppo di esperti del settore, un documento intitolato 'I sistemi di Intelligenza Artificiale come strumento di supporto alla diagnostica' che rappresenta il punto di vista del Ministero della Salute nei confronti dell'applicazione dell'Intelligenza Artificiale in ambito diagnostico.

L'AI Act, nel disciplinare l'utilizzo dell'Intelligenza Artificiale, pone grande enfasi sulla salvaguardia dei dati personali e sulla protezione della privacy dei cittadini e riconosce che i dati relativi alla sa-

lute sono particolarmente sensibili e richiedono un elevato livello di protezione. L'AI Act stabilisce norme chiare riguardo alla raccolta, all'elaborazione e all'uso dei dati sanitari nell'ambito dei sistemi di Intelligenza Artificiale. Conforme al principio di minimizzazione dei dati, l'AI Act richiede che venga utilizzata la quantità minima di dati necessaria per raggiungere gli obiettivi previsti.

In ottemperanza alle disposizioni dell'AI Act, le aziende sanitarie e gli operatori del settore devono limitare la raccolta e l'elaborazione dei dati sanitari a quanto strettamente necessario per scopi specifici come diagnosi, cura o ricerca medica. Inoltre, l'AI Act in Sanità sottolinea l'importanza di ottenere il consenso informato da parte dei pazienti per il trattamento dei loro dati sanitari. I cittadini hanno il diritto di essere adeguatamente informati su come verranno utilizzati i loro dati e devono prestare il loro consenso esplicito al trattamento di tali informazioni. Da sottolineare che l'AI Act in Sanità avrà un grande impatto sui dispositivi medici che contengono software di Intelligenza Artificiale.

L'AI Act mira anche a garantire la trasparenza nell'utilizzo dell'Intelligenza Artificiale nel campo della Salute. Le organizzazioni sanitarie che impiegano sistemi di AI per scopi sanitari devono essere trasparenti riguardo al funzionamento di tali sistemi, inclusa la logica alla base delle decisioni e i criteri utilizzati per assumerle. Infine, l'AI Act in Sanità prevede adeguate misure di sicurezza per proteggere i dati sanitari da accessi non autorizzati, perdite o utilizzi impropri. Le organizzazioni sanitarie devono adottare adeguate misure tecniche e organizzative per garantire la sicurezza e l'integrità dei dati sanitari, prevenendo qualsiasi rischio di violazione della privacy.

Va infine ricordato che già prima dell'emanazione dell'AI Act, per la precisione il 12 ottobre 2023, il Garante per la protezione dei dati personali (Calvara, 2023; Di Costanzo, 2023) ha pubblicato un decalogo contenente le regole e i principi da seguire per l'utilizzo dell'AI nell'ambito del servizio sanitario nazionale, declinando i principi che regolano il trattamento dei dati personali ai sensi del Regolamento (UE) 679/2016 (GDPR) nel caso di implementazione di tali sistemi. Nel documento il Garante sottolinea la centralità dell'accountability, che deve essere necessariamente accompagnata da un utilizzo etico dei sistemi di AI, anche tenendo conto dei doveri deontologici che gli operatori sanitari sono tenuti ad osservare. Ad esempio, potrebbero assumere particolare importanza eventuali raccomandazioni elaborate dagli ordini professionali e dai comitati etici, a riprova della

complementarità tra norme di legge e forme di autodisciplina (RUOCCO, 2023).

Particolare importanza sul tema assume la legge 23 settembre 2025, n. 132, recante 'Disposizioni e deleghe al Governo in materia di intelligenza artificiale', pubblicata nella Gazzetta Ufficiale n. 223 del 25 settembre 2025 ed entrata in vigore il 10 ottobre 2025, che rappresenta il primo intervento legislativo organico di uno Stato membro dell'Unione europea volto ad affiancare al Regolamento (UE) 2024/1689 (c.d. AI Act) una disciplina nazionale integrata in materia di sistemi di Intelligenza Artificiale.

Giova evidenziare che siffatto provvedimento legislativo trae origine dal disegno di legge n. 1146/2024. Su tale D.D.L. Il Garante aveva, in data 02 agosto 2024, espresso un parere con il quale aveva rilevato l'opportunità di modificare e integrare le disposizioni affinché potesse rappresentare una valida base giuridica per il trattamento di dati personali a fini di ricerca nella realizzazione dei sistemi di AI in ambito sanitario.

5. La disciplina italiana dell'Intelligenza Artificiale in ambito sanitario.

La legge italiana di recente approvazione si fonda su una dimensione antropocentrica dell'Intelligenza Artificiale, promuovendo un utilizzo corretto, trasparente e responsabile della tecnologia, con particolare attenzione all'impatto sui diritti fondamentali.

L'articolo 7 della legge n. 132/2025 delinea i princìpi cardine che governano l'impiego di sistemi di Intelligenza Artificiale nel settore sanitario, ponendo quale premessa fondamentale che «*l'utilizzo di sistemi di intelligenza artificiale contribuisce al miglioramento del sistema sanitario, alla prevenzione, alla diagnosi e alla cura delle malattie, nel rispetto dei diritti, delle libertà e degli interessi della persona, anche in materia di protezione dei dati personali*» (comma 1).

La norma stabilisce al comma 2 un principio fondamentale di non discriminazione, sancendo che l'introduzione di sistemi di Intelligenza Artificiale nel sistema sanitario non può selezionare e condizionare l'accesso alle prestazioni sanitarie secondo criteri discriminatori.

Tale disposizione assume particolare rilevanza costituzionale, ponendosi in diretta attuazione dell'articolo 3 Cost. e dell'articolo 32 Cost., che tutela la salute come diritto fondamentale dell'individuo e

interesse della collettività, nonché del principio di uguaglianza sostanziale nell'accesso alle cure.

Il comma 3 riconosce espressamente il diritto dell'interessato di essere informato sull'impiego di tecnologie di Intelligenza Artificiale, configurando così un obbligo informativo specifico che si pone quale presupposto essenziale per l'esercizio consapevole dell'autodeterminazione terapeutica e del consenso informato ex articolo 32, comma 2, Cost.

Tale previsione si coordina necessariamente con gli obblighi informativi previsti dal Regolamento (UE) 2016/679 (GDPR) agli articoli 13 e 14, nonché con la disciplina del consenso informato di cui alla legge 22 dicembre 2017, n. 219.

Il comma 5 dell'articolo 7 stabilisce poi un principio cardine: i sistemi di Intelligenza Artificiale in ambito sanitario costituiscono unicamente un supporto nei processi di prevenzione, diagnosi, cura e scelta terapeutica, lasciando impregiudicata la decisione finale, che è sempre rimessa agli esercenti la professione medica.

La tecnologia non può quindi sostituire il giudizio clinico professionale, ma assume una funzione meramente ausiliaria e strumentale, preservando la centralità della relazione medico-paziente e la responsabilità professionale del sanitario.

Il comma 6 introduce un obbligo di affidabilità, verifica periodica e aggiornamento dei sistemi di AI utilizzati in ambito sanitario, finalizzato a minimizzare il rischio di errori e migliorare la sicurezza dei pazienti. Tale previsione si pone in coerenza con le disposizioni dell'AI Act relative ai sistemi ad alto rischio e con gli obblighi di *quality management system* previsti dagli articoli 8-15 del medesimo Regolamento europeo.

Particolarmente interessante anche l'art. 8 della legge in discorso che introduce una disciplina specifica per i trattamenti di dati personali eseguiti nell'ambito della ricerca e sperimentazione scientifica nella realizzazione di sistemi di Intelligenza Artificiale in ambito sanitario.

La disposizione si applica ai trattamenti effettuati da una pluralità di soggetti qualificati: soggetti pubblici e privati senza scopo di lucro, Istituti di ricovero e cura a carattere scientifico (IRCCS) ex decreto legislativo 16 ottobre 2003, n. 288, nonché soggetti privati operanti nel settore sanitario nell'ambito di progetti di ricerca condotti in collaborazione con soggetti pubblici, privati non profit o IRCCS.

Il comma 1 dell'articolo 8 contiene una declaratoria di rilevante interesse pubblico per i trattamenti di dati personali finalizzati alla ricerca scientifica nella realizzazione di sistemi di Intelligenza Artificiale

per molteplici finalità: prevenzione, diagnosi e cura di malattie, sviluppo di farmaci, terapie e tecnologie riabilitative, realizzazione di apparati medicali, incluse protesi e interfacce, salute pubblica, incolumità della persona, salute e sicurezza sanitaria, nonché studio della fisiologia, della biomeccanica e della biologia umana anche in ambito non sanitario.

Il comma 2 introduce una deroga significativa in materia di consenso, stabilendo che, fermo restando l'obbligo di informativa in favore dell'interessato (che può essere assolto anche mediante un'informativa generale pubblicata sul sito web del titolare del trattamento), è sempre autorizzato l'uso secondario di dati personali privi degli elementi identificativi diretti, anche appartenenti alle categorie di cui all'articolo 9 GDPR, senza necessità di acquisire un ulteriore consenso dell'interessato ove inizialmente previsto dalla legge.

Tale previsione configura un'ipotesi di compatibilità del trattamento ulteriore ai sensi dell'art. 6, comma 4, GDPR, basata su un bilanciamento tra l'interesse alla ricerca scientifica e i diritti dell'interessato. L'anonimizzazione dei dati costituisce una garanzia fondamentale, essendo richiesta «*salvi i casi nei quali la conoscenza dell'identità degli interessati sia inevitabile o necessaria al fine della tutela della loro salute*».

Il comma 3 autorizza espressamente, previa informativa all'interessato ex art. 13 GDPR, il trattamento per finalità di anonimizzazione, pseudonimizzazione o sintetizzazione dei dati personali, anche appartenenti alle categorie particolari di cui all'articolo 9, paragrafo 1, del GDPR.

L'inclusione della 'sintetizzazione' — ossia la creazione di dataset sintetici che mantengono le proprietà statistiche dei dati originali senza contenere informazioni riferibili a individui specifici — rappresenta un elemento innovativo che riflette le più recenti evoluzioni tecnologiche nel campo del machine learning e della privacy by design.

Il comma 4 attribuisce all'Agenzia nazionale per i servizi sanitari regionali (AGENAS), previo parere del Garante per la protezione dei dati personali e tenendo conto degli standard internazionali e dello stato dell'arte tecnologico, il potere di stabilire e aggiornare linee guida per le procedure di anonimizzazione e per la creazione di dati sintetici, anche per categorie di dati e finalità di trattamento.

Tale disposizione realizza un meccanismo di soft law che consente un adattamento dinamico della disciplina all'evoluzione tecnologica.

I commi 5 e 6 prevedono un meccanismo di controllo preventivo e successivo da parte del Garante per la protezione dei dati personali: i trattamenti devono essere comunicati al Garante e possono essere avviati decorsi trenta giorni dalla comunicazione, salvo provvedimento di blocco del Garante.

Restano fermi i poteri ispettivi, interdittivi e sanzionatori dell'Autorità, garantendo così un sistema di checks and balances nella tutela dei diritti fondamentali.

Meritano un cenno anche i successivi articoli della legge nazionale di cui si discute.

L'articolo 9 prevede infatti una delega in favore del Ministro della salute per l'emanazione, entro centoventi giorni dall'entrata in vigore della legge, di un decreto che disciplini il trattamento dei dati personali anche particolari, con modalità semplificate, per finalità di ricerca e sperimentazione tramite sistemi di AI e machine learning, inclusa la costituzione di *regulatory sandbox*, sentiti il Garante, gli enti di ricerca e i presidi sanitari.

L'articolo 10 introduce nel d.l. 179/2012 un nuovo art. 12-bis che prevede l'adozione di decreti del Ministro della salute per disciplinare le soluzioni di AI nel settore sanitario e istituisce una piattaforma nazionale di Intelligenza Artificiale per il supporto all'assistenza territoriale.

La progettazione e titolarità della piattaforma sono attribuite all'AGENAS, con funzioni di supporto ai professionisti sanitari, ai medici e agli utenti delle Case della Comunità, mediante suggerimenti non vincolanti. L'AGENAS assume la titolarità del trattamento dei dati e deve specificare i tipi di dati trattati e le misure di sicurezza adottate.

Le disposizioni degli articoli 7-10 della legge n. 132/2025 delineano quindi — a opinione dello scrivente — un quadro ispirato a tre princìpi: antropocentrismo, che subordina la tecnologia alla persona; complementarità, che configura l'AI come supporto e non sostituzione del giudizio umano; proporzionalità, che richiede bilanciamento tra innovazione e tutela dei diritti. Sul piano costituzionale, la disciplina attua il combinato disposto degli articoli 2, 3, 32 e 33 Cost., declinando il diritto alla salute in una dimensione che ricomprende il diritto a beneficiare dell'innovazione tecnologica senza sacrificare l'autodeterminazione terapeutica. Il principio di uguaglianza impone che l'AI non introduca discriminazioni nell'accesso alle cure, mentre la libertà della ricerca scientifica giustifica forme semplificate

di trattamento dei dati purché accompagnate da garanzie adeguate. Resta aperta, nella fase attuativa, la sfida di tradurre questi princìpi in prassi operative coerenti attraverso l'emanazione dei decreti ministeriali e l'implementazione della piattaforma nazionale.

In prospettiva de iure condendo, appare imprescindibile l'elaborazione di protocolli standardizzati per la disclosure algoritmica, che consentano al paziente una comprensione effettiva dei meccanismi decisionali automatizzati senza sacrificare i legittimi interessi relativi alla proprietà intellettuale dei sistemi proprietari. L'equilibrio tra trasparenza tecnologica e tutela del segreto industriale costituisce infatti uno snodo essenziale per garantire la tenuta del modello costituzionale di consenso informato. Con un approccio normativo che coniughi innovazione tecnologica e primato della persona umana sarà possibile poi preservare l'essenza del diritto alla salute, evitando che l'automazione decisionale trasformi la medicina da scienza umana incentrata sulla relazione fiduciaria a mera applicazione di protocolli algoritmici impersonali.

Note bibliografiche

Adinolfi A. (2022), *L'intelligenza artificiale tra rischi di violazione dei diritti fondamentali e sostegno alla loro promozione: considerazioni sulla (difficile) costruzione di un quadro normativo dell'Unione*, in AA.VV., *Intelligenza artificiale e diritto: una rivoluzione?*, Bologna, Il Mulino.

Alpa G. (ed.) (2020), *Diritto e intelligenza artificiale*, Pisa, Pacini Giuridica.

Astromskė K.-Peičius E.-Astromskis P. (2020), *Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations*, in *AI and society*, 36/2, p. 509 ss.

Baldini V. (2015), *La concretizzazione dei diritti fondamentali. Tra limiti di diritto positivo e sviluppi della giurisprudenza costituzionale e di merito*, Napoli, Editoriale Scientifica.

Bonilini B. (2006), *Testamento per vita e amministrazione di sostegno*, in AA. VV., *Testamento biologico: riflessioni di dieci giuristi*, Milano, IlSole24Ore Edizioni.

Buzzelli D.-Palazzo M. (ed) (2022), *Intelligenza artificiale e diritti della persona*, Pisa, Pacini Giuridica.

Callipari N. (2012), *Il consenso informato nel contratto di assistenza sanitaria*, Milano, Giuffrè.

Calvara R. (2023), *Cosa pensa il Garante privacy italiano dell'uso dell'Intelligenza artificiale nel settore sanitario?*, in *Osservatorio sullo Stato digitale*, 29 novembre 2023.

Camardi V. (ed.) (2022), *La via europea per l'intelligenza artificiale. Atti del convegno del progetto dottorale di alta formazione in scienze giuridiche*, Padova, Cedam.

CARUSI D. (2009), *Tutela della salute, consenso alle cure, direttive anticipate: l'evoluzione del pensiero privatistico*, in *Rivista critica di Diritto Privato*, p. 1 ss.
CILENTO C. (2014), *Oltre il consenso informato. Il dovere di informare nella relazione medico-paziente*, Napoli, Edizioni Scientifiche Italiane.
COLLECCHIA G. (2019), *La relazione medico-paziente nel mondo digitale: una Babilonia?*, in *Recenti Progressi in Medicina*, 110, 9, p. 397 ss.
CORAGGIO C. (2008), *Il consenso informato: alla ricerca dei principi fondamentali della legislazione statale*, in *Giurisprudenza costituzionale*, 53/6, p. 4981 ss.
DASSANO F. (2006), *Il consenso informato al trattamento terapeutico tra valori costituzionali, tipicità del fatto di reato e limiti scriminanti*, Torino, Giappichelli.
DI COSTANZO C. (2023), *L'impiego delle nuove tecnologie nel settore della salute: problematiche e prospettive di diritto costituzionale*, in *Consulta Online*.
DI FEDERICO G. (2020), *Digitalizzazione dei servizi sanitari e Unione Europea: uno sguardo d'insieme sullo stato dell'arte e sulle sfide future*, in BOTTARI C. (ed.), *La salute del futuro. Prospettive e nuove sfide del diritto sanitario*, Bologna, Bononi University Press.
EYSENBACH G. (2001), *What's e-Health*, in *JMIR Publications*, 3, 2 (doi:10.2196/jmir.3.2.e20).
FAINI F. (2023), *Intelligenza artificiale e regolazione giuridica: il ruolo del diritto nel rapporto tra uomo e macchina*, in *Federalismi*, 2/23.
FARINA M. (2019), *Il cloud computing in ambito sanitario tra security e privacy*, Milano, Giuffrè.
FAZIO E. (2023), *Intelligenza artificiale e diritti della persona*, Napoli, Edizioni Scientifiche Italiane.
FERRANDO G. (2018), *Rapporto di cura e disposizioni anticipate nella recente legge* in *Rivista Critica di diritto privato*, p. 47 ss.
FIORIGLIO G. (2021), *eHealth: tecnologie, diritto e salute*, in CASADEI T.-PIETROPAOLI S. (ed.), *Diritto e tecnologie informatiche. Questioni di informatica giuridica, prospettive istituzionali e sfide sociali*, Milano, Wolters Kluwer.
FLICK G.M. (2018), *Dignità del vivere e dignità nel morire: un (cauto) passo avanti*, in VERDUCI V. (ed) *Il diritto sulla vita*, Pisa, Pacini Giuridica.
FROSINI T.E. (2022), *L'orizzonte giuridico dell'intelligenza artificiale*, in *BioLaw Journal - Rivista di BioDiritto*, 22, p. 155 ss.
MARINI G. (2010), *Il consenso*, in RODOTÀ S.-ZATTI P. (ed.), *Trattato di biodiritto Ambito e fonti del biodiritto*, Milano, Giuffrè.
GABBRIELLI E.-RUFFOLO U. (2019), *Intelligenza Artificiale e diritto*, in *Giurisprudenza Italiana*, VII, p. 1689 ss.
GRAZIADEI M. (2011), *Il consenso informato e i suoi limiti*, in LENTI L.-PALERMO FABRIS E.-ZATTI P. (ed.), *Trattato di biodiritto. I diritti in medicina*, Milano, Giuffrè.
MASCHIO F. (2021), *Autodeterminazione informativa, disposizioni anticipate di trattamento, cittadinanza digitale: evoluzione o frammentazione del*

quadro giuridico?, in *Rivista Internazionale di Diritto privato Antico e Contemporaneo*, XXV, p. 529 ss.
MITTELSTADT B., *The Impact of Artificial Intelligence on the Doctor-Patient Relationship*, Consiglio d'Europa, dicembre 2021.
MORANA D. (2025), *La salute come diritto costituzionale*, Torino, Giappichelli.
MORANA D.-BALDUZZI T.-MORGANTI F. (2022), *La salute «intelligente»: eHealth, consenso informato e principio di non discriminazione*, in *Federalismi.it*, 34.
MORANA D. (2008), *A proposito del fondamento costituzionale per il consenso informato ai trattamenti sanitari: considerazioni a margine della sent. n. 438 del 2008 della Corte costituzionale*, in *Giurisprudenza costituzionale*, 53, 6, p. 4678 ss.
NISTICÒ A. (2023), *Il consenso al trattamento è valido se le informazioni fornite sono trasparenti*, in *IUS Privacy*, 19 dicembre 2023.
PAJNO A. ET AL. (2019), *AI: profili giuridici. Intelligenza Artificiale: criticità emergenti e sfide per il giurista*, in *BioLaw Journal - Rivista di BioDiritto*, 3, p. 19 ss.
PASCUZZI G. (2025), *Il diritto dell'era digitale*, Bologna, Il Mulino.
PERLINGIERI P. (2020), *Il diritto civile nella legalità costituzionale secondo il Sistema italo-europeo delle fonti*, Vol. II, *Fonti e interpretazione*, Napoli, Edizioni Scientifiche Italiane.
POLITI F. (2021), *Libertà costituzionali e diritti fondamentali. Casi e materiali. Un itinerario giurisprudenziale*, Torino, Giappichelli.
RESCIGNO F. (2018), *L'autodeterminazione della persona: il faticoso cammino del diritto positivo*, in *Rivista Critica di diritto privato*, 1, p. 13 ss.
RICCIO M.-ZICCARDI G.-SCORZA G. (eds) (2022), *Intelligenza artificiale, Profili giuridici*, Coll. Legal Research Groups - ELSA Italia.
RODOTÀ S. (2006), *La vita e le regole: tra diritto e non diritto*, Milano, Feltrinelli.
RODOTÀ S. (2010), *Perché laico*, Roma-Bari, Laterza, 2010.
RODOTÀ S. (2015), *Il diritto di avere diritti*, Napoli, Laterza.
RUFFOLO U. (ed.) (2020), *Intelligenza artificiale. Il diritto, i diritti, l'etica*, Milano, Giuffrè Francis Lefebvre.
STANZIONE P. (2017), *Testamento biologico e autodeterminazione della persona*, in *Comparazione e Diritto Civile*, 3, p. 7 ss.
TADDEI ELMI G.-CONTALDO A. (ed) (2020), *Intelligenza artificiale. Algoritmi giuridici. Ius condendum o "fantadiritto"?*, Pisa, Pacini Giuridica.
TOZZI A.E.-CINELLI G. (2021), *Informed consent and artificial intelligence applied to RCT and Covid-19*, in *BioLaw Journal*, 2/21, p. 97 ss.
VICARELLI G.-BRONZINI M. (2019), *Sanità digitale. Riflessioni teoriche ed esperienze applicative*, Bologna, Il Mulino.

IN GUISA DI CONCLUSIONE

DIALOGO 'CONCERTATO'. QUELLO CHE L'INTELLIGENZA ARTIFICIALE NON PUÒ FARE

di *Fabrizio Von Arx* ed *Enrico Reggiani*

* *Il dialogo si apre con il Maestro VON ARX che esegue, con lo Stradivari "The Angel" - ex Madrileno 1720, il Preludio dalla Partita n. 3 in mi maggiore BWV 1006 di J.S. Bach (1720).*

REGGIANI: Buonasera a tutti. Caro Maestro, mi pare che il pubblico qui raccolto, alla fine di questa giornata faticosa ma illuminante abbia apprezzato questo frutto dell'intelligenza artificiale dell'inizio del XVIII secolo. Giusto? Non capisco perché qualcuno stia ridendo. Il Professore Della Morte, qui davanti, ride. Non è una provocazione. Adesso capirete perché questo nostro momento si intitola: *Quello che l'intelligenza artificiale non può fare*. Ed effettivamente quello che ha appena eseguito il Maestro Von Arx è qualcosa che l'intelligenza artificiale può organizzare ma non realizzare compiutamente.

Vi racconto qualcosa che fa riferimento alla presenza di un'intelligenza che, come quella artificiale di cui vi siete occupati, simula i processi mentali e li organizza giungendo, per così dire, sulla soglia di ciò che oggi noi chiamiamo 'intelligenza artificiale' (e che qualcuno definisce — a mio avviso, con qualche ragione — 'fake intelligence', contrapponendola a 'true intelligence').

Pensate a un ampio arco di tempo (pur con le ovvie differenze e specificità) tra la seconda metà del XVI secolo e la seconda metà del XVIII. Anche allora l'essere umano si preoccupava di gestire una cosa che i musicisti conoscono bene: anzi, non solo i musicisti, ma anche i giuristi, perché anche chi agisce in un'aula di tribunale sa bene quanto è importante governare la strategia retorica e gestire la ricezione (affettiva, intellettiva, ecc.) di ciò che ha appena detto, sta per dire o dirà. Soprattutto a partire dall'epoca di Bach (1685-1750), ci furono anche teorici e compositori (soprattutto nell'area di lingua tedesca) che elaborarono e praticarono sistemi retorici applicati alla

creatività musicale che potevano prevedere addirittura il lancio di dadi connessi a tali sistemi. Cosa c'è di più immaginifico, improbabile e inatteso di questa prassi? Dall'esito di tale lancio dipendeva l'organizzazione dei paradigmi musicali fondamentali (melodia, armonia, ritmo/metro, timbro) e le relative concretizzazioni compositive in brani che, certo, non potevano essere caratterizzati dal maggior tasso possibile di creatività, ma rappresentavano o imitavano, in maniera semplificata ma efficace, anche il processo mentale di colui che, secondo quell'estetica, voleva produrre una determinata esperienza in chi eseguiva e/o ascoltava.

Volete qualche nome? Se quello di Kirnberger (1721-1783) è senza dubbio sconosciuto alla maggioranza di noi, assai più noto è sicuramente quello del più famoso dei figli di Johann Sebastian Bach: Carl Philipp Emanuel Bach (1714-1788). Sono invece certo che rimarrete sorpresi di sentirmi menzionare come terzo in questa triade di esploratori di una sorta di intelligenza artificiale *ante litteram* colui che l'opinione comune considera l'incarnazione di una creatività naturale, cristallina e irripetibile: Mozart (1756-1791), al quale viene tuttora attribuito una sorta di *Gioco di dadi musicali* (catalogo: K. 516f) grazie al quale, a partire da alcune battute di musica, se ne possono determinare altre proprio mediante il lancio di dadi in una sorta di *attività compositiva fai-da-te* che non richiede competenze musicali.

Vogliamo fare un balzo acrobatico dalle nostre parti? Immagino che molti tra noi conoscano quegli incredibili musicisti e compositori che erano Emerson, Lake & Palmer (*ELP*), ai quali dobbiamo, ad esempio, una pietra miliare del *progressive rock* che si intitola *Pictures at an Exhibition* (1970-1971) e che ovviamente fa riferimento al capolavoro pianistico (1874) di Mussorgsky (1839-1881). Forse ricordate che, nelle loro esecuzioni *live* di questo album, c'era sempre un momento in cui Emerson ingaggiava una terribile lotta con il suo organo Hammond, una splendida macchina in grado di produrre materia sonora di livello straordinario, durante la quale giungeva addirittura a capovolgerlo, lanciarlo di lato sul palco e pugnalarlo, inserendo lame tra i tasti. Bene, io ho sempre attribuito alla sua lotta con il suo avanzatissimo (per quegli anni) strumento un forte valore simbolico che trovo coerente con il Convegno al quale noi, Von Arx e il sottoscritto, stiamo offrendo questo insolito contributo: era come se Emerson intendesse dire all'organo, a chi lo costruiva, alla cultura che lo aveva prodotto, all''intelligenza artificiale' che lo gestiva: «tu non mi limiti, io vado oltre. Tu puoi simulare tutto quello che ritieni,

e il tuo costruttore ti ha costruito per questo, ma sappi che al di là di una certa soglia non arrivi».

Torniamo al brano che ha aperto questo nostro incontro: il Preludio dalla Partita n. 3 in mi maggiore BWV 1006 di J.S. Bach (1720). Davvero possiamo considerarlo uno dei tanti frutti straordinari del rapporto tra Bach e il sistema tonale, cioè quello straordinario 'sistema di intelligenza artificiale', che consente di produrre musica secondo criteri estremamente codificati e che le varie epoche dell'esperienza culturale e musicale hanno interpretato e ri-codificato poi in modo differente. Il sistema tonale è quella cosa che ciascuno di noi ha nel proprio DNA: non è naturale. Riuscite a cantare insieme a me la scala di do? [*Il pubblico canta: do-re-mi-fa-sol-la-si-do*, NdA]. Perfetto. Non è detto che ciascuno di noi conosca la teoria che gestisce la distanza tra i singoli gradi di questa scala: tuttavia, questo 'sistema' — che si chiama 'sistema' perché tiene insieme tutti i paradigmi dell'esperienza musicale — l'abbiamo ormai nel nostro DNA ed è quella cosa che, prodotta all'inizio del '700, ha strutturato l'esperienza musicale fino ai nostri giorni, e oltre. Non c'è canzone di Sanremo che non sia costruita secondo la logica di questo 'sistema'.

È una logica stringente, vero, Maestro? Lo è perché ha condizionato tutto, compresa la produzione di strumenti musicali. Anche l'invenzione del pianoforte, dovuta a un italiano, Cristofori, è legata a quel sistema che si stava imponendo e stabilizzando e che poi verrà codificato, ad esempio, nel *Trattato di armonia* di Rameau (1722). Chi conosce *Il clavicembalo ben temperato* bachiano sa bene che anche il suo primo volume è stato pubblicato nel 1722. A differenza di Rameau, che fu comunque anche grande compositore, Bach non esponeva ed elaborava le sue idee in trattati 'verbali': le proponeva in opere che costituiscono veri e propri trattati 'sonori' quali, appunto, *Clavicembalo ben temperato*, *Offerta musicale*, *Arte della fuga*, in cui il sistema tonale raggiunge vertici straordinari. Queste ultime possono essere considerate l'esito personale di uno straordinario sistema di simulazione dei processi mentali in musica che emerge compiutamente all'inizio del '700, in piena temperie illuministica, nel periodo dell'enciclopedia di Diderot e D'alambert, nel momento in cui si pensa di dover ricodificare enciclopedicamente la cultura universale — esperienza sonora e musicale incluse — e si produce quel 'sistema tonale' che ciascuno di noi sperimenta tuttora nel nostro apparato uditivo e fonatorio. È d'accordo, Maestro Von Arx?

VON ARX: Buonasera a tutti. Le viole da gamba, nel corso del tempo, cedettero il passo ai violini, strumenti che proprio in quest'epoca prendono la forma che oggi conosciamo grazie al genio di Stradivari, nato intorno al 1644 e scomparso nel 1737. Una vita lunga e ricca di scoperte: ogni anno la ricerca ci svela un nuovo segreto legato alla sua arte, come la scelta di specifiche resine per la vernice o l'utilizzo esclusivo di abeti rossi provenienti dal versante sinistro delle foreste dolomitiche. Basta osservare gli strumenti di quel periodo — non solo di Stradivari, ma anche di Amati, Guarneri, Bergonzi e degli altri maestri della liuteria italiana — per riconoscere una bellezza e una raffinatezza sono tutti strumenti meravigliosi, tutti hanno una vernice fantastica ed un legno straordinario.

Dopo quarant'anni di esperienza e di variazioni sulle forme apprese dal suo Maestro Amati, Stradivari — forte di un'immensa quantità di conoscenze accumulate — rivoluziona la liuteria: allunga la cassa del violino di cinque o sei millimetri, la allarga di due, creando una cassa di risonanza che, per la sua perfezione, può essere paragonata a una sorta di intelligenza artificiale *ante litteram*, capace di restituire la complessità e la ricchezza del sistema tonale con una fedeltà quasi vocale. Non si tratta più soltanto di una traduzione lineare del suono del clavicembalo: il violino stradivariano è uno strumento espressivo, capace di incarnare molteplici personalità, quasi come se racchiudesse ventiquattro violini in uno solo: proprio quante sono le tonalità nel sistema musicale. Un violino capace di tradurre il 'carattere', l'espressione intrinseca, di ogni tonalità.

Ma qual è il vero segreto di Stradivari? In Italia, proprio in quegli anni, si verifica una straordinaria coincidenza: vivono, contemporanei di Stradivari, violinisti-compositori quali Geminiani, Locatelli, Veracini, Tartini, Corelli, Vivaldi che stanno non solo componendo di fantasia, ma stanno creando forme nuove di concerto (basti pensare all'invenzione del Concerto Grosso, da solista, di Corelli) e stanno creando una visione armonica innovativa, una nuova estetica musicale riguardante gli strumenti a corda: stanno creando un linguaggio nuovo in cui la musica strumentale diventa indipendente e non più solo di accompagnamento alla voce: diventa essa stessa 'voce'. Questi grandi compositori collaborano con i liutai della scuola cremonese per plasmare quella che diventerà poi la 'macchina perfetta': il violino di Antonio Stradivari, capace di esprimere la grande musica di questi violinisti compositori del XVIII secolo.

Chi suona il violino può suggerire all'artigiano le soluzioni armoniche, i pesi, le esigenze tecniche violinistiche; inoltre, questi violi-

nisti, essendo anche dei compositori e ragionando per grandi numeri armonici, suggeriscono al liutaio anche la risonanza tonale del suono. Il liutaio, dal canto suo, traduce queste indicazioni in materia. Si crea, così, uno strumento dal suono estremamente espressivo, caratteristica propria della scuola cremonese di liuteria. Da questo dialogo serrato tra compositore e artigiano che nasce la perfezione stradivariana: una sinergia di linguaggi e saperi, una collaborazione che trasforma l'artigiano in artista e che segna per sempre la storia della musica occidentale.

REGGIANI: Vorrei sottolineare anche che ci sono 'fattori culturali' (semiotici, preferirei dire) che vanno al di là del fatto che qualcuno può costruire materialmente uno strumento. Il punto mi pare, per dirla in sintesi, che all'inizio del '700 cambia concezione ed esperienza del suono. La matematica è fondamentale nella definizione dei rapporti sonori all'interno delle relazioni fra le note della scala che avete cantato. Quando dico che non è una cosa naturale, ma qualcosa che simula il processo che la natura realizza per produrre i suoni, diciamo che poi ci vuole uno strumento costruito in un certo modo e secondo certi criteri per esprimere qualcosa che sia coerente con tutto il processo. Quindi è ovvio che Bach, non casualmente musicista dell'area tedesca, ha caratteristiche coerenti con l'area culturale da cui proviene, ma la grande scuola italiana a cui faceva riferimento il Maestro Von Arx elabora il 'sistema tonale' condiviso secondo un'altra visione della realtà, altre necessità, altre potenzialità, altre risorse.

VON ARX: Questi violinisti-compositori italiani non componevano solo di fantasia, come detto. Vivaldi nel 1721 creava *Le quattro stagioni* per imitare la natura, ma inventava allo stesso tempo una nuova tecnica di questo strumento. Per la prima volta nella sua formula creativa trova nuove forme. Infatti, prima di Vivaldi, i violini di Amati accompagnavano solo le voci, per esempio un quartetto o un sestetto. Vivaldi porta alla massima espressione il Concerto Grosso, il concerto da solista.

* *Il Maestro VON ARX esegue un frammento dalle Quattro Stagioni vivaldiane*, NdA.

Immaginiamo Vivaldi entrare, quindi, nella bottega di Stradivari, chiedendogli, di costruire uno strumento capace di dar voce alle sue

composizioni, e alla stessa maniera Stradivari, ascoltando la musica di Vivaldi, si lasciava ispirare e a immaginare nuove forme per il suo violino. Il periodo d'oro di Stradivari va dai suoi sessanta agli ottanta anni, con continue sperimentazioni e forme di violino. Durante l'arco della sua vita, Stradivari costruì più di 450 violini e oltre 1000 strumenti, raccogliendo dati, osservando vibrazioni, quasi lavorando con algoritmi *ante litteram*, intrecciando la sua straordinaria esperienza con la fantasia creativa dei compositori.

REGGIANI: Tenete anche conto che il 'nuovo' codice tonale (nel senso simil-giuridico di sistema di regole e segni) colonizza l'esperienza musicale europea (e non solo...) nel suo complesso. Il frammento vivaldiano che avete appena ascoltato, pubblicato nel 1725 ad Amsterdam, implica in realtà non solo un nuovo modo di gestire la 'macchina violinistica', ma la mette al servizio di una rinnovata cultura del 'pensare in musica', del 'mettere insieme' in musica (cioè del 'comporre'), dell''ascoltare musica', ecc. Noi oggi abbiamo sentito un violino protagonista, ma manca tutta l'ensemble musicale. È davvero una rivoluzione tecnologica che inevitabilmente produce una rivoluzione nel gusto, una rivoluzione culturale e impatta anche l'economia della produzione musicale del tempo a 360°, quella degli strumenti, dei compositori e tutto quello che aveva a che fare con la musica. La relazione del musicista passava quasi esclusivamente attraverso la relazione con un mecenate, un committente e quindi capite anche quanto sia significativa e omnicomprensiva questa grande impresa che la cultura europea produce. Non la restringerei a una questione organologica: è davvero una rivoluzione culturale che è omologa delle cose che succedono in letteratura, in economia, nel diritto e nella società nel suo complesso. È quello il momento in cui nasce a Londra la borsa valori, che produce una trasformazione totale nelle modalità di contrattazione del mercato del tempo e rende orgogliosi i londinesi che scrivono « noi siamo il centro del mondo ». Tutto questo senza sistema tonale non sarebbe stato possibile, anche se magari non ce ne rendiamo conto.

Andiamo oltre: io so che tu hai in serbo un'altra esecuzione bachiana: l'Adagio della Sonata n. 1 in sol minore per violino, BWV 1001. Siamo sempre nel 1720. Questa è una sonata per violino solo, quella di prima era una partita: generi compositivi che implicano idee diverse, destinatari diversi, potenzialità diverse, ma lo stesso tipo di codice compositivo.

* *Il Maestro VON ARX esegue l'Adagio della Sonata n. 1 in sol minore per violino, BWV 1001*, NdA.

Suona 'bene', vero? Quanti sistemi compositivi hanno inventato gli esseri umani dalla seconda metà del '500 in avanti, quanti strumenti sono stati inventati, modificati, ecc. Tuttavia, non tutti i sistemi compositivi, radicati in sistemi culturali, né tutti gli strumenti hanno avuto la forza di imporsi come lo strumento di Stradivari.

VON ARX: È opportuno soffermarsi su quanto accadde successivamente a quel momento cruciale nella storia della musica. Nel XIX secolo, la pratica violinistica in Italia subì una significativa riduzione, principalmente a causa della predominanza del canto nel teatro d'opera. I principali solisti, tra cui Paganini, furono indotti a migrare verso altri paesi europei. Da sottolineare come importanti violinisti compositori quali Sarasate, Wieniawski, Ysaÿe e Kreisler si siano ritrovati a Parigi. La Francia divenne il fulcro di una rinomata scuola francese di liuteria.

REGGIANI: Pertanto, se interpreto correttamente, questi musicisti non si sono limitati a esibirsi, ma hanno continuato a comporre e a influenzare direttamente l'evoluzione degli strumenti. Eppure, nonostante la presenza di liutai contemporanei come Vuillaume, la loro preferenza si è orientata verso gli strumenti del Settecento cremonese. Quali sono, secondo te, le motivazioni di questa scelta?

VON ARX: Ritengo che la ragione fondamentale risieda nel fatto che gli strumenti di Stradivari, Guarneri, Bergonzi e Guadagnini sono stati concepiti con l'intento di rappresentare alla perfezione il sistema tonale. Non si tratta esclusivamente di una questione di tradizione, bensì di una risposta tecnica e strutturale alle esigenze degli interpreti. A titolo di esempio, una tastiera più lunga consente di esplorare registri più acuti, mentre l'arco con una curva differente risponde a nuove necessità espressive, come l'avvento del pianoforte. Gli strumenti della scuola cremonese sembravano anticipare tali trasformazioni, come se Stradivari avesse già intuito le evoluzioni future della prassi esecutiva e le aspettative di un diverso modo di suonare. Ma soprattutto erano nati con l'intento di esprimere le emozioni delle varie tonalità; tipica peculiarità dell'intelligenza di un compositore e degli strumenti storici italiani.

* *Il Maestro VON ARX esegue un frammento dal Concerto in re maggiore op. 35 per violino e orchestra (1878) di CAJKOVSKIJ (1840-1893)*, NdA.

REGGIANI: È stupefacente l'espressività di Čajkovskij. Quelli che menzionavi erano violinisti-compositori oppure compositori che lavoravano in *joint venture* con violinisti/virtuosi e liutai? Cosa che oggi, è piuttosto complicato rivedere perché la specializzazione porta il compositore a fare il compositore e il *performer* a fare il *performer*. Guardando al passato, pare fondamentale che Stradivari e il sistema tonale siano riusciti a essere protagonisti di un'esperienza che va dalla fine del 600 ai giorni nostri.

VON ARX: Nell'ambito della nostra attività esecutiva, affrontiamo anche opere di compositori come Šostakovič e Prokof'ev risalenti al secolo scorso, mantenendo costantemente la capacità di cogliere la tonalità, il carattere e la drammaticità propri di ciascun brano. Il processo di traduzione dell'espressione e del sentimento si configura come una funzione fondamentale dello strumento, il quale, originariamente concepito per imitare la voce umana, si avvicina in modo sorprendente a tale obiettivo. Il violino o gli strumenti musicali erano le 'intelligenze artificiali' della voce umana. La volevano imitare ad ogni costo. Ci sono andati vicino ma la materia della voce contro 'la rigida' del legno, forse vince sempre. Quindi lo strumento sviluppa un proprio linguaggio, che abbraccia il sistema tonale, quello sinfonico e quello della scuola di Mannheim.

Queste peculiarità risultano essenziali per comprendere le attuali applicazioni dell'intelligenza artificiale nell'ambito musicale e creativo, campo nel quale l'IA aspira a intervenire e ad assumere il ruolo di artista autonomo. Resta da verificare se tali sistemi riusciranno ad avvicinarsi alla sensibilità umana, ma è indispensabile instaurare un dialogo critico con queste nuove tecnologie. È opportuno sottolineare che lo strumento, nello specifico il violino della grande scuola cremonese, è in grado di tradurre con estrema precisione anche composizioni del XX secolo e sicuramente quelle future.

Un esempio: il preludio della suite per violino solo (1958) di Bloch (1880-1959) [* *appena eseguito*, NdA], in cui si evidenzia l'eccezionale equilibrio vibratorio delle corde e l'arduo gioco armonico che tende verso il nuovo. In particolare, nell'esecuzione una corda vuota tende, in strumenti convenzionali, a predominare sulla nota

suonata in bicordo, mentre negli Stradivari e in tutti i violini caratterizzati da tale equilibrio acustico, la distribuzione degli armonici risulta perfettamente bilanciata. È possibile percepire distintamente tre o quattro note simultanee, oppure una combinazione di voci (basso, tenore, baritono, soprano), ciascuna con la propria altezza, in un equilibrio armonico ideale. Da una parte per l'interprete, questa condizione rappresenta una straordinaria facilitazione esecutiva, dall'altra questi strumenti eccezionali potranno 'interpretare' tutti i capolavori della storia della musica, perché nati per riprodurre l'esperienza espressiva ed emotiva caratteristica di quei grandi violinisti-compositori che ricercavano, insieme ai liutai dell'epoca, questa specifica capacità polifonica.

REGGIANI: L'aspetto interessante è che il Preludio di Bloch è pensato secondo un differente sistema compositivo. È la scelta di qualcuno che, a partire dal proprio rapporto con il sistema tonale, prova a elaborare un sistema diverso. Giustamente ha osservato il Maestro che il tipo di strumento da lui utilizzato ha un equilibrio interno che prescinde persino dal modo in cui il violinista lo gestisce e che rende coerente la musica che viene eseguita. Noi siamo abituati a pensare la musica come esperienza soggettiva — sui piani intellettivo ed emotivo — ma non dovremmo dimenticare che la musica è anche, forse innanzitutto, esperienza culturale che, grazie al sistema che la rende possibile, dice qualcosa del mondo da cui proviene.

Questo mi consente di dire, caro Fabrizio, che è il momento di passare a quell'altro argomento connesso a questo Convegno, cioè a ciò che il nostro tempo pensa sia possibile fare affidando la musica (con il processo compositivo che la genera simulando un processo mentale e sensoriale) a una macchina concepita per riprodurre e simulare dei processi mentali umani. Questa macchina, Aiva, è il primo compositore di intelligenza artificiale. Le avevano chiesto di ''generare'' un monologo per violino. L'acronimo Aiva sta per *Artificial Intelligence Virtual Artist* e non posso fare a meno di rilevare che ci vuole una bella faccia di bronzo a usare la 'A' finale per 'Artist'. Mi sono informato in proposito: Aiva è iscritta ed è stata riconosciuta dalla società degli autori, compositori ed editori di musica francese.

VON ARX: AIVA ha dato anche un titolo a questo monologo: *Genesis* [*si esegue un frammento di Genesis*, NdA].

Reggiani: Leggo qualche nota su AIVA: « creato nel febbraio del 2016, AIVA si è specializzato nella composizione di musica classica e sinfonica ». Noi potremmo fare i pignoli e chiedere cosa vuol dire 'classica'? Del periodo classico e quindi seconda metà del diciottesimo secolo oppure 'classico' per quello che noi genericamente intendiamo con 'musica classica'? Aiva è diventato il primo compositore virtuale ad essere riconosciuto da quella società francese. Leggendo un'ampia raccolta di brani esistenti di musica classica scritta da compositori umani come Bach, Beethoven e Mozart, Aiva è capace di riconoscere regolarità nella musica e su questa base è in grado di comporre da solo. L'algoritmo è basato su *deep learning* e *reinforcement learning architectures*. Dal gennaio 2019, la società lo offre come prodotto commerciale, chiamato 'Music Engine', in grado di generare brevi (fino a 3 minuti) composizioni in vari stili: rock, pop, jazz, tango, cinematografico, cinese, eccetera. Mi piacerebbe sapere quali repertori le hanno fatto ingurgitare (studiare o leggere mi sembrano verbi impropri) per imparare. Noi qui abbiamo avuto un grande artista che si è messo al servizio della macchina. Ora, immaginate di ascoltarlo riprodotto non dal Maestro Von Arx, ma da una macchina.

Von Arx: Forse le macchine tra loro riescono a comprendersi meglio, anche se non posso esserne certo. Ciò che colpisce, tuttavia, è che nell'algoritmo che sta alla base di tutte queste composizioni, e lo si nota immediatamente, non siano mai stati inclusi, né tantomeno insegnati, i cosiddetti compositori 'minori', il brutto, l'osceno o l'errore. A mio avviso, sarebbe stato opportuno sviluppare anche un algoritmo 'invertito' che possa educare le macchine attraverso l'imperfezione. La mia impressione è che, osservando questa linearità, il risultato sia assimilabile a un esercizio scolastico, certamente privo di errori, ma senza forma, senza espressione, senza 'esperienza' viva.

Potremmo definirlo, senza esagerare, un compito di secondo anno di armonia: corretto, certamente, ma privo di innovazione. La creazione di un nuovo linguaggio, infatti, rimane prerogativa esclusiva dell'essere umano, poiché solo l'uomo, con un salto improvviso e quasi nel vuoto, sa dar vita a quella scintilla che lo porta a comporre (o a dipingere, o a scrivere) qualcosa di autenticamente nuovo e innovativo, pur partendo da solide basi storiche. In questo caso, invece, ci troviamo di fronte unicamente a una sequenza lineare, a un compito che si ripete, a un *déjà vu*. Forse, col tempo, Aiva riuscirà a sviluppare un proprio linguaggio originale; tuttavia, dal punto di vista

creativo, la sua utilità appare limitata e circoscritta solo per chi non conosce la musica di Bach, a chi non sa comporre, o semplicemente, a chi non è in grado di realizzare ciò che l'algoritmo produce. Cosa utile e giusta sarebbe senz'altro 'insegnare' all'AI il senso della storia e dell'esperienza musicale.

REGGIANI: Sarebbe interessante prendere quel brano e cercare di ricondurlo ai modelli che Wikipedia cita nella voce relativa a AIVA. In realtà, nessuno di quei modelli scriverebbe mai musica in quel modo, che peraltro all'inizio ha alcune inflessioni che non sono neppure tonali, ma sembrano far riferimento al gregoriano, oppure a certi brani dell'800 del grande virtuosismo europeo. Il linguaggio sembra modale, un linguaggio diverso. Provate a pensare anche che l'intelligenza artificiale è stata utilizzata per completare composizioni incompiute di un certo compositore (ad esempio Schubert). A quel sistema informatico è stato fatto 'leggere' tutto il repertorio di un determinato compositore e quel sistema ha provato a imitarne i 'principi creativi'. Ecco, secondo me questa pratica — per quanto limitata, perché non c'è la reale scintilla dell'intuizione creativa — può essere utile almeno perché, una volta che l'analisi ha definito i contenuti strutturali di un brano, può poi chiedere alla macchina di replicarli per consentirne una migliore esperienza e comprensione.

Avete presente la scena del film *Amadeus* in cui Salieri compone sotto dettatura di Mozart morente? Immaginate che al posto di Salieri ci sia un robot che, avendo imparato da Mozart, componga lui una conclusione alternativa di un brano di quest'ultimo. Il punto è cosa e come bisognerà 'insegnare' alle macchine, non solo nel campo della musica, ma anche nel campo della letteratura, dell'arte figurativa. È un problema anche culturale, semiotico e, penso di poter dire, anche giuridico. Perché a un certo punto potremmo anche noi completare in questo modo l'*Incompiuta* di Schubert, cominciare a eseguirla, acquisirne i diritti e imporla — perché abbiamo la forza economica — come la vera *Incompiuta*, grazie a un'efficace campagna di marketing e il sostegno di qualche musicologo compiacente.

Sono questioni che possono riguardare ogni ambito dell'esperienza umana. Ho sempre in mente i racconti di Asimov, soprattutto *Io Robot*, in cui viene rappresentato un futuro che forse potremmo perfino auspicare, almeno quando il robot comincia a sviluppare sentimenti umani, riuscendo così a sfuggire alla ragnatela tremenda della sudditanza imposta all'uomo dai robot.

DELLA MORTE: Per onorare il nostro pubblico, e gli illustri ospiti e colleghi tanto illustri, vorrei fare un gesto imprevedibile e domandare al Maestro Von Arx un'ultima esecuzione per chiudere questo confronto con una 'nota alta'.

Grazie a tutti.

Note bibliografiche (a cura di ENRICO REGGIANI)

MINCER A.-TANNOURY A. (2021), *Artificial Intelligence in Music Composition*, in MAGLOGIANNIS I.-MACINTYRE J.-ILIADIS L. (eds), *Artificial Intelligence. Applications and Innovations*, Cham, Springer.

ANESSI P. (2024), *L'intelligenza artificiale al servizio della musica*, Tricase, Youcanprint.

BALABAN M.-EBCIOGLU K.-LASKE O. (eds.) (1992), *Understanding music with AI: perspectives on music cognition*, Cambridge, AAAI Press - MIT Press.

CHRISTENSEN T. (ed.) (2002), *The Cambridge History of Western Music Theory*, Cambridge, Cambridge University Press.

CLANCY M. (ed.) (2023), *Artificial intelligence and music ecosystem*, Abingdon, Routledge.

LINCOLN H.B. (ed.) (1970), *The computer and music*, London, Cornell University Press.

MAZZA E., "*Musica e IA, perché è importante che al centro sia sempre la creatività*", in *LaRepubblica.it*, 5 ottobre 2025 (ultima consultazione: 13 dicembre 2025; *www.repubblica.it/tecnologia/2025/10/05/news/musica_e_ia_perche_e_importante_che_al_centro_sia_sempre_la_creativita-424892403/).*

MIRANDA E.R. (ed.) (2021), *Handbook of Artificial Intelligence for Music. Foundations, Advanced Approaches, and Developments for Creativity*, Cham, Springer.

SCHWANAUER S.M.-LEVITT D.A. (eds.) (1993), *Machine models of music*, Cambridge-London, The MIT Press.

UNGER H. (1941), *Die Beziehungen zwischen Musik und Rhetorik im 16.-18. Jahrhundert*, Würzburg, Triltsch; facs. HILDESHEIM, OLMS G., 1969.

RINGRAZIAMENTI

Come ogni opera collettanea, e ancor più quelle che nascono dal confronto sistematico tra discipline diverse, questo volume è il risultato di un debito intellettuale ampiamente condiviso. Esso si estende ben oltre gli autori dei contributi qui raccolti e coinvolge quanti hanno partecipato, in forme diverse, a seminari, workshop, working papers e altre occasioni di discussione scientifica intorno al progetto di ricerca pluriennale: Funzioni pubbliche / Controllo privato. Profili interdisciplinari sulla governance senza governo della società algoritmica (https:/progetti.unicatt.it/fpcp).

Se il debito è collettivo, alcuni possono tuttavia essere considerati 'creditori privilegiati', poiché senza il loro impegno la ricerca difficilmente avrebbe potuto raggiungere il grado di maturazione che trova espressione in queste pagine.

Un primo e sentito ringraziamento va ai responsabili dei sottogruppi di ricerca dedicati, oltre al diritto (gestito dal sottoscritto), alla comunicazione politica (Prof.ssa Giovanna Mascheroni), e alla sanità (Prof. Antonio Oliva).

Particolare riconoscenza è inoltre dovuta a coloro che hanno contribuito all'organizzazione dei principali momenti di confronto scientifico. Per il convegno tenutosi presso il Policlinico Gemelli di Roma il 4 giugno 2024, si desidera ricordare il Prof. Alain Dell'Osso e la Prof.ssa Anna Fagotti, la cui disponibilità e sensibilità hanno favorito un fecondo dialogo interdisciplinare. Per il convegno conclusivo del 28 e 29 ottobre 2024, ospitato presso l'Aula Magna della sede milanese dell'Università Cattolica e caratterizzato dalla partecipazione di 493 studenti, destinatari di un open badge, un ringraziamento particolare va: alla Dott.ssa Maddalena Baitieri dell'Ufficio Ricerca, per il costante supporto offerto nella gestione amministrativa del progetto; alla Dott.ssa Francesca Sironi De Gregorio e al Dott. Riccardo Ricchetti, per la preziosa collaborazione organizzativa sempre accompagnata da un sorriso; alla Dott.ssa Elena Angeleri e alla Dott.ssa Marta Cascio per la dedizione e l'entusiasmo; nonché a

tutto l'*Humane Technology Lab* che, sotto la direzione del Prof. Giuseppe Riva, si è confermato un luogo privilegiato di incontro e dialogo tra saperi.

Un pensiero riconoscente è rivolto inoltre al compianto Prof. Franco Anelli, che nel suo ruolo di Magnifico Rettore ha sostenuto con lungimiranza lo sviluppo della ricerca multidisciplinare in Ateneo. Analoga gratitudine va espressa alla Prof.ssa Elena Beccalli, che prima nell'ambito della Commissione Strategie di Ricerca e successivamente come Magnifico Rettore ha costantemente supportato il progetto.

Come sempre, eventuali errori, omissioni o imprecisioni, che certamente non mancheranno ma che oramai è possibile interpretare come una sorta di 'garanzia di autenticità' del testo, restano attribuibili esclusivamente al curatore.

GABRIELE DELLA MORTE

ISBN 978-88-28-88588-7

€ 49,00

024228603