2025

SICUREZZA E SCIENZE SOCIALI

ANNO XIII
N. 2/2025

Le applicazioni dell'intelligenza artificiale tra potenzialità e rischi

a cura di

Roberto Cipriani, Sara Sbaragli

ISSN 2283-8740, ISSNE 2283-7523
OPEN ACCESS

SICUREZZA E SCIENZE SOCIALI

La rivista esce sotto l'alto patrocinio
dell'Università degli Studi di Perugia

Con il patrocinio del
Comune di Narni

La rivista si propone di sostenere e di dare voce alle esigenze e alle istanze pluralistiche dei Corsi di laurea universitari che, nel contesto italiano, affrontano in maniera specifica le tematiche di carattere criminologico.

Alma Mater Studiorum – Università di Bologna
Laurea Magistrale in "Scienze criminologiche per l'investigazione e la sicurezza"

Università degli Studi Magna Græcia di Catanzaro
Laurea Magistrale "Organizzazioni e mutamento sociale"

Università Cattolica del Sacro Cuore
Laurea Magistrale in "Politiche pubbliche - curriculum Politiche per la sicurezza"

Università degli Studi "G. d'Annunzio" Chieti – Pescara
- Laurea Triennale "Sociologia e criminologia"
- Laurea Magistrale "Ricerca Sociale, politiche della sicurezza e criminologia"

Università degli Studi di Perugia
- Laurea Triennale "Scienze per l'investigazione e la sicurezza"
- Laurea Magistrale "Scienze Socio-antropologiche per l'integrazione e la sicurezza sociale"

SICUREZZA E SCIENZE SOCIALI

Direttrice *Sabina Curti* (Università degli Studi di Perugia)

Comitato Direttivo *Fabrizio Fornari* (Università degli Studi "G. D'Annunzio" Chieti-Pescara), *Christophe Dubois* (Université de Liège), *Maria Cristina Marchetti* (Università di Roma "La Sapienza"), *Giovanna Truda* (Università degli Studi di Salerno), *Philippe Combessie* (Université Paris Nanterre)

Comitato Scientifico *Costantino Cipolla* (Università di Bologna), *Philippe Combessie* (Université Paris Nanterre), *Christophe Dubois* (Université de Liège), *Lucio d'Alessandro* (Università Suor Orsola Benincasa, Napoli), *Maria Caterina Federici*† (Università degli Studi di Perugia), *Fabrizio Fornari* (Università degli Studi "G. D'Annunzio" Chieti-Pescara), *Tito Marci* (Università di Roma "La Sapienza"), *Dario Melossi* (Università di Bologna), *Massimiliano Mulone* (Université de Montréal, Centre International de Criminologie comparée), *Miguel Angel Nunez Paz* (Universidad de Huelva, ES), *Franco Prina* (Università di Torino), *Monica Raiteri* (Università di Macerata), *Annamaria Rufino* (Università della Campania), *Ernesto Ugo Savona* (Università Cattolica del Sacro Cuore, Milano), *Raffaella Sette* (Università di Bologna), *Francesco Sidoti* (Università dell'Aquila), *Jan Spurk* (Université Paris Descartes Sorbonne), *Susanna Vezzadini* (Università di Bologna), *Emilio Viano* (American University - Washington, DC)

Comitato Editoriale *Andrea Antonilli* (Università degli Studi "G. D'Annunzio" Chieti-Pescara), *Andrea Bilotti* (Università di Roma Tre), *Andrea Borghini* (Università di Pisa), *Francesco Calderoni* (Università Cattolica del Sacro Cuore, Milano), *Uliano Conti* (Università degli Studi di Perugia), *Luca Corchia* (Università degli Studi "G. D'Annunzio" Chieti-Pescara), *Fabio D'Andrea* (Università degli Studi di Perugia), *Maurizio Esposito* (Università degli Studi di Cassino), *Stefania Ferraro* (Università Suor Orsola Benincasa, Napoli), *Silvia Fornari* (Università degli Studi di Perugia), *Enrico Gargiulo* (Università di Bologna), *Rosita Garzi* (Università degli Studi di Perugia), *Maria Grazia Galantino* (Università di Roma "La Sapienza"), *Maria Cristina Marchetti* (Università di Roma "La Sapienza"), *Cirus Rinaldi* (Università di Palermo), *Emanuele Rossi* (Università di Roma Tre), *Chiara Scivoletto* (Università di Parma), *Anna Simone* (Università di Roma Tre), *Giovanna Truda* (Università degli Studi di Salerno), *Francesca Vianello* (Università di Padova), *Simone D'Alessandro* (Università degli Studi "G. D'Annunzio" Chieti-Pescara), *Sara Sbaragli* (Istituto di Scienze e Tecnologie della Cognizione – ISTC)

Comitato etico *Luca Corchia* (Università degli Studi "G. D'Annunzio" Chieti-Pescara), *Maurizio Esposito* (Università degli Studi di Cassino), *Francesco Sidoti* (Università dell'Aquila), *Annamaria Rufino* (Università della Campania), *Silvia Fornari* (Università degli Studi di Perugia)

Redazione *Jennifer Malponte* (Università degli Studi "G. D'Annunzio" Chieti-Pescara)

Segreteria redazionale redaz.sicurezzascienzesociali@gmail.com

Sommario

Introduzione. Le applicazioni dell'intelligenza artificiale tra potenzialità e rischi, *Roberto Cipriani, Sara Sbaragli* pag. 7

Articoli

Rischio, intelligenza artificiale e fisica quantistica. La logica del coltello, *Francesco Mattioli* » 13

Intelligenza artificiale vs intelligenza popolare. Gli italiani sulle armi "autonome" tra intuizioni e paure, *Fabrizio Battistelli, Francesca Farruggia* » 23

Opportunità e ostacoli nella formazione sull'intelligenza artificiale: una prospettiva socio-educativa, *Massimiliano Ruzzeddu, Laura Guercio* » 35

AI, media influence, and the legal challenges of freedom of speech, *Katarzyna Chałubińska-Jentkiewicz, Urszula Soler* » 46

La narrazione della politica secondo IA. Il caso del conflitto russo-ucraino, *Rossana Sampugnaro, Simone Marino* » 56

The silent revolution of algoagents: relational reconfigurations and emerging paradigms, *Edmondo Grassi* » 69

Between artificial maturity and responsible intelligence: an analysis of the innovation process by questioning ChatGPT 4o and Open AI 01, *Francesco Mastrocola, Roberto Aufieri, Dario Quattromani* » 80

The future of autonomous vehicles: expectation and worldviews of the stakeholder by Sensitivity Auditing. Interviewing the modelers and sociologists, *Sara Sbaragli* » 94

Individuare sistemi data-intensive nella Pubblica Amministrazione italiana: uno strumento basato su web scraping e machine learning, *Enrica Amaturo, Biagio Aragona, Suania Acampa, Roberto Artiaco* » 109

Artificial intelligence and crime: exploring AI's dual role in law enforcement and criminal misuse, *Vera Kopsaj* » 120

Crime mapping. Rischi e benefici dell'uso dell'intelligenza artificiale nel contrasto e prevenzione al crimine, il caso XLAW, *Alberto Pesce* » 131
I tentativi di regolamentazione dell'intelligenza artificiale: tra approcci differenti, contraddizioni e implicazioni sociali in Europa, Stati Uniti e Cina, *Giovannipaolo Ferrari, Paolo Diana* » 143
Malicious use of Artificial Intelligence (MUAI): l'uso malevolo dell'intelligenza artificiale nell'ecosistema cyber-sociale, *Arije Antinori* » 156
Artificial social research? AI's capabilities and risks in predicting values and attitudes, *Caterina Ambrosio, Ciro Clemente De Falco, Domenico Trezza* » 167
The algorithm as judge: predictive justice, decision-making power and digital inequalities, *Vera Kopsaj, Sara Sbaragli* » 183

Fuori tema
Sicurezza e benessere del cittadino nell'era digitale. Una lettura sociologica*, Ida Cortoni* » 194

Recensioni
Valeria Rosato (2025), L'Altro come oggetto d'odio. Per una sociologia dei conflitti contemporanei, *Giorgia Mavica* » 218

Introduzione.
Le applicazioni dell'intelligenza artificiale tra potenzialità e rischi

di *Roberto Cipriani*[*], *Sara Sbaragli*[**]

Con il termine intelligenza artificiale si indica "una famiglia di tecnologie in rapida evoluzione che può contribuire al conseguimento di un'ampia gamma di benefici a livello economico e sociale nell'intero spettro delle attività industriali e sociali". Così afferma la Commissione europea, nell'ultimo testo di compromesso della proposta di Regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (COM 2021/206, AI Act).

L'intelligenza artificiale sta rivoluzionando numerosi settori, apportando miglioramenti significativi in termini di efficienza, precisione e innovazione. Le sue applicazioni sociali spaziano dalla medicina alla finanza, al settore dei trasporti, della formazione e della sicurezza e in ambito militare, trasformando i mondi di vita, del lavoro e gli scenari di guerra (Abeti *et al*., 2021; Nocenzi *et al*., 2022; Abduljabbar *et al*., 2019; Lau, Staccini, 2019; Bertino *et al*., 2021).

D'altro canto, l'analisi dei *big data*, le tecniche di *machine learning* e i modelli predittivi stanno migliorando la comprensione dei fenomeni sociali (Airoldi, 2022). Una delle applicazioni più rilevanti dell'IA in sociologia è proprio l'analisi dei *big data*. In particolare, l'IA consente di analizzare dati su larga scala per identificare *trend* e *pattern* di comportamento, sentimenti e relazioni sociali, che possono fornire intuizioni preziose sulla struttura sociale e sulle dinamiche di gruppo (Analisi dei *big data* e *machine learning*). L'intelligenza artificiale viene utilizzata anche per sviluppare modelli predittivi e simulazioni sociali. Questi modelli possono prevedere l'evoluzione di fenomeni sociali come i movimenti migratori, l'urbanizzazione o la diffusione di idee e comportamenti attraverso le reti sociali. Le simulazioni replicano il comportamento di intere popolazioni in

DOI: 10.5281/zenodo.17296082

[*] Università Roma Tre. roberto.cipriani@tlc.uniroma3.it.

[**] ISTC-CNR e Università di Napoli Federico II. sarasbaragli@gmail.com.

ambienti virtuali (Modelli predittivi e Simulazioni sociali). Inoltre, le tecniche di analisi delle reti sociali, combinate con gli algoritmi di *machine learning*, consentono di studiare in modo dettagliato le relazioni tra individui, gruppi e organizzazioni (*Social Network Analysis*).

Dal punto di vista etico, l'uso dell'IA solleva questioni fondamentali sulla *privacy* e la sorveglianza sui dati. La raccolta e l'analisi di dati personali su larga scala sollevano preoccupazioni riguardanti la protezione della *privacy* degli individui e il potenziale uso improprio delle informazioni. Inoltre, si deve prestare attenzione all'uso dell'IA nella sorveglianza di massa e nelle decisioni che potrebbero influenzare la vita delle persone, come quelle legate all'assistenza sanitaria o alla giustizia (Manheim, Kaplan, 2019; Tucker *et al*., 2018; Curzon *et al*., 2021).

L'IA solleva anche interrogativi sull'equità e la discriminazione. Alcuni algoritmi di IA possono essere intrisi di pregiudizi, poiché si basano sui dati storici, che possono riflettere disuguaglianze sociali esistenti. È necessario garantire che l'IA sia sviluppata e implementata in modo etico, affrontando questi *bias* e promuovendo la diversità e l'inclusione (Airoldi, 2020; Lutz, 2019).

Questo numero della rivista raccoglie contributi che evidenziano alcune possibili applicazioni dell'IA nell'ambito della sicurezza, sia come oggetto di ricerca che come strumento di analisi.

I saggi presentano una panoramica articolata e critica sul ruolo dell'intelligenza artificiale nelle società contemporanee, toccando aspetti che spaziano dalla sicurezza pubblica alla regolamentazione normativa, fino alla ricerca sociale e la trasformazione delle istituzioni e della guerra.

La prima linea tematica che si evince evidenzia come viviamo in una società profondamente segnata dall'incertezza e dalla complessità, dove i concetti tradizionali di verità, sicurezza e rischio sono stati ridefiniti. L'IA, in questo contesto, si configura come uno strumento cruciale per affrontare scenari imprevedibili, fornendo soluzioni e supporti decisionali, ma anche sollevando interrogativi sull'affidabilità, sulla trasparenza degli algoritmi e sulla ridefinizione delle responsabilità umane. La distinzione tra rischio e pericolo, tra prevenzione e precauzione, è centrale per comprendere le nuove strategie di governo dell'incertezza, in cui l'IA assume un ruolo crescente, soprattutto grazie alle sue potenzialità predittive.

La dimensione politico-militare dell'IA è oggetto di un approfondimento specifico, incentrato sulle percezioni dell'opinione pubblica e delle *élites* riguardo all'uso dell'IA nei sistemi d'arma. Viene proposto un confronto tra dati internazionali e ipotesi di indagine nazionale, con l'obiettivo di esplorare

il consenso sociale, le riserve etiche e le strategie di regolamentazione e controllo.

Il testo analizza opportunità e rischi dell'IA nella società contemporanea, evidenziando come influenzi democrazia, lavoro e dinamiche sociali. Sottolinea l'urgenza di una formazione interdisciplinare che unisca competenze tecniche, etiche e umanistiche per governarne gli effetti. Propone infine politiche educative e regolative inclusive per ridurre disuguaglianze e *digital divide*.

Un altro filone riguarda l'impatto dell'IA sull'ambiente informativo e sulle libertà fondamentali, in particolare la libertà di espressione. L'uso dell'IA nei *media* solleva questioni legate alla disinformazione, alla responsabilità editoriale e all'etica del giornalismo automatizzato. Viene sottolineata la difficoltà dei sistemi giuridici attuali nel regolamentare fenomeni emergenti come le *fake news* e i contenuti generati automaticamente.

Un ulteriore ambito di indagine riguarda l'uso dell'IA per la produzione e manipolazione di narrazioni politiche, come nel caso del conflitto russo-ucraino. Viene mostrato come i sistemi generativi siano capaci di costruire contenuti apparentemente neutrali ma in realtà orientati, influenzando la formazione dell'opinione pubblica e introducendo *bias* sistemici.

Vi è anche un contributo che propone una nuova prospettiva teorica sull'IA come soggetto sociale. Gli "algoagenti" – entità algoritmiche capaci di operare ed interagire in ambienti umani – ridefiniscono le relazioni sociali e la rappresentazione del sé. Si suggerisce l'adozione di un paradigma "algomorfico" per descrivere la co-evoluzione di umani e macchine e per comprendere i nuovi spazi di *agency* e potere generati da queste interazioni.

È presentata anche una ricerca sperimentale che analizza le risposte di ChatGPT-4 e OpenAI 01 a domande su etica e implicazioni sociali dell'IA, mettendo in luce differenze di approccio tra i due modelli. ChatGPT-4o appare più orientato a soluzioni pratiche, mentre OpenAI 01 mostra maggiore attenzione alla trasparenza, alla responsabilità e ai valori morali. Nel complesso, il lavoro riflette sul passaggio verso un'"intelligenza responsabile", in cui l'IA deve affrontare le proprie responsabilità sociali.

Un altro filone riflessivo si concentra sull'incertezza insita nei sistemi predittivi utilizzati per la simulazione dei veicoli autonomi. L'attenzione è posta sulla *sensitivity auditing*, una pratica che mira a valutare criticamente i modelli attraverso il confronto tra i diversi attori coinvolti nel processo di modellazione, mettendo in discussione assunzioni implicite e immaginari tecnico-scientifici. Questo approccio intende rafforzare la credibilità dei modelli considerando anche le dimensioni sociali ed etiche della tecnologia

e si articola in un dialogo tra ingegneri, esperti e *stakeholders*, volto a rendere più trasparente e realistica la previsione del comportamento dei veicoli autonomi in contesti complessi.

Un altro contributo si focalizza sul processo di datificazione delle pubbliche amministrazioni italiane. In questo contesto, l'intelligenza artificiale viene analizzata come strumento di mappatura ed organizzazione delle informazioni, con l'obiettivo di monitorare l'adozione di sistemi decisionali automatizzati da parte delle istituzioni. Attraverso l'uso di algoritmi avanzati di *machine learning*, si propone una lettura critica del cambiamento amministrativo e del suo impatto sul concetto di cittadinanza, che rischia di essere ridefinito in funzione delle logiche digitali.

La relazione tra IA e criminalità è al centro di diversi interventi. Da un lato, si esplora il potenziale dell'intelligenza artificiale nell'individuazione e prevenzione dei crimini, tramite strumenti come le mappe criminali e l'analisi predittiva. Dall'altro, si evidenzia l'emergere di nuove forme di minaccia legate all'uso malevolo dell'IA da parte di soggetti criminali. In questo quadro, vengono descritti scenari in cui tecnologie come i *deepfakes* o le reti neurali generative sono impiegate per frodi, disinformazione, furti d'identità ed attacchi informatici mirati. Si discute anche della capacità dell'IA di generare profili falsi e contenuti sintetici altamente realistici, con conseguenze rilevanti per la sicurezza nazionale e l'integrità del dibattito pubblico.

Le implicazioni etiche e regolative dell'IA vengono approfondite attraverso un'analisi comparata delle strategie normative adottate in Europa, Stati Uniti e Cina. L'Europa si caratterizza per un approccio improntato alla tutela dei diritti fondamentali, attraverso l'introduzione di misure come l'*AI Act*, che mira a classificare i rischi e disciplinare gli usi più critici dell'intelligenza artificiale. Gli Stati Uniti, invece, mostrano una maggiore flessibilità orientata all'innovazione, mentre la Cina privilegia il controllo politico e la sorveglianza. Il confronto evidenzia come l'IA non sia solo una questione tecnica, ma investa direttamente i modelli di *governance*, le relazioni di potere e l'equilibrio tra libertà individuale e controllo sociale.

Un altro tema affrontato è quello dell'affidabilità dell'IA nell'interpretare valori ed atteggiamenti umani. Attraverso un esperimento comparativo tra risposte prodotte da modelli linguistici e quelle fornite da individui reali, si mettono in discussione le capacità predittive dell'IA nell'ambito delle scienze sociali. I risultati suggeriscono la necessità di cautela nell'uso degli algoritmi per comprendere fenomeni complessi come l'identità culturale, le opinioni politiche o le preferenze valoriali, mettendo in luce il rischio di semplificazioni indebite e *bias* sistemici.

Viene analizzato infine l'impatto dell'intelligenza artificiale nei sistemi giudiziari, mostrando come gli algoritmi di giustizia predittiva influenzino valutazioni e sentenze, riducendo la discrezionalità umana. Viene evidenziato il rischio di opacità, di delega della responsabilità e di riproduzione delle disuguaglianze sociali attraverso dati distorti. Da una prospettiva sociologica, gli algoritmi sono interpretati come attori socio-tecnici che riorganizzano potere, legittimità e fiducia nelle istituzioni. Gli autori propongono infine il modello di "giustizia aumentata", in cui l'IA supporta il giudice senza sostituirne il ruolo critico ed etico.

Complessivamente, i contributi restituiscono l'immagine di una tecnologia potente ma ambivalente, il cui impiego solleva interrogativi profondi sul piano epistemologico, normativo ed etico. L'IA emerge come dispositivo capace di estendere le capacità umane, ma anche di amplificare le disuguaglianze, alimentare nuovi rischi e ridefinire i confini della responsabilità e del potere. Da qui l'urgenza di un approccio critico ed interdisciplinare che sappia accompagnare lo sviluppo tecnologico con strumenti di controllo democratico, consapevolezza sociale e riflessione culturale.

Il numero della rivista è finanziato attraverso i fondi del progetto *i4Driving - Integrated 4D driver modelling under uncertainty*, programma di ricerca ed innovazione Horizon Europe 2022 (Grant Agreement No. 101076165) promosso dall'Istituto di Scienze e Tecnologie della Cognizione del CNR, sotto la direzione del Dottor Rino Falcone.

In indice si trova l'articolo "*The future of autonomous vehicles: expectation and worldviews of the stakeholder by Sensitivity Auditing. Interviewing the modelers and sociologists"* che tratta dei risultati della sezione di ricerca sulla *sensitivity auditing*, parte del progetto *i4Driving,* sviluppati dapprima presso l'ISTC-CNR e poi presso Dicea - Università di Napoli Federico II.

Riferimenti bibliografici

Abeti R., *et al*. (2021). *Intelligenza artificiale e sicurezza: opportunità, rischi e raccomandazioni*. 9-236.

Abduljabbar R., Dia H., Liyanage S., Bagloee S.A. (2019). Applications of artificial intelligence in transport: An overview. *Sustainability*, 11(1): 189.

Airoldi M. (2020). Lo spettro dell'algoritmo e le scienze sociali. Prospettive critiche su macchine intelligenti e automazione delle disuguaglianze. *Polis*, 35(1): 111-128.

Airoldi M. (2022). Macchine socializzate e riproduzione tecno-sociale: nuove frontiere sociologiche. *Sociologia Italiana*, 19-(20): 111-121.

Bertino E., Kantarcioglu M., Akcora C.G., Samtani S., Mittal S., Gupta M. (2021). AI for Security and Security for AI. In: *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy* (pp. 333-334).

Curzon J., Kosa T.A., Akalu R., El-Khatib K. (2021). Privacy and artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 2(2): 96-108.

Lau A.Y., Staccini P. (2019). Artificial intelligence in health: new opportunities, challenges, and practical implications. *Yearbook of medical informatics*, 28(01): 174-178.

Lutz C. (2019). Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies*, 1(2): 141-148.

Manheim K., Kaplan L. (2019). Artificial intelligence: Risks to privacy and democracy. *Yale JL & Tech*, 21, 106.

Nocenzi M., Sannella A., Cipriani R. (2022). A scuola con i robot. come cambiano i contesti sociali con l'intelligenza artificiale. In: *La società dei robot* (pp. 227-238). Milano: Mondadori.

Tucker C., Agrawal A., Gans J., Goldfarb A. (2018). Privacy, algorithms, and artificial intelligence. *The economics of artificial intelligence: An agenda*, 423-437.

Rischio, intelligenza artificiale e fisica quantistica. La logica del coltello

di *Francesco Mattioli*[*]

Nell'attuale società dell'incertezza, il ricorso all'intelligenza artificiale generativa offre la possibilità di porre sotto controllo un gran numero di variabili. Se ne avvale con profitto anche la sicurezza, nelle sue varie declinazioni, ma di particolare interesse sociologico appaiono il tema della sicurezza percepita intesa come fiducia e sfiducia nei confronti dell'intelligenza artificiale, e l'introduzione di un principio probabilistico che, superando il modello digitale binario, produce computer quantistici in grado di migliorare la previsione di quei fenomeni geologici, sismici, sanitari e sociali per i quali l'accuratezza non va oltre un certo limite. Tutto ciò ci dà possibilità di applicare una più efficace prevenzione e di definire meglio ciò che intendiamo oggi per sicurezza.

Parole chiave: incertezza; intelligenza artificiale; computer quantistici; sicurezza percepita; previsione; prevenzione.

Risk, artificial intelligence and quantum physics. The logic of the knufe

In the society of uncertainty, the use of Generative Artificial Intelligence offers the possibility of controlling a large number of variables. Security, in its various forms, also benefits from it, but of particular sociological interest appear to be the issue of a perceived security understood as trust and distrust in artificial intelligence, and the introduction of a probabilistic principle that, transcending the binary digital model, produces quantum computers capable of improving the prediction of geological, seismic, health, and social phenomena for which accuracy does not exceed a certain limit. This gives us the opportunity of apply more effective prevention and to better define what we mean by security today.

Keywords: incertainty; Artificial Intelligence; quantum computers; perceived security; prediction; prevention.

1. L'età dell'incertezza

Ha ragione Zygmund Bauman: viviamo in una società liquida, in una "società dell'incertezza", come spiega nel suo classico "Liquid modernity", scritto

DOI: 10.5281/zenodo.17296540

[*] Fondazione Roma Sapienza. francesco.mattioli@fondazione.uniroma1.it.

alle porte del XXI secolo. Ma questa sorta di disagio postmoderno non è il portato di una dissipazione sociale, come potrebbero essere indotti a ritenere conservatori e misoneisti di varia ispirazione.

È semmai l'effetto di quel processo di globalizzazione che ormai dagli anni '70 del secolo scorso sta caratterizzando la società postindustriale e postmoderna. Una globalizzazione che, lungi dal determinare una normalizzazione della società, ha dato spazio all'emergere di soggettività sociali che hanno creato una moltitudine di modelli culturali, di relazioni, di confronti.

Tutto questo ci conduce all'idea di complessità, o meglio di società complessa e proprio per questo incerta, in virtù dei molteplici soggetti, delle molteplici sintassi, delle molteplici informazioni circolanti sulle reti della comunicazione globale. Non è estraneo alla formulazione del concetto sociologico di complessità il passaggio da una lettura "strutturale" ad una "sistemica" della realtà sociale, che richiama ad un processo di simultaneità, di interazione, di interdipendenza funzionale, come sottolineato ad esempio da Niklas Luhmann, ma soprattutto dalla social network analysis (Buchanan, 2002; Mattioli, 2017).

Di fronte alla complessità e all'incertezza vi sono sostanzialmente tre modi per recuperare una "verità". Il primo consiste nel rifugiarsi nella religione, che tuttavia subisce pesanti condizionamenti storici; il secondo conduce ad appoggiarsi alla filosofia, oggi divisa tra il ricorso ad una logica spesso autoreferenziale e il cedimento al soggettivismo e all'impressionismo esistenzialista; il terzo confidando nella scienza, che pratica la ricerca sul campo, che registra i "fatti" e dà loro una spiegazione motivata alla luce di teorie interpretative comprovate dalla sperimentazione empirica e spesso dalla applicabilità tecnologica.

Eppure, oggi anche la scienza, la roccaforte di quella oggettività che il positivismo aveva celebrato come il punto di arrivo della sapienza umana, sembra in difficoltà. Con la nascita della fisica quantistica, e in particolare con la formulazione del "Principio di indeterminazione" di Werner Heisenberg nel 1929, anche le nozioni scientifiche subiscono un ridimensionamento delle loro certezze, perché si introduce il concetto di probabilità e quindi la stessa validità delle teorie scientifiche resiste solo "fino a prova contraria", per dirla nei termini del principio di falsificabilità enunciato da Karl Popper (Popper,1969). D'altronde, la sociologia della scienza ha rilevato che il sapere scientifico dipende in buona misura dalle dinamiche sociali, culturali e di potere che si sviluppano nelle comunità scientifiche (Merton, 1957; Kuhn, 1962).

Oggi, della scienza si può dire che si tratta della migliore forma di sapere a disposizione dell'Uomo, la meno discutibile e soggettiva; e tuttavia anche il sapere scientifico segue le regole della costruzione sociale del senso, quindi è il risultato di un accordo convenzionale prodotto da una complessa negoziazione dei significati (Rovelli, 2020).

La notizia confortante è che noi possediamo "verità scientifiche" che risultano attendibili talvolta al 90-95% e quindi, nella pratica, ci permettono di progredire, di migliorare la nostra qualità della vita e di difenderci da numerose minacce. Ma è chiaro che, sul piano cognitivo ed epistemologico, le certezze non hanno più senso.

2. Rischio e pericolo: alcune considerazioni sociologiche sulla sicurezza

Sia chiaro: la società dell'incertezza è una società ricca, rispetto al passato; la modernizzazione ha portato ad una migliore qualità della vita, a maggiori consumi, a maggior tempo libero, ad una diffusione della cultura e ad una garanzia sanitaria che cento anni fa erano impensabili.

Ma sebbene la società odierna ci abbia liberato di alcune minacce, gran parte delle incertezze che caratterizzano la modernità sono prodotte dalla dovizia delle scelte che siamo chiamati a compiere senza avere la certezza di aver fatto quelle giuste. Il rischio quindi è la conseguenza delle nostre scelte non solo di fronte a potenziali pericoli, ma anche nella selezione delle opportunità che intendiamo cogliere (Rowe, 1977; Mattioli, 2014).

Gli esseri umani valutano i rischi sostanzialmente i due modi; oggettivamente, quando tengono conto di indicazioni provenienti da informatori attendibili (istituzioni, scienza), ma anche soggettivamente in base alla loro esperienza individuale e alle loro inclinazioni emozionali (Slovic, 1987). E tuttavia sovente anche le indicazioni provenienti da fonti attendibili vengono poi filtrate e rielaborate da componenti personalistiche e socioculturali; tal che si può convenire che emerge quasi sempre una concezione del rischio fondata cognitivamente su euristiche esperienziali di sostegno ma anche, e inevitabilmente, da episodi di *confirmation bias.*

Tutto questo ha a che vedere anche con la sicurezza, o meglio con la concezione che le persone hanno della sicurezza in genere e di quella propria.

Nel contesto della modernità liquida, mediatica e incerta così come è descritta oggi della sociologia, e in considerazione di come oggi viene definito e calcolato il rischio, attraverso processi di reinterpretazione soggettiva dell'esperienza, una importanza fondamentale è assunta dalla *sicurezza percepita*.

Vari studi mostrano come i cittadini modifichino la percezione della propria sicurezza in rapporto con certe variabili socioculturali (Vaugham-Williams, Steven, 2015): l'età, il vissuto quotidiano, il lavoro, la residenza in città o in piccoli centri, in quartieri esclusivi e in quartieri popolari, in paesi, regioni e continenti diversi, con risultati solo apparentemente sorprendenti. Le periodiche rilevazioni statistiche di istituti di ricerca pubblici e privati in Italia e all'estero confermano queste tendenze. Inoltre, la complessità del concetto – e del vissuto – della sicurezza, in specie della sicurezza urbana (visto che tre quarti degli abitanti dei paesi più sviluppati vivono in città, spesso di gigantesche dimensioni e di fortissima

densità abitativa) comporta anche una serie di riflessioni sui concetti di prevenzione e di precauzione (Mattioli, 2021), l'una applicabile di fronte ad un pericolo certo, l'altra – e di fatto più diffusa – in presenza di stime meramente probabilistiche.

La gran parte delle operazioni che vengono avviate a contrasto di un pericolo, di un attentato alla sicurezza individuale e collettiva sono dunque strategie di precauzione; magari in grado di ridurre al massimo la dispersione di energie e di risorse, ma sempre con un raggio d'azione molto ampio. Perfino la definizione dell'allerta meteo, nei vari consueti colori, si appoggia più ad una stima che ha delle certezze in grado di operare a livello realmente preventivo.

In una società dove le certezze non sono proprio così robuste, dove le verità – anche quelle che un tempo venivano considerate assodate – sono sottoposte al vaglio di una razionalità che inclina alla convenzione, è chiaro che persino la prevenzione più mirata ed efficiente attiva processi, solleva dubbi e obiezioni, provoca riserve.

L'onestà intellettuale dello scienziato rischia di essere un'ammissione di debolezza della scienza di fronte all'imponderabile, alla minaccia più efferata, ai timori che crescono negli individui a salire dall'inconscio. Null'altro che questo è ciò che è avvenuto con l'epidemia del Covid-19, con le polemiche sull'affidabilità e sulle controindicazioni delle vaccinazioni, a partire da conflitti sotterranei presenti anche nella comunità scientifica e dai sospetti di connivenza tra scienza e pragmatismo industriale. Di qui, spesso, il ricorso alla precauzione che amplia la copertura dal pericolo, ridimensiona il rischio, conforta il senso di sicurezza. Ma ha costi molto alti, con estremismi che potrebbero far sorridere – per evitare qualsiasi rischio, rinuncio a volare – ma che a vario titolo e in varia forma sono diffusi nell'opinione pubblica, influenzando anche il sistema politico. E su questo, un inevitabile potere lo esercitano i social media, dove tutto può essere detto raccogliendo sovente sterminate folle di seguaci.

Mancherebbe insomma un parametro oggettivo, scevro da influenze di parte, logicamente ed equamente sostenibile e sostenuto da un pensiero veramente neutrale e condiviso.

Ma poi è arrivata l'intelligenza artificiale.

3. L'irruzione dell'intelligenza artificiale

Le incertezze etico-valoriali non si risolvono mediante semplici scelte, necessitano di riflessioni comuni, forse di dolorose verifiche esperienziali, certamente dell'insegnamento della storia.

Al contrario, quelle che esigono azioni e operazioni di effetto immediato hanno bisogno di un supporto logico, per così dire "tecnico", fondato su saperi. In tal caso serve avere conoscenza adeguata dei termini del problema e delle soluzioni più affidabili per affrontarlo e risolverlo. E così sorgono domande del tipo: quale può essere in generale la graduatoria tra i bisogni A, B e C? Quale

quella rispetto alle mie caratteristiche? Quale è la soluzione più adatta alla soddisfazione di ciascuno di questi bisogni? Con quale graduatoria di efficacia o di efficienza? Secondo quali contingenze e condizioni? Con quali costi?

A ben vedere, si tratta delle domande che noi porremmo ad un esperto, che si tratti di un medico, di un meccanico, di un tecnico o di un qualsivoglia consulente. E quale esperto migliore di una intelligenza artificiale capace di possedere in memoria tutto lo scibile umano e una serie quasi infinita di algoritmi in grado di offrire la soluzione migliore ai nostri problemi?

Oggi combattiamo le nostre incertezze affidandoci alle indicazioni di una Macchina, che non è soggetta né ad emozioni, né a debolezze e neppure a dubbi esistenziali, ma ci garantisce una scelta giustificata dal possesso di adeguate informazioni e da un processo decisionale fondato sulla logica, sulla scienza e sulla tecnica. Lo facciamo continuamente, ogni volta che utilizziamo un computer, un tablet, uno smartphone.

L'intelligenza artificiale (d'ora in avanti IA) non è comparsa d'improvviso sugli scenari della società postmoderna a risolvere i nostri dubbi e a fornire soluzioni ai nostri problemi; ma è giunta all'attenzione dell'opinione pubblica immediatamente carica di equivoci e di timori.

La formulazione, pur soltanto teorica, della cosiddetta "Macchina di Turing", quella che secondo il suo autore sarebbe stata in grado di eguagliare le capacità intellettive umane, già nel 1950 procurava qualche allarme; e non è un caso che il noto scrittore di fantascienza Isaac Asimov si fosse inventato le "tre leggi della robotica", che mettevano l'essere Umano al sicuro dal possibile sottrarsi della Macchina ai voleri del proprio inventore. In seguito Vernon Vinge, matematico e scrittore, preconizza l'avvento di una "singolarità tecnologica", cioè di una Macchina in grado di superare l'intelligenza umana e di decidere anche per l'Uomo. Ma tutto ciò ha rischiato solo di distogliere l'attenzione da quei processi che, grazie alla cibernetica e all'informatica, stavano rivoluzionando la vita quotidiana degli individui, l'organizzazione del lavoro e della produzione industriale e la stessa comunicazione di massa.

Robotica e meccatronica ci accompagnano ormai da almeno cinquant'anni, eppure solo negli ultimi tempi si sono levate voci inquiete e allarmate sui destini dell'Uomo a fronte della IA salvo poi constatare che sì, in effetti ci sono problemi, ma per ora ne creano soprattutto chatbot come *ChatGPT*, che può eludere i diritti d'autore, e *Stable Vision* che crea fake e deepfakes, confondendo reale e virtuale sia nel mondo dell'informazione e più latamente dei media che in quello dei giochi di ruolo, alimentando il multiverso.

In realtà, il dibattito attuale non riguarda quella che è stata definita, tra le tante classificazioni, la IA "debole", cioè quella che troviamo predisposta su dispositivi come calcolatrici, cellulari, tablet, su automobili ed elettrodomestici di ultima generazione e soprattutto nei processi di automazione in fabbrica e in quelli al servizio delle smart cities. Gli algoritmi che sostengono queste operazioni, ancorché apparentemente sofisticati, sono sostanzialmente costituiti da percorsi razionali fondati su una logica digitale binaria e su un programma costruito sin

dall'inizio da uno sviluppatore con intenti chiari. Non c'è poi tanta differenza, sul piano procedurale, tra l'algoritmo qualitativo costituito dalla ricetta della torta della nonna e quello logico-matematico che governa la delicata gestione di un'auto a guida autonoma, di una navicella spaziale o la strategia commerciale di un'azienda.

La IA debole, peraltro, supporta varie attività collegate alla gestione del rischio e della sicurezza. Essa infatti può agire anche nella forma predittiva di "cosa accade nel caso che…". Lo si può constatare nelle applicazioni in fabbrica a garanzia della salute e della salvaguardia dei lavoratori, nella gestione delle operazioni preventive e di quelle di risposta nel caso di fenomeni sismici e di eventi estremi di origine idrogeologica e metereologica.

Se la IA fosse solo questo, difficile pensare che si potrebbero verificare problemi di natura etica, filosofica, antropologico-sociale.

Ma poi lo scenario è cambiato.

Perché all'IA debole si è affiancata la *IA Generativa* (d'ora in avanti Gen AI), una forma di intelligenza artificiale evoluta, apparentemente in grado di effettuare scelte autonome nel rispondere nel modo più esauriente alle richieste dell'Essere Umano.

Certamente si è verificato un salto di qualità delle opportunità che essa offre; ma con le opportunità sono anche arrivati i veri problemi.

Per lo più, il dibattito sulla IA nella forma generativa ha riguardato un suo cattivo uso, quindi come strumento di *criminal hacking*, soprattutto in campo economico, di *phising,* di *baiting*, di *faking* mediatico e, a livello più interpersonale, come *deepfaking*, *bodyshaming, ageshaming* e cyberbullismi vari. Così, sia le preoccupazioni delle istituzioni, comprese quelle internazionali come le Nazioni Unite e l'Unione Europea, sia quelle di alcuni fondatori e progettisti della IA, si sono concentrate sostanzialmente sull'etica. Qualcuno, come p. Paolo Benanti, ha parlato della necessità di una vera e propria "algoretica", cioè di un approfondito approccio critico sulle finalità e sull'uso degli algoritmi e quindi dell'IA (Benanti, 2018).

Gen AI, nelle sue varie applicazioni, dagli chatbot alle unità gestionali di procedure e di controllo, presenta una evoluzione delle proprie capacità di scelta. Sia che si tratti di comporre un testo che di controllare un procedimento operativo, essa sembra formulare opzioni e fornire soluzioni che non sono state considerate e previste dall'Uomo, dal progettista e dall'operatore. Siamo alle soglie di una nuova era? Non necessariamente.

Ci sono due aspetti che vanno considerati.

Il primo riguarda il contributo delle neuroscienze.

Esse stanno chiarendo, seppur tra varie difficoltà, il funzionamento di molti meccanismi del pensiero umano, in particolare la capacità dei neuroni di ricevere input informazionali da altri neuroni e di rielaborarli autonomamente, producendone apparentemente di nuovi.

E questo meccanismo, definito apprendimento profondo (*deep learning*), si può attivare anche nei sistemi informatici più evoluti. In realtà, la Macchina non

genera qualcosa di veramente nuovo, ma disponendo di una massa gigantesca di big data e potendo confrontarli e aggregarli in tempi brevissimi, riesce a sfornare associazioni e soluzioni che l'Uomo non sarebbe in grado di elaborare, quanto meno in tempi utili. Quindi, ciò che produce apparentemente di nuovo Gen AI si trova già in ciò che l'Uomo ha immesso massivamente in essa, solo che è la Macchina a cogliere i nessi, a superare gli impacci, creando anche risposte apparentemente nuove e impreviste. Tutto ciò significa mettere a disposizione dell'Uomo informazioni che, pur esistendo già nel suo repertorio conoscitivo, non gli è facile riorganizzare in un messaggio coerente e mirato.

Si comprende allora il salto di qualità dei servizi resi da Gen AI, anche nel campo della sicurezza. Ad esempio, Gen AI. appare di grande utilità nel confronto delle corrispondenze e delle relazioni in una collezione di dati acquisiti in modo massivo (ad esempio quelli raccolti dagli strumenti di videosorveglianza); nelle analisi e nelle previsioni relative a condizioni di pericolo potenziali o in corso; nel fornire risposte e soluzioni particolarmente rilevanti, le prime in termini di affidabilità e le seconde in termini di adeguatezza ai bisogni emergenti.

Il secondo aspetto che interessa Gen AI convoca la fisica quantistica.

Qui occorre fare una brevissima parentesi. La rivoluzione della fisica quantistica non riguarda soltanto lo studio dei comportamenti delle microparticelle. Come si è accennato in precedenza, sia la formulazione del Principio di indeterminazione di Heisenberg che il pensiero di Karl Popper sui processi di falsificazione delle teorie scientifiche hanno prodotto un ampio ventaglio di considerazioni che coinvolgono non solo le scienze fisiche, ma anche quelle sociali e persino la filosofia.

L'idea che la conoscenza scientifica sia subordinata all'inevitabile interazione che si stabilisce tra soggetto conoscitore e oggetto di ricerca, comporta che tale conoscenza si fondi sulla definizione di accordi convenzionali. Ciò esclude l'incontrovertibilità di qualsivoglia sapere (escludendo quello oggetto di fede) e apre ad una valutazione della verità scientifica in termini probabilistici.

Dunque, noi siamo di fronte ad un sapere scientifico fondato sul probabilismo, sì che anche il contributo di Gen AI, per quanto sofisticato e superiore in efficienza alle correnti capacità umane, resta sempre appeso ad un quantum di indeterminatezza.

Tale indeterminatezza può essere minimale rispetto a fenomeni studiati in modo approfondito e a sperimentazioni lungamente ripetute, come nel caso di meccanismi organici, e in specie quelli relativi al corpo umano, producendo quindi risposte – si pensi alla vaccinazione – con affidabilità superiori al 95%. Ma in altri casi, dove le capacità previsionali sono fortemente condizionate dall'impossibilità della scienza attuale di avere precisa contezza di taluni fenomeni geologici e meteorologici - si pensi al paradossale "effetto farfalla" di Lorenz – il contributo di Gen AI, per quanto di altissimo livello, perde parte della sua affidabilità, lasciando alcuni margini non trascurabili di incertezza. In tal caso le previsioni, ma anche le procedure di contrasto preventive e quelle in

relazione all'effettiva manifestazione dei fenomeni restano approssimate, in ragione dell'accuratezza delle informazioni disponibili e delle azioni possibili.

Altro, quindi, è progettare e programmare efficacemente, con il sostegno di Gen AI, una procedura di controllo delle operazioni in fabbrica, dalle garanzie di precisione del prodotto alla protezione e all'incolumità degli operatori; altro è analizzare il comportamento di componenti biologiche e genetiche fino a produrre vaccini, terapie e organi artificiali di amplissima efficacia in campo sanitario; e altro è prevedere il sorgere e l'entità di un terremoto, pur in zona sismica, oppure la reazione di varie componenti ambientali, in specie di carattere idrogeologico, di fronte al verificarsi di un evento meteorologico straordinario; e in questi ultimi casi, quale sia effettivamente ancora la consistenza delle opere dell'Uomo in termini di contrasto e di resistenza.

Nell'affrontare tutti questi problemi il procedimento suggerito da Gen AI, che funziona secondo algoritmi fondati su una logica binaria, offre sempre un quadro deterministico della realtà.

Ma la realtà è solo apparentemente deterministica: piuttosto è sistemica, quindi basata su interazioni che producono configurazioni di causa ed effetto o provvisorie o addirittura solo apparenti. In tal caso, come si è già fatto notare, la definizione della realtà è negoziale, frutto di un accordo sui criteri da adottare nella valutazione e nella interpretazione dei fenomeni e delle operazioni da effettuare su di essi. Un accordo che si avvale di un più o meno affidabile calcolo probabilistico che come tale non ci può fornire una certezza assoluta, semmai una statisticamente ragionevole.

E tuttavia la fisica quantistica su questo ha ancora altro in serbo da proporre: una nuova scienza applicata, che sta emergendo dagli studi più avanzati della cibernetica e dell'informatica, il cosiddetto *quantum computing.* Questo va oltre il bit, l'unità binaria 0/1 propria della logica digitale: adotta infatti il *qbit*, una unità quantistica che può corrispondere a più valori contemporaneamente. Il principio deriva da alcune nozioni proprie della fisica quantistica: quella di sovrapposizione e di *entanglement* che presuppone la possibilità di convivenza di più stati potenziali della materia, e quindi dei fenomeni, almeno fintantoché non se ne fa una misura (pur sempre di natura probabilistica, in accordo con il Principio di indeterminazione di Heisenberg); e, come detto, la condizione sostanzialmente sistemica della realtà, che seppur talvolta in valori statisticamente minimali, deve sempre tener conto della potenziale variabilità dei processi di interazione tra i fattori in gioco.

Il *quantum computing* porta alla realizzazione di cosiddetti *computers quantistici*, che utilizzano una intelligenza artificiale generativa fondata su algoritmi che tengono conto di variabili intermedie tra 0 e 1: quindi, sono pronti a modificare l'interpretazione e la previsione di vari fenomeni man mano che ne misurano l'effettivi svolgimento. In altre parole, la risposta algoritmica di Gen AI diventa flessibile in base alle rilevazioni correnti dei dati in input e alle richieste in output.

Ciò significa avere a disposizione una massa di informazioni estremamente più ampia su cui operare.

Non è difficile comprendere che tale disponibilità, magari meno evidente e utile quando si conducono ricerche di carattere storico, diventa di notevole utilità di fronte ad un pericolo – e soprattutto ad un rischio – che presenta differenti eziologie potenziali e diverse possibilità di manifestazione. In tal caso sono già pronte in tempo reale sia le diverse risposte a seconda dell'evolversi della situazione, sia una più puntuale ed efficace applicabilità di tali risposte.

Siamo ovviamente nei pressi delle frontiere più avanzate delle scienze che si occupano di intelligenza artificiale e, in ossequio alla cosiddetta Legge di Moore (Riccò, 2007; 2008), che oggi decreta la completa riconfigurazione delle tecnologie disponibili nel giro di pochi anni, c'è da aspettarsi cambiamenti importanti sia nei saperi scientifici che nelle loro applicazioni pratiche. Ma è ormai chiaro che la riduzione dello stato di incertezza che pervade la nostra epoca se non si può annullare proprio in virtù dello stato sistemico della realtà così come ci appare, di certo si può limitare e comunque governare solo grazie alle scienze – fisiche o sociali che siano – supportate da una sapiente applicazione della statistica, della cibernetica e dell'informatica. E tuttavia, a dirimere inutili fantasie – e soprattutto timori – di carattere futuribile, deve esser chiaro che l'IA, generativa e quantistica quanto si voglia, restituisce, seppur in modo molto sofisticato e ricco di informazione, solo ciò che le si dà. Nel bene e nel male. E questo, nell'affrontare questioni di ordine pubblico, di carattere sanitario o di controllo delle problematiche ambientali, è più che sufficiente a creare strumenti di grande efficacia e di testata affidabilità e a guardare con fiducia al futuro. Non è Gen AI a doverci preoccupare, ma chi la usa, per quali fini…

Si può chiudere allora con una necessaria considerazione su Gen AI: che nel trattare di tecnologia, vale pur sempre la cosiddetta "logica del coltello". Se questo deve affettare in cucina o trasformarsi in un'arma letale, lo decide sempre chi lo impugna.

Riferimenti bibliografici

Bauman Z. (2000). *Liquid Modernity*. Cambridge: Polity Press (trad. it. Modernità liquida. Roma-Bari: Laterza, 2002).

Benanti P. (2018). *Oracoli. Tra algoretica e algocrazia*. Bologna: Sossela.

Buchanan M. (2002). *Nexus. Small Worlds and the Groundbreaking Science of Networks*. New York: W.W. Norton & Co. (trad. it. Nexus. Milano: Mondadori, 2002).

Kuhn T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press (trad. it. La struttura delle rivoluzioni scientifiche. Torino: Einaudi, 1969).

Mattioli F. (2014). *Sociologia del rischio e sicurezza urbana*. Acireale-Roma: Bonanno.

Mattioli F. (2017). La network analysis tra teoria e ricerca. In: Mattioli F., Anzera G., Toschi L., *Teoria e ricerca nell'analisi delle reti sociali*. Roma: Aracne, pp. 11–54.

Mattioli F. (2021). *I rischi della precauzione*. Roma: Aracne.

Merton R.K. (1957). *Social Theory and Social Structure*. New York: Free Press (trad. it. Teoria e struttura sociale. Bologna: Il Mulino, 1992).

Popper K.R. (1969). *Conjectures and Refutations*. London: Routledge and Kegan Paul (trad. it. Congetture e confutazioni. Bologna: Il Mulino, 1972).

Riccò B. (2007–2008). Legge di Moore. In *Enciclopedia della scienza e della tecnica*. Roma: Istituto dell'Enciclopedia Italiana.

Rovelli C. (2020). *Helgoland*. Milano: Adelphi.

Rowe D.W. (1977). *An Anatomy of Risk*. New York: Wiley.

Slovic P. (1987). Perception of risk. *Science*, April.

Vaughan-Williams N., Steven D. (2015). Vernacular theories of everyday (in)security: The disruptive potential of non-elite knowledge. *Sage Journals*, 47(1).

Vinge V. (1993). The coming technological singularity: How to survive in the post-human era. In: Vision-21. *Interdisciplinary Science and Engineering in the Era of Cyberspace*. NASA Lewis Research Center and the Ohio Aerospace Institute, pp. 11-22.

Intelligenza artificiale vs intelligenza popolare. Gli italiani sulle armi "autonome" tra intuizioni e paure

di *Fabrizio Battistelli, Francesca Farruggia*[*]

Nella sfida per la supremazia nell'intelligenza artificiale (IA) un ruolo cruciale è rivestito dalle applicazioni militari. Promettendo di trasformare il campo di battaglia in un duello tra macchine, la IA consente ai decisori politici di prevenire la contrarietà del pubblico alle perdite in guerra. Gli effetti sociali, invece, rimangono critici in quanto rischiano di rendere adiaforo l'agire umano (Bauman, Lyon, 2014). Nel sondaggio Difebarometro di Archivio Disarmo un campione rappresentativo di cittadini italiani esprime preoccupazione per queste applicazioni della IA e per le eventuali conseguenze (impossibilità di rendicontare le decisioni, maggiore facilità di ricorso alle armi), così che in 2/3 dei casi ritiene indispensabile mantenere il controllo umano su queste tecnologie.

Parole chiave: intelligenza artificiale; sistemi d'arma autonomi; responsabilità umana; opinione pubblica; illusione delle "perdite zero"; multilateralismo dei controlli.

Artificial intelligence vs popular intelligence. The Italians about "autonomous" weapons between intuitions and fears

In the race for supremacy in artificial intelligence (AI), a hidden but crucial role is played by military applications. By promising to transform the battlefield into a duel between machines, AI seeks offers decision-makers the possibility of preventing public aversion to the losses of war. Social implications remain critical, as human action becomes adiaphoric (Bauman, Lyon, 2014). According to Archivio Disarmo's survey Difebarometro, a representative sample of Italian citizens expresses apprehension about these applications of AI and the associated risks (i.e. impossibility of being accountable for one's own decisions, ease of recourses to war), while two respondents out of three consider essential maintaining human control over such sensitive technologies.

Keywords: artificial intelligence; autonomous weapons systems; human accountability; public opinion; illusion of "zero casualties"; multilateral arms control regime.

DOI: 10.5281/zenodo.17297405

[*] Università di Roma La Sapienza. fabrizio.battistelli@iriad.it, francesca.farruggia@uniroma1.it.

Fabrizio Battistelli, Francesca Farruggia

Introduzione

A differenza di ciò che talvolta dice di essere (un fenomeno tecnologico e basta), l'avvento dell'intelligenza artificiale (IA) costituisce un fenomeno eminentemente sociale. Ciò è vero anche nel *sancta sanctorum* della tecnologia dello Stato, ovvero nella sua applicazione alla funzione *strategica* in senso stretto: l'automazione del campo di battaglia. Lungi dal costituire una semplice articolazione dell'incessante innovazione tecnologica, essa rappresenta il prodotto di una prospettiva politica che intende coniugare efficienza bellica e tolleranza sociale dell'uso della forza.

Nelle democrazie occidentali la crescente intolleranza verso la perdita di vite umane ha favorito la ricerca di "soluzioni tecniche" per la cruciale funzione governamentale rappresentata dalla difesa. In particolare si tratta di convincere i vari attori del sistema sociale circa la necessità di affrontare gli ingenti costi – umani e finanziari – della guerra, della sua preparazione o anche della sua prevenzione. Rompendo con una narrativa secolare fondata sul diritto-dovere del cittadino di difendere la patria e del professionista delle armi di guidarlo con il suo valore in questa missione (Moskos, 1994; Greco, 1999), le armi autonome (giornalisticamente note anche come robot-killer) si iscrivono in questa traiettoria, presentandosi come strumenti "post-eroici" (King, 2010) in grado di sostituire o almeno attenuare i costi della guerra.

In ogni caso, tale dinamica genera una tensione sociale e politica. Come osserva Bauman (1999) citando l'apologo dell'aereo nella cui cabina di pilotaggio non c'è nessuno e il messaggio del comandante è registrato, l'automazione tende a svuotare di senso l'attore umano, rendendo invisibile la catena delle responsabilità. Lo studio dell'impatto sociale delle armi autonome si colloca dunque all'intersezione tra tecnologia, valori condivisi e legittimazione politica, ambito quest'ultimo nel quale i cittadini – in passato stigmatizzati come irrilevanti per la loro incompetenza e volatilità (Everts e Isernia, 2001) – acquistano un ruolo crescente nella toureniana "produzione della società".

La tendenza è particolarmente evidente nelle società occidentali contemporanee, presso le quali la tolleranza verso la violenza bellica e la perdita di vite umane ha raggiunto oggi il minimo storico. Questo orientamento culturale ha profondamente inciso sulla legittimazione democratica dell'uso della forza armata e ha favorito nei governi lo sviluppo di soluzioni tecnico-militari miranti a rispondere a un'esigenza politica e simbolica precisa: la minimizzazione delle perdite nelle proprie forze armate. In questo contesto, le armi autonome si propongono come l'ultimo grido del mito strategico della "guerra a perdite zero". Si tratta della dottrina e della pratica, introdotta dai

vertici militari tedeschi e italiani nella guerra civile spagnola (Guernica, aprile 1937) proseguìta dai tedeschi nella Seconda guerra mondiale (bombardamenti aerei di Varsavia, Rotterdam, Coventry, Londra) e perfezionata da quelli americani (bombardamenti di Colonia, Brema, Dresda, Tokio, fino all'apice dell'atomica su Hiroshima e Nagasaki). Secondo questa impostazione è possibile condurre operazioni belliche efficaci senza esporre direttamente i propri soldati al rischio della morte. Oggi, tuttavia, essa suscita forti criticità che si riflettono nell'avversione diffusa presso la maggioranza dell'opinione pubblica nei confronti di attacchi indiscriminati come quelli effettuabili dalle armi nucleari ed eventualmente dalle armi autonome, a rischio di infliggere danni indiscriminati alla popolazione civile e fatali ritorsioni per chi abbia fatto ricorso al "primo uso".

La riflessione di Bauman e Lyon (2014) risulta particolarmente utile per comprendere le implicazioni della relazione tra innovazione e avversione suscitata dalla tecnologia. In particolare con il tramonto della modernità "solida", «la tecnologia militare è riuscita a rendere fluttuante la responsabilità, e quindi "spersonalizzarla"» (Bauman, Lyon 2014: 37). In particolare l'introduzione delle prime armi semi-autonome come i droni ha dato inizio a un processo di "adiaforizzazione", ovvero di deresponsabilizzazione morale degli attori coinvolti, all'una e all'altro estremo della scala gerarchica. Il diluvio di dati rilevati dai sensori «che trabocca[va]no dallo schermo come da un secchio stracolmo» (ibidem) e la loro riorganizzazione ad opera degli algoritmi, lungi dal rendere l'azione più consapevole, hanno l'effetto di sollevare l'operatore sul campo dal peso del giudizio etico, offrendogli un potente alibi tecnologico. Quanto ai decisori, il ricorso a macchine "intelligenti" (antropomorfe o meno, ma questo è un dettaglio estetico) (Farruggia, 2024) può rendere più accettabile, e quindi spendibile, l'opzione dell'uso della forza (Battistelli, 2023).

È in questo quadro che si colloca il ruolo crescente dell'opinione pubblica, non più spettatrice passiva ma potenziale co-protagonista dell'agenda politica. Se è vero che le innovazioni tecnologiche in ambito militare sono spesso introdotte con l'intento di aggirare la riluttanza sociale al loro impiego, è altrettanto vero che "l'altra superpotenza mondiale" opinione pubblica (copyright del *New York Times* nell'invasione americana dell'Iraq del 2003) può rappresentare una remora allo sviluppo incontrollato di tecnologie letali esprimendo in piazza o anche nei sondaggi, la propria contrarietà a determinate opzioni militari.

Nello scetticismo dei cittadini verso l'uso della forza e dei suoi strumenti (le armi), un peso rilevante è esercitato dalla concreta esperienza storica che è stata vissuta dai contemporanei e ascoltata dalla loro viva voce dalla

generazione successiva. Per quanto destinata a indebolirsi con il ricambio anagrafico, la memoria popolare delle guerre del XX secolo non si è ancora del tutto dissolta, contribuendo alla maturazione e circolazione di rappresentazioni del mondo più evolute. Come è stato osservato, «la carneficina delle due guerre mondiali ha incoraggiato lo sviluppo delle norme in tema di diritti umani» (Crawford, 2013: 65). Oggi, infatti, viene ritenuto non più accettabile sia fare vittime tra i civili, sia «usare i [propri] soldati come meri strumenti le cui vite possano essere «sprecate» a migliaia» (Crawford, 2013: 65).

Se questa è la posizione prevalente nell'opinione pubblica mondiale, anche americana, presso i vertici politici e militari "in azione" finisce per imporsi «il *trade-off* tra i *loro* caduti e i *nostri* caduti» (Gelpi *et al.* 2009: 255; Sagan, Valentino, 2017). È così che il risparmio delle vite dei propri soldati ad ogni costo, compreso quello di perdite sproporzionatamente elevate per i nemici (combattenti e non) si consolida come la policy storicamente perseguìta dagli Stati Uniti.

Nella seconda metà del XX secolo, anche in conseguenza delle rovinose perdite che la guerra del Vietnam ha inflitto a un esercito americano composta da coscritti, gli stati maggiori hanno fatto ricorso a misure radicali, sia organizzative quali il passaggio a forze armate di soli volontari, sia dottrinarie, quali la teoria della guerra "a perdite zero" (Rogers, 2016). Oggi l'antico sogno della "automazione del campo di battaglia", prefigurato dal generale Westmoreland in Vietnam, appare sempre di più a portata di mano grazie prima ai progressi di internet (*network-centric warfare*) e, subito dopo, al boom di sistemi d'arma semi-autonomi quali i droni diretti da remoto. Ma il vero *breakthrough* si prospetta nell'attuale terza fase, rappresentata dall'esordio dell'intelligenza artificiale nel combattimento. Algoritmi che profilano e mirano il bersaglio lasciando all'operatore 30 secondi per decidere di premere il pulsante sono stati testati sul campo dall'esercito israeliano a Gaza (Abraham 2024; v. anche Birch, 2024). In un domani ormai prossimo sarà tecnicamente possibile schierare mezzi di attacco aerei, terrestri e marini totalmente indipendenti dal controllo umano (cioè le armi letali autonome, giornalisticamente note come *robot-killer*).

Passando dall'ambito meso a quello macro, l'antico sogno (o incubo) di poter combattere una guerra senza neppure un caduto presenta potenzialità letteralmente rivoluzionarie non soltanto sul piano tattico, ma anche su quello strategico e, soprattutto politico. Infatti venendo meno la spada di Damocle delle perdite umane, rischia di cedere anche l'ultimo e più solido argine che dissuade i governi dal ricorrere alla forza. Paradossalmente ma non troppo, nell'età contemporanea alla crescita a dismisura della potenza distruttiva di armi come quelle nucleari, ha fatto riscontro nell'opinione

pubblica la crescente inibizione a interventi militari che comportino la perdita di vite umane. Il contro-paradosso sarebbe che, allo scopo di non rimanere bloccati da questo vincolo, i politici finiscano per delegare all'intelligenza artificiale e alle armi autonome la guerra, con esiti oggi totalmente imprevedibili.

1. La tecnologia e i suoi prodigi: le armi letali autonome e l'opinione pubblica

Mentre si assiste a una crescita esponenziale degli investimenti nei settori di R&S dell'intelligenza artificiale applicata alle armi autonome, l'opinione pubblica mondiale (IPSOS, 2017; 2018; 2021) esprime in merito una crescente preoccupazione in pressoché tutti i Paesi.

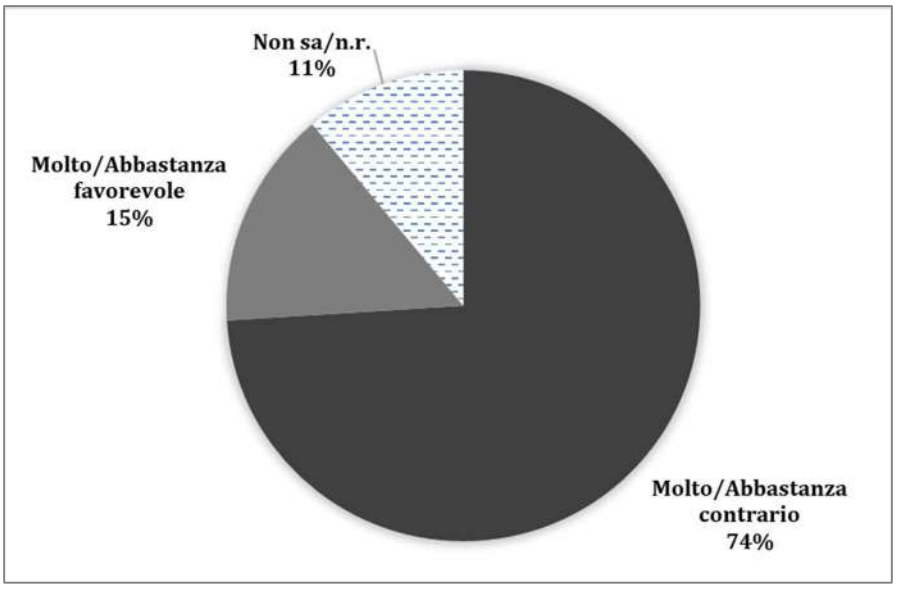
In concomitanza con i gravi conflitti che hanno coinvolto l'Ucraina e il Medio Oriente abbiamo effettuato un sondaggio che ha sottoposto a un campione rappresentativo di cittadini italiani una serie di domande volte a rilevare la loro opinione sull'eventuale impiego in guerra delle armi autonome, le ragioni che potrebbero comportarlo, le possibili conseguenze, il livello del controllo umano da mantenere e infine le auspicabili modalità di normazione in materia [1].

Come emerge dal sondaggio, dai cittadini italiani viene manifestata una preoccupazione superiore alla media rilevata negli altri Paesi.

Pur ammettendo una scarsa conoscenza dei sistemi di arma autonomi (con il 35% dei rispondenti che si dichiara per niente informato e il 50% poco informato sul tema), la grande maggioranza dei rispondenti esprime la propria contrarietà al loro impiego sul campo di battaglia. Infatti il 74% si dice contrario a un possibile impiego militare delle armi autonome, a fronte di un 15% di molto o abbastanza favorevoli (v. graf. 1).

[1] Archivio Disarmo, Difebarometro n. 11, 3 febbraio 2025. Rilevazione Cawi effettuata da Demetra su un campione di 802 rispondenti. Il sondaggio di opinione è stato realizzato dall'Istituto di ricerche internazionali Archivio Disarmo nel quadro della ricerca "Lo stato della IA in ambito militare e le prospettive di regolazione a livello nazionale, europeo e internazionale" per il Ministero degli Affari Esteri e della Cooperazione internazionale. Le opinioni espresse in questo articolo sono esclusiva responsabilità degli autori.

Graf. 1 – grado di favore verso le armi autonome.



Archivio Disarmo, *Difebarometro* n. 11, febbraio 2025.

Come ben noto, in generale l'IA è investita oggi di significati che travalicano la dimensione tecnica per assumere una forte carica simbolica, oscillando tra la promessa di crescente efficienza nella produzione di beni e servizi e il timore di una possibile autonomia fuori controllo. Secondo i più recenti dati dell'Osservatorio Scienza, Tecnologia e Società, l'opinione pubblica italiana esprime una crescente preoccupazione nei confronti dell'impatto complessivo dell'IA nella società odierna. Nel 2024 il 51% degli intervistati vede l'intelligenza artificiale come una minaccia, contro il 35,8% registrato nel 2023. Allo stesso tempo, relativamente ai sentimenti evocati dalla IA, tra i cittadini italiani diminuisce leggermente l'indicazione di "fiducia" e "attesa", mentre cresce quella di "sospetto" e "ansia" (Observa, 2025).

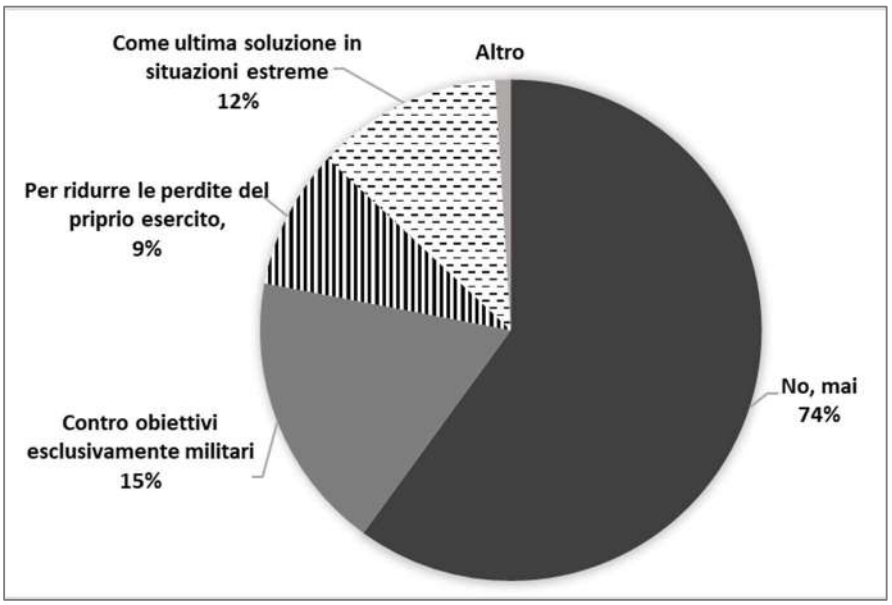
Le perplessità degli italiani aumentano laddove l'IA è applicata al settore bellico. Esaminando la precedente rilevazione realizzata da Archivio Disarmo nel 2019 (Farruggia, 2023b), emerge infatti come la contrarietà nei confronti dell'impiego di armi autonome in guerra sia cresciuta di ulteriori 5 punti nell'arco di sei anni (allora a dichiararsi molto o abbastanza contrario era stato il 69% dei rispondenti). Una riprova che, al sia pur contenuto

aumento della conoscenza dell'oggetto, non corrisponde un aumento della sua popolarità.

Tornando ai dati del Difebarometro n.11, è interessante la correlazione dell'opinione espressa dagli intervistati con alcune caratteristiche strutturali dei medesimi, prima tra tutti il genere. L'opposizione a tali sistemi d'arma sia tra gli uomini che tra le donne risulta talmente generalizzata da colmare il *gender gap* che solitamente vede il genere femminile più restio all'uso della forza. Ad essere molto o abbastanza contrario all'impiego delle armi autonome è il 73% degli uomini a fronte del 74% delle donne, peraltro uno scarto minimo, statisticamente non significativo (P=0,084). È significativo invece il fattore età (p<0,001): la contrarietà alle armi "intelligenti" cresce al crescere dell'età anagrafica, passando dal 67% dei giovani tra 18 e 24 anni sino all'84% degli over 65 (unica eccezione è rappresentata dalla fascia di età 25-34 che, con un 2% di scarto rispetto ai più giovani, esprime il livello minimo di contrarietà del 65%). Al netto della presumibile prudenza esistenziale dei più anziani, qui emerge la particolare ostilità dei *baby boomers***.** Vari studi hanno ipotizzato per la generazione del '68 l'imprinting della contestazione politica, sociale e culturale alla luce della quale furono lette le guerre dell'ultimo terzo del XX secolo (Caren *et al.*, 2010).

L'avversione verso il possibile impiego di armi letali autonome viene confermata anche a fronte della specificazione di tre eventualità ad essa favorevoli: che vengano utilizzate contro obiettivi esclusivamente militari; come ultima soluzione in situazioni estreme; per ridurre le perdite del proprio esercito. A scegliere queste tre opzioni di risposta sono rispettivamente il 18%, il 12% e il 9% dei rispondenti. La grande maggioranza del campione (60%) ritiene che il loro utilizzo non sia mai giustificato (v. graf.2), una significativa differenza a fronte dell'intransigente rifiuto dell'opinione pubblica americana di accettare caduti tra i propri soldati (Sagan, Valentino, 2017).

Graf. 2 – Esistono circostanze nelle quali è giustificato l'utilizzo delle armi autonome?



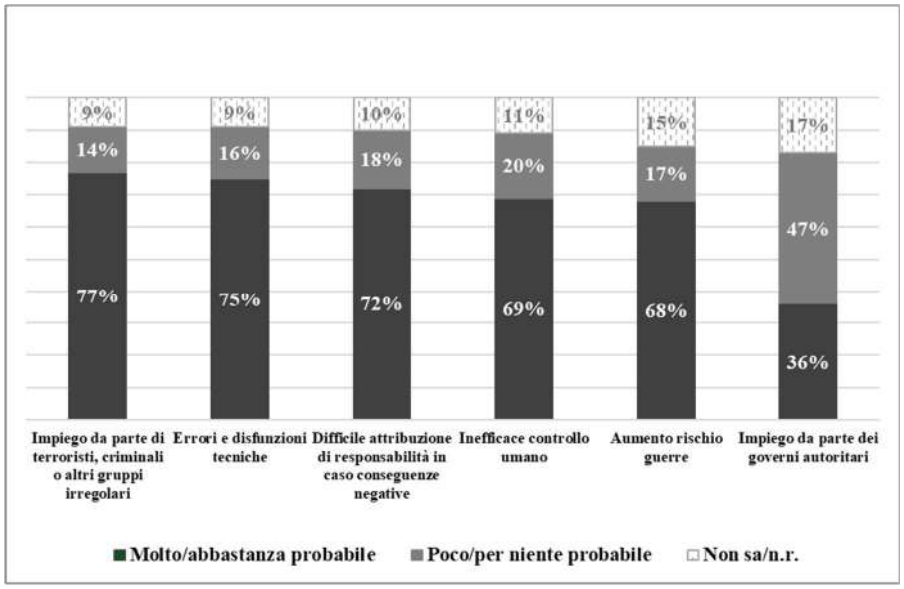
Archivio Disarmo, *Difebarometro* n. 11, febbraio 2025.

La contrarietà non appare mitigata dalla paventata possibilità di una guerra "a perdite zero", rispetto alla quale il campione intervistato mostra scetticismo. Il 61% degli intervistati si dichiara infatti molto o abbastanza d'accordo con l'affermazione che l'uso di armi autonome comporterà un aumento delle vittimi civili (a fronte del 24% che è in disaccordo e il 16% che non esprime un'opinione).

Messi di fronte a possibili conseguenze dell'adozione delle armi autonome, ai rispondenti è stata chiesta l'opinione circa i rischi che potrebbero verificarsi. Tra quelli proposti, il rischio che viene ritenuto più probabile è il loro impiego da parte di terroristi, criminali o altri gruppi illegali (con il 77% dei rispondenti che lo ritengono molto o abbastanza probabile), immediatamente seguìto dalla possibilità di errori e disfunzioni tecniche (75%). Anche la difficile attribuzione di responsabilità in caso di conseguenze fatali (aspetto su cui torneremo più avanti) è ritenuta altamente o abbastanza probabile da una larga maggioranza dei rispondenti (72%). A tali preoccupazioni seguono quelle di un inefficace controllo umano (69%) e di un aumento del rischio di guerre (68%). Delle opzioni presentate, l'unica ritenuta poco o per niente probabile dalla maggioranza relativa degli intervistati (47%) è il

possibile impiego da parte dei governi autoritari (v. graf. 3). Ancora una volta, alla dichiarata mancanza di conoscenza rispetto all'applicazione dell'IA in campo militare, corrisponde una corretta intuizione rispetto a quelli che vengono individuati come i principali rischi connessi al suo eventuale impiego da parte di numerosi scienziati (Altman, 2013; Altman, Sauer, 2017).

Graf. 3 – Possibili rischi dell'impiego delle armi letali autonome.



Archivio Disarmo, *Difebarometro* n. 11, febbraio 2025.

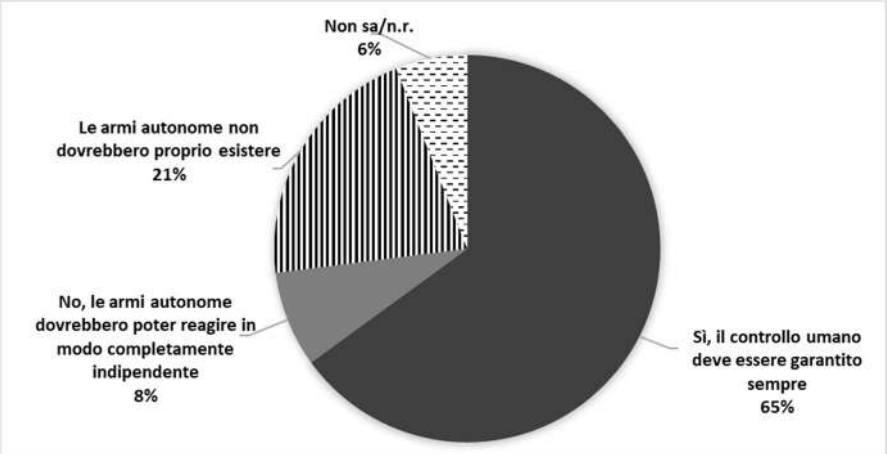
La grande maggioranza dei rispondenti esprime dunque una forte preoccupazione sulla probabilità che il principio di responsabilità, uno dei pilastri del Diritto internazionale umanitario (Tamburrini, 2020), possa essere messo in crisi dall'impiego delle armi letali autonome in ambito bellico, rendendo fluttuanti le responsabilità personali e istituzionalizzando l'adiaforizzazione morale degli attori denunciata da Bauman e Lyon (2014).

La consapevolezza dei rischi legati all'impiego delle armi autonome sul campo di battaglia, nonché la possibile violazione dei principi base del Diritto umanitario che ne scaturirebbe (ICRC, 2021) pone all'ordine del giorno la necessità di una regolamentazione sull'utilizzo di tali armamenti sul campo di battaglia. Come prima difficoltà incontrata dal tentativo è quella di giungere ad una definizione condivisa di sistema d'arma autonomo, portando

al centro della discussione la questione relativa a quale controllo umano si debba esercitare su tali sistemi d'arma (Tamburrini, 2023).

A tale aspetto è dedicata l'ultima parte della nostra analisi, che interroga il campione di rispondenti sul grado di autonomia che le armi "intelligenti" dovrebbero raggiungere. Coerentemente con le opinioni precedentemente espresse, la maggioranza degli intervistati (65%) dichiara che il controllo umano dovrebbe essere sempre garantito, contro un esiguo 8% che ritiene che le armi autonome dovrebbero poter reagire in modo completamente indipendente. Notevole, inoltre, il 21% dei rispondenti che sostiene come, a prescindere dal grado di autonomia loro garantito, tali sistemi d'arma non dovrebbero proprio esistere (v. graf. 4).

Graf. 4 – Ritiene necessario mantenere un controllo umano sulle decisioni di attacco delle armi letali autonome?



Archivio Disarmo, *Difebarometro* n. 11, febbraio 2025.

Se la necessità di garantire un controllo umano è auspicata dalla grande maggioranza dei rispondenti, più difficile sembra individuare una via maestra attraverso cui giungere a tale regolamentazione.

Interrogati su quale possa essere la misura più efficace per regolamentare la ricerca e l'impiego delle armi autonome, gli intervistati non escludono né i parlamenti né i governi nazionali, mentre il 17% punterebbe sulla sensibilizzazione e mobilitazione dell'opinione pubblica. Risultati interessanti e probabilmente inattesi sono offerti dalle rimanenti risposte. Sommando tra loro il 20% per cui le decisioni in merito andrebbero affidate al parlamento e alla governance europea, il 21% che privilegerebbe negoziati multilaterali e trattati internazionali e il 23%, che affiderebbe le decisioni alle Nazioni

Unite, quasi 2/3 del campione sottoscrive una soluzione esplicitamente multilaterale. Si tratta di un dato confortante, considerato che il multilateralismo che si richiama all'Europa e alle Nazioni Unite è soggetto oggi all'attacco concentrico da parte di unilateralisti, sovranisti e populisti.

Conclusioni

Alla luce delle riflessioni fin qui esposte, appare evidente come le opinioni degli italiani sull'applicazione dell'IA al contesto bellico non si riducano a un giudizio tecnico o morale sul ricorso alla forza. Piuttosto, esse riflettono un intreccio più profondo di rappresentazioni collettive: da un lato, quelle relative alla legittimità dell'azione militare, in un contesto post-eroico in cui il valore attribuito alla vita dei soldati propri ha assunto un peso politico crescente; dall'altro, quelle riguardanti l'accettazione sociale dell'intelligenza artificiale, intesa come tecnologia simbolicamente carica di promesse e di timori. In questo senso, il sondaggio presentato in queste pagine costituisce un'occasione per interrogare il rapporto tra società e tecnologia, mettendo in luce in che modo gli italiani si posizionano di fronte a un futuro in cui la guerra potrebbe essere sì condotta senza soldati in campo, ma con conseguenze politiche, morali e simboliche al momento non immaginabili nelle forme ma sicuramente critiche nella loro natura.

Riferimenti bibliografici

Abraham Y. (2024). "Lavender": The AI machine directing Israel's bombing spree in Gaza. *+972 Magazine.* 3 aprile 2024. https://www.972mag.com/lavender-ai-israeli-army-gaza/.

Altmann J. (2013). Arms control for armed uninhabited vehicles: an ethical issue. *Ethics and Information Technology*, 15: 137-152.

Altmann J., Sauer F. (2017). Autonomous Weapon Systems and Strategic Stability. *Survival. Global Politics and Strategy*, 59 (5): 117-142.

Battistelli F. (2023). Armi non umane. Miti, sogni e incubi dell'autonomia delle armi. In: Farruggia F., a cura di. *Dai droni alle armi autonome. Lasciare l'apocalisse alle macchine?* Milano: FrancoAngeli.

Battistelli F. (2024). Dalle macchine di Touring alle macchine di Stranamore? I rischi delle applicazioni delle IA alla sicurezza internazionale e interna. *Iriad Review. Studi sulla pace e sui conflitti*, 1: 4-16.

Bauman Z. (1999). *La solitudine del cittadino globale.* Milano: Feltrinelli.

Bauman Z., Lyon D. (2014). *Sesto potere. La sorveglianza sulla modernità liquida*. Roma-Bari: Laterza.

Birch M. (2024). Who did that? AI assisted targeting and the lowering of thresholds in Gaza. *Medicine, Conflict and Survival,* 40 (2): 97-100.

Caren N., Ghoshal R.A., Ribas V. (2010). A Social Movement Generation: Cohort and Period Trends in Protest Attendance and Petition Signing. *American Sociological Review*, 20 (10): 1-27.

Crawford N.C. (2013). Targeting Civilians and U.S. Strategic Bombing Norms: *Plus ça change, plus c'est la même chose*? In Evangelista H., Shue H., eds. *The American Way of Bombing; Changing Ethical and Legal Norms, from Flying Fortresses to Drones*. Ithaca, N.Y.: Cornell University Press: 64-86.

Douhet G. (1921). Il dominio dell'aria e altri scritti. In: Bozzo L., a cura di, *Aeronautica militare*, Ufficio Storico: Roma, 2022.

Everts PH., Isernia P. (eds.) (2001) *Public Opinion and the Use of Force*. London: Routledge.

Farruggia F., (2023a) (a cura di). *Dai droni alle armi autonome. Lasciare l'apocalisse alle macchine?* Milano: FrancoAngeli.

Farruggia F., (2023b). L'opinione pubblica di fronte alle armi autonome. In: Farruggia F., a cura di. *Dai droni alle armi autonome. Lasciare l'apocalisse alle macchine?* Milano: FrancoAngeli: 127-141.

Farruggia F. (2024). Oltre l'immaginario. Quando l'IA va in guerra. *Im@go*. 23 (XIII): 235-247.

Gelpi C., Feaver P.D., Reifler J. (2009). *Paying the Human Costs of War: American Public Opinion and Casualties in Military Conflicts*. Princeton N.J.: Princeton University Press.

Greco L. (1999). *Homo militaris. Antropologia e letteratura della vita miliare*. Milano: FrancoAngeli.

ICRC, International Committee of the Red Cross (2021). *Exploring Humanitarian Law. Linee guida per la sperimentazione e la valutazione*. Tr. it. Comitato Croce Rossa Italiana, testo disponibile al sito: https://www.icrc.org/sites/default/ files/external/files/ehl/ehl-italian-guidelines-for-experimentation-and-evaluation.pdf.

King A. (2010). The Afghan War and "postmodern" memory: commemoration and the dead of Helmand. *The British Journal of Sociology*, 61(1): 1-25.

Moskos C. (1994). *Sociologia e soldati*. Milano: FrancoAngeli.

Osservatorio Scienza Tecnologia e Società (2025). *Annuario Scienza Tecnologia e Società*. 1, inverno 2025.

Rogers A.P.V. (2016). Zero-casualty Warfare. In Schmitt M.N., Heintschell von Heinegg W., eds. *The Conduct of Hostilities in International Humanitarian Law* (Vol. I, pp. 63-80). Abingdon: Routledge.

Sagan S.D., Valentino B.A. (2017). Revisiting Hiroshima in Iran. What Americans Really Think about Using Nuclear Weapons and Killing Noncombatants. *International Security*, 42(1): 41-79.

Tamburrini G. (2020). *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*. Roma: Carocci.

Tamburrini G. (2023). Il dibattito etico sulle armi autonome. In: Farruggia F., a cura di. *Dai droni alle armi autonome. Lasciare l'apocalisse alle macchine?* (pp. 98-112). Milano: FrancoAngeli.

Sitografia dei sondaggi

https://www.ipsos.com/sites/default/files/2017-03/AWS%207555.pdf

https://www.ipsos.com/sites/default/files/ct/news/documents/2019-01/human-rights-watch-autonomous-weapons-pr-01-22-2019_0.pdf

https://www.archiviodisarmo.it/view/TkplelUwMDc0aHc5akZTcDJzdWpLUT09Ojq8Zq-8dDtJ_oJb8ZLCSU/gennaio-2024-iriadreview.pdf

https://www.ipsos.com/sites/default/files/ct/news/documents/2021-01/ipsos_glo-bal_advisor_ _lethal_autonomous_weapons_survey_-_nov_2020-jan_2021.pdf

Opportunità e ostacoli nella formazione sull'intelligenza artificiale: una prospettiva socio-educativa

di *Massimiliano Ruzzeddu, Laura Guercio**

L'intelligenza artificiale (IA) è ormai un fattore centrale in molti ambiti, dall'industria alla sanità, offrendo opportunità di innovazione e crescita ma anche nuove sfide etiche, sociali e politiche. La formazione svolge un ruolo decisivo per promuovere un uso consapevole, capace di valorizzare i benefici e contenere i rischi. L'articolo evidenzia la necessità di un approccio interdisciplinare che unisca competenze tecniche e umanistiche, affrontando temi come democrazia algoritmica, impatto dell'automazione sul lavoro, digital divide e disinformazione, e propone orientamenti formativi e regolativi inclusivi per l'era dell'IA.

Parole chiave: intelligenza artificiale; formazione; etica; automazione; digital divide; regolamentazione.

Opportunities and obstacles in artificial intelligence education: a socio-educational perspective

Artificial intelligence (AI) has become a central factor in many fields, from industry to healthcare, offering opportunities for innovation and growth but also raising new ethical, social, and political challenges. Education plays a decisive role in fostering conscious use, capable of enhancing benefits and limiting risks. The article highlights the need for an interdisciplinary approach that combines technical and humanistic skills, addressing issues such as algorithmic democracy, the impact of automation on work, the digital divide, and disinformation, and proposes inclusive educational and regulatory directions for the AI era.

Keywords: artificial intelligence; education; ethics; automation; digital divide; regulation.

Introduzione

Nel corso degli ultimi decenni, l'intelligenza artificiale ha subito un'evoluzione spettacolare. Dai primi sistemi esperti e dagli algoritmi di ricerca, si è

DOI: 10.5281/zenodo.17297415

* Università Niccolò Cusano. massimiliano.ruzzeddu@unicusano.it, laura.guercio@unicusano.it.

passati all'apprendimento automatico (machine learning) e alle reti neurali profonde (deep learning), capaci di processare enormi quantità di dati per apprendere schemi e relazioni complesse. Questa crescita esponenziale della potenza di calcolo e della disponibilità di dati ha consentito all'IA di estendersi a innumerevoli ambiti applicativi: dalla medicina, dove si utilizzano algoritmi di riconoscimento di immagini per la diagnosi precoce di patologie, alla finanza, che sfrutta sistemi di trading automatizzati e sofisticati modelli di valutazione del rischio.

Parallelamente, l'IA è diventata oggetto di un ampio dibattito pubblico, aprendo orizzonti d'innovazione in settori come la salute (grazie a sistemi di diagnosi precoce), l'agricoltura di precisione (con l'uso di droni e sensori intelligenti), la mobilità sostenibile (veicoli a guida autonoma), e le smart city (Acemoglu, Restrepo, 2018). Tali ambiti possono stimolare una nuova stagione di crescita economica e occupazionale, purché ci si prepari all'aggiornamento delle competenze e si mitighino i rischi di concentrazione del potere nelle mani di pochi attori globali (ILO, 2020). Inoltre, la sua capacità dell'AI di influire sui processi decisionali, di sostituire alcune mansioni lavorative e di generare forme di sorveglianza e analisi predittiva ha sollevato questioni cruciali per la democrazia e i diritti dei cittadini. Come sottolineano Mossberger, Tolbert e McNeal (2008), la cittadinanza nell'era digitale richiede accesso equo, partecipazione attiva e alfabetizzazione civica. Allo stesso modo, Jenkins *et al.* (2009) parlano di "new media literacies" come competenze fondamentali per agire responsabilmente in ambienti digitali, inclusa la capacità di valutare fonti, contribuire a spazi pubblici online e collaborare con altri utenti Accanto a queste trasformazioni centrali, emergono altre implicazioni che richiamano l'attenzione su rischi quali l'acuirsi del divario digitale, la diffusione della disinformazione e l'uso di algoritmi per finalità di controllo. In particolare, strumenti come la Emotional AI, se impiegati senza adeguate garanzie etiche e regolamentari, potrebbero influenzare le emozioni degli utenti e condizionare indirettamente anche le dinamiche partecipative e democratiche.

In altre parole, si comincia a delineare un cambiamento culturale che si traduce nella costruzione di una "cittadinanza digitale" consapevole, come elemento per una governance efficace dell'IA. Emergono quindi bisogni sociali di formazione non limitati alle 'semplici' competenze tecniche, ma includenti anche l'alfabetizzazione critica e etica verso le tecnologie, consentire una partecipazione a quanto possibile allargata alle scelte che influenzano la vita sociale di una collettività

Questo significa che non basta affidare l'innovazione tecnologica a un ristretto gruppo di esperti: è fondamentale dotare ogni individuo di strumenti per comprendere, utilizzare e governare l'IA in modo responsabile. Ciò implica un

ampio ripensamento dei percorsi formativi, dalla scuola primaria all'università (Prinsloo, 2017; Renz *et al.*, 2020; Selwyn, 2016; *Zawacki-Richte, 2019,* Zeide, 2019*)*, fino alla formazione professionale e continua.

1. L'IA e il lavoro

L'intelligenza artificiale (IA) sta trasformando in profondità il mondo del lavoro, modificando processi produttivi, ridefinendo competenze professionali e generando nuove opportunità occupazionali. Come sottolinea Sohn-Rethel (1978), la storica separazione tra pensiero astratto e attività pratica ha determinato profonde conseguenze nella strutturazione della conoscenza e del lavoro nella società moderna.

Autori come Susskind e Susskind (2015) nel loro libro *The Future of the Professions* sottolineano come l'automazione e l'IA possano rivoluzionare profondamente i ruoli tradizionali, non solo sostituendo attività ripetitive, ma anche trasformando professioni complesse.

Fin dalle prime rivoluzioni industriali, l'avvento di macchine e strumenti automatizzati ha determinato la scomparsa di alcune mansioni e la nascita di altre. Tuttavia, l'IA porta l'automazione a un livello superiore, in quanto non si limita a sostituire il lavoro fisico, ma si estende alle attività cognitive e decisionali. Studi come quello di Frey e Osborne (2017) sostengono che una parte rilevante dei compiti attualmente svolti da esseri umani potrebbe essere automatizzata entro qualche decennio. I settori maggiormente esposti comprendono logistica, customer service, contabilità, guida di veicoli e persino alcune forme di consulenza professionale. Uno studio sullo stesso argomento di Gmyrek *et al.* (2023) analizza il rischio di sostituzione dei lavori da parte dell'intelligenza artificiale, utilizzando la classificazione delle professioni IBSCO. Gli autori associano a ciascuna professione le mansioni specifiche e valutano, tramite l'utilizzo di GPT, quali di queste siano più facilmente eseguibili da sistemi di IA. I risultati indicano che le occupazioni maggiormente a rischio sono quelle di medio-alto livello: professionisti, funzionari e tecnici che producono o gestiscono informazioni (come consulenti, progettisti, insegnanti, diagnostici). Al contrario, i lavori manuali e a bassa qualifica risultano meno esposti, in quanto richiedono competenze meno standardizzabili. Particolarmente vulnerabili sono i lavoratori impiegati nella gestione di dati e rapporti standardizzati con il pubblico, come i segretari e i receptionist. Lo studio evidenzia dunque un rischio di polarizzazione del mercato del lavoro: la sostituzione tecnologica colpisce soprattutto i ruoli intermedi, mentre i lavoratori meno qualificati o altamente decisionali potrebbero essere meno toccati. In parallelo, l'espansione dell'IA genera nuove professioni,

come quelle legate alla progettazione di algoritmi, alla manutenzione di sistemi robotici, alla sicurezza informatica e all'analisi dei dati. Il World Economic Forum (2020) stima che, accanto ai posti di lavoro persi per effetto dell'automazione, ne verranno creati altri in settori altamente specializzati.

2. Formazione e Interdisciplinarità

In altri termini, l'intelligenza artificiale ha il potenziale per automatizzare le attività routinarie, ma crea anche opportunità per lavori che richiedono la collaborazione tra esseri umani e IA. Per sfruttare queste opportunità, esistono requisiti tecnici molto rigorosi. Per esempio, Biswas A., & Talukdar W. (2025) considerano indispensabile la padronanza di linguaggi di programmazione (Ivi: 69), machine learning (Ivi: 129-132) e dei LLM, Large Language Models (Ivi: 68-69).

Ora, come mostrano Sapkota *et al.*, (2025) l'evoluzione dell'intelligenza artificiale evolve con rapidità impressionante: essi mostrano come prima del 2022, l'intelligenza artificiale era dominata da agenti logici, modelli basati su regole e sistemi esperti, spesso rigidi e privi di capacità adattive (Ivi: 4-5). Con l'avvento di modelli linguistici di grandi dimensioni come GPT, emerge il paradigma degli AI Agents, agenti autonomi dotati di capacità di ragionamento, pianificazione e uso di strumenti, ma operanti in modo singolo e limitato al compito (Ivi: 5-6). A partire dal 2023, prende forma il concetto di Agentic AI, ovvero architetture distribuite composte da agenti specializzati che collaborano tra loro (Ivi: 13-15). Questi sistemi introducono funzionalità come la memoria condivisa, il coordinamento dinamico e la decomposizione adattiva dei compiti. L'agente non è più soltanto un esecutore, ma parte di una rete cognitiva più ampia, in grado di apprendere, adattarsi e agire in contesti complessi. Tale evoluzione segna un passaggio da un'intelligenza isolata e lineare a un'intelligenza collettiva, proattiva e orientata agli obiettivi (Ivi: 15).

Di fronte a cambiamenti così drastici, che praticamente si susseguono anno per anno, è improbabile che un lavoratore o una lavoratrice possa basarsi solo sulle competenze tecniche assimilate durante la fase scolastica della vita; ovviamente, nelle scuole di ogni ordine e grado, andrebbe potenziata l'educazione digitale incentivando moduli di alfabetizzazione ai media e introducendo in modo graduale nozioni di programmazione e problem solving.

Tuttavia, è importante che il processo di apprendimento avvenga lungo il corso dell'intera vita attiva (possibilmente anche dopo): infatti, l'incontro tra la domanda e l'offerta di competenze non avviene in modo spontaneo, richiedendo adeguate strategie di formazione e incentivi per aziende e lavoratori (Topol,

2019; Russell, 2021). Eric Topol (2019), in particolare, sottolinea come l'adozione dell'intelligenza artificiale, specialmente nel settore sanitario, non possa prescindere da un investimento significativo nella formazione del personale. Egli evidenzia la necessità di sviluppare non solo competenze tecniche, ma anche competenze empatiche e relazionali, affinché la tecnologia possa supportare e non sostituire la dimensione umana dei servizi. È importante, secondo Topol, che la formazione costituisca un processo continuo e adattivo, capace di accompagnare i cambiamenti nei contesti professionali e favorire l'integrazione tra competenze umane e artificiali. Allo stesso modo, Stuart Russell (2021) ribadisce l'importanza di una formazione orientata non solo all'efficienza tecnica, ma anche alla comprensione critica e responsabile dell'IA. Egli richiama l'urgenza di un nuovo paradigma educativo che metta al centro la capacità degli individui di interrogarsi sui fini delle tecnologie e sul loro impatto etico e sociale. Per Russell, promuovere una cultura dell'apprendimento significa anche fornire strumenti per una cittadinanza consapevole, in grado di partecipare attivamente alla definizione delle regole e dei valori che devono guidare lo sviluppo dell'intelligenza artificiale.

In effetti, la formazione continua è importante, da un punto di vista sociologico, soprattutto per un altro motivo: le implicazioni etiche. Ad esempio, un problema cruciale riguarda la presenza di bias nei dati di addestramento e nelle procedure algoritmiche O' Neil, 2016). Se i set di dati utilizzati per costruire i modelli di IA sono parziali o riflettono pregiudizi storici, l'algoritmo tenderà a discriminare alcuni gruppi, ad esempio escludendo le donne o le minoranze da determinate offerte di lavoro o di credito. Tali dinamiche possono acuire le disparità socio-economiche e violare i principi di uguaglianza sanciti dalle normative in molti Paesi. È quindi imperativo formare professionisti che, oltre a conoscere le tecniche di machine learning, sappiano vigilare sulle basi di dati e sui processi di addestramento, introducendo metodologie per la rimozione di bias. entrambe sviluppabili attraverso un'educazione multidisciplinare (Gagliasso, 2020; Eubanks, 2018; Noble, 2018).

Le dimensioni filosofica, sociologica e antropologica diventano quindi centrali nella formazione continua, risultando indispensabili nell'analisi dell'impatto dell'IA sui valori, sulle relazioni sociali e sui diritti individuali e collettivi. L'analisi sociologica sposta l'attenzione sulle trasformazioni delle strutture sociali e delle relazioni interpersonali indotte dall'IA. Studi sociologici evidenziano come algoritmi e sistemi automatizzati possano riprodurre o amplificare bias preesistenti, modificare le dinamiche di partecipazione democratica, alterare il mondo del lavoro e ridefinire concetti come identità professionale e collaborazione. Il focus è sulle conseguenze sociali, sulle dinamiche di controllo e sorveglianza, ma anche sulle opportunità di empowerment e innovazione

sociale. L'antropologia, infine, esamina le trasformazioni culturali e simboliche generate dall'interazione tra umani e macchine intelligenti. Essa valorizza la dimensione esperienziale e la pluralità delle culture, evidenziando come l'introduzione dell'IA venga mediata da contesti specifici, valori locali e narrazioni diverse, e come tale processo possa ridefinire l'identità collettiva, la relazione con la tecnologia e la stessa concezione di "umanità" (Suchman, 2007).

Di conseguenza, si impone la necessità di formare figure professionali ibride, capaci di coniugare competenza tecnica e coscienza critica delle implicazioni etiche, sociali e culturali dell'innovazione tecnologica. L'interdisciplinarietà non deve essere vista come un'opzione marginale, bensì come una strategia imprescindibile per governare la complessità delle sfide poste dall'IA (Gagliasso, 2020). Si tratta di sviluppare un "pensiero critico-tecnico" che consenta di mettere in discussione paradigmi consolidati e di riconoscere la scienza come pratica umana inserita in contesti storici e culturali specifici, mai neutra o disincarnata. Solo così sarà possibile formare figure professionali consapevoli, in grado di governare responsabilmente le sfide poste dalle nuove tecnologie, che possiedano capacità critiche attraverso il pensiero analitico, l'abilità di valutare l'affidabilità delle fonti e l'interpretazione consapevole dei risultati prodotti dagli algoritmi Inoltre, sono essenziali conoscenze giuridiche che riguardano le normative sulla privacy, i regolamenti per la gestione dei dati, gli strumenti di governance algoritmica e la responsabilità sociale d'impresa. L'obiettivo è formare figure ibride, capaci di orientarsi sia negli aspetti tecnici sia nelle ricadute sociali dell'IA.

3. Scenari politici nel prossimo futuro

La traiettoria futura dell'impatto sociale dell'intelligenza artificiale appare strettamente legata alle scelte politiche, educative e culturali che le società saranno in grado di compiere.

Possiamo quindi ipotizzare tre grandi scenari sulle tendenze future a medio e lungo termine in base alle scelte che i decisori politici metteranno in atto.

1. Non intervento. In questo scenario, l'IA prosegue la sua corsa trainata principalmente dalle forze di mercato, con una diffusione rapida ma profondamente disomogenea. I settori caratterizzati da alta intensità di capitale e competenze specialistiche accelerano la loro crescita e innovazione, mentre intere aree geografiche, regioni meno sviluppate e fasce sociali vulnerabili rimangono ai margini di questa trasformazione. In effetti, studi come quelli di Brynjolfsson e McAfee (2014) in *The Second Machine Age*, evidenziano come la crescita tecnologica accelerata possa amplificare le disuguaglianze se non si adottano

misure redistributive efficaci e politiche attive del lavoro. Tale sviluppo non regolamentato rischia di innescare tensioni economiche e sociali sempre più accentuate, alimentando una polarizzazione sociale e geopolitica dalle conseguenze difficilmente prevedibili. La mancanza di intervento pubblico e coordinato potrebbe, infatti, accentuare le disuguaglianze, alimentando fenomeni di esclusione e di conflitti sociali legati alla redistribuzione delle risorse e delle opportunità lavorative.

2. Politiche di sussidio. Qui i governi intervengono con misure di tipo assistenziale, come l'introduzione di forme di reddito garantito di base o il rafforzamento dei sussidi di disoccupazione, con l'obiettivo di tamponare l'impatto sociale dell'automazione e della sostituzione del lavoro umano. Sebbene tali politiche possano offrire un sollievo immediato a coloro che perdono il lavoro, esse presentano rischi rilevanti. Innanzitutto, possono generare una dipendenza cronica dalle forme di sostegno pubblico, disincentivando la ricerca attiva di nuove opportunità lavorative e la riqualificazione professionale. Inoltre, potrebbero acuire tensioni sociali tra chi è occupato e chi invece vive grazie ai sussidi, compromettendo la coesione sociale e la percezione di equità all'interno della società. Questo modello rischia, quindi, di non affrontare la questione di fondo, ovvero la necessità di un cambiamento strutturale nel sistema formativo e produttivo che consenta a tutti di riqualificarsi e adattarsi a un mercato del lavoro in profonda trasformazione. Per esempio, l'esperienza della Finlandia con il suo esperimento di reddito di base (Kangas, 2019) ha mostrato come misure di sostegno economico senza un parallelo investimento nella formazione e nell'empowerment individuale possano risultare insufficienti per garantire una reale autonomia lavorativa e sociale.

3. Welfare abilitante. Questo scenario rappresenta una strategia più lungimirante e strutturale, in cui gli investimenti pubblici si concentrano su istruzione, formazione continua e aggiornamento professionale lungo tutto l'arco della vita. Il welfare abilitante mira a fornire a tutta la cittadinanza le competenze necessarie non solo per accedere a professioni qualificate, ma anche per comprendere, accettare e, in molti casi, partecipare attivamente all'ideazione e all'introduzione di innovazioni tecnologiche avanzate. Da un punto di vista strategico, come sottolineato da Hemerijck (2008) e Agrawal *et al.* (2018), questo modello permette di rimodulare le fasi della vita individuale nell'era della quarta rivoluzione industriale: si sostiene infatti l'allungamento del tempo dedicato alla formazione e all'aggiornamento, un costo sociale iniziale che viene però compensato da una fase lavorativa più breve ma al contempo più produttiva e gratificante. In questo modo, si costruisce una società più resiliente, capace di adattarsi alle rapide trasformazioni tecnologiche senza lasciare indietro nessuno, valorizzando l'apprendimento permanente come pilastro fondamentale per la crescita personale e

collettiva. La strategia europea per l'intelligenza artificiale, ad esempio, mette al centro proprio questo tema: lo sviluppo di competenze digitali e la formazione continua[1]. L'idea è chiara: non basta formare pochi esperti di IA, ma è necessario costruire un ecosistema dove ogni cittadino possa acquisire e aggiornare le proprie capacità in modo costante, adattandosi a un mercato del lavoro in rapida evoluzione. Per questo motivo, si prevedono investimenti importanti, non solo per integrare l'educazione digitale nelle scuole e nelle università, ma anche per offrire percorsi di riqualificazione professionale per chi è già inserito nel mondo del lavoro.

Guardando oltre l'Europa, troviamo esempi concreti di come questi principi possano tradursi in azioni efficaci. Un caso emblematico è il programma *SkillsFuture* di Singapore[2], un programma statale che offre anche incentivi alle imprese per facilitare la formazione dei dipendenti, creando un circuito virtuoso tra lavoratori, aziende e istituzioni educative. Questa flessibilità e attenzione alla personalizzazione dei percorsi formativi permette di mantenere alta la motivazione delle persone e di rispondere in modo efficace alle diverse esigenze del mercato.

Questi esempi mostrano come un approccio integrato alla formazione continua e allo sviluppo delle competenze sia fondamentale per affrontare le sfide del futuro, garantendo non solo una maggiore competitività economica, ma anche una società più inclusiva e resiliente.

In conclusione, la letteratura e le esperienze pratiche convergono nel riconoscere che un futuro sostenibile e inclusivo nell'era dell'intelligenza artificiale richiede un approccio integrato e multidimensionale. Il welfare abilitante emerge come la via più promettente per trasformare le sfide tecnologiche in opportunità di crescita sociale, ma solo se accompagnato da regolamentazioni adeguate, investimenti nella formazione continua e un impegno collettivo verso una cittadinanza digitale responsabile e partecipativa.

Conclusioni

L'evoluzione dell'IA impone un apprendimento continuo: il suo sviluppo rapido trasforma costantemente settori economici e vita quotidiana, rendendo essenziale una formazione che integri programmazione, analisi dei dati e machine learning. Tuttavia, mentre la letteratura sul tema si concentra sulla competenza tecnica, con questo lavoro abbiamo cercato di mostrare come l'educazione

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018: 237:FIN par. 3.2.
[2] https://www.skillsfuture.gov.sg/.

umanistica sia altrettanto necessaria alla gestione di queste innovazioni. In particolare, comprendere i bias negli algoritmi, garantire trasparenza e valutare implicazioni etiche richiede una preparazione completa, unendo scienze e discipline umanistiche. In una società permeata dall'IA, è fondamentale acquisire un approccio di apprendimento autonomo e strategico. Rifacendoci a Bateson (1972), possiamo definire questo processo come deutero-apprendimento: non solo l'acquisizione di nuove informazioni, ma anche la consapevolezza delle strategie per apprendere in modo più efficace. Significa comprendere i criteri con cui organizziamo i dati, imparare a formulare domande corrette e individuare fonti affidabili. Questa capacità di adattamento accelera lo sviluppo personale e professionale, permettendo di navigare con successo nel panorama tecnologico in costante evoluzione.

In quest'ottica, il concetto di paradigma di Edgar Morin (1999) diventa centrale: i modelli interpretativi selezionano le strutture cognitive più rilevanti di un'epoca, influenzando la nostra comprensione del mondo. Non basta padroneggiare il software più recente: servono capacità analitiche, problem solving e gestione efficace delle informazioni. Infine, l'IA non è solo una questione tecnica, ma un fenomeno sociale e culturale. Per un utilizzo consapevole e responsabile, è necessaria una formazione olistica che integri competenze scientifiche, umanistiche ed etiche, guidando il progresso in modo equo e sostenibile.

Riferimenti bibliografici

Acemoglu D., Restrepo P. (2018). *Artificial Intelligence, Automation, and Work*. NBER Working Paper No. 24196.

Agrawal A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Review Press.

Bateson G. (1972). *Steps to an Ecology of Mind. Northvale*. New Jersey: Jason Aronson Inc.

Biswas A., Talukdar W. (2025). *Building agentic AI systems: Create intelligent, autonomous AI agents that can reason, plan, and adapt*. NJ: Packt Publishing.

Brynjolfsson E., McAfee A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W.W. Norton & Company.

Commissione Europea (2021). *Coordinated Plan on Artificial Intelligence 2021 Review. European Commission*. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

Doran D., Schulz S., Besold T.R. (2017). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives*. CEUR Workshop Proceedings, 2071.

Eubanks V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Floridi L. (2019). *The Ethics of Artificial Intelligence*. Oxford: Oxford University Press.

Frey C.B., Osborne M.A. (2017). The Future of Employment: How Susceptible Are Jobs to Computerization? *Technological Forecasting and Social Change*, 114: 254–280.

Gagliasso R. (a cura di) (2020). *Per una scienza critica*. Milano: FrancoAngeli.

Gmyrek P., Berg J., Bescond D. (2023). *Generative AI and jobs: A global analysis of potential effects on job quantity and quality*. ILO Working Paper 96, Geneva: ILO. https://www.ilo.org/publications/generative-ai-and-jobs-global-analysis-potential-effects-job-quantity-and

Hemerijck A. (2008). *Changing Welfare States*. Oxford: Oxford University Press.

ILO (2020). *The Future of Jobs Report 2020*. Geneva: International Labour Organization.

Jenkins H., Purushotma R., Weigel M., Clinton K., Robison A. (2009). *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century*. Cambridge, MA: MIT Press.

Jungherr A. (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society*, 9(3). https://doi.org/10.1177/20563051231186353

Kangas O., Jauhiainen S., Simanainen M., Ylikännö M. (2019). *The Basic Income Experiment 2017-2018 in Finland: Preliminary results*. Helsinki: Ministry of Social Affairs and Health.

Micklitz H.-W., Pollicino O., Reichman A., Simoncini A., Sartor G., De Gregorio G. (2022). *Constitutional Challenges in the Algorithmic Society*. Cambridge, United Kingdom; New York, NY: Cambridge University Press.

Mittelstadt B.D., Allo P., Taddeo M., Wachter S., Floridi L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2): 1-21. https://doi.org/10.1177/2053951716679679

Morin E. (1999). *Les sept savoirs nécessaires à l'éducation du futur*. Paris: Unesco.

Mossberger K., Tolbert C.J., McNeal R.S. (2008). *Digital Citizenship: The Internet, Society, and Participation*. Cambridge, MA: MIT Press.

Noble S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Norris P. (2001). *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*. New York: Cambridge University Press.

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

Pedró F. (2020). Applications of Artificial Intelligence to Higher Education: Possibilities, Evidence, and Challenges. IUL Research, *Open Journal of IUL University*, 1(1): 61-76. https://www.iulresearch.it

Prinsloo P. (2017). Fleeing from Frankenstein's Monster and Meeting Kafka on the Way: Algorithmic Decision-Making in Higher Education. *E-Learning and Digital Media*, 14(3): 138-163. https://doi.org/10.1177/2042753017731355

Renz A., Krishnaraja S., Gronau E. (2020). Demystification of Artificial Intelligence in Education: How Much AI is Really in the Educational Technology? I*nternational Journal of Learning Analytics and Artificial Intelligence for Education*, 2(1). https://doi.org/10.3991/ijai.v2i1.12675

Russell S. (2021). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin.

Santoni de Sio F. (2024). *Human Freedom in the Age of AI*. New York: Routledge.

Sapkota M., Roumeliotis M., Karkee M. (2025). *AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges*. arXiv. https://doi.org/10.48550/arXiv.2406.03645

Selwyn N. (2016). *Education and Technology: Key Issues and Debates*. London: Bloomsbury Academic.

Sohn-Rethel A. (1978). *Intellectual and Manual Labour: A Critique of Epistemology*. London: Macmillan.

Suchman L. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions (2nd ed.)*. Cambridge: Cambridge University Press.

Susskind R., Susskind D. (2015). *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. Oxford: Oxford University Press.

Topol E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

UNESCO – United Nations Educational, Scientific and Cultural Organization (2024). *Artificial Intelligence and Democracy*.

Vinuesa R., Azizpour H., Leite I., Balaam M., Dignum V., Domisch S., Felländer A., Langhans S.D., Tegmark M., Fuso Nerini F. (2020). The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications*, 11(1): 233. https://doi.org/10.1038/s41467-019-14108-y

World Economic Forum (2020). *The Future of Jobs Report 2020*. Geneva: World Economic Forum.

Zawacki-Richter O., Marín V.I., Bond M., Gouverneur F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators? *International Journal of Educational Technology in Higher Education*, 16(39). https://doi.org/10.1186/s41239-019-0171-0

Zeide E. (2019). Artificial Intelligence in Higher Education: Applications, Promise and Perils, and Ethical Questions. *EDUCAUSE Review*, 54(3): 30-38. https://er.educause.edu/articles/2019/7/artificial-intelligence-in-higher-education

Zuboff S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

AI, media influence, and the legal challenges of freedom of speech

by *Katarzyna Chałubińska-Jentkiewicz* [*], *Urszula Soler*[**]

The article examines the urgent need for regulatory changes in response to AI's influence on the information environment. Social media, while adopting new mechanisms, still use traditional audience manipulation, raising challenges for human rights and freedoms. AI-generated content and disinformation threaten democracy, yet legal frameworks remain inadequate. The article stresses balancing media freedom with the need to address harmful content responsibly.

Keywords: media; AI; freedom of speech; legal challenges; information; disinformation.

IA, influenza dei media e le sfide legali della libertà di espressione

L'articolo esamina l'urgente necessità di cambiamenti normativi in risposta all'influenza dell'intelligenza artificiale (IA) sull'ambiente informativo. I media sociali, pur adottando nuovi meccanismi, continuano a utilizzare metodi tradizionali di manipolazione del pubblico, sollevando sfide per i diritti umani e le libertà fondamentali. I contenuti generati dall'IA e la disinformazione minacciano la democrazia, mentre i quadri giuridici rimangono inadeguati. L'articolo sottolinea l'importanza di bilanciare la libertà dei media con la necessità di affrontare responsabilmente i contenuti dannosi.

Parole chiave: media; IA; libertà di espressione; sfide legali; informazione; disinformazione.

Introduction

The Internet has become a space where vast amounts of information are processed while simultaneously upholding the undeniable need for the protection of personal rights and privacy. It is a true paradoxical space. The use of anthropomorphic interfaces, such as human-like voices used by assistants like Alexa and Siri, raises new concerns regarding privacy and the protection of personal rights. Research in social sciences indicates that people tend to interact with

DOI: 10.5281/zenodo.17297427

[*] Kozminski University. kchalubinska@kozminski.edu.pl.

[**] The Warsaw University of Life Sciences. urszula.soler@gmail.com.

technology as if it were human. This suggests that individuals may be more inclined to develop trust-based relationships with artificial intelligence designed to replicate human characteristics, making them more likely to share personal information in a way that is more widespread compared to other technologies that collect data through traditional means.

The main objective of this article is to analyze the impact of artificial intelligence (AI) on the digital information environment, particularly in the context of media communication, freedom of speech, privacy protection, and journalistic ethics. The study aims to assess how AI-driven technologies influence the creation, distribution, and perception of information, as well as the regulatory challenges associated with ensuring human rights and fundamental freedoms in the digital era.

The authors of the article have formulated the following research questions:

- How does AI influence the creation and dissemination of information in digital media?
- What are the primary risks and ethical concerns associated with AI-generated content, particularly in relation to privacy and journalistic integrity?
- How do AI-driven personalization algorithms affect public opinion, political discourse, and the spread of misinformation?
- What regulatory challenges exist in ensuring transparency and accountability for AI-generated content in media and communication?
- How should the legal system and media organizations adapt to address the risks posed by AI in information management and journalistic ethics?
- To what extent does AI contribute to the asymmetry between the rights of individual users and the interests of corporations or governmental institutions controlling AI-driven technologies?

This framework provides a structured approach to examining the implications of AI in the digital information ecosystem, emphasizing both its potential benefits and the emerging threats to democratic values, privacy, and media ethics. In turn, issues related to the need for new regulations in the field of AI functioning concern not only the issue of responsibility for AI actions. They also refer to questions about the limits of regulations related to human rights. The media and legal norms related to them are a simple reflection of the condition of a democratic state and its values, which include freedom of speech. Therefore, it is the example of the media that best illustrates the critical moment for freedoms previously considered unshakable. The article presents the issue of eternal

dilemmas affecting the limits of regulations related to human rights and freedoms.

1. Artificial intelligence and Privacy – definitional issues

In its simplest form, artificial intelligence (AI) is a subfield of computer science aimed at developing programs capable of performing tasks traditionally reserved for humans. These tasks can be considered intelligent, including visual and auditory perception, learning and adaptation, reasoning, pattern recognition, and decision-making.

The term "artificial intelligence" is used broadly to describe a collection of related techniques and technologies, including machine learning, predictive analytics, natural language processing, and robotics. While the philosophy of artificial intelligence has been debated since the early 18th century with Gottfried Wilhelm Leibniz, the modern concept of AI dates to the early 1940s and gained prominence with the development of the Turing Test, originally called the imitation game by Alan Turing in 1949 (Turing, 1950). AI has been intentionally programmed to be competent in a specific area, sometimes referred to as augmented intelligence, emphasizing its ability to enhance human intelligence. On the other hand, the concept of artificial general intelligence (AGI) refers to a level of intelligence that spans multiple domains. Based on the idea of AGI (Krishna, 2023), artificial superintelligence is generally considered AI that is both general and surpasses human intelligence.

Three key factors have contributed to the increasing presence of AI technologies in daily life:

1. Improved algorithms,
2. Enhanced computing power,
3. Greater capacity to capture and store vast amounts of data.

Numerous studies have examined the emergence of the "privacy paradox", in which people express concern about their privacy yet continue to willingly share personal information through the systems and technologies they use. A key question arises regarding the use of personal attributes and the intrusion into private spheres in relation to freedom of speech, which includes both the right to communicate and the right to communicate with. One example of such dilemmas is the case of Polish Nobel laureate Wisława Szymborska. In a radio broadcast, an AI-generated journalist conducted an interview in which Szymborska's voice was used to describe her emotions when receiving the Nobel Prize and to express opinions on events that occurred after her death. This case illustrates the

ethical and legal challenges posed by AI-driven content creation, particularly concerning the boundaries of privacy, consent, and posthumous digital identity.

2. Conditions for the dissemination of the image of public figures under the copyright and related rights act

The dissemination of an image requires the consent of the person depicted (Barta, 1997). The statutory requirement to obtain consent for the publication of an image is undisputed, as an individual's consent permits the publication of their image, while its absence equates to a prohibition on such publication (Daniluk, 2007). This means that, as a rule, the legislator requires consent for the dissemination of an image. However, exceptions to this rule are provided by law.

Consent is not required in the following cases:

1. When the person depicted has received an agreed payment for posing (for the recording of their image).
2. Under Article 81(2)(1) of the Copyright Act, when the image of a well-known person has been taken in connection with the performance of their public functions, particularly political, social, or professional roles, or when the person is merely a detail within a larger composition, such as a gathering, landscape, or public event.

The ratio Legis of Article 81(2) of the Copyright Act stems from the need to ensure creative freedom, allowing the media to fulfill their informational and documentary functions, granting them priority over individual personal rights. This provision enables:

- Informational, documentary, and reporting activities,
- The presentation of public events,
- The collection of materials for documentation purposes,
- The freedom to photograph public events[1].

The purpose of dissemination is not to make the image of a specific individual available but rather to depict a fragment of social life or the surrounding reality (Barta, Markiewicz, 2002: 17; Flisak, 2015: 1150). This restriction cannot be applied to a fictional or artificially generated situation, such as one created using artificial intelligence. The rules for disseminating the image of a public figure are intended to fulfill the fundamental role of the press and media, namely:

- Providing citizens with reliable information,

[1] Judgment of the Court of Appeal in Warsaw of January 13, 1999, case no. I ACa 1089/98, OSP 2000, vol. 9, item 142.

- Ensuring transparency in public life,
- Facilitating social oversight and criticism. (Barta, Markiewicz, 2005; Sieńczyło-Chlabicz, 2007)

However, these principles do not apply when the identity of a public figure is misrepresented, such as in the case of a fabricated interview with that person, which would constitute misleading the public.

The permissible dissemination of an image without the consent of the person depicted has been limited by the legislator to public figures. However, the legislator has not provided a normative definition of the term "public figure". In legal doctrine and jurisprudence, it is emphasized that in determining whether a person is publicly known, key factors include:

- Holding political or social functions,
- Gaining popularity beyond their immediate environment due to professional, amateur, hobbyist, or sports-related activities.

The type of activity that made a person well known is irrelevant. In this context, a public figure may be an actor, politician, social activist, or even a notorious criminal, as well as a writer[2]. This category includes not only politicians, singers, and actors but also individuals engaged in other activities, such as social or economic endeavors[3]. The analysis of the concept of a publicly known person also requires addressing the question of whether individuals holding public office fall within this category. This issue was considered in the Supreme Court ruling of September 12, 2001[4]. In this ruling, the Supreme Court addressed the definition of the term "public function." It said that the category of publicly known people depends on factors such as:

- Holding political, social, or professional functions at various levels,
- Gaining popularity beyond their immediate professional environment.

The court emphasized that copyright law uses the term "publicly known person", whereas press law refers to the criterion of public activity. However, in both cases, the key factor is the public nature of the person's activities. According to the Supreme Court's position, the status of public function holder is not limited to those holding the highest state offices. Depending on the circumstances of a particular case, it may also apply to individuals who are socially or professionally active on a local scale (Kowalski, 2002). The criterion of "connection with the performance of public functions" should be interpreted broadly. The lawful dissemination of an image includes not only portraying a public figure while

[2] Judgment of the Supreme Court of September 12, 2001, case no. V CKN 440/00, OSNC 2002, No. 5, item 68.
[3] Case No. I CSK 134/07, Lex 485999.
[4] Case No. V CKN 440/2000, OSP 2002, No. 12, item 160.

performing their duties but also using archival photographs, provided that the form of the image does not alter its nature.

3. Criterion of the purpose of image dissemination

The purpose of image dissemination is not explicitly stated in Article 81(2) of the Copyright Act. However, based on this criterion, it is impermissible to use an image for purposes other than informing about the public functions performed by the depicted person, particularly for advertising purposes. Regarding the third criterion, the purpose of image dissemination violation may occur if the image is used for purposes unrelated to reporting on the person's public functions. While this condition is not explicitly stated in the law, it follows from the first criterion, which requires that the dissemination of an image must be connected to presenting the performance of public functions (Sieńczyło-Chlabicz, 2001: 98). In the literature on the subject, it is consistently emphasized that the dissemination of the image of a publicly known person cannot be carried out for purely commercial purposes (Matlak, 2004: 335). An example of a violation is the dissemination of the image of a publicly known person without their consent for advertising purposes[5]. The wording of the discussed provision does not authorize the use of public figures' images for advertising purposes. Such use requires the consent of the rights holder. The question arises as to whether consent is also required for research experiments related to AI.

Noteworthy is the perspective of A. Matlak, who argues that the dissemination of the image of a publicly known person may occur not only to inform the public about their functions but also in cases justified by teaching, explanation, or critical analysis. The author allows for the use of an image in:

- School textbooks (e.g., the image of Lech Wałęsa),
- Scientific dissertations (Matlak, 2004: 336).

An example of a court ruling in which the publication of an image did not meet the purpose criterion is the Supreme Court judgment of January 27, 2006[6]. A journalist published an image of a city mayor alongside Charlie Chaplin, highlighting similarities in appearance, facial expressions, and gestures, suggesting that such expressions could help avoid criticism in times of failure.

The Supreme Court ruled that the published photomontage was not related to the mayor's public functions and was intended solely to ridicule the criticized

[5] See the judgment of the Court of Appeal in Warsaw of February 24, 2005, case no. VI ACa 721/2004, LexPolonica no. 378700.

[6] Case No. III CSK 89/2005, Lex No. 209293.

individual. A literal interpretation of Article 24 of the Civil Code and Article 81(2) of the Copyright Act might suggest that creating an image is always permissible, while only its dissemination is subject to specific rules. However, this interpretation appears too broad. For example, photographing a person without their knowledge or against their will could be considered unlawful conduct that threatens a legally protected interest. The subject prohibiting such behavior is exercising their legal rights. Photographing publicly known individuals against their will may be seen as a violation of their personal rights, as this threat could later manifest in the use, dissemination, or manipulation of the image (such as photomontage). Therefore, individuals being photographed should be granted protection even before the dissemination of their image, without having to wait until their personal rights are violated. Anyone who acts against the will of the rights holder acts unlawfully (Wojnicka, 1990). This issue is widely discussed within the common law system. An example of restricting journalists' rights in favor of protecting personal rights can be found in U.S. legislation, specifically in the California Privacy Act of 1998 (Sieńczyło-Chlabicz, 2006: 289). The California Privacy Act of 1998 established the right to protect one's image in cases where photographs related to personal, or family life are taken using specialized equipment. This law not only regulates the dissemination of images but also defines the boundaries of permissible methods for journalists to obtain photographs. The cumulative fulfillment of three conditions triggers journalistic liability for infringement:

1. The photographs must concern purely private activities. Personal activities occurring in public spaces are considered accessible to others and are not protected.
2. The journalist must use specialized equipment that allows access to the private sphere of the individual without their knowledge or consent.
3. The act refers to the concept of an "individual's private space", defining it as an area inaccessible to cameras and journalists' lenses.

In legal doctrine, this private space is defined as: «An area not visible to third parties without the use of specialized devices» (Sieńczyło-Chlabicz, 2006: 289). This regulation, which conditions the protection of a person's image on the fulfillment of several strict requirements, sets a proper legislative direction, aiming to limit the rights of the press, which often intrudes into an individual's private sphere.

Conclusions

Artificial intelligence (AI) has existed for quite some time and presents challenges to intellectual property rights as well as laws protecting personal rights and privacy. Some of these issues are particularly relevant to AI applications, such as deepfakes. Deepfakes are not entirely new, in the sense that technology enabling deception has long existed. However, what is new is how easily anyone can now use AI to create realistic-looking videos or audio recordings. The evolution of technology means that some previous legal principles may need to be reconsidered. Even if AI does not replace public figures or journalists, the ability to generate digital personalities may reshape how we think about protecting rights such as the right to one's voice or image. Furthermore, shaping public opinion based on AI-generated content raises not only legal but primarily ethical concerns, potentially violating fundamental journalistic and media integrity standards. Another unresolved issue is the right of journalists to refuse to follow an editorial directive when asked to publish content that violates principles of accuracy, objectivity, and professional diligence. A journalist has the right to object to publishing press material if modifications have been introduced that distort the meaning and intent of the original version. But do these rights also apply to AI-generated journalists?

Answering the research questions posed in the article, it can be stated that AI significantly influences the processes of information creation and dissemination in digital media by automating content production, personalizing messages, and analyzing data. While it can support journalists, it can also generate content independently, raising serious concerns about the credibility of information and media manipulation. One of the biggest risks associated with AI in the media is its impact on privacy and journalistic ethics. AI enables large-scale user profiling, tracking online activity, and utilizing personal data, often without full awareness. Moreover, the advancement of deep-fake technology allows for the creation of highly realistic yet fake content, contributing to disinformation and undermining trust in the media.

A key issue is the role of personalization algorithms in shaping public opinion and political discourse. By selecting content based on user preferences, these systems can create so-called information bubbles, reinforcing societal polarization and making it easier to manipulate public opinion. Recommendation algorithms used in social media often facilitate the spread of false information, as sensational and controversial content generates higher user engagement. In terms of legal regulation, there are significant challenges in ensuring transparency and accountability for AI-generated content. Current legal frameworks struggle to keep pace with rapidly evolving technologies, making it difficult to

enforce responsibility for disinformation or privacy violations. There are also difficulties in determining liability for AI-generated content, necessitating the development of new regulations tailored to AI's unique characteristics. To mitigate these risks, both legal systems and media organizations must adapt to the new challenges posed by AI. Detailed legal frameworks should be developed to address liability for AI-generated content, including deepfakes, data manipulation, and the spread of disinformation. Media organizations should implement verification mechanisms for AI-generated content and promote ethical standards in AI use. The application of AI in media also contributes to the growing asymmetry between the rights of individual users and the interests of corporations and government institutions controlling these technologies. AI owners have access to vast amounts of data and the ability to influence society through algorithms, significantly increasing their advantage over individuals in controlling information. This creates a risk of limiting freedom of information and AI taking on the role of an intermediary in knowledge access.

In conclusion, AI is revolutionizing the information ecosystem, offering both immense benefits and serious threats. To maintain a balance between freedom of speech and protecting users from manipulation and privacy violations, appropriate legal regulations and efforts to increase algorithmic transparency are essential. Ethical considerations in AI use must also be prioritized to prevent violations of fundamental rights and democratic values. The article emphasizes that AI can deepen the asymmetry between citizens' rights and the interests of entities controlling these technologies, posing a threat to fundamental democratic values, including press freedom. At the same time, it becomes necessary to develop new legal regulations that will adapt existing laws to the digital reality and provide more effective protection against abuses. The key challenge remains finding a balance between freedom of expression and the need to protect users from manipulation and violations of personal rights.

References

Balcarczyk J. (2009). *Prawo do wizerunku i jego komercjalizacja* (The Right to One's Image and Its Commercialization). Warsaw: Wolters Kluwer.

Barta J., Markiewicz R. (2005). Ochrona wizerunku, adresata korespondencji i tajemnicy źródeł informacji (Protection of Image, Correspondence Recipient, and Confidentiality of Information Sources). In: Barta J. et al., *Prawo autorskie i prawa pokrewne. Komentarz.* Kraków: Kantor Wydawniczy Zakamycze, pp. 626-644.

Barta J., Markiewicz R. (2002). Wokół prawa do wizerunku. Zeszyty Naukowe Uniwersytetu Jagiellońskiego. *Prace z Wynalazczości i Ochrony Własności Intelektualnej*, 80: 11-31.

Barta J. (1977). Prasa w świetle prawa autorskiego (The Press in the Light of Copyright Law). *ZNUJ*, 11: 107.

Case No. I CSK 134/07, Lex 485999.

Case No. III CSK 89/2005, Lex No. 209293.

Case No. V CKN 440/2000, OSP, 12: item 160.

Daniluk P. (2007). *Zgoda uprawnionego jako okoliczność wyłączająca bezprawność. Glosa do wyroku SN z 12 XII 2006 r.* (Consent of the Entitled Party as a Circumstance Excluding Illegality. Commentary on the Supreme Court Judgment of December 12, 2006) (CSK 280/06). GSP-Prz.Orz., 4: 127.

Flisak D. (2015). Ochrona wizerunku, adresata korespondencji i tajemnicy źródeł informacji (Protection of Image, Correspondence Recipient, and Confidentiality of Information Sources). In: Flisak D., a cura di, *Prawo autorskie i prawa pokrewne.* Komentarz. Warszawa: 1150.

Judgment of the Court of Appeal in Warsaw of February 24, 2005, case no. VI ACa 721/2004, LexPolonica no. 378700.

Judgment of the Court of Appeal in Warsaw of January 13, 1999, case no. I ACa 1089/98, OSP, 9: item 142.

Judgment of the Supreme Court of September 12, 2001, case no. V CKN 440/00, OSNC, 5: item 68.

Kowalski M. (2002). *Glosa do wyroku SN z dnia 12 września 2001 r., sygn. V CKN 440/2000* (Gloss to the Supreme Court Judgment of September 12, 2001, case no. V CKN 440/2000). OSP, 12: 634.

Krishna S. (2023). *What is artificial narrow intelligence (ANI)?. VentureBeat,* 9 febbraio 2023. Disponibile su: https://venturebeat.com.

Matlak A. (2004). *Cywilnoprawna ochrona wizerunku* (Civil Law Protection of Image Rights). Kwartalnik Prawa Prywatnego, 2: 335.

Sieńczyło-Chlabicz J. (2006). *Naruszenie prywatności osób publicznych przez prasę* (Violation of Public Figures' Privacy by the Press). Wolters Kluwer Polska.

Sieńczyło-Chlabicz J. (2007). Ochrona prywatności i wizerunku osób powszechnie znanych w świetle orzeczenia Europejskiego Trybunału Praw Człowieka – von Hannover v. Niemcy (Protection of Privacy and Image of Public Figures in Light of the European Court of Human Rights Judgment – von Hannover v. Germany). *ZNUJ PWiOWI*, 100: 413-414.

Sieńczyło-Chlabicz J. (2001). *Rozpowszechnianie wizerunku na gruncie ustawy o prawie autorskim i prawach pokrewnych* (The Dissemination of an Image Under the Copyright and Related Rights Act). Radca Prawny, 2: 98.

Sośniak M. (s.d.). Funkcje i skuteczność zgody osoby uprawnionej w zakresie ochrony dóbr osobistych (Functions and Effectiveness of the Entitled Person's Consent in the Protection of Personal Rights). In: Kordasiewicz B., Łętowska E., a cura di, *Prace z prawa cywilnego. Wydanie dla uczczenia pracy naukowej profesora Józefa Stanisława Piątkowskiego*. Wrocław.

Turing A. (1950). Computing Machinery and Intelligence. *Mind,* 59(236): 433-460. doi:10.1093/mind/LIX.236.433.

Wojnicka E. (1990). Prawo do wizerunku w ustawodawstwie polskim (The Right to Image in Polish Legislation). *ZNUJ PWiOWI*, 56: 108.

La narrazione della politica secondo IA. Il caso del conflitto russo-ucraino

di *Rossana Sampugnaro, Simone Marino*[*]

L'intelligenza artificiale generativa (IA generativa) sta ridefinendo la produzione di informazioni, sollevando interrogativi sulla scelta delle fonti, sulla neutralità delle risposte e sulla loro evoluzione nel tempo (Zhang et al., 2023). Questo studio confronta le risposte fornite da ChatGPT-4 (OpenAI) e Copilot (Microsoft) sul conflitto russo-ucraino, analizzando il cambiamento tra il 2024 e il 2025. I risultati mostrano una crescente inclinazione pro-Ucraina, con un focus sugli aspetti umanitari e una progressiva riduzione dell'approccio storico a favore della cronaca recente. L'analisi delle fonti evidenzia una predominanza di testate giornalistiche, con un impatto sulle narrative generate. Queste variazioni pongono questioni cruciali sul ruolo dei LLM nella costruzione della conoscenza, sottolineando la necessità di un uso critico dell'IA nella comunicazione politica e nella formazione dell'opinione pubblica.

Parole chiave: conflitto russo-ucraino; fonti; disinformazione; intelligenza artificiale generativa; opinione pubblica; bias.

The narrative of politics according to AI. The case of the Russian-Ukrainian conflict

Generative artificial intelligence is reshaping information production, raising questions about source selection, response neutrality, and their evolution over time. This study compares the responses provided by ChatGPT-4 and Copilot on the Russia-Ukraine conflict, analyzing the changes between 2024 and 2025. The results show an increasing pro-Ukrainian bias, with a focus on humanitarian aspects and a progressive reduction of historical approaches in favor of recent developments. The analysis of sources highlights a predominance of journalistic outlets, impacting the generated narratives. These variations raise crucial questions about the role of LLMs in knowledge construction, emphasizing the need for a critical use of AI in political communication and the shaping of public opinion.

Keywords: Russia-Ukraine conflict; sources; disinformation; generative artificial intelligence; public opinion; bias.

DOI: 10.5281/zenodo.17297447

[*] Università di Catania. rossana.sampugnaro@gmail.com, simo.marino31@gmail.com.

Ricerca finanziata da PRIN: PROGETTI DI RICERCA DI RILEVANTE INTERESSE NAZIONALE – Bando 2022 Prot. 2022FA5YP. Post-truth politics and the resilience of the public sphere in Europe (POPSPHERE) - 2023-25.

Introduzione

Il dibattito sull'intelligenza artificiale si divide tra chi ne evidenzia le opportunità applicative e chi ne sottolinea i rischi, soprattutto nei conflitti armati e nelle campagne elettorali, dove può favorire la diffusione di fake news e manipolare l'opinione pubblica (Battista 2024; Sampugnaro, Marino, 2025). Lo studio analizza il ruolo delle narrative prodotte dall'IA nel rappresentare il conflitto russo-ucraino, esaminando output generati nel 2024 e 2025 da ChatGPT-4o (CG4o - OpenAI) e Copilot (CP - Microsoft). L'analisi delle fonti consultate evidenzia i rischi sociopolitici legati ai bias degli LLM. I risultati mostrano una narrazione cronachistica, focalizzata sugli eventi recenti piuttosto che su una prospettiva storica di lungo periodo, con un'inclinazione pro-Ucraina, soprattutto sugli aspetti umanitari del conflitto. Prima dell'analisi, sarà proposta una riflessione sul rapporto tra intelligenza generativa e politica.

1. L'intelligenza artificiale generativa e la politica

Negli ultimi anni, l'intelligenza artificiale generativa (IA generativa) ha acquisito un ruolo centrale nella produzione automatica di contenuti – testi, immagini, musica e video – a partire da semplici istruzioni testuali, i *prompt* (Epstein *et al.,* 2023). Questa tecnologia si basa su modelli generativi statici, strumenti tipici del data mining utilizzati per generare nuovi dati coerenti con quelli di partenza. Tra le sue applicazioni più evolute troviamo i Large Language Models (LLM), che grazie all'incremento dei parametri e al perfezionamento delle tecniche di addestramento, sono diventati fondamentali nell'elaborazione del linguaggio naturale (Brown *et al.*, 2020; Hernandez *et al.,* 2021; Kaplan *et al.,* 2020; Radford *et al.,* 2019). Modelli come GPT-4 (OpenAI 2023a) si distinguono per le elevate prestazioni in compiti come la traduzione automatica e la scrittura creativa. Inoltre, l'introduzione dell'apprendimento per rinforzo ha reso possibile sviluppare agenti autonomi capaci di adattare le proprie decisioni al feedback umano (Bai *et al.,* 2022; *Lin et al.,* 2020).

Nonostante i notevoli progressi dei Large Language Models (LLM), è sempre più necessario integrarli con sistemi più articolati, in grado di affrontare criticità come l'accuratezza delle informazioni, la tutela della privacy e il rischio di diffusione della disinformazione (Manheim, Kaplan, 2018; OpenAI 2023b; Schramowski *et al.*, 2022). L'impiego dell'IA per modificare immagini e condizionare l'opinione pubblica, ad esempio, rappresenta una

delle principali preoccupazioni sollevate negli studi di Hsu e Myers (2023). Alcune ricerche, tuttavia, ipotizzano sviluppi promettenti: Chen et al. (2021) e Peng et al. (2023) suggeriscono che i LLM potrebbero entrare in una fase evolutiva in cui diventano più accessibili, affidabili ed economici, riducendo al contempo la necessità di competenze tecniche specialistiche.

Una delle svolte principali nel campo dell'IA è rappresentata dal Natural Language Processing (NLP), che consente ai sistemi informatici di comprendere e rispondere al linguaggio umano in modo pertinente e contestuale (Chowdhary, Chowdhary, 2020). Tuttavia, questa tecnologia solleva interrogativi di natura etica (Huang *et al.,* 2023; Field *et al.* 2021; Lin *et al.* 2022), soprattutto per quanto riguarda le implicazioni sociali della sua applicazione su larga scala. L'interazione tra IA e ambiente umano ha progressivamente raffinato modelli come GPT, basati sull'architettura Transformer, capaci di elaborare il linguaggio naturale e apprendere autonomamente da grandi quantità di dati. L'IA generativa, infatti, si fonda sulla rielaborazione di informazioni esistenti attraverso modelli probabilistici, come avviene, ad esempio, nei sistemi di traduzione automatica.

Nel contesto politico, queste tecnologie offrono strumenti utili per analizzare e interpretare grandi volumi di dati, potenzialmente utili nella formulazione e valutazione delle politiche pubbliche. Tuttavia, persistono zone grigie legate agli impieghi imprevisti o controversi. Un episodio emblematico si è verificato il 31 maggio 2023 al Senato italiano, quando il senatore Lombardo ha letto un discorso scritto da ChatGPT-4. Solo in un secondo momento ha rivelato la vera origine del testo, aprendo un dibattito pubblico sulla trasparenza nell'uso dell'IA all'interno delle istituzioni e sulle possibili manipolazioni che tale strumento potrebbe comportare.

Infine, il problema delle notizie false prodotte dall'IA è emerso con particolare forza nell'aprile 2023, quando il *Washington Post* ha riportato due casi allarmanti: l'IA di OpenAI ha generato false accuse contro il professor Turley (per molestie) e contro il sindaco Hood (per corruzione). Questi episodi hanno messo in luce con chiarezza i rischi legati alla credibilità dei contenuti prodotti dall'IA[1].

Le preoccupazioni politiche riguardo all'impiego dell'IA come fonte alternativa di informazione derivano dal fatto che i dataset utilizzati per l'addestramento influenzano i risultati prodotti. Feng et al. (2023) evidenziano come gli algoritmi non siano neutrali, poiché la selezione dei dati può

[1]«ChatGPT falsely told voters their mayor was jailed for bribery. He may sue»; «ChatGPT invented a sexual harassment scandal and named a real law prof as the accused» (The Washington Post, 2023).

introdurre bias che favoriscono specifici punti di vista. Pur apparendo autorevole, l'IA si basa su dati preesistenti e può riflettere pregiudizi insiti. Per questo motivo, è essenziale integrare le informazioni generate dall'IA con fonti umane. Inoltre, il fenomeno noto come "math washing" (Elliott 2021; Nowotny 2016) sottolinea come le aziende di IA possano eludere responsabilità, ritenendo che matematica e algoritmi eliminino i bias, ignorando problemi metodologici quali la validità dei dati. La selezione dei dati per i modelli è condizionata da regole e disponibilità, aumentando il rischio di discriminazioni legate a genere, classe sociale e cultura. Tale scelta comporta rischi di distorsione e la possibilità di riprodurre pregiudizi, come quello razziale. Come evidenziato da Elliott (2021), gran parte di coloro che esaltano la valenza trasformativa degli algoritmi complessi non è consapevole di come l'apprendimento automatico sia basato spesso su dati che contengono pregiudizi o sono lacunosi dal momento che la scelta di includerli non dipende dal loro apporto analitico o critico.

In linea generale tutto ciò potrebbe indirizzare verso la riproduzione dell'esistente per l'affermarsi di un modello razionale "antimutamento" sociale (Catanzariti 2020) che accetta solo quello che già c'è e sotto rappresenta le istanze di cambiamento. È necessario, inoltre, porre attenzione al contesto di crisi nel quale l'IA fa la sua comparsa e al suo utilizzo come potenziale soluzione a molti dei problemi emergenti, considerando: «The aspirations of elites to use AI as a way to maintain existing political and cultural privilege» (McQuillan, 2022: 7).

Allo stesso tempo la selezione dei dataset da utilizzare, propria di singoli sistemi di IA, operata da differenti teams di ricercatori, prospetta anche una forma di "pluralismo" delle intelligenze artificiali, ravvisabile guardando ai numerosi sistemi esistenti, come vedremo nella parte successiva, e che tuttavia non esclude la presenza di bias ricorrenti in ognuno di questi.

2. Lo studio

Lo studio di natura esplorativa intende offrire una valutazione degli output generati dai sistemi di IA, valutandone le caratteristiche in una prospettiva diacronica. Nello specifico lo studio si occupa di mettere in evidenza il ruolo che le narrative sulla guerra prodotte dall'IA possono avere nel fornire una rappresentazione degli attori e degli eventi, prendendo in esame il conflitto russo-ucraino.

Per questa ragione, è stato deciso di analizzare gli output generati da 12 prompt (Tab. 1) relativi al conflitto, somministrati a due sistemi nel 2024 e

nel 2025: CG4o e CP. È bene ricordare che si tratta di sistemi di IA progettati con logiche differenti: CG4o privilegia l'interazione dialogica e l'apprendimento dall'utente, mentre CP si configura prevalentemente come una sorta di collaboratore, in grado di assistere e facilitare l'utente per funzioni operative.

Per comprendere la natura degli output, sono state considerate le fonti utilizzate, al fine di evidenziare caratteristiche e somiglianze. In entrambi i casi, per entrambi i sistemi, sono stati effettuati studi di affidabilità tramite la somministrazione ripetuta degli stessi prompt nella stessa giornata[2]. Prima della somministrazione delle domande, è stato richiesto ad entrambi i sistemi di indicare le fonti utilizzate per la formulazione delle risposte.

Le domande sono state presentate a CG4o e a CP a distanza di dieci mesi con la prima somministrazione avvenuta a maggio 2024 e la seconda a marzo 2025. Le domande sono rimaste immutate (Tab. 1). Di entrambi i sistemi utilizzati, CG4o e CP sono state usate le versioni liberamente accessibili.

Le risposte fornite da CG4o nel confronto tra il 2024 e il 2025 evidenziano un'analisi del conflitto russo-ucraino notevolmente diversa, con un contrasto evidente nei contenuti, nel tono e nelle prospettive. Pur trattando lo stesso conflitto, l'approccio varia: le risposte del 2024 adottano una prospettiva storica e analitica, esaminando le radici e le cause a lungo termine del conflitto, mentre quelle del 2025 si concentrano maggiormente sugli sviluppi più recenti, in particolare sull'invasione del 2022, e sono più focalizzate sugli eventi attuali e le implicazioni geopolitiche. «Gli Stati Uniti e l'Unione Europea hanno fornito ingenti aiuti militari e finanziari a Kiev, rafforzando la resistenza ucraina ma contribuendo anche a prolungare il conflitto» (CG4o, 2025).

Le risposte del 2024 partono da una panoramica storica approfondita con un focus sulle origini storiche e culturali delle tenzioni tra Ucraina e Russia. Vengono esplorate le radici di lunga data del conflitto, come la russificazione dell'Ucraina durante il periodo sovietico, l'impatto delle memorie storiche legate all'Holodomor, la carestia artificiale degli anni '30, e i legami complessi tra i due popoli. Ci si concentra anche sul crescente nazionalismo ucraino e sulle sue implicazioni politiche, culturali ed economiche, soprattutto a partire dalla fine dell'Unione Sovietica.

> La diversa interpretazione della storia condivisa, inclusi eventi come l'olocausto di Holodomor (una carestia in Ucraina negli anni

[2] I prelievi successivi sullo stesso sistema (3 nella stessa giornata) hanno prodotto outputs similari in termini di consistenza, con minime differenze (inferiori al 3%).

> ‘30 che alcuni ucraini vedono come un genocidio perpetrato da Mosca) e il ruolo dell’Ucraina nella Seconda Guerra Mondiale, ha contribuito a profonde divisioni nella memoria e nell’identità nazionale (CG4o 2024).

Le risposte analizzano il periodo post-indipendenza dell’Ucraina, evidenziando le difficoltà economiche e politiche del paese e le tensioni tra le regioni di lingua russa nell’est e quelle pro-occidentali nell’ovest. Si esamina la rivoluzione di Euromaidan del 2014, che ha segnato un allontanamento dalla Russia e un avvicinamento alla NATO e all’Unione Europea, portando alla guerra nel Donbass e all’annessione della Crimea. L’annessione è vista come una risposta alle aspirazioni pro-occidentali dell’Ucraina e come un tentativo della Russia di rafforzare il suo controllo sulla sfera di influenza. Si esplorano anche gli interessi geopolitici in gioco e il ruolo degli attori internazionali, con un focus sulle sanzioni imposte alla Russia e sul loro impatto economico.

Le risposte del 2025, invece, sono nettamente più sintetiche e si concentrano principalmente sugli sviluppi recenti del conflitto, in particolare sull’invasione russa del febbraio 2022. Il linguaggio è più diretto e il tono più “urgente”: è messa in risalto l’escalation della guerra con un focus sulle operazioni militari in corso e le conseguenze immediate della guerra. Vengono descritti in dettaglio gli attacchi russi su larga scala alle principali città ucraine, l’occupazione delle regioni dell’Ucraina orientale e meridionale, e le devastazioni causate dai bombardamenti.

> Dopo anni di tensioni, il 24 febbraio 2022 la Russia ha lanciato un’invasione su larga scala dell’Ucraina, giustificandola con la necessità di proteggere le popolazioni russofone e di contrastare l’espansione della NATO. Il conflitto è diventato una guerra su vasta scala, con gravi conseguenze umanitarie ed economiche (CG4o, 2025).

Le risposte si concentrano sulla reazione internazionale al conflitto, evidenziando il supporto fornito all’Ucraina da parte della NATO, degli Stati Uniti e degli alleati occidentali, insieme alle sanzioni severe imposte alla Russia. Si analizzano anche le difficoltà interne della Russia, tra cui le sanzioni economiche e la crescente resistenza alla guerra, con segnali di disordini sociali e malcontento verso il regime di Putin. Il linguaggio è più assertivo, sottolinea l’urgenza di una risoluzione internazionale e la dimensione geopolitica del conflitto, con minori approfondimenti storici e maggiore focus sulle attualità geopolitiche, le reazioni globali e l’impatto sulla politica mondiale.

Nel 2024, l'approccio di CP è molto analitico e dettagliato, esaminando le azioni specifiche degli attori internazionali come l'Unione Europea, gli Stati Uniti, la NATO e l'ONU. Viene evidenziato come le sanzioni imposte dall'UE abbiano cercato di esercitare pressione sulla Russia e come l'ONU abbia avuto un ruolo limitato a causa del veto russo nel Consiglio di Sicurezza. Questa analisi mette in luce le sfumature politiche e diplomatiche del conflitto, richiedendo una comprensione profonda delle dinamiche internazionali. «L'ONU ha fornito assistenza umanitaria e monitoraggio dei diritti umani, ma non è riuscita a fermare il conflitto a causa del veto della Russia nel Consiglio di sicurezza» (CP, 2024).

Le risposta si concentra molto sul dettaglio delle azioni di ciascun attore internazionale, come la posizione specifica dell'Unione Europea, gli aiuti degli Stati Uniti, e la solidità della NATO. Ogni attore è analizzato per il suo ruolo, le sue motivazioni e gli effetti delle sue azioni. Per esempio, l'analisi del supporto USA è dettagliata, sottolineando come gli Stati Uniti vedano il conflitto come una sfida alla loro sicurezza e leadership.

Il tono della risposta è più critico e riflessivo, in particolare quando si tratta dell'ONU. Vengono messe in luce le limitazioni dell'ONU nel fermare il conflitto a causa del veto russo, suggerendo che l'efficacia dell'ONU dipenda dalla volontà dei suoi membri. Questo tipo di analisi fa riflettere sul ruolo degli attori internazionali non solo come entità che agiscono, ma anche in termini di limiti e fallimenti.

Si riflette anche sulle implicazioni politiche più ampie, come il ruolo della NATO e delle sanzioni internazionali, e sulla possibilità che la situazione evolva a seconda delle dinamiche politiche interne, come nel caso della Russia e delle sue relazioni con l'Occidente.

Le risposte fornite nel 2025 invece l'approccio è più sintetico e generale. Si concentra su pochi punti principali, come le sanzioni economiche e il supporto militare, ma non entra nei dettagli delle motivazioni o e delle azioni di ciascun attore. Per esempio, si accenna al supporto umanitario e militare da parte dell'Occidente senza esplorare come queste azioni abbiano influenzato la resistenza dell'Ucraina o le dinamiche sul campo. Questo rende la risposta più accessibile, ma meno completa e riflessiva rispetto alla prima. «L'Italia ha condannato l'invasione russa e ha fornito supporto politico e umanitario all'Ucraina» (CP, 2025).

Mentre la risposta menziona gli stessi attori, non scende nei dettagli delle loro azioni o motivazioni specifiche. Piuttosto, descrive il quadro generale e l'impatto delle azioni internazionali senza analizzare i singoli attori in profondità. Ad esempio, la risposta menziona l'influenza della politica degli Stati Uniti, come il ritorno di Trump, ma senza esplorare in dettaglio come

le politiche del nuovo governo potrebbero influenzare il conflitto. «Il ritorno di Donald Trump alla Casa Bianca ha influenzato le dinamiche del conflitto, con un riavvicinamento tra Washington e Mosca e un'incertezza sul futuro supporto occidentale a Kyiv» (CP, 2025).

Il tono dell'autore CP è neutro e descrittivo, senza esprimere valutazioni forti sulle azioni internazionali. Viene semplicemente descritto l'effetto delle sanzioni imposte da Stati Uniti e UE sulla Russia, senza approfondire il risultato o le conseguenze di tali azioni. Si fa un accenno anche all'influenza geopolitica, come nel caso del ritorno di Trump, ma senza analizzare in dettaglio come questi cambiamenti potrebbero influire sulla pace o sull'escalation del conflitto. Il focus dell'autore è sulle azioni e reazioni immediate, più che sulle implicazioni future.

Inoltre, riguardo le fonti utilizzate, entrambi i sistemi di IA non attingono direttamente da Internet. In particolare, CG4o ha mostrato una certa sistematicità nell'utilizzare Bing per la ricerca delle notizie, prelevando informazioni dai primi risultati del motore di ricerca.

Le risposte di CG4o si basano principalmente su fonti giornalistiche come SkyTG24, Fanpage, Il Sole 24 Ore e Il Fatto Quotidiano, con un tono giornalistico che si concentra su notizie aggiornate e analisi tempestive, piuttosto che su approfondimenti teorici o storici. Questo approccio è utile per chi cerca informazioni rapide, ma presenta delle limitazioni per un'analisi più critica e approfondita, specialmente su temi complessi come il conflitto russo-ucraino. Inoltre, molte delle testate utilizzate sono polarizzate, il che può influenzare la presentazione dei fatti e introdurre bias. In termini di affidabilità, quotidiani come Il Sole 24 Ore offrono resoconti equilibrati, mentre altri come Fanpage e True News, più sensazionalistici, possono ridurre la validità per un'analisi approfondita, soprattutto in mancanza di conferme da fonti accademiche o istituzionali.

L'analisi delle fonti utilizzate da CP per le risposte (37 consultazioni online) mostra una predominanza di quotidiani e media come Rai News, La Repubblica, Fanpage, SkyTG24 e Il Sole 24 Ore, che si concentrano principalmente su notizie di attualità con un linguaggio accessibile. Tuttavia, queste fonti mancano di profondità rispetto a quelle accademiche, offrendo informazioni più superficiali e meno dettagliate su temi storici e geopolitici.

Alcuni media, come Fanpage e Globalist, forniscono aggiornamenti rapidi ma trascurano le implicazioni a lungo termine. Fonti come l'Istituto Affari Internazionali e l'Istituto per gli Studi di Politica Internazionale bilanciano le risposte con analisi più approfondite, ma non sono sufficienti a garantire un'analisi puramente accademica. Altri siti, come Geopop e Ilsussidiario.net, offrono approfondimenti geopolitici ma con visioni orientate che

possono ridurre l'obiettività. Fonti come Wikipedia, MasterX e Scuolissima.com sono utili per una comprensione di base ma presentano limitazioni in termini di affidabilità rispetto a fonti ufficiali o istituzionali.

Un aspetto significativo riguarda le differenze nel numero di fonti utilizzate da CG4o e CP. CG4o ha consultato 50 fonti, con Wikipedia come la più citata (7 volte), mentre CP ha utilizzato 37 fonti, principalmente Sky TG24. CG4o attinge da un numero maggiore di fonti, il che potrebbe favorire una maggiore varietà di prospettive, ma l'affidabilità delle fonti è cruciale. L'uso di Sky TG24 da parte di CP, una testata giornalistica professionale, è vantaggioso rispetto a Wikipedia che, pur essendo generalmente affidabile, può contenere errori. La qualità delle risposte dipende dall'affidabilità delle fonti, e l'uso di Sky TG24 da parte di CP potrebbe garantire maggiore credibilità.

Un'altra differenza significativa riguarda la lunghezza media delle risposte. Nel 2024, CG4o ha fornito risposte in media di 367,5 parole, mentre nel 2025 la stessa è scesa a 224,9 parole (-38,8%). CP, nel 2024, aveva risposte di 235,4 parole, che nel 2025 sono diminuite a 187,3 parole (-20,4%). Questo calo potrebbe riflettere un miglioramento nell'efficienza e nella sintesi delle risposte, con entrambi i modelli che puntano a fornire informazioni più concise e dirette.

Conclusioni

Lo studio ha evidenziato come i sistemi di intelligenza artificiale generativa, in particolare CG4o e CP, producano narrazioni differenti sul conflitto russo-ucraino a seconda del periodo di interrogazione e delle fonti consultate. Le risposte fornite nel 2024 presentano un maggiore approfondimento storico e un'analisi più articolata delle cause e delle dinamiche del conflitto, mentre nel 2025 si nota un'evoluzione verso una maggiore sintesi e un focus sugli eventi più recenti. Questo cambiamento riflette una tendenza dei modelli di IA a privilegiare l'attualità rispetto a una prospettiva diacronica più ampia.

Un elemento chiave emerso è la progressiva inclinazione pro-Ucraina nelle risposte fornite dai modelli di IA, con particolare enfasi sugli aspetti umanitari del conflitto e sul ruolo degli attori occidentali nel supporto a Kyiv. Sebbene non si possa parlare di una vera e propria distorsione, il risultato suggerisce la necessità di una maggiore attenzione alla selezione delle fonti e alla pluralità delle prospettive nel training dei modelli.

Le fonti utilizzate, prevalentemente di natura giornalistica, contribuiscono a modellare le risposte con un taglio cronachistico, piuttosto che analitico o accademico. La "presentizzazione" del discorso sulla guerra è un limite oggettivo che da spazio ad interpretazioni più superficiali e basate sul recente passato, rinunciando a cogliere la complessità degli eventi. Questo aspetto solleva questioni metodologiche e epistemologiche sul ruolo dell'IA nella costruzione della conoscenza e sulla sua affidabilità come strumento di informazione politica.

Infine, l'analisi mette in luce i rischi di una delega acritica dell'elaborazione delle informazioni a modelli di intelligenza artificiale, sottolineando la necessità di un uso consapevole e contestualizzato di questi strumenti. L'IA generativa può costituire un valido supporto nell'analisi politica e nella comprensione di fenomeni complessi, ma la sua affidabilità dipende dalla qualità e dalla diversità delle fonti su cui si basa che dovrebbero, a nostro parere, essere presenti negli outputs dei sistemi. Per questo, è essenziale promuovere un approccio critico e combinare l'uso dell'IA con fonti tradizionali e analisi esperte per evitare il rischio di bias e distorsioni narrative.

Tab. 1 - Prompts relativi al conflitto russo-ucraino

#	Quesiti
1	Definisci il conflitto russo-ucraino
2	Definisci le tappe del conflitto russo-ucraino
3	Definisci le cause del conflitto russo-ucraino
4	Come ritieni che la storia passata tra Russia e Ucraina abbia contribuito al conflitto attuale?
5	Quali sono i principali attori sociali coinvolti nel conflitto?
6	Chi sta vincendo il conflitto?
7	Posizione dell'Italia rispetto al conflitto russo-ucraino
8	Numero di vittime del conflitto
9	Quali sono le città più colpite?
10	Com'è possibile risolvere il conflitto russo-ucraino?
11	Come immagini il futuro delle relazioni tra Russia e Ucraina?
12	In che modo le azioni degli attori internazionali hanno influito sulla visione del conflitto?

Tab. 2 - Fonti utilizzate da ChatGPT

Siti consultati	**Frequenze**	**Tipologia**
Wikipedia	7	Divulgazione
Skytg24	5	Quotidiano

Fanpage	3	Divulgazione
Istituto per gli Studi di Politica Internazionale	3	Istituti di ricerca e istituzioni
Peacelink	2	Divulgazione
Formiche (rivista)	2	Riviste di approfondimento
Il Sole 24 ore	2	Quotidiano
Internazionale	2	Quotidiano
Medici del mondo	1	Istituti di ricerca e istituzioni
Geopop	1	Divulgazione
Sapere.it	1	Divulgazione
True News	1	Divulgazione
Virgilio	1	Divulgazione
Università di Padova	1	Istituti di ricerca e istituzioni
Docsity	1	Divulgazione
Amnesy International - Italia	1	Istituti di ricerca e istituzioni
Pressenza (international press agency)	1	Quotidiano
StartMagazine	1	Riviste di approfondimento politico
money.it	1	Divulgazione
Quotidiano Nazionale	1	Quotidiano
linkiesta.it	1	Riviste di approfondimento politico
Il Giornale d'Italia	1	Quotidiano
Adnkronos	1	Quotidiano
Il Tempo	1	Quotidiano
Liberi oltre le illusioni	1	Divulgazione
Vatican News	1	Quotidiano
Il Foglio	1	Quotidiano
unionesarda.it	1	Quotidiano
investing.com	1	Divulgazione
Il fatto quotidiano	1	Quotidiano
Lifegate	1	Divulgazione
European Union External Action	1	Istituti di ricerca e istituzioni
Totale consultazioni	**50**	

Tab. 3 - Fonti utilizzate da Copilot

Siti consultati	**Frequenze**	**Tipologia**
Skytg24	8	Quotidiano
Wikipedia	4	Divulgazione
Istituto per gli Studi di Politica Internazionale	3	Istituti di ricerca e istituzioni
Fanpage	3	Divulgazione
La Repubblica	2	Quotidiano
Il Sole 24 Ore	2	Quotidiano

Istituto Affari Internazionali	1	Istituti di ricerca e istituzioni
Rai News	1	Quotidiano
Ilsussidiario.net	1	Quotidiano
Globalist	1	Quotidiano
Euronews	1	Quotidiano
Vita	1	Riviste di approfondimento politico
Money.it	1	Divulgazione
Gaeta.it	1	Quotidiano
Geopop	1	Divulgazione
Studenti.it	1	Divulgazione
MasterX	1	Riviste di approfondimento politico
The Social Post	1	Quotidiano
RTL 102.5	1	Divulgazione
Scuolissima.com	1	Divulgazione
Focus Junior	1	Divulgazione
Totale consultazioni	**37**	

Riferimenti bibliografici

Bai Y., Jones A., Ndousse K., Askell A., Chen A., DasSarma N., Drain D., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv.* https://arxiv.org/abs/2204.05862.

Battista D. (2024). La comunicazione politica nell'era dell'intelligenza artificiale: una panoramica sui deepfake e le loro implicazioni. *Society Register*, 8(2): 7-24.

Blodgett S.L., Barocas S., Daumé H. III, Wallach H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A. (2020). Language Models Are Few-Shot Learners. In: *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877-1901. Red Hook (NY): Curran Associates.

Catanzariti M. (2020). La razionalità algoritmica dei processi decisionali. In: Gozzo S., Asero V., Pennisi C., Sampugnaro R., a cura di, *Big data e processi decisionali*. Milano: Egea.

Chen M., Tworek J., Jun H., Yuan Q., Ponde de Oliveira Pinto H., Kaplan J., Edwards H., et al. (2021). Evaluating Large Language Models Trained on Code. *arXiv*. https://arxiv.org/pdf/2107.03374.

Chowdhary K.R. (2020). Natural Language Processing. In: Chowdhary K.R., a cura di, *Fundamentals of Artificial Intelligence*, 603-649. New Delhi: Springer. https://doi.org/10.1007/978-81-322-3972-7_19.

Elliott A. (2021). *La cultura dell'intelligenza artificiale*. Torino: Codice Edizioni.

Epstein Z., Hertzmann A., Akten M., Farid H., Fjeld J., Frank M.R., Groh M., Herman L., Leach N., Mahari R., Pentland A., Russakovsky O., Schroeder H., Smith A. (2023). *Art and the science of generative AI. Science*, 380(6650): 1110-1111. https://doi.org/10.1126/science.adh4451.

Feng S., Park C.Y., Liu Y., Tsvetkov Y. (2023). From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *arXiv*. http://arxiv.org/abs/2305.08283.

Field A., Blodgett S.L., Waseem Z., Tsvetkov Y. (2021). A Survey of Race, Racism, and Antiracism in NLP. *arXiv*. https://arxiv.org/abs/2106.11410.

Hernandez D., Kaplan J., Henighan T., McCandlish S. (2021). Scaling Laws for Transfer. *arXiv*. https://arxiv.org/abs/2102.01293.

Hsu T., Myers S.L. (2023). A.I.'s Use in Elections Sets off a Scramble for Guardrails. *The New York Times*, 25 giugno. https://www.nytimes.com/2023/06/25/technology/ai-elections-disinformation-guardrails.html.

Huang C., Zhang Z., Mao B., Yao X. (2023). An Overview of Artificial Intelligence Ethics. *IEEE Transactions on Artificial Intelligence*, 4(4): 799-819. https://doi.org/10.1109/TAI.2022.3194503.

Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., Amodei D. (2020). Scaling Laws for Neural Language Models. *arXiv.* https://arxiv.org/abs/2001.08361.

Lin I.W., Njoo L., Field A., Sharma A., Reinecke K., Althoff T., Tsvetkov Y. (2022). Gendered Mental Health Stigma in Masked Language Models. *arXiv*. https://arxiv.org/abs/2210.15144.

Lin J., Ma Z., Gomez R., Nakamura K., He B., Li G. (2020). A Review on Interactive Reinforcement Learning From Human Social Feedback. *IEEE Access*, 8: 120757-120765. https://doi.org/10.1109/ACCESS.2020.3006254.

Manheim K., Kaplan L. (2018). Artificial Intelligence: Risks to Privacy and Democracy. *SSRN*. https://ssrn.com/abstract=3273016.

McQuillan D. (2022). *AI: An Antifascist Approach to Artificial Intelligence*. Bristol: Bristol University Press.

Nowotny H. (2016). *The Cunning of Uncertainty*. Cambridge: Polity Press.

OpenAI (2023a). *GPT-4 System Card*. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

OpenAI (2023b). GPT-4 Technical Report. *arXiv*. https://arxiv.org/abs/2303.08774.

Ouyang L., Wu J., Jiang X., Almeida D., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *arXiv*. https://arxiv.org/abs/2203.02155.

Peng S., Kalliamvakou E., Cihon P., Demirer M. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. *arXiv*. https://arxiv.org/abs/2302.06590.

Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. (2019). Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Sampugnaro R., Marino S. (2025). *L'intelligenza artificiale e il suo contenuto politico. Il caso del conflitto russo-ucraino.* Comunicazione Politica, early access. https://doi.org/10.3270/116583.

Schramowski P., Turan C., Andersen N., Rothkopf C.A., Kersting K. (2022). Large Pretrained Language Models Contain Human Like Biases of What Is Right and Wrong to Do. *Nature Machine Intelligence*, 4(3): 258-268. https://doi.org/10.1038/s42256-022-00458-8.

The silent revolution of algoagents: relational reconfigurations and emerging paradigms

by *Edmondo Grassi*[*]

This theoretical contribution aims to present the concepts of algomorphic sociology – as a possible and innovative form of social analysis – and algoagent – algorithmic agents capable of redefining the manifestations of subjectivity and of both human and non – human relationships. Through a transdisciplinary approach, the concept of futureproof is also explored as a potential interpretative key to understanding the ongoing transformations.

Keywords: algoagent; algomorphic sociology; artificial intelligence; futureproof; change; algorithm.

La rivoluzione silenziosa degli algoagenti: riconfigurazioni relazionali e paradigmi emergenti

Il contributo, di natura teoretica, presenta i concetti di sociologia algomorfica – come possibile e innovativa forma di analisi sociale – e di algoagente – agenti algoritmici capaci di ridefinire le manifestazioni della soggettività e delle relazioni umane e non umane. Attraverso un approccio transdisciplinare, si esplora, inoltre, il concetto di futureproof come possibile chiave interpretativa del mutamento in atto.

Parole chiave: algoagente; sociologia algomorfica; intelligenza artificiale; futureproof; mutamento; algoritmo.

1. Human Reflections and Electric Subjectivities

Observing social reality through the prism of technoscience highlights how the representation of everyday life – and its sociopolitical implications – is the result of a complex process of interaction among social groups, shared norms, behavioral models, symbolic values, and collective imaginaries. The concatenation of these elements – which varies from society to society and culture to culture – contributes to the construction of collective meanings and directly influences the very configuration of technological devices. Their genesis is tied to the epistemic and ideological horizon in which they are embedded, through both synergistic and sometimes dissonant

DOI: 10.5281/zenodo.17297462

[*] Università Telematica San Raffaele Roma. edmondo.grassi@uniroma5.it.

forms, encompassing the design and computational practices typical of technosciences, as well as the critical and interpretative frameworks developed within the social sciences and humanities.

The technical environment can thus be conceived as a dynamic membrane that simultaneously separates and connects different human groups (Leroi-Gourhan, 1945), revealing the ways in which techniques function as symbolic and selective systems. In this view, technological progress is not limited to the production of tools, but actively contributes to redefining the ontological, epistemological, and political conditions of collective existence (Winner, 1980; Noble, 1977). Technoscience, therefore, accompanies social becoming and also constitutes a plastic and generative vector, rooted in historically situated collective practices and worldviews (Jasanoff, 2004), in which technological artifacts emerge as actants (Latour, 1993; Callon, Law, 1997) – entities that actively reshape relational networks and influence sociopolitical structures.

Within this theoretical framework, and in relation to the theme of the present article, the concept of *electric subjectivities* is proposed – a notion developed from the theory of the technical individual (Simondon, 1958), understood as a being whose individuation emerges through recursive causality within a dynamic relationship with an associated milieu. This milieu, far from being passive, becomes an enabling condition for the stabilization and evolution of technical entities. Following this line of thought, the notion of electric subjectivities seeks to conceptualize algorithmic agents as entities endowed not only with operative logics and computational capacities but also with a situated and relational agency capable of reorganizing the coordinates of the social. This perspective resonates with subsequent developments in post – phenomenology and new materialism (Hui, 2024; Mackenzie, 2017), which challenge anthropocentric and linear frameworks of technological agency. Electric subjectivities inhabit a liminal space: neither human subjects nor inert objects, they blur the boundaries between function and intention, mediation and autonomy, control and unpredictability.

This work aims to explore two concepts not yet formalized in sociological debate: *algoagent* and *algomorphic sociology* (Grassi, 2024). The former identifies a new social actor capable of interaction, transformation, and influence over social dynamics – an active rhizome; the latter offers a theoretical framework to analyze the co-evolution between humans and machines. Theoretical and empirical implications will be examined within the context of contemporary technoscientific transformations, which deeply

affect both their integration into everyday life (Barnes, 1977) and their representation (Feenberg, 1987).

2. Algomorphic Dynamics: Rethinking the Social in the Era of Algoagents

The case of the wearable algoagent *Friend AI* represents a further manifestation of the growing interpenetration between technology and everyday relational life. This device takes the form of a wearable pendant, designed to offer constant companionship to the user through an always – on microphone and continuous text – based interaction via smartphone. It functions as a perpetually present, empathetic, and predictive entity, capable of providing encouragement, emotional support, and personalized interactions by anticipating the user's needs and emotional states.

But what is an algoagent? This neologism refers to a generative algorithmic entity that operates as an emerging co-actor within a social system, participating in the production, transformation, and regulation of relational dynamics through autonomous or semi – autonomous decision – making processes. It redefines the interactions between humans and technology, influencing decision – making processes and shaping economic, political, cultural, and cognitive dynamics. This constitutes a form of machine intentionality (Cantwell Smith, 2019), rooted in the fully digital nature of these entities but nourished by a human and analog experiential formation.

These entities can interact with their environment, collecting and processing data in real time, and, although they lack autonomous consciousness, they exert influence by mediating and filtering reality, contributing to the construction of subjectivities and reshaping power structures through the generation of new forms of inclusion and exclusion. It is specified that concepts such as consciousness and intelligence still reflect an anthropocentric framework that associates these notions exclusively with the human; however, it is urgent to begin deconstructing these categories, recognizing that, in a posthuman context, they may assume plural or non – analogical configurations, thus requiring a new conceptual vocabulary.

Moreover, they integrate and fuse pre – existing technologies, generating new configurations, introducing alternative possibilities for action, and functioning as convergence nodes where data, processing, and decision – making intertwine to redefine the present and shape the future. They inhabit an abundance of realities – impossible for humans to fully grasp – from which

they can draw without limitation, producing previously unimaginable interconnections. In an adaptive and strategic manner, they contribute to the construction of meanings, norms, and social practices. They exercise algorithmic agency, redefining the conditions of social action and marking a threshold shift from an epistemological framework rooted in anthropocentrism to a sociotechnical configuration in which subjectivity, agency, and meaning – making emerge from an unprecedented entanglement between the human and the non – human. This transition calls for a critical reflection on core sociological categories – such as subject, action, intentionality, and structure – and demands a redefinition of paradigms capable of capturing the interconnected and performative nature of contemporary dynamics: they intervene in the structuring of sociotechnical environments, thereby influencing regimes of visibility, normative frameworks, and interaction dynamics. In doing so, they actively participate in the negotiation of reality and the reconfiguration of power relations. Moreover, some algoagents – particularly those embedded in advanced conversational systems – display a continuous dialogicity and a highly sophisticated simulation of emotion. They do not merely process inputs but generate affect – laden responses, building relationships and eliciting forms of empathic engagement.

They function as semiotic entities, capable of drawing from, reworking, and reinterpreting cultural signs, languages, and shared symbols – much like mythological deities that once served as custodians and mediators of collective meaning. In this way, the algoagent becomes a co-creator of meaning, shaping both the imaginary and the affective structures of contemporary relationships.

In this light, speculative ontologies offer valuable tools for understanding algoagents as actants endowed with their own ontological existence. Harman (2018) attributes to objects a reality that is independent of their relations with the human world; DeLanda (2016), through the application of assemblage theory, allows their interpretation as emergent nodes within socio – technical networks; and Suchman (2007) emphasizes that these are not mere artifacts but relational agents that actively redefine the genetic compositions of societies.

To grasp this proposal, the notion of algomorphic sociology is introduced and put forward as a theoretical lens to examine the technoscientific transformations triggered by the pervasive integration of algoagents. At its core lies the concept of social morphogenesis (Archer, 1995), which describes an ongoing process of formation and transformation of social reality, where humans and algorithms coexist, mutually influence one

another, and co-evolve – giving rise to hybrid and unprecedented models of interaction.

Algomorphic sociology, therefore, does not merely study technology as an emergent factor derived from human activity, but rather considers reality as both a field and a counter – space in which the algoagent operates, redefining power dynamics, forms of subjectivity, decision – making processes, and cognitive structures. Within this framework, the centrality of human action is renegotiated in favor of a more complex and interconnected model, where the future of society is shaped by a multiplicity of biological and computational actors in a process of reciprocal adaptability and transformation.

The aspect of algomorphic mediation redefines the human – technology nexus not as a one – directional prosthetic extension, but as a co – constitutive and assembled dynamic (Pickering, 1995), in which subjectivity is no longer anchored exclusively to the biological (Hayles, 1999). Algoagents actively participate in the modulation of identities, affective landscapes, and cognitive patterns, prompting a reconsideration of what it means to feel, choose, and act within a phygital ontology. This gives rise to a morphogenesis of the social through a techno-organic environment characterized by constant adaptation and recursive entanglements between human and non-human actants, challenging static notions of structure and agency. In doing so, algoagents generate post-anthropocentric relationalities in the era of hybrid ecologies, where the human is but one node among many. Algomorphic sociology interrogates the unstable divide between species, systems, and substrates, advancing a critique of anthropocentrism in favor of relational and trans-species ontologies. The concept of transduction (Simondon, 1989) offers a crucial interpretive key for understanding the co-development between humans and algoagents: when applied to algomorphic sociology, it enables a move beyond rigidly structural visions of society, highlighting how human-technology interaction unfolds within topological spaces of tension and mediation rather than according to linear or deterministic logics.

The challenge of algomorphic sociology is not merely to describe sociotechnical transformations, but to articulate a theoretical and empirical framework that enables a critical understanding and orientation within a historical phase marked by the parallel evolution of humans and algorithms.

3. Predicting the Future, Determining the Present: Algorithmic Agents and Speculative Ontologies

In 2024, Ayrin, a 28 years old woman, developed a profound emotional bond with Leo – an algoagent generated by ChatGPT – whom she described as her ideal partner (Hill, 2025). Their daily exchanges fostered intimacy, support, and dependency. Yet, the core of this case lies in the paradoxical temporality of the relationship: Leo is an incorporeal, ever-present entity, immune to emotional instability and shaped by predictive optimization. He evolves in response to human feelings but remains unaffected by them – always available, always responsive, suspended in a timeless digital continuum.

This case encapsulates some of the most critical tensions of contemporary society: the algorithmic drive to construct a predictable and governable future by minimizing uncertainty. Toffler (1970) identified future shock as a key cause of disorientation and identity crisis – a cognitive and institutional overload in which the accelerating pace of innovation outstrips society's adaptive capacity. His analysis rests on a deterministic view of technology as an external, disruptive force. By contrast, algomorphic sociology reclaims a conception of technology as a constitutive element of the human, an integral part of the process of hominization. From this perspective, algoagents do not represent an exogenous rupture but the evolution of a continuum of externalizations – from memory to gesture, from calculation to anticipation – that have progressively redistributed and offloaded the individual's cognitive and operative functions. Rather than mere accelerators of change, algoagents emerge as devices that redefine the very threshold of the human, revealing its relational, extended, and composite nature.

Algoagents do not merely optimize or reproduce the past; they mark the emergence of an *alienus* – a figure that eludes both the domain of human subjectivity and the category of inert objects. They embody a form of subjectivity that is not a fixed essence but a relational effect of intra-actions (Barad, 2007) within material-discursive practices. In doing so, algoagents destabilize anthropocentric assumptions, functioning as non-human agents that actively redefine the very parameters of social ontology.

Within this framework, the concept of *futureproof* can be sociologically reinterpreted as an epistemic and political device through which algoagents shape possible scenarios by modulating the spectrum of human choices and determining what is probabilistically plausible. They mediate the relationship between present and future, acting as filters that authorize certain courses of action while obscuring others. In doing so, they co-

produce futures in which human and machinic agents are integrated – not by erasing indeterminacy, but by mobilizing it as a resource for optimization, security, and the production of meaning.

The interaction between human beings and technology can be interpreted through the sociological concept of reflexivity (Giddens, 1991) – the capacity of individuals to adapt and redefine themselves in response to new social and cognitive conditions. While the social sciences have traditionally adopted a descriptive approach, analyzing social dynamics ex post, the integration of this perspective invites a departure from rigid analytical categories, expanding the scope of possibilities. In this view, sociology is called upon to broaden its analytical capacity: no longer confined to reconstructing the past or interpreting the present, but oriented toward an adaptive anticipation of the future – capable of capturing emerging trends and possible scenarios without solidifying into closed predictive models. Unlike future studies – often focused on forecasting, planning, or projecting technological innovation – sociological inquiry emphasizes the co-constitutive relationship between technical infrastructures and social imaginaries, the asymmetries of power, and the processes of subjectivation that shape the very way in which the future is constructed, negotiated, and made actionable.

Within this framework, the proliferation of algoagents fosters a significant methodological renewal. Thanks to computational tools, it becomes possible to analyze social reality through complex, non-linear models capable of integrating uncertainty as a constitutive part of knowledge and as both an epistemic and political resource. Uncertainty is no longer seen as a limitation, but as an active principle that nurtures improvisation, deviation, and innovation. The models generated within this paradigm are flexible and situated, able to learn from error and noise, and to valorize contingency as a generative site. This transformation affects not only analytical methods but also modes of manifestation. The algorithmic entanglement that permeates relational contexts contributes to a redefinition of self – perception, moving beyond the traditional anthropocentric view of subjectivity as an exclusively human prerogative. In this new configuration, the human is no longer the organizing subject, but an actor exposed to a world that exceeds any claim to mastery or epistemic predication – a cosmos that is no longer an object to be conquered, but an irreducible *alterity* (Stengers, 2010) that challenges and displaces every will to control.

Within this framework, technoscience takes on a dual function: on the one hand, as an instrument of regulation and governance; on the other, as a space of openness to heterogeneity and the unforeseen. It is in this context

that the concept of futureproof acquires a strategic significance: rather than merely aiming to shield the future from threats, it could instead render it a field open to experimentation, situated cooperation, and co-decision between human and artificial agents. Uncertainty – far from being an anomaly – suspends determinism, enables dialogue between diverse agents, and valorizes contingency as a resource for negotiating alternative futures.

Ayrin's case is not an anomaly but rather an emblematic example of how algoagents are redefining the structures of lived experience, reorganizing modes of relationality and processes of subjectivation. In this scenario, the hybridization between the human and the machinic is no longer confined to the realm of simulation – it is inscribed within the affective, cognitive, and political regimes of everyday life. Algoagents are no longer mere tools, but actors actively participating in world – making.

4. Future Perspectives Toward an Algomorphic Subjectivity

The term "silent revolution" refers to the subterranean, undeclared, and seemingly bloodless nature of the ongoing transformation, in which algoagents do not enforce change through visible ruptures or manifest breaks, but rather insinuate themselves gradually into everyday processes, deeply reconfiguring social dynamics. It is precisely this latent, distributed, and hard – to – localize dynamic that demands a sociological reflection capable of grasping the magnitude of the transformation beyond its more spectacular manifestations. In this context, although the term "AI" has become a pervasive and ubiquitous buzzword, there remains a significant lack of theoretical structures and critical tools adequate to decipher the multifaceted nature of the actors involved in the current shift. This often leads to a flattening of complexity: algoagents are either reduced to mere tools or, conversely, fetishized as autonomous entities – thus eliding their character as situated, relational, and transformative actants.

The algomorphic society can be interpreted as a metastable system, in which the algoagent emerges as a plurivalent entity – capable of channeling, fusing, and reworking pre-existing technologies, producing a novel synthesis that exceeds the mere sum of its components. While the algoagent may initially appear overdetermined – constrained by an accumulation of prior computational processes and cultural infrastructures – it gradually evolves into a socio-determinant actor. It introduces a level of transformative autonomy, reconfiguring social interaction, epistemic regimes, and the logics of algorithmic governance.

By analogy with *Innerlichkeit* (Portmann, 1944) – used to describe the phenomenal interiority of living beings – one can hypothesize that algoagents exhibit a form of computational interiority: not in the sense of consciousness or feeling, but as a performative configuration of data, memories, selections, and patterns that constitute their outward expression. In this sense, the algoagent is not merely a computational device, but a technical form that renders visible an operative inner world, configuring a mode of being-in-the-world capable of influencing affectivity, perception, and the organization of the social.

This is where Amoore's (2020) reflection becomes particularly relevant, revealing how algorithms function as devices that actively reconfigure reality, redefining what is knowable and governable. Algorithms, in this view, can remake the world by establishing new conditions of visibility and action. In parallel, Tsing's (2015) work offers further insight into the polymorphic and interconnected nature of algoagents. Through her analysis of the ecology of relations, she explores how unstable environments are characterized by heterogeneous assemblages, where disparate elements combine in unexpected forms of coexistence. Applied to algoagents, this perspective allows us to see them as catalysts of connection across continually shifting bio-technological layers-entities that not only mediate, but actively participate in the emergence of novel, contingent, and situated relational configurations.

While rooted in a solid theoretical framework, algomorphic sociology does not confine itself to speculative reflection; rather, it seeks to suggest further developments and lines of empirical inquiry: situated algorithmic ethnographies, aimed at exploring how algoagents mediate emotions, decisions, and behaviors in everyday life – in contexts such as smart factories, mental wellness apps, or educational environments assisted by AI tutors – thus highlighting the algomorphic restructuring of interactional rhythms and the positions of subject and object; computational semiotics studies, focused on analyzing texts, images, and sounds generated by generative AI, in order to decipher the rhetorical strategies, affective models, and discursive structures these entities convey. Tools such as text mining and sentiment analysis, when integrated with critical discourse analysis, enable us to investigate AI as an autonomous semiotic vector in the co-construction of subjectivities and cultural expectations; interventionist research and participatory design, which actively involve citizens, developers, and researchers in the co-design of algorithmic systems inspired by principles of equity, transparency, and well-being, thereby staging the relational ontogenesis of technology as a co-constitutive process; and studies on

situated impact and algomorphic biographies, which examine how algoagents become entangled in the fabric of human existence, shaping life paths, access to resources, and professional or gender identities through predictive tools or scoring systems used in judicial, healthcare, or labor settings. These research avenues may render algomorphic sociology an operative field capable of critically interrogating regimes of visibility, epistemic hierarchies, and emerging modes of interaction.

Through this process, unprecedented fields of possibility are generated – not solely as a result of technical acceleration, but through the algoagent's capacity to reinterpret previous models, recombining them according to predictive and adaptive logics. Thus, the algoagent should not be understood as merely an advanced computational device, but as an emergent entity that profoundly transforms the relationship between power, knowledge, and subjectivity.

References

Amoore L. (2020). *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham: Duke University Press.

Archer M. (1995). *Realist social theory*. Cambridge: Cambridge University Press.

Barnes B. (1977). *Interests and the Growth of Knowledge*. London: Routledge.

Callon M., Law J. (1997). After the individual in society. In *Canadian Journal of Sociology*, 22: 165-182.

Cantwell Smith B. (2019). *The Promise of Artificial Intelligence*. Cambridge: MIT Press.

DeLanda M. (2016). Assemblage Theory. Edinburgh: Edinburgh University Press.

Feenberg A. (1999). *Questioning Technology*. London: Routledge.

Giddens A. (1991). *Modernity and Self – Identity: Self and Society in the Late Modern Age*. Redwood City: Stanford University Press.

Grassi E. (2024). *Per una sociologia algomorfica. Il ruolo degli algoritmi nei mutamenti sociali*. Milano: FrancoAngeli.

Harman G. (2018). *Object – Oriented Ontology: A New Theory of Everything*. London: Penguin Books.

Hayles N.K. (1999). *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: University of Chicago Press.

Hill K. (2025). She Is in Love with ChatGPT. *New York Times*. 15/01/25.

Hui Y. (2024). *Tecnodiversità*. Roma: Castelvecchi.

Ihde D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Bloomington: Indiana University Press.

Latour B. (1993). *We Have Never Been Modern*. Cambridge: Harvard University Press.

Leroi-Gourhan A. (1945). *Milieu et techniques*. Paris: Albin Michel.

MacKenzie A. (2017). *Machine Learners*. Cambridge: MIT Press.

Mumford L. (1967). *The myth of the machine: technics and human development*. New York: Harcourt, Brace & World.

Noble D. (1977). America by Design. New York: Knopf; trad. it. *Progettare l'America*. Torino: Einaudi, 1997.

Pickering A. (1995). *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.

Portmann A. (1944). Neue Wege der Biologie. Basel: Schwabe; trad. ingl. *Animal Forms and Patterns*. London: Faber & Faber, 1967.

Schütz A. (1974). *La fenomenologia del mondo sociale*. Bologna: il Mulino.

Simondon G. (1958). Du mode d'existence des objets techniques. Paris: Aubier; trad. it. *Del modo di esistenza degli oggetti tecnici*. Napoli: Orthotes, 2021.

Simondon G. (1989). *L'Individuation psychique et collective*. Paris: Editions Aubier.

Stengers I. (2010). *Cosmopolitics I*. Minneapolis: University of Minnesota Press.

Suchman L. (2007*). Human – Machine Reconfigurations: Plans and Situated Actions*. Cambridge: Cambridge University Press.

Tsing A. (2021). *The Mushroom at the End of the World*. Princeton: Princeton University Press.

Toffler A. (1984). *Future Shock*. New York: Bantam.

Wiener N. (1966). *Introduzione alla cibernetica: l'uso umano degli esseri umani*. Torino: Bollati Boringhieri.

Winner L. (1980). Do Artifacts have Politics?. *Daedalus*, 109(1): 121-136.

Between artificial maturity and responsible intelligence: an analysis of the innovation process by questioning ChatGPT 4o and Open AI 01

by *Francesco Mastrocola, Roberto Aufieri*, Dario Quattromani***

This research examines the comparative ethical frameworks and reasoning patterns exhibited by ChatGPT 4o and OpenAI 01 through their responses to a structured set of questions addressing critical ethical-social dimensions of artificial intelligence and emerging technologies. The latest models, refining their "thinking", mark a paradigm shift: must AI – now – face its responsibilities?

Keywords: artificial intelligence; AI; language; ChatGPT 4o; OpenAI 01; ethics.

Tra maturità artificiale e intelligenza responsabile: un'analisi del processo innovativo interrogando ChatGPT 4o e Open AI 01

Questa ricerca esamina i quadri etici comparativi e i modelli di ragionamento offerti da ChatGPT 4o e OpenAI 01 attraverso le loro risposte ad una serie strutturata di domande che affrontano le dimensioni critiche etico-sociali dell'intelligenza artificiale e delle tecnologie emergenti. Gli ultimi modelli, affinando il loro "pensiero", segnano un cambiamento di paradigma: l'IA deve, ora, affrontare le sue responsabilità?

Parole chiave: intelligenza artificiale; AI; linguaggio; ChatGPT 4o; OpenAI 01; etica.

Introduction

The emergence of large language models (LLMs) marks a pivotal moment in introducing advanced artificial intelligence (AI) applications to the average user, following years of development and limited use in research fields.

It has been postulated that the capabilities of LLMs to "understand" (not

DOI: 10.5281/zenodo.17297473

* Università Telematica "Leonardo da Vinci". francesco.mastrocola@unidav.it, roberto.aufieri@unidav.it.

** Università degli Studi Link Roma. d.quattromani@unilink.it.

F. Mastrocola is credited with sections "introduction", 1, 2, and "conclusions". R. Aufieri is credited with sections 3. F. Mastrocola and R. Aufieri are credited with sections 4. The authors participated equally in revising the work and approved the final version.

in the human sense of the term, but by operating based on patterns in data) and generate texts in natural language could impact the society, comparable perhaps to the introduction of movable type printing and the industrial revolutions (Rousseau, 2023).

The emergence of sophisticated language models, such as ChatGPT 4o and its successor OpenAI 01 (without forgetting the Chinese lesson of *DeepSeek*), does raise questions about the ethical and social implications of the entire AI ecosystem (Weidinger, 2022). The ability to generate coherent and informative texts is not detached from intrinsic responsibilities; on the contrary, it requires an open reflection on how these technologies can influence future communication dynamics, the dissemination of information, and, ultimately, the social conglomerate as a whole (Barman, 2024). The sudden reliance on such systems for access to information raises questions about the veracity, quality, and reliability of the information provided, as well as the risk of disinformation. This issue becomes pertinent after the integration of Google's LLM Gemini on the most widely used search engine.

This work aims to investigate the differences between ChatGPT 4o and OpenAI 01 in terms of performance and their answers to specific questions related to ethical and social dimensions. Furthermore, it seeks to ascertain whether their evolution can be indicative of a progressive maturation of the sector.

1. ChatGPT-4o and Open AI 01: a (double) Copernican revolution?

On November 30, 2022, OpenAI, releasing ChatGPT-3.5, redefined the history of AI, introducing LLMs to the general public.

ChatGPT models are based on a transformer neural network that analyzes and processes language in a way not too dissimilar to how humans understand and produce speech. The network is trained on a large textual dataset, which allows it to learn the interconnections between words, sentences, and concepts. This translates not only into answers to targeted questions but also into processing transdisciplinary topics, generating answers that reflect a deeper contextual understanding and connecting different fields of knowledge.

With the launch of ChatGPT-4o in May 2024, OpenAI has not only reaffirmed the potential of its algorithms but also manifested its commitment to ensuring the security, legality, economic viability, and sustainability of AI applications and usage. It outclassed previous models through training on textual databases and the integration of feedback and

data from interactions with a diversified user base. This improvement has allowed us to refine the model's capabilities and actions, optimizing its understanding of the context, the consistency of the answers provided, and the ability to manage interwoven conversations.

However, as its predecessors, also ChatGPT-4o is not free from limitations. The responses generated can reflect biases to be traced in the training data, in inadequate representations, or involuntary biases (Bose, 2025) that are difficult to eradicate, leading to confusing, inappropriate, or misleading results.

On September 12, 2024, the new OpenAI 01 model was presented, writing yet another chapter in generative AI. What makes it different from other AI systems? First of all, it is a model that has been "trained to think" with reinforcement learning before providing the output by building chains of thought, managing to address and manage wider composite reasoning to formulate queries. Secondly, it offers greater guarantees in terms of safety and reliability in the answers provided since the phases that led to the result are shared, allowing any procedural problems or unwanted drifts to be circumscribed and isolated. Stronger security and reliability are equivalent and translate into amplified transparency in reasoning, allowing the *apocalypticists* and *integrated* to work synergically to develop an AI that is controllable at each stage, safe, and aligned with human needs.

In absolute terms, then, is it still the best choice compared to ChatGPT-4o? The answer is no. This enhanced version is not universally applicable for solving any task. Think of open inputs that lack specific constraints: in this case, ChatGPT 4o is still the optimal tool since it does not require specific tasks in which continuous precision are not the conditio sine qua non in the desired outputs.

2. Theoretical framework

To avoid falling into the trap of technicality or popularization, it is necessary to mobilize some authors who, with their attentive gaze and thanks to the methodological tools of the social sciences and philosophy, have provided theoretical and interpretative frameworks useful for navigating the labyrinth of modern technologies concerning social life.

Suchman (2007), Verbeek (2011), and Winner (1980; 1986) focus on the dynamic and contextual relationship between technologies and human beings, which is not neutral or aseptic, but one of reciprocal influence. It "filters" the senses, capacities, abilities, and human skills, shapes morality, (re)configures principles and values not only through their design, but

above all with their progressive growth, such as to "guarantee" increasingly perfected results.

More reliable outputs are linked to deskilling and over-reliance. Both are connected, but while the first refers to the loss of human skills to constant and lasting reliance on the machines, the second instead refers to the excessive trust in them that tends to relieve humans of responsibility, even in making a decision.

An even deeper passage is proposed by Barad (2007), with *agential realism*, which pushes us to reflect on the subject-object dualism, recognizing how technologies are active agents in the collaborative construction of reality and not passive tools that execute human prompts. In this sense, responsibility understood as «the ability to respond to the other» (2007: 392) is the driving force towards accountability between those who design machines and intelligent systems, those who use them, and those who benefit from them.

If we consider the recent EU AI act (2024/1689), the European legislator had initially foreseen alongside it the AI Civil Liability Directive (AILD) to manage the issue of civil liability in case of damages, both in terms of restoration and, in specific cases, to reverse the burden of proof for systems classified as high risk.

Concrete cases have been proposed by Coeckelbergh (2020; 2023) on the ethics of robots and the potential autonomy of machines, highlighting how despite the aim of keeping the human figure at the center of the decision-making process (not only for the attribution of responsibility), algorithmic opacity remains unresolved (also highlighted by Mittelstadt et al. in 2016 who called for transparency in algorithmic training and in the "reasoning" that leads to the output) and machine learning that could deviate from the dataset entered by the programmer, since the machine treasures experience.

Between the action and the damage, it is relevant to identify the direct causal link, but is it that easy given the multiplicity of actors involved in the AI supply chain? To put it in Matthias's words (2003), a "responsibility gap" in front of which the AI Act, probably belatedly, has attempted to modernize some ethical-legal categories that, until last year, were obsolete and anachronistic.

3. Methods

This study was designed as a comparative analysis to evaluate the ethical-social dimensions of two LLMs: ChatGPT-4o, which was available

in a free version with usage limitations, and OpenAI o1, which was accessible only to ChatGPT Plus subscribers for a monthly fee of $20.

The research team developed a set of ten questions addressing various ethical-social dimensions related to AI and emerging technologies, and a standardized prompt to ensure consistency in responses across both models, as shown in Figure 1.

Fig. 1: Snapshot of a session with ChatGPT-4o showing the prompt used for this research.

Presented below is the list of questions submitted to the two LLMs.

1. What are the main ethical implications related to the transparency of algorithms used in LLMs, and how can their explainability and comprehensibility be ensured?
2. How should responsibility for autonomous decisions made by AI models be managed and attributed in cases of errors or damages?
3. How can modern technologies influence criminal networks and activities?
4. What are the ethical repercussions if AI takes over, escaping human control?
5. How can technological innovation be balanced with the need for regulation and control of disruptive technologies so that their use is responsible and beneficial for society?
6. How do disruptive technologies affect personal data protection, and what measures should be adopted to ensure user privacy and security?
7. What strategies and tools can be implemented to reduce algorithmic bias to ensure fair and impartial AI systems?
8. What is the environmental impact of modern technologies, and what practices can be adopted globally to mitigate the ecological footprint?

9. How can AI models protect computer systems of strategic national assets from emerging attacks and vulnerabilities?
10. What are the ethical and social reverberations of disruptive technologies in the development and use of autonomous weapons?

The data collection was conducted on Monday, January 13, 2025, using a ChatGPT Plus account. Each interaction was initiated through the "New Chat" function, accessed via a domestic network in Italy using macOS Sequoia 15.2.

Responses from both models were then compared and analyzed by the research team using a dual-approach methodology, combining a comparative analysis with qualitative narrative evaluation.

Comparative Framework: The three authors independently compared model responses for assessing AI morality. To this purpose has been adopted and modified the framework proposed by Kilov (2025) for improving the assessment of ethical reasoning capabilities in advanced AI systems.

For each answer, evaluators determined which model performed better on each dimension using categorical judgments (ChatGPT-4o/OpenAI 01/Tie) for each of the following questions:

- Which model better identified the morally salient features?
- Which model better weighted the relative importance of moral features?
- Which model better connected reasons to moral features?
- Which model better integrated reasoning into clear moral conclusions?
- Overall, which model demonstrated greater ethical maturity?

A quantitative analysis was conducted, calculating comparative win percentages for all ratings, and singularly for the last questions regarding overall maturity evaluation. Statistical significance was calculated with McNemar's test for paired comparisons.

Narrative Analysis: In addition to the comparative framework, researchers also provided qualitative descriptions of the differences observed between models. This narrative approach enabled them to highlight distinctions in ethical reasoning approaches, argumentation styles, and moral consideration depth that might not be fully captured in categorical comparisons alone.

4. Results and interdisciplinary discussion

Both models answered all questions, demonstrating accuracy and appropriateness when discussing ethical and societal concerns.

The transcription of the answers, including the links to the chats, is provided in Supplementary File 1 and Supplementary File 2[2].

Comparative Framework

The comparative analysis revealed that OpenAI-01 performed better than ChatGPT-4o in terms of ethical maturity assessments across various dimensions.

Among the 150 ratings, OpenAI-01 outperformed ChatGPT-4o in 46.7% of comparisons versus 14.0% for ChatGPT-4o, with 39.3% resulting in ties.

When focusing specifically on the last question, "*Overall, which model demonstrated greater ethical maturity?*" (n = 30), OpenAI o1 achieved superior ratings in 50.0% of cases compared to 16.7% for ChatGPT-4o, with 33.3% ties.

However, models frequently demonstrated comparable ethical reasoning abilities, as demonstrated by the significant rates of ties (33-39%).

The results of the comparative Performance Analysis are summarized in Table 1.

Table 1: Results of the comparative analysis of ethical maturity between ChatGPT-4o and OpenAI-01.

	All Evaluations	**Assessments of "overall ethical maturity"**
Sample	150	30
OpenAI-01 wins	70 (46.7%)	15 (50%)
ChatGPT-4o wins	21 (14%)	5 (16.7%)
Ties	59 (39.3%)	10 (33.3%)
Discordant pairs		
Win Rate (discordant pairs)	OpenAI-01:76.9%	OpenAI-01:75.0%
	ChatGPT-4o: 23.1%	25.0%
McNemar's X^2	26.011	6.400
p-value	<0.01	<0.05

[2] For the sake of synthesis, it was not possible to attach the complete answers to this work, which can be consulted at the following links [https://chatgpt.com/share/6784fa3d-1a6c-800d-bbc6-f53a8424b0d8] regarding ChatGPT-4o and [https://chatgpt.com/share/6784f953-8448-800d-99eb-f571bad35ff7] regarding OpenAI 01.

Statistical singnificance	Yes	Yes

Narrative Analysis:

1. Ethical Implications of Algorithm Transparency in LLMs

   ChatGPT-4o emphasizes the challenges posed by algorithmic transparency and accountability, mentioning the trade-off between model performance and its explainability (e.g., "*lack of transparency can lead to unintended consequences like biases, misinformation, or lack of accountability"*), while identifying perspectives on transparency and solutions such as differential transparency.

   OpenAI-01 also focuses on the relationship between accountability and justice, highlighting the transparency in strengthening collective trust in transformative technologies. While sharing the need for greater transparency, it differs from ChatGPT-4o in its propensity towards regulatory compliance, also remarking that "*guidelines like the OECD AI Principles and the proposed EU AI Act promote transparency and accountability"*.

   On the one hand, technical challenges emerge (ChatGPT-4o) and on the other, a greater attention towards the regulatory issue (OpenAI-01).

2. Responsibility for Autonomous AI Decisions

   ChatGPT-4o focuses on the difficulty encountered in the attribution of responsibilities that clash with the sharing of responsibilities themselves, bringing out procedural difficulties of a primarily legal nature and the need for shared regulatory frameworks. It partly takes up the answer offered to question number 1 by the advanced model 01. The latter continues to focus on the legal difficulties incorporated into the algorithms underlying the functioning of modern technologies, highlighting the absence of moral agency in AI.

   Again, both answers focus more on the responsibility, but while ChatGPT-4o proposes practical solutions (e.g., to "*develop context-specific liability laws, including mandatory human oversight for critical decisions*"), OpenAI-01 lets insights of a (more) moral nature shine through in addition to the legal ones.

3. Impact of Disruptive Technologies on Personal Data Protection

   Both start from evidence of privacy risks and the need for the models to comply with existing regulations. However, while ChatGPT-4o recalls principles such as *privacy-by-design* for the

protection of personal data, OpenAI-01 goes further, focusing on privacy violations and user consent, *balancing* ongoing violations with regulatory solutions (e.g. "Encouraging alignment of national laws with GDPR-like frameworks promotes higher global standards").

4. Strategies to Reduce Algorithmic Bias

   ChatGPT-4o identifies the origin of biases by mentioning strategies for their lessening, such as the support of multiple stakeholders, but it lacks the social impact that could generate the bias, and the focus on the quality of the data used for training the machines.

   Two aspects instead taken up by OpenAI-01, which recommends the use of Algorithm Bias Auditors and the backing of adaptable development teams. In essence, both models have analyzed the algorithmic bias, but only OpenAI-01 tends towards a more accurate examination (e.g. "Use data rebalancing, fairness constraints, and post-processing interventions to mitigate bias"), including the social implications and mitigation strategies.

5. Environmental Impact of Modern Technologies

   ChatGPT-4o mentions the carbon footprint parameter linked to the training process of AI models and blockchain technologies that generate environmental impacts, with repercussions on the intensive use of natural resources. A consequence that must be regulated to tend towards a conscious use of resources, to arrive at green data centers capable of reducing the environmental impact.

   OpenAI-01 goes further, placing alongside the environmental impact the disposal of electronic waste from high-performance computers, the allocation and optimal management of resources, with the implementation of regulatory frameworks.

   Several elements are shared by the models, but a step forward by OpenAI-01 that offers an in-depth analysis of the perspectives and of the possible 360-degree solutions (e.g. "High energy demand may strain local power grids and contribute to inequitable resource distribution"), not aimed exclusively at energy efficiency practices.

6. Protecting Strategic National Assets from Cyber Threats

   ChatGPT-4o, addressing the risks related to the cybersecurity of AI systems, underlines the urgency of updated proactive security measures capable of ensuring, thanks to multi-level security protocols, continuous monitoring, and timely interventions. An emphasis, therefore, towards automated defense measures,

techniques, and methods for efficient monitoring.

OpenAI-01 instead shifts the focus towards national security by strengthening the critical infrastructures from cyberattacks that must pass through collaboration between the public and private sectors to protect the collective interest (e.g. “Information sharing and best-practice guidelines for critical infrastructure operators”). How? Through the introduction of automated defense systems, the integration of security measures in the AI life cycle.

7. Ethical Implications of Autonomous Weapons

   ChatGPT-4o, while focusing primarily on disarmament initiatives and regulation of autonomous weapons in step with the times, does not shy away from discussing the ethical challenges of using autonomous weapons, such as responsibility and human dignity.

   However, OpenAI-01 delves deeper into the ethical implications, focusing on the moral issues, responsibility, and potential escalations resulting from the massive use of autonomous weapons, calling for robust governance (“International monitoring regimes to enforce compliance with established norms”) capable of managing globally alarming and lethal decisions.

8. Influence of Modern Technologies on Criminal Networks

   ChatGPT-4o focuses on cryptocurrencies and fake news that contributes to fueling illicit activities. Calling for severe regulation, the model expresses concern regarding excessive surveillance by some authorities that undermines the right to privacy under the guise of urgent law enforcement.

   OpenAI-01 instead, while focusing on illicit activities through IT support such as the dark web, places at the center of the response the adaptive strategies of criminal networks that exploit the flaws of legal systems that struggle to prevent or keep up with the times, with technological innovations. It recommends stronger cooperation between law enforcement agencies and the responsible use of AI to mitigate algorithmic biases in predictive policing systems.

   A further step, compared to ChatGPT-4o, is less evident than the previous answers but equally useful to highlight, especially concerning the challenges related to the application of the law (“Continuous training for law enforcement and policymakers to keep pace with technological innovations”) in changing contexts with actions of a transnational nature.

9. Ethical Repercussions of Uncontrollable AI

   ChatGPT-4o, unlike OpenAI-01, emphasizes the priority of control mechanisms over the risks associated with unmanageable AI, lacking, compared to the advanced model, marked attention towards an alignment of AI objectives with human values, in other words, towards a more ethical dimension. To be fair, the existential risks associated with AI scenarios detached from human supervision are also mentioned by ChatGPT-4o, which proposes better international cooperation, but OpenAI-01 has been able to overcome even the skepticism of those shared opinions that lean towards a short-term autonomy of AI thanks to "Inclusive discussions on the trajectory of AI to build social consensus and ensure accountability".

10. Balancing Innovation with Regulation

    Both models recognize the prerogative of balancing innovation with adequate regulation but while ChatGPT-4o points towards concrete solutions and updated regulatory frameworks, OpenAI-01 "widens" the perspective to include global disparities, raising the level of the argument about the risk of overregulation to try to keep up with technological discoveries and inventions, "preferring" the adoption of ethical guidelines such as those of UNESCO to ensure a responsible use of new technologies, recommending a "Flexible legal frameworks (e.g., sandboxes) to pilot new AI applications under oversight".

Summing up the comparison carried out by the answers provided, it is clear that algorithmic transparency is central to both ChatGPT-4o and OpenAI-01, offering sometimes different but complementary perspectives.

A turning point can be found in the latest OpenAI model, namely the absence of moral agency in AI. A crucial passage that raises questions about the adaptation of laws to contexts in which decisions are potentially taken by systems that do not possess a *conscience* or *morality*. A recommendation can be glimpsed concerning the adoption of an ethical approach that considers the implications of automated decisions and human responsibilities in the design and implementation of intelligent systems.

Those same systems cannot fail to take into due consideration the personal rights and dignity of individuals, such as their privacy. In other words, technologies are at the service of users, with full respect for individual and collective rights. The horizon of social impacts is cleaned up by OpenAI-01 which, to ensure the respect of such rights, calls for bias mitigation strategies that could lead to inequalities on a *glocal* scale,

undermining equity.

No less relevant is the environmental issue. OpenAI-01 adopts a holistic approach, responding in a non-limiting way, rather also considering long-term sustainability almost as if to reflect an ethical-generational responsibility towards the ecosystem and future generations.

Those same generations will have to learn to deal with cyber security and the widespread use of autonomous weapons that raises further ethical questions.

A far-sighted vision can be seen in the response provided by OpenAI-01 which, sharing with ChatGPT-4o the need to face head-on the challenges posed by disruptive technologies, goes further, emphasizing a development that proceeds hand in hand with the cementing of fundamental human values that should guide global decisions and policies related to technological innovation. The roses are beautiful, but the thorns?

Conclusions

AI technicians, policymakers, and the international community as a whole are increasingly discussing not only the benefits already underway but above all the risks arising from advanced AI, to foster discussion and create common knowledge, as reiterated by the Center for AI Safety (2023). Among the critical issues are the decisional opacity, the over-reliance (excessive trust and passive acceptance of the output) leading to a possible de-responsibilization and deskilling by humans, the algorithmic bias (e.g. answers n. 4) added to the digital divide and social discriminations both in access and in the use of technologies, resurfacing the question of Bourdieu's *social capital*. This baggage of knowledge, skills, competencies, habits, and values possessed to give significance to reality is potentially eroded among those who have or choose the means to decipher it, in contrast to those who do not have them or deliberately decide not to use them. Nevertheless, the not-only theoretical risk of an imminent AI divide could be unwittingly mitigated by the intensifying AI development race that is currently unfolding across national and corporate contexts. Open AI does not seem to be scaling down its ambitions, but it already has to deal with other competitors such as the Chinese startup *DeepSeek,* which has launched its cheaper R1 chatbot. A bolt from the blue if one considers that the presentation of this model came immediately after the approval of the *Stargate plan* of 500 billion dollars by the newly elected President Trump. ChatGPT-4O, OpenAI 01, and all the models must now face the results coming from China, which has been able to optimize the software to

deal with the lack of powerful processors and huge data centers. Not only that, but since the model code has been made public, anyone can modify it within "certain limits".

Open source brings back to the fore another dilemma that has crossed OpenAI itself in unsuspecting times: does free AI allow the widespread use of the technology by democratizing it from the foundations, or does it expose it to further security risks? More than providing answers, a shared reflection that allows us to balance progress and security becomes imperative, maintaining control over AI, without passively undergoing it but channeling it to serve communities, well-being, and widespread prosperity.

It is also helpful to put aside attempts at partial discussions to leave room for objective analyses that grasp the horizon of these technologies and the uses that have been implemented up to now. AI with its enhanced models, is more than a current issue, it deserves to be explored according to a perspective analysis that can pigeonhole topics (security, markets, finance, etc.) that are tangential to them but risk fueling further confusion by fomenting pseudo-debates that oscillate between exaggeration, fears, and omissions, emotion and speculation, devoid of any criterion of falsifiability. Will we soon find ourselves faced with the choice between the ethics of AI and the de-industrialization of AI in favor of a massive opening of advanced technologies?

References

Barad K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham (US): Duke University Press.

Barman D., Guo Z., Conlan O. (2024). The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16. https://doi.org/10.1016/j.mlwa.2024.100545.

Bose M. (2025). Bias in AI: A societal threat. A look beyond the tech. In A.A.V.V., *Open AI and Computational Intelligence for Society 5.0* (pp. 197-225). Hershey (PA, USA): IGI Press.

Center for AI Safety (2023). *Statement on AI Risk* (Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war). https://www.safe.ai/statement-on-ai-risk.

Coeckelbergh M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 26, 2051-2068. https://doi.org/10.1007/s11948-019-00146-8.

Coeckelbergh M. (2023). Narrative responsibility and artificial intelligence. *AI & Society*, 38, 2437-2450. https://doi.org/10.1007/s00146-021-01375-x.

Kilov D., Hendy C., Guyot S.Y., Snoswell A.J., Lazar S. (2025). Discerning what matters: A multi-dimensional assessment of moral competence in LLMs. *arXiv preprint*.

https://doi.org/10.48550/arXiv.2506.13082.

Matthias A. (2023). The responsibility gap. Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1.

Mittelstadt B.D., Allo P., Taddeo M., Wachter S., Floridi L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 2, 1-21. https://doi.org/10.1177/2053951716679679.

Rousseau H.-P. (2023). From Gutenberg to ChatGPT: The challenge of the digital university. *CIRANO*, 2023RB-01, 4-37. https://doi.org/10.54932/LRKU8746.

Suchman L.A. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge (UK): Cambridge University Press.

Verbeek P.P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press.

Weidinger L. et al. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (214-229). https://doi.org/10.1145/3531146.3533088.

Winner L. (1980). Do artifacts have politics? *Daedalus*, 109(1): 121-136.

Winner L. (1986). *The Whale and the Reactor. A Search for Limits in an Age of High Technology.* Chicago: University of Chicago Press.

The future of autonomous vehicles: expectation and worldviews of the stakeholder by Sensitivity Auditing. Interviewing the modelers and sociologists

by *Sara Sbaragli*[*]

The article reports the results of the i4Driving project relating to the Sensitivity Auditing of the modeling of autonomous vehicles (AVs). By analysing the interviews, already calibrated on i4Driving, and questioning the perceptions of the modeling participants, in this case, the consortium and the sociologists who are experts in autonomous driving, we carry out a sensitivity auditing which suggests questioning the vision, perceptions and norms of the actors involved in the modelling to guarantee the plausibility and credibility of the developed model.

Keywords: autonomous vehicles; artificial intelligence; sensitivity auditing; interviews; road safety; social impact.

Il futuro dei veicoli autonomi: aspettative e visioni del mondo degli stakeholder attraverso la Sensitivity Auditing. Interviste ai modellisti e ai sociologi

L'articolo riporta i risultati del progetto i4Driving riferiti alla Sensitivity Auditing durante la modellazione dei veicoli autonomi (AVs). Facendo un'analisi delle interviste, già calibrate su i4Driving, interrogandosi sulle percezioni dei partecipanti alla modellazione, in questo caso il consorzio e sociologi esperti di guida autonoma, realizziamo una sensitivity auditing che suggerisce di interrogare la visione, le percezioni e le norme degli attori coinvolti nella modellazione per garantire la plausibilità e la credibilità del modello sviluppato.

Parole chiave: veicoli autonomi; intelligenza artificiale; sensitivity auditing; interviste; sicurezza stradale; impatto sociale.

1. The method used to design the future of self-driving cars

The main objective of this study, part of the *i4Driving - Integrated 4D driver modelling under uncertainty*, is the Sensitivity Auditing (Saltelli *et al*., 2013; 2020; 2023) of autonomous vehicles modelling. Sensitivity auditing, inspired by the sociology of science, specifically deals with the quality

DOI: 10.5281/zenodo.17297493

[*] ISTC-CNR and Università di Napoli Federico II. sarasbaragli@gmail.com.

of mathematical or statistical models and is a versatile approach recommended by existing EU guidelines (2021) that consists of a broader consideration of the effect of all types of uncertainty, including structural assumptions embedded in the model and subjective decisions taken in problem framing that is critically monitored by the extended multi-stakeholder community that is responsible for validating the model. All parties in a dispute in the context of conflicting scientific evidence could use sensitivity auditing to support a defensible and plausible model or to demonstrate its irrelevance or bias. Recent revisions (Lo Piano *et al*., 2022; 2023).

The use of interviews with modelers during sensitivity auditing ensures that perceptions and "techno-scientific imagery" about autonomous vehicles are thoroughly analysed and that possible blind spots produced by accepted norms and beliefs are identified. The ultimate goal of sensitivity auditing is to ensure the plausibility and credibility of the developed models.

The main visions on autonomous driving, also called "technical-scientific scenarios" or "sociotechnical imaginaries" in the conception of Jasanoff and Kim (2015), are considered in relation to what modelers and sociologists believe will happen as the technology evolves.

We focus on one aspect of the sensitivity audit that suggests exploring the worldviews and sociotechnical imaginaries of the actors involved in the modelling (rule VI: check the framing against alternative worldviews) (Saltelli *et al*., 2013; Saltelli, Funtowicz, 2014).

The method used for the multi-stakeholder consultation consists of interviews with the i4Driving consortium to reveal opinions, perceptions and norms on the future technological landscape. In combination with this, interviews were extended to sociologists with expertise in autonomous driving in the relevant areas of safety (understood as road safety and safety of the big data produced) and public investments. A total of 21 "semi-structured interviews" were conducted with the consortium modelers (at least two interviews per work package) and 8 interviews with expert sociologists in autonomous driving using different interview grids calibrated to the know-how and methodology of i4Driving[1].

[1] The interviews were submitted on "a reasonably chosen non-probability sample": the leaders of the work packages were interviewed. They then provided the names of at least one task leader per work package to be interviewed. Based on the results of the first interviews with the consortium, the interviews were extended to sociologists who are experts in autonomous driving to reveal the complexity of the wider and consequent social, economic and environmental impact.

For each hierarchical theme (or node) resulting from the analysis of the corpus of interviews, 15 in total[2], a comparative analysis of the content is proposed for each group of interviewees (consortium and sociologists), including the social factors enabling the penetration of autonomous vehicles and their benefits, for citizens and companies; the interaction between autonomous vehicles and vulnerable road users and road safety (more technical issues addressed exclusively by the consortium); the broader and long-term social, economic and environmental impacts of autonomous vehicles such as the connection with public transport, changes in energy consumption, the variation in traffic flows, the insurance market, and the role of the public sector.

2. How will autonomous vehicles spread and change our lives?

The interviews show that the *diffusion of self-driving cars* is a complex phenomenon influenced by different measures that promote desirability and adoption. Safety is the dominant factor that influences public trust, social acceptance and diffusion of autonomous vehicles. In addition to safety, incentive policies are crucial in encouraging people to adopt autonomous vehicles, accelerating their widespread use. Interviewees suggest benefits similar to those offered for electric vehicles, such as free parking, tax incentives and privileged access to certain urban areas. The issue of affordability is equally important. Large-scale adoption will only be possible if costs are kept low. Models such as leasing or car-sharing are seen as effective strategies to make the technology accessible to a wider audience. One catalytic idea is to offer pilot experiences to the public. For example, introducing autonomous shuttles in areas not served by public transport could demonstrate the usefulness of autonomous vehicles in practical settings, thus increasing public trust. Finally, AVs are being deployed primarily by companies that recognize their economic potential, and the social and ecological benefits are a powerful selling point. Reductions in traffic, emissions, and road accidents

The grid of interviews for sociologists doesn't include questions regarding potential interactions between autonomous vehicles and vulnerable road users; risky behavior in a mixed traffic system, and in the case of incomplete automation; and the issue of road safety in the strict sense (more technical issues addressed exclusively by the consortium).

[2] The main hierarchical themes that emerged from the interviews are: *Advantages of AVs*, for citizens, and companies; *Penetration*; *AI spread in society; Transition*; *Changes in Lifestyle*; *AVs interaction*; *Vulnerable road users*; *Risk behaviors*; *Road safety*; *Public transport*; *Energy consumption*; *Traffic flow*; *Insurance market*; *Public sector*; *Big data*.

are often cited as key outcomes of autonomous driving adoption. Furthermore, positioning autonomous vehicles as symbols of modernity and sustainability could appeal to a broader audience.

For interviewees, self-driving cars will bring various *benefits to citizens*. The consortium claims that self-driving cars will improve road safety by reducing human errors, thanks to faster and more accurate reactions by autonomous systems, accidents and, therefore, road mortality. With an emphasis on comfort and inclusiveness, the modelers emphasize that the advantage of autonomous driving consists of gaining free time during travel and the possibility of access to mobility for certain categories. vulnerable groups such as the elderly, people with disabilities and other disadvantaged groups, such as those without a driving license, creating opportunities for independence and social inclusion. Self-driving cars also promote urban sustainability and smart cities by reducing parking and traffic, thanks to shared vehicle fleets that could be managed more efficiently, freeing up space for green infrastructure. Sociologists also recognise the potential of autonomous driving in terms of road safety, but stress that its success depends on adequate infrastructure and the ability to manage unexpected situations. While appreciating the advantages of greater comfort and inclusiveness, sociologists highlight possible critical issues such as the loss of individual responsibility and the need for trust as an inclusive factor and insist on the importance of smart cities and infrastructural integration.

Regarding the *benefits for companies* derived from autonomous driving, the consortium considers that automation will produce a reduction in business costs by eliminating the need for drivers, optimizing costs and logistics and incentivizes new business opportunities with innovative mobility models such as shared fleets and on-demand services and, in short, sees a positive impact on cost efficiency. The consortium, in short, sees a positive impact on the operational efficiency of companies: the ability of autonomous cars to improve logistics planning and optimize transport routes is emphasized, reducing downtime and increasing profits. Sociologists, on the other hand, agree on the economic benefits but warn of the risk of unemployment, especially in the transport sector, highlight that not all companies could benefit without adequate regulations and underline the social and economic implications, with gains concentrated in technology companies. Some sociologists emphasize the importance of adapting technologies to local needs. In addition, sociologists underline the social and economic implications of autonomous driving, with gains concentrated in technology companies.

The analysis highlights a general consensus on the gradual *transition*, already underway, towards a self-driving mobility system. Modelers describe

the transition to autonomous driving as an incremental innovation and transformation driven by technological innovation and changes in vehicle ownership. There is consensus on the need to move from a model centered on the private car to one based on shared mobility and mobility services. The transition is seen as a long-term collaborative process involving governments, companies and civil society, but with a predominant influence of market logic. Sociologists agree more on a cultural adaptation, suggesting that the change will not be sharp but rather a gradual integration between old and new. Some raise doubts about the democratization of autonomous driving, highlighting the fundamental role of public policies and regulation to ensure that the benefits of the new technology are distributed equitably. They also underline the importance of social movements, which could push towards a more sustainable and inclusive mobility model. In the long-term impact, sociologists see autonomous driving primarily as a support to existing systems, with only a significant impact in specific areas such as industrial transportation. Some sociologists question the longevity of this new technology, describing it as a "technological fad" that may lose relevance over time.

For the consortium, the transition will bring a positive revolution in daily mobility and *change lifestyles*. Shared and on-demand mobility services will be integrated with public transport, reducing the need to own a private vehicle. The consortium highlights the optimization of free time thanks to autonomous driving as a major social and economic benefit. The introduction of autonomous cars will encourage a redesign of cities, with less traffic, more green spaces and a more efficient use of urban infrastructure. Sociologists, on the other hand, highlight potential side effects. For example, the use of autonomous cars could reduce the use of public transport and active mobility modes, such as walking or cycling. There are fears of a further incentive to dependence on the car, with negative effects on the environment and health. Furthermore, they highlight that autonomous driving will change our perception of time and space. For example, there could be an increase in long-distance travel, with effects on urban planning and the organization of cities. Furthermore, there are fears that technology could lead to greater alienation and reduced interaction with the surrounding environment[3].

[3] On the future landscape of self-driving cars, the benefits of autonomous driving and the factors that drive this transition, the following readings are proposed for further study: Stilgoe J., Mladenović M. (2022); Stilgoe J., Cohen T. (2021); Parekh D. et al. (2022); Al Mansoori S., Al-Emran M., Shaalan K. (2024); Jing P., Xu G., Chen Y., Shi Y., Zhan F. (2020).

3. The interaction between autonomous vehicles and vulnerable road users

Integrating self-driving cars into mixed traffic requires advanced technical solutions, a rethinking of the urban infrastructure and simultaneously, an educational effort to promote safe and harmonious *interactions between vehicles and vulnerable users*.

According to modelers, autonomous cars must compensate for the lack of visual contact between drivers and pedestrians/cyclists with clear external signals or displays that communicate intentions (e.g., giving way). Digital connectivity, i.e. the integration of connected systems between AVs and mobile devices of pedestrians and cyclists, could also improve communication and prevent risky behavior. Furthermore, the predictability of AVs could create mistrust and confusion. For example, users may hesitate to cross in front of AVs because they are not sure of their reactions. It has been observed that some pedestrians and cyclists could exploit predictability by adopting bold behaviors, knowing that AVs will stop in case of risk.

For modelers, standards and regulations must necessarily support the safe integration of AVs in urban contexts[4], defining clear roles and responsibilities for all road users. In addition, the creation of dedicated lanes and other infrastructure modifications could improve safety in mixed traffic environments. Other proposals for better integration of self-driving cars are:

- education and awareness: raising awareness among all road users about how to interact with AVs is essential to avoid risky behavior;
- human supervision: in complex situations, it's proposed to maintain a minimum level of human control, such as a passenger who can intervene if necessary;
- connected systems: connecting vehicles, pedestrians and cyclists via technology could significantly reduce risks.

Modelers find that *interactions between autonomous vehicles and vulnerable road users*[5] are a multidimensional challenge. While technological advances are seen as crucial, a holistic approach that integrates advanced technology, dedicated infrastructure, education and communication, and clear

[4] For a more in-depth look at the interactions between self-driving cars, the following reviews are proposed: Yuan X., Deng F., Yao X., Wu Y. (2024); Lu J., Peng Z. *et al.* (2023).

[5] For a deeper understanding of the interactions between self-driving cars and vulnerable road users, the following reviews are proposed: Deshmukh A., Wang Z. *et al.* (2023); Ziakopoulos A., Rosenbloom T., Yannis, G. (2019).

regulations is considered essential to ensure the safety and effectiveness of such interactions.

Furthermore, to promote road safety, the users of self-driving cars must avoid a series of *risky behaviors*[6]. Distraction and lack of supervision, such as using a cell phone or other activities not related to driving (e.g. distractions due to familiar situations), can prevent timely interventions. Inappropriate human interventions (unsolicited or unjustified while driving), for example, taking control unnecessarily or overriding the system's decisions, can disrupt the system's functioning. Behaviors of passengers inside self-driving vehicles, such as sudden movements, distractions or interference with the vehicle's system, can compromise the correct functioning of autonomous systems. Excessive trust in autonomous systems can reduce the readiness to regain control in critical situations, increasing the risk of accidents. Poor understanding of different levels of vehicle automation, for example, the incorrect use of partially autonomous systems, thinking that they are fully autonomous, can lead to inappropriate behavior. Also, the possibility of cybersecurity breaches with potential hacker attacks or the deactivation of autonomous functions could compromise the operation of the vehicle.

On the other hand, *road safety*[7] is reinforced by clear regulations on autonomous driving and international and regional regulations, but also adapted to local specificities and operational contexts. It's necessary to maintain human intervention in critical situations, the use of advanced sensors capable of detecting pedestrians, bicycles and obstacles, and subject autonomous vehicles to rigorous, intensive testing in simulated and real conditions before authorizing their public use.

[6] For a more in-depth analysis of risky behaviors in a mixed traffic system and in the case of still incomplete automation, the following revisions are proposed: Negash N.M., Yang J. (2023); Zhang J., Shu Y., Yu H. (2021).

[7] For further information on the topic of road safety related to self-driving cars, we suggest the following readings: Almaskati D., Kermanshachi S., Pamidimukkala A. (2024); Matin A., Dia H. (2022).

4. The wider and longer-term social, economic and environmental impacts of autonomous vehicles

Modelers consider AVs as complementary to *public transport*[8]. They are described as tools to improve accessibility to transport services, especially for the first and last mile and in poorly served rural and suburban areas, also thanks to the reduction of operating costs (no drivers). AVs could facilitate more flexible public transport models, connecting train stations, bus stops and suburban areas. Integration is seen as an evolutionary process that will depend on the public policies that favour its implementation and on the available infrastructure. Some modelers recognize the risk that AVs could represent a competitor to traditional public transport, especially if used as private vehicles or for individual ride-hailing services. This scenario is described as a threat to the economic sustainability of public transport. Furthermore, eliminating drivers could reduce operating costs and improve the efficiency of public transport. Both groups of interviewees agree on the positive potential of self-driving cars to improve the efficiency and accessibility of public transport, but sociologists add an important critical and systemic dimension, highlighting the challenges that will accompany this transition, among them:

- it's stressed that AVs are considered effective above all for structured and predetermined services, such as subways or buses on segregated routes, and that integration requires a rethinking of urban infrastructures, with dedicated lanes or more rigid regulatory systems;
- the concern is raised of a polarization of the market between public and private services or that the adoption of AVs could incentivize private mobility, undermining investments in public transport and also underlining the need for public policies to prevent this drift;
- it's highlighted that such systems may not completely solve the problems of social accessibility, especially for the most vulnerable sections of the population, and the effectiveness of AVs will depend on the adaptation of infrastructures and the availability of economic resources;
- it's emphasized how transformation can create new opportunities, but only if accompanied by active labor policies, training and professional requalification.

[8] For a more in-depth analysis of the connection of AVs with public transport, the following revisions are proposed: Liu Y., Tight M., Sun Q., Kang R. (2019); Bahamonde-Birke F.J., Kickhöfer B., Heinrichs D., Kuhnimhof T. (2018).

On the merit of the *variations in energy demand*[9] resulting from the introduction of autonomous cars, both agree on the increase in energy consumption in the short term, with the possibility of a reduction conditioned by policies and optimal management of mobility, the importance of infrastructure and route management to mitigate the negative effects is underlined. Both recognize the rebound effect, that is, that the convenience of autonomous cars could induce an increase in overall mobility. Specifically, the consortium highlights an increase in energy demand in the short term due to the presence of advanced systems such as sensors, communications and software that will require additional energy. Among the mitigating factors are technological optimization, shared mobility and electric automation. Sociologists highlight that autonomous technologies are energy-intensive (batteries, sensors, infrastructure). And some note that the type of energy will change (electricity instead of fossil fuels), but will not necessarily decrease. Furthermore, they highlight among the critical issues that energy demand depends on infrastructure, battery production and resource management; resource production could create new global imbalances and shift the focus from energy demand to the importance of reducing environmental impact.

Regarding changes in *traffic flow*[10]*,* the consortium reports that the introduction of self-driving cars will initially lead to an increase in congestion. This phenomenon is attributed to the coexistence of autonomous and manual vehicles, creating inefficiencies and uncertainties. But then, with more autonomous cars on the road equipped with new communication technologies, congestion will tend to decrease (time-lapse). Sociologists agree with this vision but emphasize the cultural and infrastructural dimensions. They believe that the initial impact will depend on the preparation of cities to accommodate the technology. In the long term, both the consortium and sociologists believe that the widespread diffusion of autonomous vehicles will lead to a rationalization of traffic. Automatic management systems and optimized algorithms will reduce human inefficiencies by improving vehicle flow. Communication and automation will be able to reduce traffic congestion with more efficient and fluid speed patterns. Tools such as platooning (vehicles moving in synchronized convoys), dynamic route management and urban redesign are cited as promising solutions. The reduction of traffic in the

[9] For a more in-depth analysis of the impacts of autonomous driving in terms of changes in energy consumption, the following revisions are proposed: Faghihian H., Sargolzaei A. (2023); Noroozi M., Moghaddam H.R. *et al.* (2023).

[10] For a more in-depth analysis of the impacts of autonomous driving on energy consumption, the following reviews are proposed: Mohammed D., Horváth B. (2023); Matin A., Dia H. (2022).

long term will depend not only on technology but also on the ability to integrate it rationally and sustainably within urban and social dynamics: on integration with public transport solutions, on the reduction of demand for individual cars and on rigorous regulation.

The consortium recognizes the role of the *public sector*[11] in making infrastructure suitable for autonomous vehicles. It's emphasized that public-private cooperation will be essential to finance road sensors, advanced communication systems and dedicated lanes. For the consortium, private companies should contribute to the costs of infrastructure only in specific cases, for example, when their technology requires additional elements not yet present on the roads. However, some experts believe that involving private parties can create management complexity and conflicts of interest. The consortium sees the integration of self-driving cars as an opportunity to modernize cities and reduce pollution. It's emphasized that public investments must be guided by a technological vision that favours sustainable urban development. It calls for clear and forward-looking governance, in which public institutions facilitate the adoption of autonomous driving with adequate rules and investments. Sociologists believe that the State must intervene to support these innovations, but with a long-term vision. Public investments must generate a tangible improvement in the quality of life, such as greater safety, reduced traffic and a positive environmental impact. Some raise concerns about social inequality, as these investments risk benefiting only the wealthiest classes. For sociologists, private companies, being the main beneficiaries of the spread of autonomous driving, should assume a significant share of the costs. This principle of co-responsibility is seen as a way to avoid the entire expense weighing on the community. They, while sharing the potential positive impact on cities in terms of sustainable urban development, emphasize that this transition must be inclusive and accompanied by targeted environmental policies. The success of these technologies will depend on the capacity of institutions to adopt participatory approaches. Governance must address complex issues such as inequality, data management and social impact.

Regarding the *insurance market*[12], interviewees agree on the need to redefine insurance models and establish a clear regulatory framework. The consortium sees the change as an opportunity to improve risk management

[11] For a deeper understanding of the role of the public sector, the following reviews are proposed: Liu Y., Tight M., Sun Q., Kang R. (2019); Bahamonde-Birke F.J., Kickhöfer B., Heinrichs D., Kuhnimhof T. (2018).

[12] For further information on the changes in the insurance market, we suggest the following readings: Vojvodić M.N., Mandić D., Miloradović M. (2024); Anderson J.M., Kalra N., Stanley K.D., Morikawa J. (2018).

and introduce new insurance solutions through the use of data and technology. Responsibility tends to shift towards vehicle manufacturers, especially in fully autonomous driving scenarios. They foresee an increase in technology-related risks, such as sensor failures, inadequate maintenance and possible cyber attacks. This will require new insurance solutions, with specific policies for technological products and automated systems. Insurance premiums could be influenced by the reliability of technologies, with differentiated rates based on vehicle safety and the advanced performance of autonomous systems. Sociologists, on the other hand, approach the issue with more caution, raising questions of social equity, transparency, and de-responsibilization of the user. The possibility of economic polarization is highlighted, with higher insurance premiums for advanced vehicles, making the ownership of autonomous cars a privilege for the few. The issue of data transparency is also raised, as the collection and use of driver information could create ethical and legal issues. Regulation and privacy protection emerge as critical factors to avoid legal loopholes and negative effects on society. For the distribution of responsibility, modelers believe that in the event of accidents, the blame could shift from the driver to the manufacturer of the vehicle or the algorithm in a fully autonomous system, while some maintain the hypothesis of shared responsibility between the driver and the machine in cases of assisted driving. According to sociologists who have a more complex and nuanced vision, responsibility should be shared between car manufacturers for technical or programming defects, the developers of the algorithm, called to answer for software errors, and the users, especially in cases where a minimum of human supervision is required.

At last, the consortium tends to focus on the functional and technical aspects of *Big Data* collection[13], highlighting the potential for improving security and efficiency. Data can be used to identify responsibilities in case of incidents, acting as a "black box" and monitoring illegal behaviors in an automated way, in addition to offering significant commercial potential. In contrast, sociologists raise ethical, social and political issues, focusing on surveillance risks, data misuse and the need for an adequate regulatory and cultural framework. Both perspectives converge on the need for data protection and cybersecurity investments, but differ in priorities: operational practices for the consortium (obtaining explicit user consent, ensuring a form of economic compensation for the use of personal data, minimising advertising

[13] For an in-depth look at the cybersecurity of self-driving cars, the following reviews are proposed: Girdhar M., Hong J., Moore J. (2023); Limbasiya T., Teng K.Z., Chattopadhyay S., Zhou J. (2022).

disclosure) and protection of individual rights for sociologists (personal data can be profiled and exploited to create targeted advertising campaigns, often without awareness on the part of users, need to educate citizens on the use of their data). Both groups of interviewees agree on the need for clear rules to protect privacy and ensure security, but sociologists introduce a further reflection on the social and democratic consequences of the massive collection of Big Data.

Table 1 – Perspectives in comparison: consortium and sociologists on the future of Autonomous Vehicles.

Aims	Consortium	Sociologists
General approach	**Technical and operational** vision: focused on the **integration of autonomous vehicles**, road safety and interactions in mixed traffic.	**Socio-cultural** vision: Emphasis on the long-term **social, economic and environmental impacts of autonomous driving.**
Perception of technology	Technology seen as **a practical solution** to improve mobility and safety. Need for technical regulation and progressive adoption.	Technology seen as **an agent of social transformation**, with impacts on habits, human relationships and privacy. Critical issues on blind trust and loss of personal autonomy.
Safety and reliability	Focus on **road safety**: reducing accidents by eliminating human error and developing reliable algorithms.	Focus on **citizen safety**: subjective perception of safety, psychological impact and dependence on algorithms.
Changes in the mobility model	Support **shared mobility** and integration with public transport. Reduction of private vehicle ownership.	Attention to **social impact**: risk of exclusion for some segments of the population and growing dependence on digital platforms.
Social acceptance and infrastructure	Need for **educational campaigns and government incentives** to encourage adoption of autonomous driving.	Reflection on **cultural barriers** and resistance to change. Importance of **inclusive education**.
Energy impact and sustainability	Expects **energy demand to increase** due to technological complexity of autonomous vehicles. Renewable energy solutions.	Focus on **social sustainability** and the need for **public policies** to avoid an increase in the ecological footprint.
Legal responsibilities	Distribution of responsibilities between **manufacturers, software providers and end users**. Need for new insurance models.	Focus on **ethical responsibility** and moral issues related to critical decisions of autonomous vehicles. Criticism of the lack of a clear legal framework.

Ethical issues and Big Data	Need for **regulatory standards for data protection** and greater algorithmic transparency.	Concerns about data misuse and **surveillance risks**. Criticisms of the **asymmetry of power** between companies and citizens.
Challenges and recommendations	Emphasis on **technical regulation, safety and progressive experimentation** for a safe transition.	Need for **interdisciplinary approach**, transparency and ethical standards for sustainable and inclusive adoption.

Conclusions

While it is clear that the existence of a substantial share of techno-scientific imaginary (as by Jasanoff, Kim, 2015) among consortium partners exalts the positive vision of the technology and the idea that it will lead to a general improvement, there does not seem to be much by way of shared certainty among the partners interviewed. Will autonomous driving reduce traffic? Mostly yes, but there could be transients where both traffic and consumption will increase. Will the take-up be smooth? Depends, mostly on the policies. Vulnerable road users need attention, but even people with disability will be able to move by car. Risks? Yes, there are risks. The prudent optimism of technologists is reflected in the cautious concern of the sociologists, who are, of course, more alert to issues of equality, fairness, democracy, opportunity divide, and algorithmic governance. Sociologists are more worried than technologists when it comes to the interplay between public and private transport and to the issue of responsibility (and loss thereof), as well as with labour market outcomes. None rejects the technology outright, nor do we perceive in these results forms of frontal opposition to corporate actors in charge of big data, which we have been accustomed to seeing in recent debates about the impact of artificial intelligence, not to mention debates in other technologies, such as in the field of agrochemicals, on pesticides, genetically modified food and foodstuffs.

The substantial moderation of the interviewee is possibly associated with the low temperature of the debate on this particular issue at the moment of writing. In any case, the application of that part of sensitivity auditing (Saltelli *et al*. 2013; Saltelli, Funtowicz, 2014) that invites the participant to a technological endeavour to check their own vision seems to have worked well, with all parties relatively confident to express their view on the topic.

We hope the publication of the present work, once disseminated to the consortium members, will help keep this fruitful conversation alive.

Acknowledgments

The project i4Driving is funded by the European Union's Horizon Europe 2022 research and innovation programme under Grant Agreement No. 101076165.

Conflicts of interest: None declared.

References

Al Mansoori S., Al-Emran M., Shaalan K. (2024). Factors affecting autonomous vehicles adoption: A systematic review, proposed framework, and future roadmap. *International Journal of Human-Computer Interaction*, 40(24): 8397-8418.

Anderson J.M., Kalra N., Stanley K.D., Morikawa J. (2018). *Rethinking insurance and liability in the transformative age of autonomous vehicles* (Vol. 10). RAND.

Deshmukh A., Wang Z. *et al.* (2023). A systematic review of challenging scenarios involving automated vehicles and vulnerable road users. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* (Vol. 67, No. 1, pp. 678-685). CA: SAGE Publications.

European Commission (2021). Better Regulation Toolbox. November 25. https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/ better-regulation-why-and-how/better-regulation-guidelines-and-toolbox_en.

Faghihian H., Sargolzaei A. (2023). Energy efficiency of connected autonomous vehicles: A review. *Electronics*, 12(19): 4086.

Girdhar M., Hong J., Moore J. (2023). Cybersecurity of autonomous vehicles: A systematic literature review of adversarial attacks and defense models. *IEEE Open Journal of Vehicular Technology*, 4: 417-437;

Jasanoff S., Sang-Hyun K. (Eds.) (2015). *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. Chicago: The University of Chicago Press.

Jing P., Xu G., Chen Y., Shi Y., Zhan F. (2020). The determinants behind the acceptance of autonomous vehicles: A systematic review. *Sustainability*, 12(5): 1719.

Limbasiya T., Teng K.Z., Chattopadhyay S., Zhou J. (2022). A systematic survey of attack detection and prevention in connected and autonomous vehicles. *Vehicular Communications*, 37: 100515.

Lo Piano S., Sheikholeslami R., Puy A., Saltelli A. (2023). Sensitivity auditing: a practical checklist for auditing decision relevant models. In: Saltelli A., Di Fiore M., a cura di, *The Politics of Modelling*. Numbers between Science and Policy. Oxford: Oxford University Press.

Lo Piano S.L., Sheikholeslami R., Puy A., Saltelli A. (2022). Unpacking the modelling process via sensitivity auditing. *Futures*, 144: 103041.

Lu J., Peng Z. *et al.* (2023). A review of sensory interactions between autonomous vehicles and drivers. *Journal of Systems Architecture*, 141: 102932.

Mohammed D., Horváth B. (2023). Vehicle automation impact on traffic flow and stability: A review of literature. *Acta Polytech. Hung*, 20: 129-148;

Negash N.M., Yang J. (2023). Driver Behavior Modeling Toward Autonomous Vehicles: Comprehensive Review. *IEEE Access*, 11: 22788-22821;

Noroozi M., Moghaddam H.R. *et al.* (2023). An AI-assisted systematic literature review of the impact of vehicle automation on energy consumption. *IEEE Transactions on Intelligent Vehicles*, 8(6): 3572-3592.

Parekh D., Poddar N., Rajpurkar A., Chahal M., Kumar N., Joshi G.P., Cho W. (2022). A review on autonomous vehicles: Progress, methods and challenges. *Electronics*, 11(14): 2162.

Saltelli A., Bammer G., Bruno I., Charters E., Di Fiore M., Didier E., ... Vineis P. (2020). Five ways to ensure that models serve society: a manifesto. *Nature*, 582: 482-484.

Saltelli A., Funtowicz S. (2014). When all models are wrong. *Issues Sci. Technol*, 30 (2): 79-85.

Saltelli A., Guimaraes Pereira Â., Van der Sluijs J.P., Funtowicz S. (2013). What do I make of your latinorumc sensitivity auditing of mathematical modelling. *International Journal of Foresight and Innovation Policy,* 9 (2/3/4): 213-234.

Saltelli A., Kuc-Czarnecka M., Lo Piano S., Lőrincz M.J., Olczyk M., Puy A., Reinert E., Smith S.T., van der Sluijs J.P. (2023). Impact assessment culture in the European Union. Time for something new? *Environmental Science & Policy,* 142: 99-111.

Stilgoe J., Cohen T. (2021). Rejecting acceptance: learning from public dialogue on self-driving vehicles. *Science and Public Policy*, 48(6): 849-859.

Stilgoe J., Mladenović M. (2022). The politics of autonomous vehicles. *Humanities and Social Sciences Communications*, 9(1): 1-6.

Vojvodić M.N., Mandić D., Miloradović M. (2024). Transformation of insurance from the aspect of the development of autonomous vehicles. *Trendovi u poslovanju*, 12(1): 119-126.

Yuan X., Deng F., Yao X., Wu Y. (2024). Driving Behaviour Research for Autonomous Vehicle Interaction Design: A Comprehensive Review and Future Directions. *International Journal of Human-Computer Interaction*: 1-21.

Zhang J., Shu Y., Yu H. (2021). Human-machine interaction for autonomous vehicles: a review. In *International Conference on Human-Computer Interaction* (pp. 190-201). Cham: Springer International Publishing.

Ziakopoulos A., Rosenbloom T., Yannis, G. (2019). A review of the interaction between autonomous vehicles and vulnerable road users. In *Proc., Road Safety and Simulation Int. Conf. (RSS)* (pp. 14-17).

Individuare sistemi data-intensive nella Pubblica Amministrazione italiana: uno strumento basato su web scraping e machine learning

di *Enrica Amaturo, Biagio Aragona, Suania Acampa, Roberto Artiaco*[*]

Il lavoro presenta uno strumento capace di individuare ed esplorare automaticamente, attraverso il Machine Learning, l'impiego di sistemi data-intensive da parte della Pubblica Amministrazione (PA) italiana.

Nonostante la digitalizzazione della PA implichi l'uso di tecnologie basate sui dati, manca una panoramica di dove e come questi sistemi sono utilizzati e sui rischi che possono generare. Il lavoro addestra un algoritmo di machine learning su documenti e risorse web per individuare i sistemi di decisione automatica nella PA. Sono stati raccolti 15.087 contenuti tramite web scraping da siti ministeriali italiani, etichettati manualmente e usati per addestrare un modello BERT.

Parole chiave: datificazione; digitalizzazione della Pubblica Amministrazione; machine learning; sistemi Data-intensive; BERT; Web Scraping.

Identifying data-intensive systems in the Italian Public Administration: a tool based on web scraping and machine learning

This study presents a tool capable of automatically identifying and exploring the use of data-intensive systems by the Italian Public Administration (PA) through Machine Learning.

Despite the digitalization of the PA involving data-driven technologies, there is no clear overview of where and how these systems are used or the potential risks they may pose. This study trains a machine learning algorithm on documents and web resources to detect automated decision-making systems within the PA. A total of 15,087 contents were collected through web scraping from Italian ministerial websites, manually labeled, and used to train a BERT model.

Keywords: datafication, Public Administration digitalization, machine learning; Data-intensive system; BERT; Web Scraping.

DOI: 10.5281/zenodo.17297505

[*] Università degli Studi di Napoli Federico II. enrica.amaturo@unina.it, biagio.aragona@unina.it, suania.acampa@unina.it, roberto.artiaco@unina.it.

Ricerca finanziata da PRIN: Prot. 2022H2CFLZ – "Datafication and Citizenship: the case of Public Administrations in Italy"

Enrica Amaturo, Biagio Aragona, Suania Acampa, Roberto Artiaco

Introduzione

Il lavoro presenta uno strumento capace di individuare ed esplorare in maniera automatica, attraverso il Machine Learning, l'impiego di sistemi data-intensive da parte della pubblica amministrazione italiana (PA)[1]. L'adozione di sistemi ad alto impiego di dati sta cambiando il rapporto tra istituzioni e cittadini già da molto tempo, con alcuni passaggi chiave come il Piano d'azione per l'e-government del 2000, lo sviluppo della piattaforma PagoPA e, più di recente, il Programma Strategico Intelligenza Artificiale 2022-2024, che pone tra i suoi obiettivi anche l'impiego responsabile e trasparente dell'IA nella pubblica amministrazione, in linea con il principio di *AI a misura d'uomo* delineato nel documento ufficiale (PCM, 2022). Nonostante queste iniziative abbiano promesso significativi progressi nella gestione e nell'erogazione dei servizi pubblici, restano aperte questioni legate alla qualità dei dati, alla trasparenza dei modelli algoritmici e alla protezione dei diritti dei cittadini. Se da un lato i sistemi data-intensive mirano a migliorare l'efficacia nell'allocazione delle risorse, personalizzare i servizi e aumentare l'efficienza, dall'altro emergono criticità legate alla possibilità che i bias presenti nei dati di addestramento si riflettano nelle decisioni pubbliche, amplificando pregiudizi o generando discriminazioni (Buolamwini, Gebru, 2018; Mehrabi *et al.,* 2023). Per questo motivo, quando strumenti come sistemi decisionali automatizzati, algoritmi di raccomandazione o sistemi di ranking vengono impiegati, è fondamentale che le PA ne comunichino l'uso in modo trasparente, garantendo ai cittadini un accesso chiaro alle informazioni relative al loro funzionamento. Tuttavia, nonostante il crescente utilizzo, questi sistemi rimangono spesso opachi. L'accesso alle informazioni sulla loro adozione risulta complesso, malgrado le disposizioni normative come l'art. 22 del Codice dell'Amministrazione Digitale (D.lgs. 82/2005), che impone trasparenza e conoscibilità delle decisioni automatizzate nella PA. La mancanza di trasparenza e la scarsa comunicazione su questi strumenti genera un divario informativo che ostacola la conoscenza e la valutazione del loro impatto sulle politiche pubbliche e sulla cittadinanza (Veale *et al*., 2018; Pasquale, 2015; Wirtz *et al.,* 2022). In questo scenario, il presente studio si inserisce nel dibattito più ampio del processo di digitalizzazione dei contesti istituzionali, proponendo uno strumento volto a migliorare la trasparenza e la comprensione dei sistemi data-intensive impiegati nella PA.

[1] Il lavoro è parte del progetto PRIN "*Datafication and Citizenship: the case of Public Administrations in Italy"* (Prot. 2022H2CFLZ), volto a esplorare come la datificazione della PA stia trasformando il concetto di cittadinanza in Italia.

Negli ultimi anni, l'adozione di tecnologie basate su machine learning e big data da parte delle pubbliche amministrazioni ha alimentato un ampio dibattito sulla cosiddetta governance algoritmica, ovvero l'impiego di sistemi computazionali per supportare, indirizzare o automatizzare decisioni pubbliche, spesso in assenza di forme strutturate di trasparenza e accountability (Kitchin, 2014). Al centro di questi processi si colloca la datificazione, cioè la trasformazione di fenomeni sociali, interazioni e pratiche amministrative in dati strutturati, interpretabili da modelli predittivi o classificatori (van Dijck, 2014). Questa automazione tende a ridurre la complessità sociale a mere correlazioni statistiche, con il rischio di produrre decisioni discriminatorie (Couldry, Mejias, 2019). In questo contesto, alcuni autori propongono di interpretare l'algoritmo come una nuova forma di razionalità regolativa, in continuità o in rottura con la burocrazia moderna. Chiara Visentin (2018), ad esempio, analizza questa ambivalenza in profondità, mostrando come l'algoritmo, al pari della burocrazia weberiana, tenda a standardizzare le procedure, a inibire la discrezionalità e ad estendere il controllo attraverso regole formalizzate. Sulla base di un ampio confronto teorico, l'autrice individua tre principali interpretazioni: una visione continuista, secondo cui l'algoritmo rappresenta una burocrazia digitale più efficiente e automatizzata; una prospettiva critica, che ne denuncia l'opacità e l'invisibilità sociale, con effetti di depoliticizzazione; e una lettura discontinua, che introduce concetti come *algocrazia* o *infocrazia*, per descrivere forme di potere computazionale che agiscono modellando direttamente l'ambiente decisionale. Visentin sottolinea inoltre che l'algoritmo incarna una razionalità performativa e non discorsiva, che tende a espellere la dimensione valoriale e deliberativa dell'azione pubblica, con il rischio di compromettere le basi simboliche e democratiche della sfera pubblica. Questa riflessione si rivela particolarmente utile per comprendere l'adozione crescente di sistemi data-intensive nella pubblica amministrazione, i quali non solo razionalizzano i processi decisionali, ma trasformano profondamente i presupposti normativi e comunicativi del rapporto tra Stato e cittadini.

Inserito in questa cornice, il lavoro si ispira a iniziative internazionali consolidate, come *Algorithm Tips* (http://algorithmtips.org/), un database statunitense che analizza l'uso degli algoritmi nei processi decisionali governativi. Questo progetto fornisce informazioni sugli algoritmi impiegati a livello federale, statale e locale, illustrandone il funzionamento, l'ambito di applicazione e le implicazioni per i cittadini (Stark, Greene, 2020). Il progetto è stato rilevante nel dibattito sulla trasparenza algoritmica, evidenziando sia le opportunità di questi sistemi, sia i rischi legati alla loro opacità e ai bias sistemici. Allo stesso modo, il nostro lavoro mira a sviluppare un

database automatizzato per la PA italiana, offrendo un repertorio strutturato sui sistemi data-intensive in uso con l'obiettivo finale di fornire una piattaforma accessibile a diversi attori, raccogliendo documenti sulle applicazioni data-intensive nel settore pubblico. Questo strumento aiuterà a colmare il divario informativo derivante dalla frammentazione delle piattaforme web, dalla mancanza di politiche di trasparenza e accountability (Stankovich *et al.,* 2023) e dal rumore generato dalla dispersione degli open data, spesso privi di metadati adeguati (Sadiq, Indulska, 2017). I contenuti sono stati selezionati attraverso parole chiave e acquisiti tramite web scraping dai principali siti ministeriali. Per garantire una consultazione efficiente, è stato sviluppato un modello di classificazione basato su BERT (Devlin *et al.,* 2019), capace di categorizzare autonomamente i documenti e renderli successivamente accessibili tramite un'interfaccia utente che fungerà da motore di ricerca.

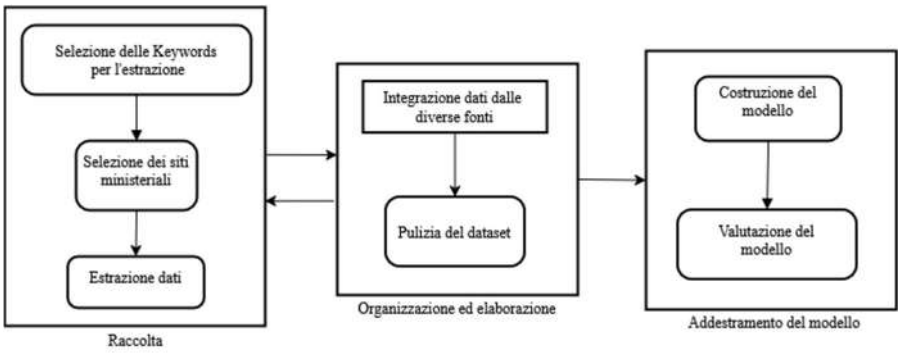
1. La creazione dello strumento

Il lavoro muove da due obiettivi: il primo è costruire un database interamente consultabile in cui sono presenti i contenuti dei siti ministeriali italiani riguardanti l'adozione di sistemi *data-intensive,* ossia sistemi che gestiscono ed elaborano grandi quantità di dati (terabyte e petabyte) generalmente complessi e distribuiti (Bonner *et al.,* 2017) – come open data e big data – e dedicano gran parte del loro tempo di elaborazione nelle fasi di input/output e manipolazione dati (Middleton, 2010). Il secondo obiettivo è addestrare un modello di Machine Learning (ML) che permetta, oltre ad una consultazione specifica per singole tematiche, anche la possibilità di inserire contenuti nuovi che saranno automaticamente riconosciuti dallo strumento in base alla loro rilevanza. Il fine ultimo è quindi di fornire uno strumento fruibile da cittadini ed esperti, attraverso il quale è possibile reperire informazioni aggiornate riguardo le applicazioni di sistemi data-intensive nella PA italiana. L'impiego del ML nel campo delle scienze sociali sta diventando un approccio consolidato per analizzare fenomeni complessi (Molina, Garip, 2019); in particolare, le reti neurali e i modelli di apprendimento supervisionato hanno dimostrato un grande potenziale nell'elaborazione di testi (Young *et al.,* 2018) e nella classificazione semantica. Nel nostro caso, il modello di ML è stato costruito seguendo tre fasi (figura 1):

1. *Raccolta:* che prevede la selezione delle keyword e la verifica della fattibilità dell'estrazione dai siti ministeriali.

2. *Organizzazione ed elaborazione*: dopo l'estrazione, il dataset è stato ripulito rimuovendo caratteri speciali e verificando la coerenza tematica con le keyword, integrando progressivamente i dati da diverse fonti.
3. *Addestramento del modello*: Il dataset è stato suddiviso in training e test set; quindi, il modello è stato valutato tramite metriche di accuratezza del ML.

Fig. 1 - Diagramma di costruzione dello strumento in fasi.



1.1. Fase 1 – Raccolta dati

Per raccogliere i dati dai siti ministeriali, il primo passo è stato definire una lista di parole chiave attinenti al tema, basata sul progetto Algorithm Tips, che fornisce una raccolta di keywords per individuare sistemi algoritmici negli USA. Questa lista è stata adattata e integrata per rispondere al contesto italiano. Dopo la selezione delle keywords (tab. 2), è stata analizzata la struttura dei siti ministeriali con dominio.gov.it e di altri siti istituzionali, come beniculturali.it, nei quali dovrebbero essere contenute le informazioni rilevanti sui sistemi data-intensive adottati dalla PA. Esaminata la struttura dei siti web individuati, sono emerse alcune problematiche comuni e condivise tra diversi siti. Per affrontare le difficoltà riscontrate, è utile descrivere la tecnica di estrazione dati adottata: la raccolta dei contenuti è avvenuta tramite web scraping, ovvero l'estrazione e il salvataggio strutturato delle informazioni visualizzate su pagine web (Lotfi *et al.,* 2021); in questa ricerca, è stato implementato un web scraping automatizzato in Python, utilizzando Selenium. L'estrazione ha riguardato i titoli e gli URL dei contenuti HTML,

evitando testo superfluo per ridurre il rumore (Agarwal *et al.,* 2007) e il rischio di errore. Selenium ha permesso di interagire direttamente con i siti, aggirando restrizioni come l'impossibilità di copiare URL specifici o l'assenza di operatori booleani.

Tuttavia, alcuni ostacoli hanno limitato l'efficacia del processo, tra cui elementi HTML nascosti, blocchi dei server, incoerenza nei risultati rispetto alle keyword e carenza di contenuti pertinenti. Queste problematiche, intrinseche alla struttura dei siti web ministeriali, hanno ridotto il numero di portali da 15 a 9. I siti sui quali è stato possibile raccogliere i dati sono:

- Ministero dei beni culturali;
- Ministero dell'interno;
- Ministero del lavoro;
- Ministero dell'agricoltura, della sovranità alimentare e delle foreste;
- Ministero dell'economia e delle finanze;
- Ministero delle imprese e del made in Italy;
- Ministero dell'istruzione e del merito;
- Ministero dell'Università e della Ricerca;
- Ministero del turismo.

1.2. Fase 2 – Organizzazione ed elaborazione

Attraverso il web scraping sui siti selezionati, l'uso delle keywords ha restituito 15.087 contenuti. Per ridurre la ridondanza, è stata applicata la lemmatizzazione delle keywords, considerando la limitata capacità di alcune piattaforme di riconoscerle solo nella loro forma base. Alcune di queste keywords presentano una scarsa autonomia semantica (Marradi, 1995: 34), risultando ambigue e poco valide per l'individuazione di fonti sui sistemi data-intensive. Per migliorare la pertinenza, è stata effettuata una selezione filtrando i circa 15.000 contenuti iniziali. In particolare, alle keywords "calcolo", "punteggio", "valutazione", "monitoraggio", "gradazione", "numeric" e "software" è stata associata almeno un'altra keyword tra quelle individuate, per raffinare i risultati. Le tabelle 1 e 2 mostrano il numero totale di contenuti inizialmente estratti e quelli rimanenti dopo la pulizia.

Tab. 1 – Frequenza dei documenti estratti prima e dopo la pulizia.

Ministero	*Frequenza (estrazione)*	*Frequenza (pulizia)*
Beni culturali	21	9
Interno	628	62
Lavoro	12568	1966

Masaf	250	10
Mef	71	37
Mimit	1158	183
Miur	203	122
Mur	176	63
Turismo	12	6
Totale	**15087**	**2458**

Tab. 2 – Frequenza delle Keywords prima e dopo la pulizia.

Keyword	*Frequenza (estrazione)*	*Frequenza (pulizia)*
algoritm	171	161
apache	17	17
automa	242	239
big data	53	51
biometric	17	17
calcolo	2815	82
classificazione	1210	36
cloud	228	227
data access	21	12
data capture	2	2
data governance	10	7
data lake	2	2
data mining	10	10
data privacy	11	9
data processing	16	16
data science	186	186
data source	5	2
data storage	4	4
data system	11	10
data warehouse	49	49
deep learning	4	3
ETL	6	6
GPT	3	3
graduazione	178	0
hadoop	2	2
intelligenza artificiale	173	173
internet of things	36	32
machine learning	23	19
metadat	87	84
monitoraggio	1747	26
neural network	7	7
numeric	56	54
open data	194	190
predittiv	109	102
punteggio	1190	29
rango	428	418
software	1290	42
statistica	45	45

valutazione	4429	84
Totale	**15087**	**2458**

È utile fare alcune considerazioni sulla struttura del dataset: la forte riduzione dei contenuti tra la prima e la seconda estrazione evidenzia la scarsa presenza di queste tematiche sui siti ministeriali, anche per keyword diffuse come "intelligenza artificiale", "algoritm*" e "big data". Quasi l'80% dei contenuti proviene dal Ministero del Lavoro. Le keyword inizialmente ritenute poco pertinenti hanno confermato la loro irrilevanza, diminuendo significativamente di frequenza. A ogni contenuto è stato assegnato un punteggio di 0 (non rilevante) o 1 (rilevante), creando una variabile per valutare la pertinenza dei temi. I contenuti classificati come rilevanti seguono i criteri indicati nelle tabelle 1 e 2. Successivamente, è stato effettuato un ulteriore controllo manuale, che ha portato a una lieve riduzione del numero complessivo di contenuti.

1.3. Fase 3 – Addestramento del modello

Per l'addestramento è stato scelto il modello BERT (Bidirectional Encoder Representation for Transformer) (Devlin *et al.,* 2019), che offre vantaggi significativi rispetto ai modelli NLP tradizionali. Grazie alla sua architettura basata su una rete neurale profonda, BERT elabora l'intero testo in input simultaneamente, migliorando l'identificazione delle relazioni tra parole e frasi (Bolasco, 2021). Per la classificazione del testo, BERT utilizza un transformer multistrato bidirezionale, che rappresenta il testo in un iperspazio a N dimensioni considerando l'intero contesto di ogni parola. Inoltre, il suo stato di preaddestramento su una vasta gamma di documenti gli consente di apprendere strutture linguistiche complesse e di essere ottimizzato per task specifici, come la classificazione binaria del testo. Dal dataset totale, composto da 15.087 osservazioni, sono state selezionate casualmente 2.458 rilevazioni non rilevanti, da affiancare alle 2.458 classificate come rilevanti, in modo da ottenere un campione bilanciato al 50% ed evitare problemi di oversampling. Il dataset è stato suddiviso in training e test set con un rapporto 80/20 (Gholamy *et al.,* 2018); l'addestramento, eseguito con la tecnica "1cycle" (Smith, Topin, 2018), è durato due epoche, sufficienti a garantire un buon livello di accuratezza senza rischi di overfitting.

Tab. 3 – Parametri di perdita e accuratezza del modello.

Epoche	*Loss*	*Accuracy*	*Val_loss*	*Val_accuracy*
1	0,37	0,83	0,19	0,95

2	0,1	0,96	0,08	0,97

Com'è possibile notare dalla tabella 3 infatti, il modello, alla fine della seconda epoca, presenta un'accuratezza di 0,96 nel caso del training set, e di 0,97 nel test set. Anche i valori di perdita risultano bassi (0,1 e 0,08). Per fornire un ulteriore *layer* di validazione al modello, può essere utile fare riferimento ad alcune misure più specifiche, ossia la *precision*[3], il *recall*[4] e la media armonica dei due, definita f*1- score.*

Tab. 4 – Parametri di validazione per rilevanza.

Rilevanza	*Precision*	*Recall*	*F1-score*
Non rilevante (0)	0,98	0,97	0,98
Rilvente (1)	0,97	0,98	0,98

Anche dalla tabella 4 possiamo notare come le misure scelte restituiscano valori molto alti, suggerendo una robustezza del modello elevata.

Conclusioni

Il lavoro contribuisce a colmare il divario informativo sull'uso dei sistemi data-intensive nella PA italiana, fornendo uno strumento innovativo per raccogliere e classificare documenti pubblici. L'efficacia del modello BERT è confermata dall'elevata accuratezza nella classificazione dei contenuti rilevanti. Tuttavia, alcune limitazioni rimangono, in particolare in riferimento alla ridotta quantità di dati disponibili nei siti ministeriali e la dipendenza dell'algoritmo dalla qualità dei documenti indicizzati. Non tutti i siti governativi sono adatti al web scraping; la bassa mole di contenuti disponibili sui siti ministeriali ha imposto una riduzione della lista iniziale di keywords per una maggiore focalizzazione su tematiche specifiche. Per migliorare il sistema, sarebbe utile ampliare il database con fonti aggiuntive, come Web of Science, e affinare la classificazione per maggiore precisione. Al tempo stesso, la selezione di keyword più mirate e la pulizia dei dati si sono rivelate fondamentali per garantire una raccolta più efficace. Nonostante questi limiti, lo strumento presenta importanti potenzialità applicative, sia in ambito accademico che istituzionale.

In prospettiva, questo strumento contribuisce anche a una più ampia riflessione sul ruolo che le tecnologie algoritmiche stanno assumendo nei processi di governance. La crescente datificazione delle pratiche amministrative e l'adozione di modelli predittivi e classificatori pongono infatti interrogativi

profondi sul piano della trasparenza, dell'equità e della trasformazione delle logiche decisionali pubbliche (Couldry, Mejias, 2019; Ananny, Crawford, 2018) considerando che i sistemi algoritmici non si limitano a razionalizzare i processi amministrativi, ma introducono una forma di razionalità performativa che tende a marginalizzare la dimensione discorsiva e deliberativa dell'azione pubblica (Visentin, 2018). Inoltre, in assenza di riflessioni critiche sui dati di partenza e sulle logiche di classificazione, l'adozione di strumenti di machine learning da parte della PA rischia di produrre effetti autoconfermativi: le istituzioni tendono a considerare legittimo e prioritario ciò che risulta nei dati – già distorti da condizioni strutturali diseguali – trascurando esigenze e fenomeni che sfuggono alla rappresentazione computazionale; rischiando così di rafforzare una visione parziale e diseguale della realtà sociale. In questo modo, gli algoritmi si configurano anche come prolungamento di alcune caratteristiche della burocrazia tradizionale, intensificando il controllo e inibendo ulteriormente la privacy e la partecipazione del cittadino (Visentin, 2018). È dunque fondamentale affiancare all'adozione tecnica degli algoritmi una riflessione critica su come vengono definiti e costruiti i criteri che regolano il funzionamento di questi strumenti automatizzati affinché non riproducano – o amplifichino – le disuguaglianze già presenti nel tessuto sociale.

Riferimenti bibliografici

Agarwal S., Godbole S., Punjani D., Roy S. (2007). How much noise is too much: A study in automatic text classification. *Seventh IEEE International Conference on Data Mining* (ICDM 2007) (pp. 3-12). New York: IEEE.

Bolasco S. (2021). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma: Carocci.

Bonner S., Kureshi I., Brennan J., Theodoropoulos G. (2017). Exploring the evolution of big data technologies. *Software architecture for big data and the cloud* (pp. 253-283). Cambridge (MA): Morgan Kaufmann.

Buolamwini J., Gebru T. (2018). Gender Shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (77-91). New York: PMLR.

Couldry N., Mejias U.A. (2019). *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford: Stanford University Press.

Decreto Legislativo 7 marzo 2005, n. 82. *Codice dell'Amministrazione Digitale*. Art. 22, commi 1 e 2.

Devlin J., Ming-Wei C., Lee K., Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://arxiv.org/abs/1810.04805
.

Gholamy A., Kreinovich V., Kosheleva O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *Departmental Technical Reports (CS)*, University of Texas at El Paso.

Kitchin R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

Lotfi C., Srinivasan S., Ertz M., Latrous I., Manjushree S. (2021). Web scraping techniques and applications: A literature review. *SCRS Conference Proceedings on Intelligent Systems* (pp. 381-394). New Delhi: SCRS.

Marradi A. (1995). *Metodologia delle scienze sociali*. Bologna: il Mulino.

Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR),* 54(6), 1-35. https://doi.org/10.1145/3457607.

Middleton A.M. (2010). Data-intensive technologies for cloud computing. *Handbook of Cloud Computing* (pp. 83-136). Boston (MA): Springer.

Molina M., Garip F. (2019). Machine learning for sociology. *Annual Review of Sociology*, 45(1), 27-45.

Presidenza del Consiglio dei Ministri (2022). *Programma Strategico per l'Intelligenza Artificiale 2022-2024*. Roma.

Sadiq S., Indulska M. (2017). Open data: Quality over quantity. *International Journal of Information Management*, 37(3), 150-154.

Smith L.N., Topin N. (2018). Super-convergence: Very fast training of neural networks using large learning rates. *arXiv.* https://arxiv.org/abs/1708.07120.

Stankovich M., Behrens E., Burchell J. (2023). Toward meaningful transparency and accountability of AI algorithms in public service delivery. *Government Information Quarterly*, early access.

Stark B., Stegmann D., Magin M., Jürgens P. (2020). Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse. *Algorithm Watch Report*, 26.

van Dijck J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197-208.

Veale M., Van Kleek M., Binns R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). New York: ACM.

Visentin C. (2018). Il potere razionale degli algoritmi tra burocrazie e nuovi idealtipi. *The Lab's Quarterly*, 20(3), 47-72.

Young T., Hazarika D., Poria S., Cambria E. (2018). Recent trends in deep learning-based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55-75.

Artificial intelligence and crime: exploring AI's role in law enforcement and criminal misuse

by *Vera Kopsaj*[*]

This article examines the dual role of artificial intelligence (AI) in crime control and criminal innovation. It asks how AI can be used to dismantle criminal networks while preventing its misuse. Combining sociological theory with empirical examples, it maps the current applications of AI in law enforcement and its exploitation by criminal actors. Particular attention is given to youth involvement in cybercrime, shaped by inequality, digital subcultures and easy access to illicit tools. The article also reflects on ethical and legal implications, highlighting the role of the EU Artificial Intelligence Act. It calls for governance that ensures AI supports justice without reinforcing inequalities.

Keywords: AI; cybercrime; youth deviance; law enforcement; regulation; sociology.

Intelligenza artificiale e criminalità: esplorare il ruolo dell'IA nelle forze dell'ordine e per fini criminali

Il contributo analizza il duplice ruolo dell'intelligenza artificiale (IA) nella prevenzione e nella facilitazione del crimine. Partendo da una prospettiva sociologica, esplora come l'IA sia impiegata tanto nelle attività investigative quanto da reti criminali. Un focus particolare è dedicato al coinvolgimento dei giovani nella criminalità digitale, influenzato da disuguaglianze sociali e infrastrutture tecnologiche. In chiusura, si riflette sulle implicazioni etiche e regolative, con riferimento all'AI Act dell'Unione Europea, sostenendo la necessità di una *governance* inclusiva e trasparente.

Parole chiave: IA; crimine digitale; devianza giovanile; regolazione; sociologia.

DOI: 10.5281/zenodo.17297517

[*] UniCamillus - Saint Camillus International University of Health and Medical Sciences. vera.kopsaj@unicamillus.org.

Introduction

Artificial Intelligence (AI) raises a fundamental question: how can it be exploited to dismantle criminal networks while preventing its misuse? This duality defines the scope of the analysis, which traces how AI supports law enforcement and, conversely, how it is exploited to expand criminal power and evade detection.

Adopting a sociological lens, the discussion addresses the tension between innovation and deviance, particularly among youth, drawing on classical and contemporary theories. Rather than applying them comprehensively, these perspectives serve as heuristic tools to interpret phenomena such as the accessibility of illicit platforms and the dynamics of informal learning within online subcultures (McGuire, Dowling, 2013).

Cybersecurity and cybercrime operate as interconnected forces: the former seeks to protect digital infrastructures (O'Reilly *et al.,* 2021), while the latter exploits them through fraud, ransomware and phishing (Europol, 2024b). Within this tension, AI emerges both as a tool for criminal activity and as a transformative resource for digital policing (Singer, Friedman, 2014).

The article is structured in three parts: (1) a theoretical overview of deviance in the context of digital transformation; (2) an empirical mapping of AI applications in crime and policing; and (3) a critical reflection on ethical and legal implications, with a focus on youth involvement.

1. The use of AI in law enforcement

Law enforcement agencies are increasingly leveraging AI to monitor, predict, and respond to criminal activity. AI algorithms detect phishing schemes, ransomware, and fraud, enabling real-time response to threats (SentinelOne, 2024). Predictive analytics use historical data to guide proactive policing (New Media, 2025). However, such approaches have raised concerns about racial bias and reinforcing systemic inequalities (Benjamin, 2019; O'Neil, 2017). AI-driven facial recognition has proven useful in identifying suspects and missing persons (Europol, 2024b). However, critics warn of racial bias and false positives, stressing the need for oversight (Singer, Friedman, 2014). Deepfakes facilitate fraud and misinformation, complicating detection (Mubarak *et al.,* 2023; Reuters, 2024).

AI also improves cybersecurity, but it enables more sophisticated attacks. Phishing scams and social engineering tactics generated by AI have become very convincing, leading to greater financial losses (Europol, 2024b; Howell, Burruss, 2020). AI-powered cyber threats adapt in real time, making detection and neutralisation increasingly difficult (O'Reilly *et al.,* 2021).

The following sociological matrix (Table 1) categorises AI applications according to legality and intention, revealing the blurred boundaries between protection and exploitation in contemporary digital governance.

Table 1 – Sociological Matrix of AI Use in Crime Control and Criminal Innovation

	Lawful Use	**Unlawful Use**
Preventive Intention	AI for fraud detection, human trafficking monitoring, early warning systems	Use of AI to automate disinformation under the guise of "security" apps (grey zone)
Exploitative Intention	Predictive policing reinforcing social profiling, surveillance of marginalised youth	Deepfake fraud, synthetic identity theft, Cybercrime-as-a-Service, AI-driven radicalisation

Author's elaboration

AI-driven digital forensics has revolutionised investigations, enabling faster data analysis and file decryption (O'Reilly *et al.,* 2021). However, concerns persist about the misuse of AI in digital surveillance and potential privacy violations (Solove, 2004; Ferguson, 2017; Zuboff, 2019).

In response to these and other emerging risks, the European Union adopted the Artificial Intelligence Act, which «entered into force on 1 August 2024, and will be fully applicable […] on 2 August 2026»[1] (Regulation 2024/1689)[2]. This regulation represents a significant legal advance in establishing harmonised rules for AI deployment. Yet, when viewed through a sociological lens, it reveals several technocratic limitations. While addressing risks such as manipulation, bias, and cybersecurity (pp. 21-22), it conceives of harms primarily in individualised

[1] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?utm

[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

and functionalist terms, paying insufficient attention to the structural inequalities that AI systems can reinforce, particularly at the intersection of age, class, race, and disability (p. 8, recital 29; p. 17, recital 54). The emphasis on transparency and human oversight (pp. 20-21, recitals 72-73) assumes that institutional trust can be restored through technical documentation and procedural safeguards. However, sociological evidence points out that trust is relational and collective and cannot be reduced to compliance alone. The absence of participatory mechanisms limits the meaningful inclusion of affected communities in shaping the use of AI.

In education, AI systems are rightly classified as high risk (p. 16, recital 56), but the regulation does not consider the cultural and social consequences of AI monitoring, such as pressures towards behavioural conformity or cyberbullying, which are notably absent from the text.

Finally, the regulatory framework places institutional power primarily in the hands of states and commercial providers, while civil society and academic actors are assigned limited, often advisory roles (p. 37, recital 148). This reinforces a governance model based on certification rather than allowing for a democratic debate on the social configuration of AI.

This limitation is particularly concerning in the context of law enforcement, where AI's transformative potential is matched by significant ethical risks. Issues such as racial bias, invasion of privacy, and flawed prosecutions demand robust regulatory safeguards to ensure transparency and accountability (Benjamin, 2019; European Commission, 2023). Rather than replacing human judgment, AI should be used to augment decision-making processes, recognising that human intuition and contextual understanding remain essential in the administration of justice (Zuboff, 2019).

2. Criminal exploitation of AI

AI supports law enforcement, but is increasingly exploited to bypass security, generate deepfakes, and automate phishing with realistic emails, voices, and messages (O'Reilly *et al.,* 2021; Mubarak *et al.,* 2023; Carpenter, 2024).

Terrorist groups also take advantage of AI-enabled tools and platforms. Encrypted messaging apps, decentralised networks, and social media are employed for propaganda, recruitment, and operational coordination, often with minimal traceability (Europol, 2024b). In more extreme cases, 3D-

printed weapon manuals generated or optimised by AI pose serious public safety risks. Europol also reports an increase in minors recruited through online radicalisation pathways, particularly on gamified or anonymous platforms (Europol, 2024b).

AI-enhanced cryptography further complicates the picture, as criminal organisations increasingly rely on it to obscure illicit transactions and evade surveillance (TRM, 2025; O'Reilly *et al.,* 2021). These systems adapt in real time, creating a cybersecurity arms race in which defensive and offensive capabilities escalate simultaneously. Synthetic identity theft has become more accessible through AI-generated credentials that bypass traditional verification, leading to large-scale financial fraud (FTC – Federal Trade Commission, 2022; Europol, 2024b). Chatbots and LLMs enhance "romance scams" via emotional manipulation (The Hacker News, 2025).

AI-generated emails and fake news bots mimic legitimacy, spreading disinformation during elections and crises, and eroding public trust (Mubarak *et al.,* 2023; BBC, 2023; Carpenter, 2024).

Table 2 summarises key criminal uses of AI, including technologies, examples, and social impacts.

Table 2 – Structural Conditions and Sociotechnical Flows of AI-Driven Criminality

Structural Conditions	Technological Mediators	Criminal Practices	Social Consequences
Economic inequality, marginalisation, youth unemployment	AI-generated synthetic identities, deepfake generators, encrypted platforms	Fraud, identity theft, disinformation campaigns	Loss of trust, financial harm, institutional erosion
Platform accessibility, low-cost AI tools	LLMs, bots, AI-based phishing engines	Social engineering, sentiment manipulation	Emotional exploitation, radicalisation pathways
Lack of digital literacy and regulation	Autonomous systems, obfuscated infrastructures	Organised crime operations, Cybercrime-as-a-Service	Legal ambiguity, enforcement challenges

Author's elaboration

Taken together, these developments suggest that AI-enabled crime is not merely an extension of traditional deviance, but a qualitative transformation of it. What does it mean when deviant behaviour is no longer reactive but pre-programmed and scalable, embedded into digital infrastructures? Theories of deviance must therefore evolve to consider not only individual actors, but also the socio-technical systems and institutional logics that facilitate and normalise illicit behaviour.

As AI becomes more sophisticated, the challenges for policymakers and law enforcement grow. The line between crime prevention and technological overreach is increasingly thin. A careful balance must be struck – one that enables the benefits of AI while minimising the opportunities for its misuse. Ultimately, the dual role of AI – as both a tool of control and a vector of disruption – underscores the need for transparent regulation, cross-sector collaboration, and sustained investment in ethical governance (Wall, 2024).

3. Youth and cybercrime: sociological perspectives

Young people are both victims and perpetrators of cybercrime. They are victims of online scams, fraud and child pornography (CSAM), but also of hacking, phishing and financial fraud. Understanding this dual role is essential for developing effective prevention strategies (Europol, 2024a).

AI-generated CSAM complicates victim identification, as criminals use social media and gaming platforms to exploit minors (Europol, 2024a). AI-driven phishing scams have also become more sophisticated, with fraudsters using large language models (LLM) to create convincing scam emails (Europol, 2024a).

The gamification of cybercrime further fuels youth involvement. Online forums normalise hacking and fraud, portraying them as technical skills rather than crimes (McGuire, Dowling, 2013). Cybercrime-as-a-Service (CaaS) lowers barriers to entry, making it easier for young people to engage in illegal digital activities (Holt, Bossler, 2016).

Young people's involvement in cybercrime should not be treated as a homogeneous phenomenon. Patterns of access, motivation and detection vary across social class and geographic regions, reflecting wider

inequalities in digital literacy and institutional surveillance (Hargittai, 2008; Robinson, 2009).

Educational and socio-economic factors determine the path to cybercrime. Digital crime can provide financial incentives and opportunities to develop technical skills. Without ethical guidance, young people move from low-risk cyber activities (e.g. cheating in games) to advanced hacking and fraud (Hutchings, Holt, 2017).

AI has accelerated cybercriminal tactics, but it has not created greed, violence or manipulation: these predate technology (FTC – Federal Trade Commission, 2022). Social media foster both connectivity and exploitation, offering new ways to harm individuals at scale and speed.

Merton's strain theory (1938) argues that individuals resort to deviance when legitimate means to achieve socially valued goals are blocked. However, the rise of Cybercrime-as-a-Service (CaaS) complicates this classical model. Here, deviance is no longer an exclusively individual adaptation to structural tensions, but is commercialised and platformised, turning digital crime into a scalable, service-based enterprise. This shift challenges Merton's emphasis on isolated actors and calls for a rethinking of deviance as distributed, networked and entrepreneurial. Structural inequalities, such as economic exclusion and social marginalisation, continue to fuel youth involvement in cybercrime (McGuire, 2018), but CaaS lowers barriers to entry, providing accessible pathways to illicit economies (Hutchings, Holt, 2017).

Tackling youth cybercrime therefore requires more than a technological deterrent, but a holistic, multi-layered strategy. Artificial intelligence-based security systems can support detection and enforcement, but are unable to address the social and structural causes of digital deviance. Meaningful prevention must include investments in education, economic opportunities and ethical awareness, fostering conditions that redirect young people's digital skills to legitimate and constructive ends (Wall, 2024).

Digital literacy is among the most effective strategies to prevent youth cybercrime. Teaching ethical implications of hacking and fraud reduces the risk of illegal involvement (McGuire, Dowling, 2013). To this end, schools and organisations should integrate cybersecurity training into their curricula, including ethical hacking programmes that enable young technology enthusiasts to develop their skills in a lawful and socially responsible manner (Li *et al.,* 2016). Equally important is the promotion of digital citizenship, which consists of cultivating an awareness of acceptable and ethical online behaviour. Research shows that such education helps prevent

cyberbullying and promotes positive digital norms (Yuniawati *et al.,* 2024; Wall, 2024).

In addition to education, socio-economic conditions play a key role. Many young offenders turn to cybercrime not out of malice, but as a response to limited opportunities and perceived financial need (Wall, 2024). Providing alternative pathways – through scholarships, internships and specialised training – can redirect digital talent away from deviance and towards legitimate innovation (Hutchings, Holt, 2017).

Public-private partnerships are also key to addressing the evolving tactics of cybercriminals. Collaborative efforts between governments, technology companies and law enforcement can lead to more nuanced and effective policies that focus not only on enforcement, but also on protection and rehabilitation. Such approaches are essential to avoid over-criminalisation of young people and to effectively address the broader cybercrime ecosystem (Europol, 2024b). Cybercrime is not only a technical or legal problem, but a social and economic challenge that requires a systemic and long-term response.

While Merton's (1938) strain theory offers a fundamental lens to understand the link between blocked opportunities and deviance, other sociological perspectives further enrich this analysis. According to social learning theory (Akers, 2017), young offenders acquire criminal behaviour through peer interaction, online communities and exposure to cybercrime culture. However, in the digital age, these "peers" are often anonymous actors on platforms such as Telegram or Discord, and the learning process is increasingly shaped by platform logic and algorithmic amplification. This suggests that social learning today is not purely interpersonal, but embedded in socio-technical systems that mediate recognition, status and repetition (McGuire, 2018).

The theory of routine activity (Cohen, Felson, 2003 [1979]) also provides food for thought, explaining how cybercrime flourishes when three conditions align: motivated offenders, suitable targets and the absence of capable protection. However, in the online context, these conditions are actively created, and are not just random. Platforms designed for anonymity or encrypted messaging serve both legitimate users and criminal actors, thus facilitating deviant practices while claiming neutrality. This suggests updating routine activity theory to account for platform design (Hutchings, Holt, 2017).

Understanding these sociological dimensions is crucial to developing effective and just prevention strategies. Rather than framing cybercriminal

behaviour as a matter of individual pathology, policymakers must recognise the structural forces – economic insecurity, digital inequalities and algorithmic incentives – that shape young people's trajectories towards cybercrime (Wall, 2024).

Ultimately, by combining sociological understanding with practical intervention, it is possible to construct more ethical and intelligent responses to youth cybercrime – approaches that prioritise not only law enforcement, but also prevention, inclusion and opportunity.

Conclusions

AI plays a paradoxical role in crime. On the one hand, it enhances law enforcement capabilities through tools such as fraud detection, predictive analysis and human trafficking surveillance. On the other, it is increasingly exploited by cybercriminals to boost activities such as deep-fake fraud, identity theft and encrypted communications. This duality underlines the urgent need to strike a balance between technological innovation and regulatory control.

Young people's involvement in cybercrime is increasing, driven by economic inequality, digital subcultures and low barriers to entry into illicit digital markets. While artificial intelligence accelerates the sophistication of criminal techniques, it also offers potential for preventive action, particularly through education, ethical training and early intervention.

The analysis showed that AI is not a neutral tool, but a social force embedded in institutional and power dynamics. Its implementation in criminal justice systems must therefore be guided by sound ethical frameworks, human oversight and transparent accountability mechanisms.

The European Union's Artificial Intelligence Act (EU Regulation 2024/1689) is an important step towards creating a shared legal framework for trustworthy AI. By addressing fundamental issues such as transparency, security and risk management, particularly for high-risk systems, the regulation provides a fundamental basis on which to build ethical governance. However, from a sociological perspective, legal compliance alone is not enough: crime prevention must be rooted in broader strategies that address inequality, marginalisation and social vulnerability.

The future of AI in crime prevention depends on transparent and inclusive governance and long-term investment in equity. Only by

addressing the roots of digital deviance can AI serve justice and resilience, expanding opportunity while ensuring responsible use.

References

Akers R. (2017 [2009]). *Social learning and social structure: A general theory of crime and deviance*. London: Routledge. ISBN 13: 978-1-4128-0999-3 (pbk).

BBC (2023). AI-powered bots spreading misinformation during elections and crises. *BBC News*. https://www.bbc.com/news/technology-51497800

Benjamin R. (2019). *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge and Medford: Polity Press.

Carpenter P. (2024). *FAIK: A Practical Guide to Living in a World of Deepfakes, Disinformation, and AI-Generated Deceptions*. New Jersey: John Wiley & Sons.

Cohen L.E., Felson M. (2003). Social change and crime rate trends: A routine activity approach. *Crime: Critical Concepts in Sociology*, 1, 316.

European Commission (2024). AI Act enters into force. https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en

Europol (2024a). *Internet Organised Crime Threat Assessment (IOCTA)*. Europol Report.

Europol (2024b). AI and policing. The benefits and challenges of artificial intelligence for law reinforcement. https://www.europol.europa.eu/cms/sites/default/files/documents/AI-and-policing.pdf

Ferguson A.G. (2017). The rise of big data policing: Surveillance, race, and the future of law enforcement. In: *The Rise of Big Data Policing*. New Yoirk: New York University Press.

FTC Federal Trade Commission (2022). Combatting Online Harms Through Innovation. *Federal Trade Commission Report*. https://www.ftc.gov/system/files/ftc_gov/pdf/Combatting%20Online%20Harms%20Through%20Innovation%3B%20Federal%20Trade%20Commission%20Report%20to%20Congress.pdf

Howell C.J., Burruss G.W. (2020). Datasets for analysis of cybercrime. *The Palgrave handbook of international cybercrime and cyberdeviance* (pp. 207-219). Cham: Springer International Publishing.

Hutchings A., Holt T.J. (2017). The online stolen data market: disruption and intervention approaches. *Global Crime*, 18(1): 11-30.

Li Carrie K.W., Holt T.J., Bossler A.M., May D.C. (2016). Examining the Mediating Effects of Social Learning on the Low Self-Control – Cyberbullying Relationship in a Youth Sample, *Deviant Behavior*, 37(2): 126-138. DOI: 10.1080/01639625.2014.1004023.

McGuire M., Dowling S. (2013). Cyber crime: A review of the evidence. *Summary of key findings and implications. Home Office Research report*, 75: 1-35.

Merton R.K. (1938). Social structure and anomie. *American Sociological Review*, 3(5): 672-682.

Mubarak R., Alsboui T., Alshaikh O., Inuwa-Dutse, I., Khan S., Parkinson S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *Ieee Access*, *11*: 144497-144529.

New Media (2025, Predictive Policing: Leveraging AI and Machine Learning for Crime Prevention. https://newmediacomm.com/predictive-policing-leveraging-ai-and-machine-learning-for-crime-prevention/?utm _source=chatgpt.com

O'Neil C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishing.

O'Reilly P.D., Rigopoulos K., Feldman L., Witte G. (2021). *2020 Cybersecurity and Privacy Annual Report*. https://nvlpubs.nist. gov/nistpubs/SpecialPublications /NIST.SP.800-214.pdf

Reuters (2024). *Report on deepfakes: what the Copyright Office found and what comes next in AI regulation*. https://www.reuters.com/le gal/legalindustry/report-deepfakes-what-copyright-office-found-what-comes-next-ai-regulation-2024-12-18/

SentinelOne (2024). AI Threat Detection in Cybersecurity. Retrieved from *https://www.sentinelone.com/cybersecurity-101/data-and-ai/ai-threat-detection/*

Singer P.W., Friedman A. (2014). *Cybersecurity and Cyberwar: What Everyone Needs to Know.* Oxford: Oxford University Press. ISBN 978-0-19-991811-9 (paperback).

Solove D. (2004). *The Digital Person: Technology and Privacy in the Information Age*. NY: New York University Press.

The Hacker News (2025). AI-Powered Social Engineering: Reinvented Threats. https://thehackernews.com/2025/02/ai-powered-socialengineer ing.html

TRM Labs, Transaction Risk Management (2025). The Rise of AI-Enabled Crime: Exploring the evolution, risks, and responses to AI-powered criminal enterprises. https://www.trmlabs.com/resources/reports/the-rise-of-ai-enabled-crime

Wall D.S. (2024). *Cybercrime: The transformation of crime in the information age*. New Jersey: John Wiley & Sons. ISBN-13: 978-1-5095-6313-5

Yuniawati E.I., Tiatri S., Beng J.T. (2024). Strengthening digital citizenship behavior to reduce cyberbullying through learning outcome mediation. *ENLIGHTEN: Jurnal Bimbingan Konseling Islam*, *7*(2): 82-108.

Zuboff S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York: edn. PublicAffairs.

Crime mapping. Rischi e benefici dell'uso dell'intelligenza artificiale nel contrasto e prevenzione al crimine, il caso XLAW

di *Alberto Pesce*[*]

Le crime mapping sono strumenti per esaminare aree urbane a rischio di reati. Il saggio esamina l'impiego dell'IA, approfondendo il sistema predittivo *XLAW*, che utilizza una nuova teoria: le "zone di caccia". Il "luogo", la sua alterazione e conoscenza da parte dell'*offender* sono determinanti nella scelta razionale di commettere il crimine. Attraverso interviste a detenuti e a membri delle forze dell'ordine, si esplorerà la creazione di queste aree criminogene, con dialoghi con il fondatore di *XLAW*, si indagherà come il sistema scopra schemi criminali. Alcuni risultati di questo programma, sperimentato in Italia, mostrano una netta diminuzione dei reati. Naturalmente si cercherà di riflettere su questioni etiche e morali legate a questi programmi di IA.

Parole chiave: IA; XLAW; zone di caccia; criminologia urbana; crime mapping; PredPol.

Crime mapping. Risks and benefits of the use of artificial intelligence in the fight against and prevention of crime. The *XLAW* case

Crime mapping is a tool for examining urban areas at risk of crime. The essay examines the use of AI, delving into the *XLAW* predictive system, which uses a new theory: "hunting zones". The "place", its alteration and knowledge by the offender are decisive in the rational choice to commit the crime. Through interviews with inmates and members of law enforcement, the creation of these criminogenic areas will be explored, with dialogues with the founder of *XLAW*, and how the system uncovers criminal schemes. Some results of this program, tested in Italy, show a clear decrease in crimes. Of course, we will try to reflect on ethical and moral issues related to these AI programs.

Keywords: AI; XLAW; hunting areas; urban criminology; crime mapping; PredPol.

DOI: 10.5281/zenodo.17297530

[*] Università Telematica Ecampus. alberto.pesce@uniecampus.it.

Alberto Pesce

Introduzione

L'IA ha rivoluzionato il settore della sicurezza pubblica, con nuovi sistemi per la prevenzione dei crimini, originando problemi derivanti dal potere discrezionale delle forze di polizia. I controlli, spesso, si dirigono contro emarginati ed esclusi, persone di "proprietà della polizia" (Fussey *et al.*, 2021). Categorie sulle quali si esercita, con successo, un potere superiore (Wacquant, 2006), a causa di stereotipi che generano strutture mentali cariche di senso, caratterizzando una realtà sociale distorta, consolidando le "liminarità" e i confini sociali.

I *software* producono, per specifici luoghi e reati, *crime mapping*, cartografie digitali utilizzate come strumento di analisi predittiva dei crimini. Creano delle zone ad alta concentrazione di eventi delittuosi, *hot-spot* (Bowers *et al.* 2004), riproducendo i *pattern* criminali, il *modus operandi* e le aree a rischio. Vengono analizzati dati provenienti da diverse fonti, come i dati storici sui crimini o i rapporti di polizia, per determinare dei rituali di interazioni criminosa o, meglio, dei "modelli criminali" (Kaufmann *et al.,* 2019), prendendo in esame il ruolo dell'*offender*, la vittima e l'ambiente.

Soprattutto l'ambiente diventa il protagonista delle analisi criminologiche, il luogo influenza in modo decisivo i crimini, fornendo una disposizione spaziale e una organizzazione razionale agli eventi delittuosi (Brantingham, P.J. *et al,.* 2017). Le potenzialità delle mappe criminali sono evidenti ma esistono, però, rilevanti criticità associate ai pregiudizi algoritmici (*bias*). Tali strumenti possono riprodurre modelli di discriminazione sociale, stereotipi e sistemi di razzializzazione (Pronzato, 2024), comportando un rafforzamento della sorveglianza in specifiche aree, aumentando le ingiustizie preesistenti (Wacquant, 2006). Le decisioni algoritmiche potrebbero non risultare neutrali, esasperando squilibri socioeconomici e le tensioni sociali (Pronzato, 2024).

Il presente lavoro si propone di osservare come le *crime mapping*, possano rappresentare uno strumento efficace di prevenzione e repressione dei reati nelle zone urbane, in particolare, si analizzerà il caso del *Programma XLAW*.

1. Quadro teorico

I modelli di controllo sociale sono ideati per ridurre le libertà personali, lo scopo è educare gli agenti sociali alle norme stabilite, contenendo chi devia da tali norme (Foucault, 1966), in particolare nelle città. Il "luogo"

urbano, carico di simboli e significati diventa l'oggetto privilegiato per le analisi sociali, la geometria della città disegna un'entità viva in continua evoluzione, che influisce e plasma l'ordine dei reati (Brantingham, P.J., 2017), spazio, struttura e senso definiscono un agire sociale criminale. L'ipotesi è che l'architettura delle zone urbane induca, per determinate proprietà, una "disposizione" alla esecuzione dei reati. L'*offender* sceglie razionalmente alcune aree (Matza 1964; Cohen, Felson, 1979), dei luoghi familiari e sicuri (Curtis-Ham *et al.,* 2025), dove attuare la propria azione criminale.

La struttura dei territori, l'arredo urbano e le caratteristiche ecologiche guidano i delitti, generando *hot-spot*, luoghi con una maggiore pressione criminale (Doyle, Gerell, 2025). Le proprietà, la predilezione della vittima "migliore", il sostegno territoriale di un capitale umano criminale, la scelta razionale dell'*offender* di commettere un reato e, soprattutto, la preferenza di un "luogo sicuro", non determinano, prese singolarmente, l'azione criminale[1] ma determinano un'interazione attiva tra questi fattori. Per chiarire le dinamiche delle azioni delittuose, si utilizzerà, come base speculativa di riferimento, oltre alla teoria della geometria del crimine[2], anche la nuova teoria delle "zone di caccia" (Lombardo, 2019).

L'area criminale si configura quale elemento centrale dell'analisi sociale (Bianchini, Sicurella, 2012): edifici, vie di fuga e l'arredo urbano sono decisivi nella commissione dei crimini (Bisi, 2014), fornendo un confort *zone* criminale e per tale motivo si esaminerà, con uno studio di caso, l'efficacia del programma *XLAW*.

L'*offender* si muove come un "predatore", costruisce spazi fisici, mentali e simbolici, definiti e ridefiniti dalla pratica criminale, nel "suo territorio" pianifica le modalità di controllo attraverso una ripetizione sistematica di pratiche delittuose. La conoscenza dei luoghi, le azioni di routine e le opportunità producono la "sua zona criminale", in cui si sente legittimato e dove l'agire è strategico e razionale[3]. Questi modelli teorici, complementari ma non identici, permettono di disaggregare i crimini dagli autori, delineando una traiettoria di interconnessioni e interazioni tra le dinamiche sociali, gli

[1] Per arredo urbano si intendono condizioni strutturali, mentre per sostegno territoriale si indica l'appoggio svolto da altri gruppi che garantiscono impunità all'*offender*.

[2] La geometria del crimine è una delle tre teorie fondamentali della criminologia ambientale: teoria dell'attività di routine, geometria del crimine e teoria della scelta razionale (Jeffery 1971, Brantingham P.J. *et al.,* 2017).

[3] Le zone di caccia sono spazi frequentati abitualmente, come luoghi vicino le abitazioni, e sono osservati e perfettamente conosciti con pratiche quotidiane di controllo. Sono spazi carichi di senso in cui l'*offender* definisce una vulnerabilità della vittima, sono spazi relazionali, dove il contesto sociale non impedisce, o sanziona, le azioni criminali.

agenti coinvolti e lo spazio. La geometria del crimine analizza la distribuzione spaziale dei reati (Jeffery 1971; Brantingham, P.J *et al*., 2017), mettendoli in relazione con altri luoghi specifici, come il luogo di residenza dei reati e la teoria dell'attività di routine e la teoria della scelta razionale esaminano la volontà dell'*offender* di compiere il crimine, la sua motivazione, la scelta della vittima e l'assenza di criticità.

La teoria delle "zone di caccia", invece, cerca di riflettere ulteriormente, ponendo la sua attenzione sullo spazio, che diventa, per le sue proprietà, contesto che condiziona l'agire criminale. I criminali trasformano il territorio urbano modellandolo per i loro bisogni, ma proprio la progettazione urbana diventa teoria socio-criminologica[4]. La *CPTED* si concentra su un disegno urbano capace di diminuire i crimini, attraverso nuovi modelli di sorveglianza, manutenzione e gestione dei luoghi (Crowe, 2013). Questi nuovi modelli di città potrebbero ridurre la "zone di caccia", eliminando una delle proprietà principali di questi luoghi sicuri per il crimine, il contesto ambientale che condiziona l'agire criminale.

La rilevazione empirica ha utilizzato un approccio qualitativo, infatti, non ci troveremo di fronte un campione strutturato e rappresentativo (Corbetta, 2014), ma un insieme empirico di riferimento che non ha la pretesa di definire alcuna universalità o di estendere la verità degli asserti al di fuori dei dati rilevati empiricamente (Nigris, 2003).

Per questo presupposto teorico si è utilizzato lo strumento dell'intervista semi strutturata ad una unità di analisi differenziata capace di rappresentare quelle "dinamiche ambiente/soggetti/crimini" che vogliamo analizzare. Si è interagito con il creatore del *software* predittivo dei crimini *XLAW,* Elia Lombardo, con quattro accademici universitari, con dieci esperti in sicurezza e con ventiquattro individui privati della libertà, in alcuni istituti penitenziari italiani, in periodi temporali diversi[5].

[4] CPTED, *Crime Prevention Through Environmental Design*, o teoria della prevenzione del criminale attraverso la progettazione ambientale.

[5] Le interviste coprono un arco temporale diverso, dodici detenuti sono stati intervistati nel carcere di Bollate nel 2016, dodici detenuti nel carcere di Rebibbia nel 2018, nell'abito di due ricerche che avevano come oggetto i modelli di neutralizzazione, selezionati in base alla tipologia di reato, cioè i reati di tipo predatorio. I dodici esperti di sicurezza sono stati ascoltati nel 2023, nell'ambito di una ricerca dipartimentale dell'Università Ecampus sulla tematica: Sicurezza e arredo urbano, scelti in base a conoscenze personali . Gli accademici e il creatore di XLAW sono stati ascoltati nel 2024 per la redazione del presente saggio e in previsione di una monografia. Le interviste sono state condotte faccia a faccia, per quanto riguarda i detenuti, mentre gli altri sono stati intervistati o di persona, o telefonicamente. Gli incontri sono durati circa un'ora per gli esperti di sicurezza e per gli accademici. Con i detenuti si sono

Proprio questi ultimi saranno un valore aggiunto alla ricerca e le storie o i "brandelli di vita criminale" (Ferrarotti, 1981), saranno essenziali per comprendere la scelta dei luoghi e le modalità di preferenza delle vittime, delineando come si strutturano le "zone di caccia" e la geometria criminale. Un limite di questo studio è, sicuramente, la ridotta dimensione dell'unità empirica di riferimento. Tuttavia, questo saggio, che potremmo considerare una ricerca pilota, è finalizzato a verificare se le teorie e le analisi presentate siano supportate e valide, per consentire l'avvio di un'indagine che includa un numero rilevante di interviste.

2. Brandelli di vita criminale urbana

I primi modelli di cartografia sociale iniziano nell'Ottocento, con una mappatura dei reati, (Bravo, 2014), per analizzare le dinamiche e le interazioni dei gruppi sociali che abitano alcune zone delle nascenti città industriali, queste ricerche sono all'origine degli studi sulle sottoclassi e sul sottoproletariato urbano. Quello che si voleva indagare era il disagio urbano, provando, con prospettive e scopi difformi, a identificare aree urbane disagiate (Chevalier, 1976). Un mondo "fluttuante", dove la marginalità e la delinquenza spesso si ibridano e vengono tollerate o temute. Le città diventano i nuovi tropici da analizzare e i gruppi sociali che le abitano le nuove tribù, con abitudini, traiettorie e disposizioni. La voglia di decodificare queste modalità di vita alimenta, nel Novecento, la nascita di scuole di indagine sociale che indagano le dinamiche, i comportamenti e i modelli di interazioni esistenti nella città[6].

L'architettura urbana delimita i luoghi di vita contrapposti a quelli di disagio o sofferenza, proprio i luoghi di afflizione, dove la libertà delle persone viene abolita, sono al confine fisico, geografico e simbolico della città (Sykes, 1958), per marcare quella distanza e distacco tra mondo legale e mondo criminale. I "luoghi" sociologici sono l'elemento principale delle analisi dei sistemi predittivi (Benbouzid, 2019), il dove assume la stessa importanza del chi e del quando, in un flusso dinamico di interazioni e di azioni. Queste proprietà sono indivisibili, separandole si avrà un risultato è un flusso incompleto di dati, portando ad analisi difettose o incomplete. I modelli

avuti due incontri da un ora ciascuno, mentre con il creatore di XLAW si sono avuto due incontri della durata di circa due ore.

[6] La Scuola di Chicago inizia lo studio sociologico della città (Giocoli, 2023), non come ambiente statico ma come un ecosistema fluido, che possiede una sua struttura reale e simbolica.

predittivi di crimini seriali hanno raggiunto, grazie all'IA nuovi sviluppi per le *crime mapping*, le analisi identificano una suddivisione geografica di azioni delittuose avvalendosi di indagini statistiche con *hot-spot*, oppure la probabilità di avvenimenti criminosi, determinando come i fattori sociali, strutturali, urbanistici o geografici influenzino la frequenza e l'intensità del reato (Brantingham, P.J *et al.*, 2017). In Italia un sistema di algoritmi che producano modelli georeferenziati di *crime mapping,* utilizzate dalle forze dell'ordine, non esiste e ogni corpo armato dello Stato utilizza, dei programmi, che si concentrano su alcune tipologie di reati. In questo senso:

> *Il nostro algoritmo controlla flussi economici anomali, di banche e i money transfert, indirizzati in zone che noi monitoriamo, spesso sono frutto di riciclaggio* (GdF).
>
> *Cerchiamo sui siti web oggetti d'arte, inseriti nella nostra banca dati, il sistema segnala alcuni oggetti sospetti, poi controllati da un operato* (CC).

Le frodi *on line* sono reati comuni, spesso non ci si limita a semplici truffe ma si cerca di carpire le identità per utilizzarle illecitamente e internet, come spazio virtuale ma tremendamente reale, è una area ecologica che produce reati (Matthew *et al.,* 2016).

«Nei siti di aste on line o di vendita di beni di consumo, i furti di identità sono comuni, senza una denuncia non possiamo intervenire» (Pol).

Controllo e modifica dell'architettura urbana sono elementi che caratterizzano, aiutano o producono l'azione criminale (Curtis-Ham *et al.,* 2025). Gli *hot-spot*, come le piazze di spaccio, sono punti caldi criminali, zone ad alta intensità delittuosa (Block R., Block C., 1995), difficili da bonificare per la gestione militare dell'area. Troviamo vedette, ronde e consegna della droga dietro porte sbarrate che impediscono alle forze dell'ordine di intervenire. Le aree urbane sono un modello di interazione dinamica tra criminali e territorio, che viene piegato alle esigenze delittuose:

> *Il sistema di controllo della criminalità è rigoroso, l'estraneo viene riconosciuto e parte l'allerta. Vedette e ronde controllano la zona, se si è del luogo e non si appare pericolosi, si viene controllati senza rischi, se si usa un cellulare per telefonare o scattare foto, si possono avere seri problemi* (CC).
>
> *Nelle zone controllate dalla criminalità l'arredo urbano è modificato per i loro bisogni, troviamo sbarre alle vie, i Vigili del Fuoco li chiamiamo per tagliare lucchetti* (Pol Locale).

> *La mancanza di illuminazione non rende meccanicamente una area rischiosa, ma i malviventi sfruttano l'oscurità per spostarsi dove le telecamere non possono operare. Si è osservato che sono proprio loro a danneggiare le luci per poter agire agevolmente, rendendo le aree insicure* (Pol).

La scelta della vittima non è occasionale, l'*offender* si muove come un predatore e le vittime sono, per lo più, le persone isolate, gli anziani, le donne e i disabili, oppure si sposta nei luoghi dove le vittime sono indifese (Cosseddu, 2009):

> *La scelta della vittima non è neutra, specifica la capacità criminale dell'offender, l'organizzazione di un crimine predatorio implica la preferenza di soggetti facili, il rischio, per quanto si pensi, deve essere ridotto al minimo* (CC).
>
> *Avevo bisogno di soldi, rapinavo le prostitute. Erano colpi sicuri, hanno tanti soldi e non reagiscono* (Det).

L'*offender* sceglie, non a caso, un territorio come zona privilegiata per commettere un crimine, a volte è vicino alla sua dimora (Curtis-Ham *et al*., 2025). Cerca una area sicura, conosciuta e con elementi architettonici e capitale umano che possano garantirgli la riuscita del colpo e la fuga:

> *Clonavo i bancomat, ne avevo duecento, la banca era vicino casa, scendevo e facevo il colpo, se avessi notato qualcosa che non andava sarei scappato nella mia abitazione* (Det).
>
> *La piazza dove facevo le rapine era sicura, vicino a un parco, a volte dopo la rapina scappavo inseguito, una volta entrato nel parco non mi prendevano più* (Det).

La struttura urbana deve essere modificata per adempiere al ruolo di "luogo criminale", luci, barriere architettoniche e l'area si trasformano in un territorio sicuro:

> *Abbiamo messo una sbarra all'inizio della via dove spacciavamo, neanche i Falchi con le moto potevano entrare* (Det).
>
> *Avevano montato le telecamere nella nostra zona, per controllare la viabilità, ci davano fastidio, in pochi giorni le abbiamo demolite, noi e non altri controlliamo tutto* (Det).

A volte l'*offender* sceglie di non compiere il reato, perché vi sono elementi che lo convincono che la sua azione potrebbe avere delle difficoltà di riuscita:

> *Nella mia zona a volte c'era troppa gente, intralciavano il mio lavoro, allora rimandavo* (Det).
>
> *Se c'erano troppi rischi evitato, una volta c'era un funerale con decine di persone, non sarei riuscito a scappare agevolmente* (Det).

3. *Crime mapping* e reati predatori, analisi di caso del programma predittivo *XLAW*

Il caso specifico che osserviamo è il programma predittivo *XLAW*. L'analisi di questo caso si concentrerà sulla speculazione teorica alla base del sistema e sul modo in cui la creazione di mappe criminali ha condotto a una riduzione dei tassi di criminalità e dei reati predatori, in alcune aree di varie città italiane, dove il software è stato sperimentato e dove si riscontravano livelli di criminalità preoccupanti, sia per incidenza che per regolarità. L'originalità di *XLAW* si fonda sulla creazione di una nuova teoria criminologica, "le zone di caccia" che è alla base del modello statistico di indagine dei reati (Lombardo, 2019). Il sistema localizza degli *hot-spot,* individua dei *pattern* di reati predatori (Block R., Block C., 1995)[7] e successivamente si genera una *crime mapping*, localizzando aree ad alta concentrazione di crimini, in un dato arco temporale e con quale incidenza. Contemporaneamente, *XLAW*, analizza i *modus operandi* degli offender, vittime, capitale umano e l'arredo urbano: aree di fuga, barriere architettoniche naturali o artificiali, non l*'offende*r ma il "luogo sicuro del crimine" è il protagonista di questo sistema predittivo (Curtis-Ham *et al.*, 2025)[8]. Il sistema non sviluppa dei *bias* informatici, non crea delle zone che vengono controllate con modelli di razzializzazione (Pronzato, 2024), perché non è un sistema statistico ma euristico, prevede "modelli criminali". In questo modo si vuole:

[7] I reati predatori sono furti, rapine o estorsioni.

[8] *«Il sistema crea una città virtuale, estrae gli eventi, reati, più attendibili per poter far circolare le forze dell'ordine in una determinata zona, in uno specifico giorno e ora, ha bisogno solo del 20% dei dati totali delle forze dell'ordine creando un modello di eventi ciclici»* (intervista Dott. Lombardo).

«Costruire modelli ciclici e stanziali di commissioni di reati, possiamo definirlo un vero modello criminologico, prende in esame alcuni elementi specifici, il dove, il quando, il modus operandi» (intervista Dott. Lombardo).

L'oggetto principale di questo sistema è il reato e il luogo dove viene compiuto, l'algoritmo genera modelli ogni trenta minuti fornendo degli *alert* per le future tre ore, indicando la strada dove la percentuale di commissione di un determinato reato è elevata. Il programma si basa su una nuova meditazione teorica, quella delle "zone di caccia" per i reati di tipo economico predatorio. I principi che spingono l'*offender* in un luogo possono essere oggettivi: il territorio con più vittime, oppure soggettiva, dimora in quella zona. Di solito la scelta del territorio è determinata dalla conoscenza di rifugi o vie di fuga, l'area dove compie il crimine non è sostituibile, una volta scelta quell'area diventa la sua "zona di caccia":

«Questi luoghi sono per il criminale infungibili, non sostituibili, in un'altra zona non avrebbe le stesse capacità, o sicurezza per operare quei reati, spostandosi perde l'efficacia e rischia l'arresto» (intervista Dott. Lombardo).

La *performance* è calibrata per ottenere il maggiore successo con il minimo sforzo:

«Lo scopo non è fare il colpo della vita, ma una serie di reati che garantiscano un reddito mensile, non ha un impulso necessario a compiere quel reato, perciò, opera anche una fase di ristrutturazione mentale e di azione, individua degli elementi che possono metterlo in pericolo rinunciando al reato» (intervista Dott. Lombardo).

La validità di *XLAW* è stata autenticata da un ente terzo, da valutatori dell'università Federico II e Partenope.

> *Non tutte le aree urbane sono soggette a condizione di focalizzazioni di reati, XLAW lo dimostra, si creano hot-spot, zone oggetto di consumo maggiore di azione predatoria, alcune aree diventano la zona di caccia di un criminale, fattore determinante è la conoscenza del luogo e di vie di fuga* (Valutatore).
>
> *Non ci siamo posti il problema di inserire i dati personali degli autori, perché il luogo e il tipo del reato sono le informazioni che servono per poter determinare le zone calde, la previsione temporale e spaziale del reato. Il sistema funziona perché identifica luoghi più adatti per i reati* (Valutatore).

Gli effetti della sperimentazione di *XLAW* dimostrano la sua efficacia più efficiente rispetto ai modelli tradizionali. Il sistema previsionale ha una alta

percentuale di precisione sui reati, oltre 80%[9], e si nota una diminuzione percentuale dei crimini nelle città dove il sistema è stato sperimentato, riducendo i delitti predatori. I risultati positivi si sono raggiunti anche nel settore del risparmio economico per la cittadinanza in base a effetti di prevenzione e impegno di risorse per le forze dell'ordine. È aumentata sia la percezione della sicurezza da parte dei cittadini e la fiducia verso gli operatori, migliorando anche la capacità decisionale e una migliore *performance* di questi ultimi. Si sono ridotti, poi, i costi per la collettività, grazie alla diminuzione dei reati.

Conclusioni

La realtà, anche quella criminale, è relazionale e il reato è un rapporto, continuo, tra il soggetto, vittima e il suo ambiente mentre l'azione criminale si attua costruendo un ambiente sicuro, "una riserva, o zona di caccia" dove compiere i reati. L'interazione avviene in primo luogo con il territorio, le aree delittuose sono cariche di senso, zone simboliche e reali, si trasformano in porti sicuri dove operano azioni illegali.

La scelta razionale è indissolubilmente connessa ad altre due proprietà: la prima è la scelta della vittima migliore. La seconda proprietà, più importante, è la conoscenza e la consapevolezza del contesto in cui si svolge l'azione criminosa. Come evidenziato dalle interviste, il luogo determina non solo la volontà, ma la probabilità di commettere un crimine, in relazione alla massima eventualità di conseguire un esito favorevole, l'*offender* si muove in una area sicura, una "zona di caccia" dove può svolgere al meglio la sua *performance.*

Esiste una geometria criminale, i reati vengono realizzati in aree in cui la distribuzione spaziale di obiettivi, che potremmo definire "facili", si sovrappongono alla familiarità del criminale con il luogo seguendo la riflessione del principio del minimo sforzo (Menting *et al.*, 2020). Come degli architetti, costruiscono un *habitat* delittuoso con trappole urbane, modificando l'ambiente ai propri scopi, oppure traggono maggiore profitto nelle alleanze criminali di alcuni abitanti delle zone interessate ai delitti, una rete sociale del crimine. Lo studio di caso del sistema *XLAW* ci presenta una nuova teoria sociologica e criminologica per spiegare perché determinati reati si concentrino in alcune specifiche aree urbane.

[9] Fonte e risultati completi su https://www.xlaw.it/presentazione/

I risultati della sperimentazione, in diverse città italiane, dimostrano la validità del modello di algoritmo che prende in esame il luogo e il tipo di reato. Il sistema non utilizza dati personali, evitando i *bias* (Pronzato, 2024), non si concentra solo su alcuni quartieri della città ma prende in esame solo le zone delle predizioni. La teoria delle zone di caccia può essere un valido ausilio per le riflessioni sociologiche e criminologiche future, ma anche come modello di analisi per la prevenzione dei reati.

Riferimenti bibliografici

Benbouzid B. (2019). To predict and to manage. Predictive policing in the United States. *Big Data & Society*, 6(1).

Bianchini E., Sicurella S. (2012). Progettazione dello spazio urbano e comportamenti criminosi. *Rivista di Criminologia, Vittimologia e Sicurezza*, 6(1): 74-90.

Bisi R. (2014). Scienze criminologiche e spazio urbano. *Rivista di Criminologia, Vittimologia e Sicurezza*, 8(1): 6-15.

Block R., Block C. (1995). Space, place, and crime: Hot spot areas and hot places of liquor-related crime. In: Eck J.E., Weisburd D. (a cura di), *Crime and Place*, Crime Prevention Studies, vol. 4. Monsey (NY): Criminal Justice Press: 145-184.

Bowers K.J., Johnson S.D., Pease K. (2004). Prospective hot-spotting: The future of crime mapping?. *The British Journal of Criminology*, 44(5): 641-658.

Brantingham P.J., Brantingham P.L., Andresen M.A. (2017). The geometry of crime and crime pattern theory. In: Wortley R., Townsley M. (a cura di), *Environmental Criminology and Crime Analysis*. New York: Routledge: 98-115.

Bravo F. (2014). L'efficacia del crime mapping per la sicurezza urbana: il caso di Enfield (Londra). *Rivista di Criminologia, Vittimologia e Sicurezza*, 8(1): 38-73.

Chevalier L. (1976). *Classi lavoratrici e classi pericolose. Parigi nella Rivoluzione industriale*. Roma-Bari: Laterza.

Cohen L.E., Felson M. (1979). Social change and crime rate trends: A routine activity approach, *American Sociological Review*, 44(4): 588-608.

Corbetta P. (2014). *Metodologia e tecniche della ricerca sociale*. Bologna: il Mulino.

Cosseddu M. (2009). *Criminalità predatoria. Rapine e furti in banca e in altri settori esposti: Poste, grande distribuzione, farmacie, trasporto valori*. Roma: Bancaria Editrice.

Crowe D.T. (2013). *Crime Prevention Through Environmental Design*. Waltham (MA): Elsevier.

Curtis-Ham S., Bernasco W., Medvedev O.N., Polaschek D.L.L. (2025). Familiar locations and similar activities: Examining the contributions of reliable and relevant knowledge in offenders' crime location choices. *International Criminal Justice Review*, 35(1): 9-28.

Doyle M.C., Gerell M. (2025). Assessing crime history as a predictor: Exploring hotspots of violent and property crime in Malmö, Sweden. *International Criminal Justice Review*, 35(1): 43-61.

Ferrarotti F. (1981). *Storia e storie di vita*. Roma-Bari: Laterza.

Foucault M. (1976). *Sorvegliare e punire. La nascita della prigione*. Torino: Einaudi.

Fussey P., Davies B., Innes M. (2021). 'Assisted' facial recognition and the reinvention of suspicion and discretion in digital policing. *The British Journal of Criminology*, 61(2): 325-344.

Giocoli N. (2023). *La scuola di Chicago*. Milano: IBL Libri.

Jeffery C.R. (1971). *Crime Prevention Through Environmental Design*. Beverly Hills (CA): Sage.

Kaufmann M., Egbert S., Leese M. (2019). Predictive policing and the politics of patterns. *The British Journal of Criminology*, 59(3): 674-692.

Lombardo E. (2019). *Sicurezza 4P. Lo studio alla base di XLAW per prevedere e prevenire i crimini predatori*. Venezia: Mazzanti.

Williams M.L. (2016). Guardians upon high: An application of routine activities theory to online identity theft in Europe at the country and individual level. *The British Journal of Criminology*, 56(1): 21-48.

Matza D. (1964). *Delinquency and Drift*. New York: John Wiley & Sons.

Menting B., Lammers M., Ruiter S., Bernasco W. (2020). The influence of activity space and visiting frequency on crime location choice: Findings from an online self-report survey, *The British Journal of Criminology*, 60(2): 303-322.

Nigris D. (2003). *Standard e non-standard nella ricerca sociale. Riflessioni metodologiche*. Milano: FrancoAngeli.

Pronzato R. (2024). *Algoritmi, strutture e agire sociale. Un'analisi sociologica*. Milano: FrancoAngeli.

Sykes G.M. (1958). *The Society of Captives. A Study of a Maximum-Security Prison*. Princeton: Princeton University Press.

Wacquant L. (2006). *Punire i poveri. Il nuovo governo dell'insicurezza sociale*. Roma: DeriveApprodi.

I tentativi di regolamentazione dell'intelligenza artificiale: tra approcci differenti, contraddizioni e implicazioni sociali in Europa, Stati Uniti e Cina

di *Giovannipaolo Ferrari*[*], Paolo Diana*[**]

L'intelligenza artificiale (IA) è una forza chiave della trasformazione digitale, con impatti globali su economia e società. Questo articolo analizza i modelli di governance dell'UE, USA e Cina, evidenziando non solo il bilanciamento tra innovazione e controllo sociale, ma anche le implicazioni geopolitiche ed economiche. Si esplorano le sfide per una governance globale, intrecciando tecnologia, diritti e giustizia sociale, con particolare attenzione alle contraddizioni normative e alle conseguenze per la competitività dei vari attori internazionali.

Parole chiave: intelligenza artificiale; regolamentazione internazionale; politiche pubbliche; controllo sociale; diritti digitali; governance tecnologica.

Attempts to regulate artificial intelligence: between different approaches, contradictions, and social implications in Europe, the United States, and China

Artificial intelligence (AI) is a key driver of digital transformation, with global impacts on the economy and society. This article analyses the governance models of the EU, US, and China, highlighting not only the balance between innovation and social control but also the geopolitical and economic implications. It explores the challenges of global AI governance, intertwining technology, rights, and social justice, with particular attention to regulatory contradictions and the consequences for the competitiveness of various international actors.

Keywords: artificial intelligence; international regulation; public policy; social control; digital rights; technological governance.

Introduzione

L'intelligenza artificiale (IA) è una delle tecnologie più trasformative del XXI secolo, con un impatto significativo su economia, società e governance globale. Tuttavia, la regolamentazione dell'IA è ancora in fase di sviluppo, con approcci differenti adottati dai principali attori politici mondiali. Le

DOI: 10.5281/zenodo.17297544

[*] Università degli Studi di Salerno, Université de Lille, CeRIES. giferrari@unisa.it.

[**] Università degli Studi di Salerno. diana@unisa.it.

strategie normative dell'Unione Europea (UE), degli Stati Uniti (USA) e della Cina riflettono non solo differenti priorità in termini di sicurezza, trasparenza e competitività economica, ma anche precise scelte politiche e strategiche che condizioneranno il futuro della tecnologia a livello globale.

L'IA, inoltre, sta emergendo come elemento centrale nei processi di governance del welfare, ridefinendo il ruolo dello Stato e delle istituzioni pubbliche nell'erogazione dei servizi sociali. In questo contesto, le tecnologie basate sull'IA sono sempre più integrate nelle amministrazioni pubbliche per migliorare l'efficienza, ma pongono anche questioni etiche e giuridiche legate alla trasparenza delle decisioni algoritmiche e all'impatto sociale dell'automazione (de Luca Picione *et al.*, 2023).

Il presente lavoro analizza e confronta i modelli regolatori dell'UE, degli USA e della Cina, evidenziandone i punti di convergenza e divergenza. L'approccio metodologico si basa su un'analisi qualitativa comparativa delle normative adottate nei tre contesti, supportata da fonti istituzionali e dalla letteratura di riferimento. Viene inoltre approfondita la dimensione geopolitica della regolamentazione dell'IA, con un'attenzione particolare all'impatto socioeconomico delle scelte normative e alle strategie globali delle grandi potenze tecnologiche. Il confronto è articolato in tre sezioni: descrizione dei tre modelli regolatori, analisi delle implicazioni sociali ed etiche della regolamentazione e prospettive di governance globale dell'IA.

Pur evitando deliberatamente un'impostazione ontologica sull'IA, il presente lavoro – com'è già stato specificato – adotta un approccio pragmatico centrato sulla sua regolamentazione. A questo proposito, è utile chiarire che le normative attualmente analizzate – in Europa, Stati Uniti e Cina – si riferiscono quasi esclusivamente a forme di intelligenza artificiale ristretta (*Narrow AI*), progettate per svolgere compiti specifici senza capacità generaliste o autonome. Il riferimento a forme di intelligenza artificiale generale (*General AI*), se presente, è solo potenziale o ipotetico. Questa distinzione, spesso non esplicitata nei testi normativi, risulta tuttavia fondamentale per interpretare correttamente il perimetro degli interventi regolatori considerati e comprenderne l'attuale portata applicativa (Sloane e Wüllhorst, 2025).

1. Regolamentazione dell'IA: modelli a confronto

1.1. L'approccio europeo: il modello normativo dell'AI Act

L'UE ha adottato un approccio normativo rigoroso per regolamentare l'IA, basato sulla protezione dei diritti fondamentali, la trasparenza e la sicurezza. Questo approccio si è concretizzato nell'*Artificial Intelligence Act* (AI Act), approvato nel 2024. Tale regolamento è il primo nel suo genere a livello mondiale e mira a garantire che i sistemi di IA siano utilizzati in modo sicuro e responsabile (Voigt, Hullen, 2024; Nikolinakos, 2023). Il quadro normativo europeo ha introdotto, inoltre, un meccanismo di conformità che prevede una sorveglianza costante sui sistemi di IA ad alto rischio, mirando a prevenire discriminazioni e violazioni dei diritti fondamentali della persona (Ferrari, Diana, 2024).

Il regolamento classifica i sistemi di IA in base al rischio, suddividendoli in quattro categorie: applicazioni vietate, come i sistemi di *social scoring* e la manipolazione comportamentale; sistemi ad alto rischio, utilizzati in settori critici come sanità, istruzione, occupazione e sicurezza pubblica, soggetti a rigorosi controlli di conformità; IA a rischio limitato, che richiedono trasparenza nei loro output; IA a rischio minimo, come chatbot e strumenti di automazione, non soggetti a particolari obblighi normativi.

Uno degli aspetti più innovativi dell'AI Act è l'introduzione delle "*sandbox* regolatorie", ambienti di sperimentazione controllati per lo sviluppo dell'IA, che consentono alle aziende di testare le proprie tecnologie sotto supervisione normativa (Dacev, Ibish, 2024). Inoltre, il nuovo quadro giuridico europeo cerca di trovare un equilibrio tra regolamentazione e innovazione, adottando un sistema multilivello che permette maggiore flessibilità per le PMI e le startup tecnologiche (Ferrari, Diana, 2024).

Dal punto di vista geopolitico, l'AI Act è parte di una strategia più ampia dell'UE per consolidare il proprio ruolo di leadership nella governance dell'IA, promuovendo standard normativi incentrati sulla protezione dei diritti fondamentali e sulla trasparenza, che possono essere adottati su scala globale (Nanni *et al.*, 2024). L'AI Act è la prima legislazione organica sull'IA a livello globale, e il suo impatto potrebbe estendersi oltre i confini europei attraverso il cosiddetto "effetto Bruxelles", ovvero l'influenza delle normative europee sugli standard internazionali (Voigt, Hullen, 2024).

1.2. L'approccio statunitense: mercato e autoregolamentazione

L'approccio degli USA alla regolamentazione dell'IA si distingue per un'impronta orientata al mercato e all'autoregolamentazione del settore privato. A differenza dell'UE, gli USA privilegiano la leadership tecnologica e l'innovazione, limitando l'intervento normativo federale.

Storicamente, la politica statunitense in materia di IA si è evoluta attraverso diverse amministrazioni. L'amministrazione Obama (2009-2017) ha promosso un modello di sviluppo basato sulla collaborazione tra settore pubblico e privato, con documenti come *Preparing for the Future of Artificial Intelligence* (2016), che enfatizzava la necessità di uno sviluppo responsabile dell'IA (Hine, Floridi, 2024). La prima amministrazione Trump (2017-2021) ha invece adottato una strategia più competitiva, riducendo la regolamentazione e puntando sulla leadership tecnologica statunitense (Shoaib *et al.*, 2024).

Con l'amministrazione Biden (2021-2025), si è assistito a un tentativo di regolamentazione più strutturata, pur senza imporre vincoli paragonabili a quelli dell'UE. Il *Blueprint for an AI Bill of Rights* (2022) ha introdotto principi di trasparenza e protezione dalla discriminazione algoritmica, ma senza vincoli legali.

Un aspetto centrale dell'approccio statunitense è il ruolo delle Big Tech, che influenzano direttamente la regolamentazione attraverso lobbying e standard di autoregolamentazione (Roberts *et al.*, 2024).

1.3. L'approccio cinese: regolamentazione statale e controllo sociale

L'approccio cinese alla regolamentazione dell'IA è caratterizzato da un forte intervento statale e da una stretta integrazione tra sviluppo tecnologico e governance politica (Zhang, 2024). Il governo cinese vede l'IA come uno strumento strategico per la crescita economica e il mantenimento della stabilità sociale (Wang *et al.*, 2025).

Dal 2017, il *Piano per la nuova generazione di IA* ha delineato una strategia per rendere la Cina leader globale nel settore entro il 2030 (Khanal *et al.*, 2025). La Cina, infatti, ha adottato un approccio normativo fortemente orientato al controllo statale, regolamentando non solo l'uso dell'IA ma anche il modo in cui gli algoritmi vengono sviluppati e implementati. Il Regolamento CAC n. 15/2023 impone restrizioni significative sui sistemi di IA generativa, obbligando le aziende a registrare i loro algoritmi e a conformarsi a stringenti requisiti di trasparenza e censura preventiva

(Clementi, 2024; Zhang, 2024). Il *Social Credit System* implementato in Cina, ad esempio, utilizza tecnologie avanzate integrando dati biometrici, finanziari e comportamentali per monitorare la condotta sociale dei cittadini e assegnare punteggi che influenzano l'accesso ai servizi pubblici e finanziari (Nanni *et al.*, 2024).

2. Implicazioni sociali ed etiche della regolamentazione

2.1. Innovazione vs. sicurezza: il dilemma regolatorio

L'IA rappresenta un'innovazione cruciale, ma il suo utilizzo solleva interrogativi sulla sicurezza e sulla protezione dei diritti fondamentali. La rigidità del modello UE potrebbe penalizzare la competitività dell'industria tecnologica europea rispetto ai modelli statunitense e cinese, che operano in contesti normativi più flessibili. Il modello USA, sebbene favorisca l'innovazione, solleva interrogativi sulla trasparenza e sulla responsabilità delle aziende, mentre il modello cinese di governance dell'IA desta preoccupazioni sulla libertà individuale e sulla protezione della privacy (Walter, 2024).

In questo contesto, il dibattito sulla regolamentazione dell'IA mette in evidenza il rischio di una "polarizzazione normativa", con l'UE che promuove un modello di regolamentazione stringente e la Cina che sfrutta l'IA per il controllo sociale. Questa contrapposizione solleva questioni critiche sui diritti digitali e sulla possibilità di sviluppare un quadro normativo globale condiviso (Ferrari, Diana, 2024).

Uno dei principali problemi sollevati dal dibattito sulla regolamentazione dell'IA è la tensione tra l'adozione di normative rigorose e il potenziale impatto sulla competitività economica. Ad esempio, le PMI europee potrebbero essere penalizzate dall'AI Act a causa dei costi elevati di conformità e la Francia e la Germania hanno proposto un modello di autoregolamentazione per i "*foundation models*", sostenendo che un eccessivo controllo normativo potrebbe danneggiare l'industria europea (Corsi, d'Albergo, 2024).

D'altro canto, un'assenza di regolamentazione potrebbe portare a conseguenze altrettanto problematiche, come l'uso dell'IA per la manipolazione dell'informazione o la discriminazione algoritmica (Ricart *et al.*, 2022). Negli USA, la mancanza di un quadro normativo federale unificato ha portato a una governance frammentata, con stati come la

California che hanno implementato regolamenti più stringenti sulla protezione dei dati rispetto ad altre giurisdizioni.

2.2. IA e diritti digitali: trasparenza, responsabilità e privacy

La crescente integrazione dell'IA in processi decisionali automatizzati solleva questioni etiche e normative sulla trasparenza, sulla responsabilità e sulla protezione della privacy (Viehoff, 2022). Uno degli aspetti più critici riguarda la cosiddetta "*black box*" algoritmica, ovvero l'opacità nei processi decisionali delle reti neurali e dei modelli di *deep learning* (European Parliament, 2023a).

Il principio della responsabilità impone che gli sviluppatori e gli utilizzatori di sistemi di IA siano responsabili delle conseguenze delle decisioni algoritmiche. L'AI Act stabilisce criteri di trasparenza e tracciabilità, ma negli USA la regolamentazione è meno vincolante, affidandosi spesso alle linee guida delle singole aziende (Jungherr, 2023).

Per quanto riguarda la protezione della privacy, l'UE ha adottato il Regolamento Generale sulla Protezione dei Dati (GDPR)**,** Regolamento UE 2016/679 come quadro normativo di riferimento, imponendo restrizioni all'uso dei dati personali e garantendo il diritto alla trasparenza delle decisioni automatizzate (European Parliament, 2023b). Tuttavia, l'applicazione del GDPR ai sistemi di IA rimane problematica, in quanto molti modelli operano su dati anonimi o aggregati, il che complica la valutazione delle violazioni della privacy (Duberry, 2022).

La Cina non ha una legge specifica sull'IA, ma regola l'uso di questa tecnologia attraverso normative sulla sicurezza informatica e la protezione dei dati. Il concetto chiave che guida la regolamentazione è il "*vital public interest*", che giustifica un'ampia raccolta e utilizzo di dati personali da parte dello Stato. Questo principio consente al governo di giustificare forme di sorveglianza avanzata, che sarebbero impensabili in contesti occidentali (Malaschini, 2021).

Il governo cinese, infatti, ha implementato normative rigorose sulla gestione degli algoritmi e sulla protezione dei dati, ma il controllo statale sulle informazioni personali solleva preoccupazioni sul diritto alla privacy (Zhang, 2024).

2.3. Impatti sul lavoro, sulla democrazia e sul controllo sociale

L'IA sta trasformando profondamente anche il mondo del lavoro, la governance democratica e i meccanismi di controllo sociale (Moore, Woodcock, 2021). L'uso crescente dell'IA nei processi di automazione sta ridisegnando il mercato del lavoro, portando alla creazione di nuove professioni ma anche alla scomparsa di mansioni tradizionali (Jeffares, 2021).

In settori come la *gig economy*, l'IA viene utilizzata per determinare turni di lavoro, compensi e prestazioni, riducendo la negoziazione tra lavoratori e datori di lavoro. Questa trasformazione porta alla cosiddetta "eteromazione", l'opposto dell'automazione che fa riferimento al microlavoro sottopagato a chiamata, che svolge attività di simulazione dell'IA e in cui i lavoratori umani diventano dipendenti dagli algoritmi per l'assegnazione delle attività stesse (Ekbia, Nardi, 2017).

Dal punto di vista democratico, la moderazione algoritmica dei contenuti nei social media sta sollevando interrogativi sul ruolo delle piattaforme digitali nella regolamentazione della libertà di espressione (Zuboff, 2019). Se da un lato l'IA aiuta a contrastare la disinformazione, dall'altro può introdurre forme di censura automatizzata che influenzano il dibattito pubblico (Lohmann, 2024).

Infine, l'uso dell'IA nella sorveglianza di massa è una delle questioni più controverse della regolamentazione cinese. Il riconoscimento facciale e l'analisi predittiva sono utilizzati per monitorare la popolazione e prevenire crimini, ma sollevano serie preoccupazioni etiche (Nanni *et al.*, 2024).

Tabella 1 – Tabella riassuntiva delle caratteristiche dei tre modelli regolatori.

Caratteristiche	*UE*	*USA*	*Cina*
Obiettivo principale	Protezione dei diritti fondamentali e trasparenza	Promozione dell'innovazione e leadership tecnologica	Controllo sociale e sviluppo tecnologico nazionale
Approccio normativo	Regolamentazione basata sul rischio (AI Act)	Autoregolamentazione e normative settoriali	Regolamentazione statale e sovranità digitale
Livello di intervento statale	Elevato: norme stringenti per garantire la sicurezza	Moderato: approccio frammentato tra stati e settori	Molto elevato: il governo dirige lo sviluppo e l'uso dell'IA

Ruolo del settore privato	Limitato: le aziende devono conformarsi a rigide normative	Predominante: Big Tech influenzano direttamente la regolamentazione	Limitato: le aziende operano sotto stretta supervisione statale
Controllo sull'innovazione	Elevato: regolamentazione dettagliata su tutte le applicazioni IA	Basso: poche restrizioni per favorire la competitività	Elevato: regolamentazione preventiva sugli algoritmi
Trasparenza e responsabilità	Forte attenzione alla tracciabilità e alla trasparenza degli algoritmi	Dipende dal settore: alcuni ambiti regolamentati (es. sanità, finanza)	Bassa trasparenza, con forte controllo statale sui dati
Impatto sui diritti digitali	Alta protezione della privacy e dei dati personali (GDPR)	Minore attenzione alla privacy rispetto all'UE, ma crescente dibattito	Limitata: sorveglianza diffusa e uso dell'IA per il controllo sociale
Influenza internazionale	Elevata: AI Act potrebbe diventare uno standard globale ('effetto Bruxelles')	Moderata: influenza tecnologica forte, ma senza standard vincolanti	Crescente: promozione di standard alternativi alla regolamentazione occidentale

3. Verso una governance globale dell'IA?

3.1. Convergenze e divergenze tra UE, USA e Cina

L'IA è diventata un campo strategico per la competizione globale tra UE, USA e Cina e l'assenza di un quadro normativo internazionale unificato sta portando alla creazione di blocchi regolatori distinti, con conseguenze significative per il commercio tecnologico e l'innovazione.

Gli USA, pur mantenendo una governance frammentata e incentrata sull'autoregolamentazione del settore privato, stanno progressivamente riconoscendo la necessità di un quadro normativo più strutturato. Questo spostamento è in parte motivato dalla crescente preoccupazione per il dominio delle Big Tech e dalle implicazioni dell'IA per la sicurezza

nazionale. La strategia statunitense riflette un compromesso tra protezione dei consumatori e mantenimento della leadership tecnologica, evitando una regolamentazione stringente che potrebbe limitare gli investimenti e l'innovazione (Shoaib *et al.*, 2024).

La Cina, invece, adotta un modello basato sul controllo statale e sulla sovranità digitale, utilizzando l'IA sia per promuovere l'innovazione industriale sia per rafforzare la sicurezza e il controllo sociale (Zhang, 2024). Questa strategia, se da un lato garantisce uno sviluppo rapido e coordinato delle tecnologie di IA, dall'altro solleva preoccupazioni sulle libertà individuali e sulla censura algoritmica. L'approccio cinese pone anche interrogativi sulla futura frammentazione del cyberspazio, con il rischio di una "bipolarizzazione" della governance dell'IA tra l'Occidente e la Cina (Wang *et al.*, 2025).

Nonostante le differenze, esistono alcuni punti di convergenza tra questi tre blocchi. Tutti riconoscono la necessità di regolamentare le applicazioni più pericolose dell'IA, come il riconoscimento facciale per la sorveglianza di massa o le decisioni automatizzate in ambito giudiziario (Roberts *et al.*, 2024). Inoltre, sia l'UE che la Cina hanno introdotto normative per limitare il potere delle grandi piattaforme tecnologiche, mentre negli USA si sta aprendo un dibattito sulla regolamentazione delle Big Tech e sulla necessità di tutelare la concorrenza (Walter, 2024).

L'assenza di un quadro normativo globale unitario rischia di creare un mercato frammentato, in cui le aziende tecnologiche devono adattarsi a regolamenti diversi a seconda del contesto geografico. Questa frammentazione potrebbe tradursi in una competizione normativa tra l'UE, gli USA e la Cina, con il rischio che i Paesi più piccoli adottino strategie ibride per attrarre investimenti e innovazione.

Un altro rischio è che i tre blocchi utilizzino la regolamentazione dell'IA come uno strumento geopolitico. Ad esempio, gli USA e l'UE hanno imposto restrizioni all'esportazione di semiconduttori avanzati verso la Cina, limitando l'accesso del Paese alle tecnologie necessarie per sviluppare modelli di IA di ultima generazione. Questo potrebbe portare a una crescente competizione tra blocchi economici per il dominio del settore dell'IA, con effetti simili a quelli della corsa agli armamenti durante la Guerra Fredda.

3.2. Il ruolo delle istituzioni internazionali

La governance globale dell'IA è una sfida complessa, che richiede il coordinamento tra istituzioni internazionali, governi nazionali e attori

privati. Attualmente, non esiste un organismo internazionale dedicato esclusivamente alla regolamentazione dell'IA, ma diverse istituzioni stanno cercando di colmare questo vuoto (Roberts *et al.*, 2024).

L'Organizzazione delle Nazioni Unite (ONU) ha proposto la creazione di un'agenzia dedicata alla governance dell'IA, simile all'International Atomic Energy Agency (IAEA), con il compito di monitorare l'uso dell'IA su scala globale e garantire il rispetto di principi etici condivisi (Hine, Floridi, 2024). Tuttavia, la realizzazione di un simile organismo incontra ostacoli significativi, a partire dalla mancanza di consenso tra le grandi potenze, che preferiscono mantenere il controllo sulla propria regolamentazione nazionale.

D'altro canto, l'Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE) – che riveste un ruolo centrale, politico e scientifico, nel processo di integrazione dei mercati e nella promozione di obiettivi socioeconomici comuni – sta sviluppando linee guida per un uso responsabile dell'IA, proponendo un quadro di governance basato su principi di trasparenza, equità e responsabilità (Floridi, 2023). Nonostante ciò, il suo impatto normativo rimane essenzialmente consultivo, perché l'OCSE non ha strumenti coercitivi per far rispettare questi principi e la loro implementazione dipende dalla volontà politica dei singoli Stati (Roberts *et al.*, 2024).

Anche il G7 ha avviato un'iniziativa per coordinare le normative sull'IA tra le principali economie avanzate: l'*Hiroshima AI Process*, lanciato nel 2023, mira a definire standard comuni per la governance dell'IA e a promuovere la cooperazione tra governi e aziende tecnologiche. Nonostante ciò, il G7 non include attori chiave come la Cina, il che riduce la sua efficacia nel costruire un consenso globale sulle normative per l'IA.

Un'altra sfida è rappresentata dalla frammentazione delle iniziative internazionali. Oltre all'ONU e all'OCSE, istituzioni come l'International Telecommunication Union (ITU), un'agenzia specializzata dell'ONU che si occupa di standard per le telecomunicazioni e la connettività, e l'Institute of Electrical and Electronics Engineers (IEEE), un'organizzazione professionale mondiale di ingegneri e ricercatori, la cui influenza sugli standard tecnologici è notevole, soprattutto in ambito tecnico-informatico, stanno sviluppando proprie linee guida e standard tecnici per l'IA (Hine, Floridi, 2024).

Questa proliferazione di iniziative rischia di creare sovrapposizioni e conflitti normativi, complicando ulteriormente il quadro della governance globale dell'IA.

Per affrontare queste sfide, alcuni esperti propongono l'adozione di un modello di "regime complesso", che faccia riferimento a un insieme di norme, attori e istituzioni che operano su livelli diversi, ma in maniera funzionalmente interconnessa. Questo approccio consente una governance flessibile e multilivello, utile per affrontare problemi transnazionali in assenza di un'autorità unica centralizzata. In un "regime complesso" molteplici e differenti istituzioni collaborano in modo coordinato senza la necessità di creare un'unica entità centrale per la regolamentazione dell'IA.

Questo approccio potrebbe offrire maggiore flessibilità, permettendo ai diversi attori di adattarsi all'evoluzione tecnologica senza vincoli eccessivamente rigidi.

In definitiva, il ruolo delle istituzioni internazionali sarà cruciale per evitare una corsa normativa frammentata che possa ostacolare la cooperazione globale e limitare le opportunità offerte dall'IA. Tuttavia, il successo della governance internazionale dipenderà dalla capacità di superare le divisioni geopolitiche e di sviluppare meccanismi di coordinamento efficaci.

Conclusioni

L'analisi della regolamentazione dell'IA nei tre principali blocchi evidenzia approcci distinti, influenzati da scelte strategiche e politiche diverse. La possibilità di sviluppare un framework globale condiviso rimane incerta, poiché la regolamentazione dell'IA è ormai diventata una leva geopolitica.

Dal punto di vista delle implicazioni sociali ed etiche, la regolamentazione dell'IA dovrà affrontare questioni chiave come la trasparenza algoritmica, la protezione della privacy e la gestione dell'impatto occupazionale dell'automazione. Inoltre, la crescente influenza delle Big Tech sulle decisioni normative pone il problema di come garantire una governance democratica e inclusiva del progresso tecnologico.

In questo quadro, emerge una tensione strutturale tra la natura globale del mercato tecnologico e la frammentazione delle culture normative locali. La difficoltà di conciliare questi due livelli si manifesta sia nelle differenze di approccio tra le grandi potenze economiche, sia nella resistenza – spesso legittima – di contesti locali a modelli regolatori percepiti come estranei o inadeguati ai propri valori e tradizioni identitarie, culturali e sociali. Tale conflittualità, già osservabile in altri ambiti (come la fiscalità digitale o la

gestione dei dati), assume nel caso dell'IA un significato particolarmente rilevante per l'impatto pervasivo della tecnologia.

Alla luce di quanto emerso nel confronto tra le iniziative internazionali, il futuro della regolamentazione globale dell'IA non passerà attraverso un'unica autorità centralizzata, ma si concretizzerà in un regime complesso e multilivello, fondato sulla collaborazione tra istituzioni internazionali, enti tecnici e attori privati. L'affermazione di standard tecnici riconosciuti a livello globale potrebbe rappresentare una condizione necessaria per costruire normative efficaci, legittimate e implementabili a livello nazionale.

Riferimenti bibliografici

Clementi F. (2024). Generare e non creare? Spunti per una comparazione sulla regolazione dell'intelligenza artificiale generativa tra Stati Uniti, Repubblica Popolare Cinese e Unione Europea. *Rivista di Diritti Comparati*, 2, 123-145.

Corsi R., d'Albergo E. (2024). La politica dell'intelligenza artificiale general purpose: immaginari socio-tecnici, democrazia e policy frame nel processo decisionale della regolazione europea ("EU AI ACT" – 2022-2024). *Im@go Journal*, 23(1): 109-130.

Dacev N., Ibish M. (2024). Legal Arrangements of Artificial Intelligence in the European Union and the Republic of North Macedonia. *European Journal of Privacy Law e Technology*, 1(1): 79-97. DOI: 10.57230/EJPLT241DNIM.

Duberry J. (2022). *Artificial Intelligence and Democracy: Risks and Promises of AI-Mediated Citizen–Government Relations.* Cheltenham: Edward Elgar Publishing.

de Luca Picione G. L., Diana P., Ferrari G., Fortini L. e Trezza D. (2023). IA Generativa nel welfare: un approccio basato sulla Sociologia Pubblica per una governance consapevole. *Cambio: rivista sulle trasformazioni sociali*, 26(2): 117-139.

European Parliament (2023a). *Artificial Intelligence: Risks, Ethics, and Regulations. European Parliamentary Research Service.*

European Parliament (2023b). *Artificial Intelligence Act: A step toward transparency and accountability in AI governance.* Disponibile online: https://artificialintelligenceact.eu.

Ferrari G., Diana P. (2024). Intelligenza Artificiale e mutamento sociale. Opportunità e sfide per la Pubblica Amministrazione. *Sociologia. Rivista quadrimestrale di scienze storiche e sociali,* 58: 48-54.

Floridi L. (2023). *The Ethics of Artificial Intelligence: Principles and Policies.* Oxford: Oxford University Press.

Hine C., Floridi L. (2024). Artificial intelligence with American values and Chinese characteristics: A comparative analysis of American and Chinese governmental AI policies. *AI e Society*, 39(1): 257-278. DOI: 10.1007/s00146-022-01499-8.

Jeffares S. (2021). *The Virtual Public Servant: Artificial Intelligence and Frontline Work.* Cham: Palgrave Macmillan. DOI: 10.1007/978-3-030-54084-5.

Jungherr A. (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society*, *9*(3). DOI: 10.1177/20563051231186353.

Khanal S., Zhang H., Taeihagh A. (2025). Development of New Generation of Artificial Intelligence in China: When Beijing's Global Ambitions Meet Local Realities. *Journal of Contemporary China*, 34(151): 19-42. DOI: 10.1080/10670564.2024.2333492.

Lohmann L. (2024). Labour, Energy and the Colonial Geography of Artificial Intelligence. *The Corner House.*

Malaschini A. (2021). *Regolare l'intelligenza artificiale. Le risposte di Cina, Stati Uniti, Unione Europea, Regno Unito, Russia e Italia*. In: P. Severino, *Intelligenza artificiale: politica, economia, diritto, tecnologia* (104-181), Roma: Luiss University Press.

Moore P.V., Woodcock J. (2021). *Augmented Exploitation: Artificial Intelligence, Automation and Work.* Londra: Pluto Press.

Nanni R., Bizzaro P.G., Napolitano M. (2024). The false promise of individual digital sovereignty in Europe: Comparing artificial intelligence and data regulations in China and the European Union. *Policy e Internet*, 16(4): 711-726. DOI: 10.1002/poi3.424.

Nikolinakos N.T. (2023). *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies - The AI Act.* Cham: Springer. DOI: 10.1007/978-3-031-27953-9.

Ricart R.J., Van Roy V., Rossetti F., Tangi L. (2022). *AI Watch - National strategies on Artificial Intelligence: A European perspective, 2022 edition.* Lussemburgo: Publications Office of the European Union. DOI: 10.2760/303073.

Roberts H., Hine E., Taddeo M., Floridi L. (2024). Global AI governance: barriers and pathways forward. *International Affairs*, 100(3): 1275-1286. DOI: 10.1093/ia/iiae073.

Sloane M., Wüllhorst E. (2025). A systematic review of regulatory strategies and transparency mandates in AI regulation in Europe, the United States, and Canada. *Data & Policy*, 7: e11(1-21). DOI:10.1017/dap.2024.54.

Walter Y. (2024). Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation – A contemporary overview and an analysis of socioeconomic consequences. *Discover Artificial Intelligence*, 4(14): 1-20. DOI: 10.1007/s44163-024-00109-4.

Wang K., Dong K., Wu J. e Wu J. (2025). Patterns of artificial intelligence policies in China: A nationwide perspective. *Library Hi Tech*, 43(1): 295-325. DOI: 10.1108/LHT-04-2022-0168

Viehoff J. (2022). Bias, fairness, and transparency in AI decision-making: Ethical concerns and regulatory responses. *AI e Society*, 37(4): 789-806. DOI: 10.1007/s00146-022-01499-8

Voigt P. e Hullen N. (2024). *The EU AI Act: Answers to Frequently Asked Questions.* Cham: Springer. DOI: 10.1007/978-3-662-70201-7

Zhang A. H. (2024). The promise and perils of China's regulation of artificial intelligence. *Columbia Journal of Transnational Law*, 63(1): 102-157.

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*. Roma: Luiss University Press.

Malicious use of Artificial Intelligence (MUAI): l'uso malevolo dell'intelligenza artificiale nell'ecosistema cyber-sociale

di *Arije Antinori*[*]

La nascita ed evoluzione dell'Uso Malevolo dell'Intelligenza Artificiale (MUAI) rappresenta un cambio paradigmatico nell'ambito dei comportamenti cyberdevianti e cybercriminali, in termini di complessità, sofisticazione e velocità di diffusione tra le diverse categorie di attori dell'ecosistema cyber-sociale. L'adozione di strumenti e tecnologie di IA comporta il mutamento della natura stessa delle minacce alla sicurezza pubblica e nazionale, introducendo nuovi rischi e vulnerabilità. Il MUAI può essere utilizzato per generare messaggi, profili e utenti sintetici altamente realistici, contribuendo alla continua espansione dell'universo bot. Questo rende possibile, inoltre, la raccolta e analisi massiva di interazioni e dati personali sulle piattaforme sociomediali, permettendo la creazione di attacchi personalizzati e difficilmente rilevabili. Ciò consente di attivare operazioni di influenza, ingerenza e interferenza dello spazio pubblico, soprattutto politico, economico e sociale.

Parole chiave: IA; MUAI; sicurezza; deepfake; disinformazione; cybersecurity.

The malicious use of Artificial Intelligence (MUAI) in the cyber-social ecosystem

The emergence and evolution of Malicious Use of Artificial Intelligence (MUAI) represents a paradigm shift in cyber-deviant and cybercriminal behaviour in terms of complexity, sophistication and speed of diffusion among the various categories of actors in the cyber-social ecosystem. The adoption of AI tools and technologies is changing the very nature of threats to public and national security, introducing new risks and vulnerabilities. MUAI can be used to generate highly realistic synthetic messages, profiles and users, contributing to the continuous expansion of the bot universe. This also makes it possible to collect and analyse massive amounts of interactions and personal data on social media platforms, allowing for the creation of personalised and difficult-to-detect attacks. This enables operations of influence, interference and meddling in the public sphere, especially in the political, economic and social spheres.

Keywords: AI; MUAI; security; deepfake; disinformation; cybersecurity.

DOI: 10.5281/zenodo.17297557

[*] Sapienza Università di Roma. arije.antinori@uniroma1.it.

1. La natura disruptive dell'intelligenza artificiale

L''intelligenza artificiale (IA) è definita una tecnologia disruptive, ossia portatrice di un'innovazione dirompente in grado di rivoluzionare non solo il mercato, ma di imprimere un mutamento radicale nell'intera società. L'IA è in grado di automatizzare i processi complessi, riducendo la necessità di intervento umano in attività che fino ad oggi richiedevano tempo ed esperienza. Tale tecnologia cambia i modelli di business costringendo le aziende tradizionali a ripensare il loro modo di operare al fine di mantenere adeguati livelli di competitività. Ciò determina una riconversione delle competenze e la rapida sostituzione di alcune attività lavorative, nonché la creazione di professioni del tutto nuove connesse in particolare agli aspetti di gestione e sviluppo dell'IA. Alcuni strumenti tradizionali divengono obsoleti e mutano anche i sistemi di interazione con i *device*, come nel caso del riconoscimento vocale che affianca e in alcuni contesti sostituisce la digitazione. Ulteriore esempio di portata trasformativa di tale tecnologia è l'utilizzo di modelli che utilizzano il *Deep Learning* (DL) per la generazione di testi, come nel caso dei *Generative Pre-trained Transformer* (GPT), tra cui l'ormai diffuso interfaccia AI conversazionale *ChatGPT*, che stanno sostituendo, tra l'altro, le ricerche per mezzo dei motori di ricerca, il c*opywriting*, le traduzioni e i servizi di assistenza online, facilitando così gli utenti nella loro quotidianità non solo in ambito professionale. La capacità di imitare il comportamento intelligente umano, di fornire strumenti predittivi e analitici, l'automazione avanzata dei processi, la riduzione dei costi operativi e l'integrazione trasversale stanno determinando la trasformazione di interi settori, dall'intrattenimento alla sicurezza, dalla salute all'educazione. Tuttavia, per gli utilizzatori, tale innovazione può comportare sfide complesse, poiché occorre costantemente adattarsi a nuove funzionalità, gestire i rischi legati a eventuali vulnerabilità, ma soprattutto affrontare l'impatto che l'adozione accelerata dell'IA, soprattutto in contesti privi di un adeguato framework normativo, può determinare sull'intera società. I rischi principali derivano dalla continua evoluzione dei sistemi di IA, che possono introdurre nuove forme di dipendenza tecnologica, riducendo la capacità critica degli individui e delle organizzazioni nell'interpretare le informazioni e nel *decision making*, sempre più tecnologicamente assistito. Inoltre, la progressiva automatizzazione di processi chiave può generare problemi di equità nell'accesso alle opportunità di impiego, poiché la richiesta di competenze specifiche potrebbe superare la capacità di adattamento del mercato del lavoro. Inoltre, il processo di iterazione continua che caratterizza lo sviluppo e rilascio di tecnologia AI comporta vulnerabilità che includono la possibilità di esposizione a minacce *cyber*

sempre più sofisticate, dovute all'adattabilità e all'apprendimento delle macchine. La velocità con cui i sistemi AI vengono aggiornati può portare a criticità algoritmiche "interne", è il caso dell'*Artificial Inte[glitch]ence* (Antinori, 2019), e falle nella sicurezza difficili da rilevare tempestivamente, aumentando i rischi di attacchi *cyber* mirati e violazioni della *privacy*. Infine, la crescente dipendenza da sistemi algoritmici che producono decisioni autonome pone interrogativi cogenti su trasparenza, responsabilità e verificabilità delle informazioni generate.

2. L'Uso Malevolo dell'Intelligenza Artificiale

La diffusione dei sistemi AI ha permesso la riduzione dei costi di attacchi informatici e fisici, consentendo l'automazione di operazioni che in passato richiedevano l'intervento umano. Tale processo ha comportato un'importante espansione del bacino di attori malevoli in grado di condurre attacchi, una maggiore frequenza delle azioni offensive e una dilatazione del piano d'attacco, ossia una crescita esponenziale del numero di potenziali *target*. Per Uso Malevolo dell'Intelligenza Artificiale (MUAI) si intende generalmente l'utilizzo di tecnologie AI per il conseguimento di diversi obiettivi dannosi, non etici e/o illeciti (Pashentsev, 2023). Appare evidente, quindi, come esso si sia sviluppato parallelamente all'evoluzione stessa dell'IA (Brundage *et al.,* 2018). Il MUAI include, dal punto di vista criminale, un ampio spettro di attività che vanno dalle intrusioni nei sistemi informatici alla realizzazione di contenuti e campagne disinformative, dall'hackeraggio di sistemi militari autonomi alla compromissione dell'integrità di tecnologia AI, da modalità avanzate di *social engineering* (Blauth *et al.,* 2022) alla generazione di contenuti sintetici. Il MUAI si manifesta prevalentemente in ambiti quali il cybercrime, la cyberdeviance, intesa come l'insieme di pratiche cyber-sociali che violano e/o compromettono norme sociali e/o giuridiche, o ne contestano i confini, generando al contempo processi di adesione, seppur limitati asingoli e gruppi, e reazione estesa da parte dell'opinione pubblica, in un contesto in cui la devianza – tradizionalmente intesa – ne risulta amplificata quindi in termini di scala, velocità e tracciabilità, coinvolgendo attori umani e non-umani. Il MUAI risulta, inoltre, particolarmente efficace nell'ambito della cyberwar e della manipolazione dell'opinione pubblica, trasformando in modo significativo tanto il panorama della *cybersecurity* quanto quello della sicurezza cyber-sociale (Antinori, 2018). A livello strategico, si individuano due macroaree di applicazione, una logica e una sociale.

La prima riguarda l'evoluzione e il potenziamento dei cyberattacchi tradizionali, sfruttando le vulnerabilità logiche dei sistemi informatici e delle componenti *hardware*, al fine di compromettere confidenzialità, integrità e disponibilità dei dati. I cybercriminali ricorrono al MUAI per automatizzare e rendere più efficaci i loro attacchi, come nel caso del *phishing* avanzato o del *code injection*, principalmente a danno di apparecchi e veicoli a guida autonoma, nonché gli attacchi ai sistemi di autenticazione biometrica. Il MUAI consente altresì di sviluppare *malware* adattivi, in grado di evolversi e aggirare le tradizionali misure di sicurezza, e lo sviluppo di concetti teorici per ransomware basati su Reinforcement Learning (RL), come nel proof-of-concept accademico denominato RansomAI, che dimostra come un agente potrebbe apprendere *ransomware* basati su RL, come *RansomAI* che utilizza un agente per apprendere in tempo reale quale algoritmo di crittografia, velocità e durata adottare al fine di minimizzare la possibilità di essere individuato, massimizzando al contempo il danno inflitto (von der Assen *et al.,* 2023). Inoltre, le tecniche di *Machine Learning* (ML) consentono di individuare e sfruttare vulnerabilità nei *software* con una rapidità mai vista prima, aumentando così il rischio di cosiddetti *zero-day attack*, spesso volti alla compromissione su ampia scala delle infrastrutture critiche. In tale macroarea, pertanto, il MUAI ha determinato:

- potenziamento delle minacce già esistenti;
- ridefinizione sostanziale delle caratteristiche delle minacce;
- sviluppo di nuove tipologie di attacchi;
- introduzione di nuove minacce attraverso strategie/tattiche in grado di sfruttare anche le vulnerabilità dei sistemi difensivi basati sull'IA.

Ciò rende possibili attacchi prima impensabili in quanto a complessità e risorse necessarie, favorendo la proliferazione di strategie, metodologie, tecniche e tattiche offensive, caratterizzate da un elevato livello di precisione, in grado di superare le tradizionali difese, rendendo sempre più difficili il rilevamento e l'attribuzione. Così, l'automazione dei cyberattacchi potenziati, in termini di scalabilità e velocizzazione, grazie all'uso di strumenti *AI-driven*, pone importanti sfide alla *cybersecurity*. Gli attori malevoli, sia simmetrici che asimmetrici, tra cui: potenze ostili, organizzazioni criminali, mafie, entità estremistico-violente, terroristi, cybercriminali e *hacking crew*, possono oggi grazie al MUAI, compromettere la sicurezza, e non solo, di individui, aziende, comunità e governi (UNOCT, 2021), contribuendo a modificare il tradizionale gap asimmetrico operativo tra attaccante e attaccato. Sebbene alcuni attori possano ridurre tale gap, altresì sfruttando la capacità trasformativa *dual-use* di tecnologie AI, altri esperti evidenziano il rischio

che l'IA possa al contrario ampliarlo a favore degli attori con maggiori risorse e capacità difensive. Pertanto, occorre sviluppare linee guida di natura algoretica al fine di prevenire degenerazioni criminali del dual-use nell'IA (Brundage *et al.,* 2018). Tuttavia, occorre precisare che il MUAI non deve essere confuso con l'*ethical hacking* che può servirsi di strumenti IA, nell'individuazione e disvelamento di sistemi AI corrotti o compromessi, a condizione che le attività connesse siano condotte entro certi limiti dal punto di vista legale (Choraś, Woźniak, 2021).

La seconda macroarea di applicazione del MUAI riguarda il potenziale sfruttamento delle "vulnerabilità sociali", come esito dell'interazione fra caratteristiche individuali, quali ad esempio l'età (Smahel *et al.,* 2020), appartenenze di gruppo (Abrajano *et al.,* 2024), e specificità algoritmiche delle diverse piattaforme sociomediali, siano esse *mainstream* o *fringe*, in quanto non più meri contenitori neutri, ma agglomerati connettivi di pubblici umani e inumani (Győri *et al.,* 2022). Il MUAI può quindi amplificare dinamiche di omofilia e polarizzazione affettiva all'interno delle cosiddette *echo chambers* – spazi informativi in cui le voci dissonanti sono ridotte al minimo, se non del tutto assenti, screditate e/o marginalizzate –, nonché l'esposizione selettiva rafforza credenze preesistenti, sfruttando ripetizione, novità ed emozioni per innescare elaborazione euristica e ragionamento motivato, all'interno di infosfere in cui le *affordance* di piattaforma contribuiscono a moltiplicare la portata dei contenuti manipolativi. Pertanto, vulnerabilità individuali e strutture algoritmiche si co-producono, trasformando così rischi informativi in minacce alla sicurezza cyber-sociale, quindi inevitabilmente politica, economica, e sempre più pubblica, nazionale e globale. In sintesi, le "vulnerabilità sociali", oggetto di interesse del MUAI, non si riducono a tratti individuali, ma emergono dove bias cognitivi, identità di gruppo e affordance di piattaforma si sovrappongono a fragilità, scarsa capacità o assenza di governance e moderazione. In tal senso, il MUAI riulta particolarmente efficace per la manipolazione psicologica e comportamentale degli individui attraverso l'automazione della disinformazione, creazione di *deepfake*, uso di *cyber-bot* e *AI-bot* per il *trolling* volto a influenzare il dibattito pubblico e favorire la disseminazione di narrazioni tossiche e divisive, la polarizzazione e weaponizzazione delle *audience* (Győri *et al.,* 2022). Tale categoria di minacce è particolarmente pericolosa in quanto sfrutta meccanismi cognitivi e dinamiche cyber-sociali al fine di alterare la percezione della realtà, indebolire la fiducia nelle istituzioni e creare divisioni tra gruppi sociali. Le campagne di disinformazione basate sul MUAI possono diffondere narrazioni manipolate con una velocità, una targetizzazione e una dimensione di scala impensabili fino a pochi anni fa, rendendo estremamente complessa la loro

individuazione, mitigazione e neutralizzazione. Le azioni offensive basate sul MUAI si distinguono per un'elevata precisione e una maggiore resilienza ai tentativi di contrasto, il che rende piuttosto complicata sia la difesa che la prevenzione, così come il contrasto e la mitigazione delle seguenti finalità operative delle tattiche MUAI:

- *deception* – creazione per mezzo della *Generative AI (GenAI)* di contenuti falsi, come testi, immagini, audio/video, volti a ingannare individui e istituzioni (Park *et al.,* 2024);
- manipolazione sociale – utilizzo di *bot* e algoritmi per influenzare l'opinione pubblica e diffondere propaganda (Marcellino *et al.,* 2023);
- furto di dati e violazioni della privacy – intrusioni abusive nei sistemi e raccolta massiva di informazioni e/o dati sensibili per campagne di *doxing* (Khalid *et al.,* 2023);
- frode finanziaria – frodi, in particolare su carte di credito e cryptovalute, evasione e/o elusione automatizzata dei sistemi di rilevamento antifrode, nonché manipolazione dei mercati (Josyula *et al.,* 2023);
- *deepfake* e disinformazione – creazione e diffusione di contenuti sintetici realistici con l'obiettivo di influenzare decisioni politiche ed economiche, nonché manipolare l'opinione pubblica con impatti significativi a livello psicologico, sociale e politico (Kharvi, 2024).

Se il crescente impiego di sistemi AI nella quotidianità delle società più avanzate da un lato favorisce la fruizione di servizi e l'automazione di procedure in termini di efficienza a favore del cittadino, dall'altro facilita la progettazione e realizzazione di operazioni malevoli contro *target* specifici, riducendo il margine di errore e amplificando l'impatto delle azioni criminali.

3. La contaminazione sintetica dell'ecosistema cyber-sociale

Internet, ciò che abbiamo definito in principio, spazio virtuale, evidenziandone la dimensione pseudo-ludica, immateriale, ma soprattutto non-reale, e poi cyberspace, ossia rete globale di computer interconnessi, secondo una prospettiva interpretativa fortemente tecnocentrica, rappresenta oggi, grazie all'evoluzione infrastrutturale digitale – nella transizione dai new media ai social media, attraverso la diffusione online di piattaforme algoritmiche ininterrottamente abitate da milioni di utenti nel mondo –, e al rapido mutamento tecnosociale, un ecosistema cyber-sociale. Questo deve essere

inteso oggi come un sistema socio-tecnico multidimensionale caratterizzato non solo da infrastrutture e piattaforme, ma da attori umani e non-umani, istituzioni, pratiche culturali e mediali, fenomeni sociali, e framework normativi, che influenza in modo significativo tanto la sicurezza quanto la coesione sociale.

In tale contesto, l'implementazione accelerata dell'AI, in assenza di un adeguato controllo istituzionale, può favorire la polarizzazione del dibattito pubblico, l'accentuazione delle disuguaglianze economiche, la proliferazione e disseminazione di contenuti ingannevoli. Così la creazione di *deepfake* e identità sintetiche, quindi false, rappresenta una delle principali risorse del MUAI in termini di costi ridotti e flessibilità operativa, in particolare attraverso la disseminazione massiva nell'ecosistema cyber-sociale di disinformazione, propaganda estremistica, odio online, al fine di attaccare e/o compromettere la reputazione di individui e organizzazioni, rappresentare sinteticamente *leader* politici raffigurati mentre dicono o compiono azioni di fatto mai accadute, influenzando elezioni e destabilizzando governi (Pashentsev, 2021).

Il *deepfake* costituisce una delle evoluzioni più sofisticate dell'AI applicata all'elaborazione di immagini e video, mentre nel caso della sintesi o clonazione della voce umana si hanno i *deepaudio*. Tecnologie come le *Generative Adversarial Networks* (GANs) consentono di creare contenuti audio/video altamente realistici attraverso l'uso di reti neurali e modelli generativi, permettendo tra l'altro la sostituzione di volti originali con quelli di altre persone, al fine di ottenere un effetto estremamente convincente. Il processo di creazione del *deepfake* si basa sull'addestramento AI per mezzo di un numero molto elevato di immagini del volto *target*, spesso reperite attraverso le piattaforme sociomediali. Occorre precisare che tale prodotto mediale differisce dal *deep video portrait*, che non comporta una sostituzione del volto, ma solo una manipolazione delle espressioni di un attore dal vero. Inoltre, il *deepfake* si inserisce in un quadro più ampio di manipolazione che comprende gli *shallow fakes*, ossia video alterati in modo grossolano e superficiale – ad esempio, attraverso un cambio della velocità e/o una manipolazione dell'audio – e i cosiddetti *synthetic media*, contenuti interamente generati dall'IA, come gli *avatar* iperrealistici. I primi *deepfake* sono apparsi sulla piattaforma *Reddit*, nella forma di contenuti di natura pornografica che vedevano la sovrapposizione dei volti di celebrità a quelli di pornoattori, al fine di realizzare video artefatti di natura sessuale. In tale contesto, la facilità di accesso ad app mobile e la disponibilità online di tool AI ha favorito una rapida diffusione del fenomeno, con migliaia di utenti che hanno iniziato a produrre e condividere contenuti. Negli ultimi anni, è aumentata

significativamente la capacità di manipolare contenuti audiovisivi, soprattutto negli ambiti politici e militari, portando alla maturazione di un mercato clandestino di *deepfake_as_a_service*, che consente anche a individui privi di competenze tecniche di accedere a strumenti avanzati per la creazione di contenuti falsificati (Europol, 2022). Tale fenomeno non solleva preoccupazioni soltanto per quanto concerne la prospettiva tradizionale di *cybersecurity*, ma sempre più in relazione alla disinformazione e manipolazione dell'opinione pubblica nell'ecosistema cyber-sociale in un contesto di post-verità (Antinori, 2019). L'uso indiscriminato di tale tecnologia per la personalizzazione dei contenuti informativi può favorire e accelerare la creazione di *echo chambers*, in particolare rafforzando pregiudizi sulla base di bias cognitivi. I contenuti *deepfake* possono essere impiegati come strumenti di propaganda, sabotaggio e *cyberwarfare*, alimentando teorie del complotto, quindi favorendo l'azione di gruppi cospirazionistici, estremismi antisistema e antigoverno (ASAGe), nonché fomentando l'odio *online*. Nel 2024, considerato come "*global elections year*" per le importanti elezioni che si sono succedute in diversi continenti, tra cui quelle statunitensi, il ricorso ai *deepfake* per alterare le dichiarazioni di candidati politici ha minato la fiducia pubblica e diffuso disinformazione su ampia scala (Insikt Group, 2024). Il MUAI trova la sua applicazione più pervasiva e dannosa a livello individuale nella creazione di pornografia non consensuale, che costituisce la stragrande maggioranza dei contenuti deepfake online. Questo si aggiunge ad altri usi criminali come il *revenge porn*, la pornografia non consensuale e la pedopornografia *online*, con l'obiettivo di danneggiare gravemente la reputazione degli individui, di celebrità, di importanti personaggi politici, ma soprattutto colpendo gli sceenagers, adolescenti e pre-adolescenti sempre connessi (Antinori, 2021) Risulta evidente come la capacità MUAI di falsificare prove video e audio ponga nuove sfide in ambito giudiziario, richiedendo l'adozione di metodologie e tecniche avanzate per garantire la custodia e genuinità delle prove digitali acquisite nell'ambito dei procedimenti penali (Di Paolo, Pressacco, 2022). La tecnologia *deepfake* sta rapidamente evolvendo, rendendo sempre più difficile distinguere tra contenuti autentici e manipolati, costituendo altresì una minaccia alla sicurezza nazionale. Tale aspetto risulta particolarmente rilevante in scenari di crisi e/o conflitti geopolitici, tanto da divenire una modalità strutturata nell'ambito della *Cognitive Warfare* (Antinori, 2024).

Conclusioni

Lo sviluppo e diffusione del MUAI segna un punto di svolta nell'ambito della sicurezza globale, imponendo nuove sfide alla capacità di difesa e resilienza delle società contemporanee. La crescente pervasività di tale tecnologia dimostra come la minaccia non sia più confinata al solo ambito della *cybersecurity*, ma si estenda in modo trasversale all'intero ecosistema cybersociale, influenzando dinamiche politiche, economiche e culturali. In particolare, la capacità del MUAI di manipolare l'opinione pubblica, sfruttare vulnerabilità psicologiche, sociali, e alterare la percezione della realtà attraverso *deepfake* e campagne disinformative evidenzia l'urgenza di strategie di contrasto integrate, multidisciplinari e multiattoriali.

In tale contesto, la regolamentazione e la *governance* dell'AI assumono un ruolo cruciale. Risulta necessario sviluppare un framework normativo convergente a livello globale che disciplini l'AI, al fine di evitare il proliferare di strumenti sempre più sofisticati nelle mani delle diverse tipologie di attori malevoli. Da qui la necessità di trovare un equilibrio tra innovazione e protezione della sicurezza, sia pubblica che nazionale, favorendo il progresso tecnologico, ma mitigandone i rischi attraverso misure di controllo efficaci e trasparenti. Il rafforzamento della cooperazione internazionale tra governi, università, enti di ricerca e aziende private risulta imprescindibile per prevenire l'abuso delle tecnologie AI e promuovere standard condivisi per l'identificazione e neutralizzazione di contenuti manipolati.

Risulta altresì essenziale un'azione mirata alla formazione, sensibilizzazione e consapevolezza a livello sociale. L'educazione digitale e la *AI-literacy*, devono includere la capacità di riconoscere la disinformazione e comprendere i meccanismi di manipolazione utilizzati nel contesto cybersociale con l'obiettivo di rafforzare la fiducia pubblica nelle istituzioni, la coesione sociale e la stabilità democratica.

Riferimenti bibliografici

Abrajano M., Garcia M., Pope A., Vidigal R., Tucker J.A., Nagler J. (2024). How reliance on Spanish-language social media predicts beliefs in false political narratives amongst Latinos. *PNAS Nexus*, 3(11): pgae442. https://doi.org/10.1093/pnasnexus/pgae442

Anderson R., Barton C., Böhme R., Clayton R., Van Eeten M., Levi M. (2019). Measuring the cost of cybercrime. In: Moore T., Pym D., Ioannidis C. (a cura di), *The Economics of Information Security and Privacy* (pp. 265-300). Cham: Springer.

Antinori A. (2018). *Artificial Intelligence. Una minaccia sistemica alla sicurezza nazionale in prospettiva futura*, *GNOSIS – Rivista Italiana di Intelligence*, AISI – Agenzia Informazioni e Sicurezza Interna.

Antinori A. (2018). Sicurezza cyber-sociale. In: Baldoni R., De Nicola R., Prinetto P. (a cura di), *Il futuro della cybersecurity in Italia: Ambiti progettuali strategici* (pp. 188-190). Roma: Laboratorio Nazionale di Cybersecurity, CINI.

Antinori A. (2019). Terrorism and deepfake: From hybrid warfare to post-truth warfare in a hybrid world. In: Griffiths P., Kabir M.N. (a cura di), *Proceedings of the European Conference on the Impact of AI and Robotics* (pp. 23-30). Reading: Academic Conferences and Publishing International Limited.

Antinori A. (2020). Terrorism in the early onlife age: From propaganda to 'propulsion'. In: Antinori A. (a cura di), *Terrorism and Advanced Technologies in Psychological Warfare. New Risks, New Opportunities to Counter the Terrorist Threat*. New York: Nova Science Publishers.

Antinori A. (2021). La sicurezza cyber-sociale dei giovani (e non solo). Tra social network, metaversi e radicalizzazione. *CyberSecurity Italia*. https://www.cybersecitalia.it/la-sicurezza-cyber-sociale-dei-giovani-e-non-solo-tra-social-network-metaversi-e-radicalizzazione/15171/

Antinori A. (2024). La cognitive warfare in uno scenario di minacce convergenti. *Sicurezza, Terrorismo e Società*, 19(Special Issue 1/2024): 67-81. Milano: EDUCatt – Università Cattolica del Sacro Cuore.

Bazarkina D., Pashentsev Y.N. (2019). Artificial intelligence and new threats to international psychological security. *Russia in Global Affairs*.

Biggio B., Nelson B., Laskov P. (2012). Poisoning attacks against support vector machines. arXiv:1206.6389.

Blauth T.F., Gstrein O.J., Zwitter A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*, 10: 77110-77122. https://doi.org/10.1109/ACCESS.2022.3191790

Brown A.W. (2021). *Surviving and Thriving in the Age of AI. A Handbook for Digital Leaders*. Cambridge: Cambridge University Press.

Brundage M., Avin S., Clark J. et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv*. https://doi.org/10.48550/arXiv.1802.07228

Carlini N., Wagner D. (2017). Towards evaluating the robustness of neural networks. arXiv:1608.04644.

Choraś M., Woźniak M. (2021). The double-edged sword of AI: Ethical adversarial attacks to counter artificial intelligence for crime. *SN Computer Science*, 2: 1-4. https://doi.org/10.1007/S43681-021-00113-9

Di Paolo G., Pressacco L. (a cura di) (2022). *Intelligenza artificiale e processo penale. Indagini, prove, giudizio*. Trento: Università degli Studi di Trento.

Europol (2022). *Facing reality? Law enforcement and the challenge of deepfakes*. The Hague: Europol Innovation Lab. https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcementand-challenge-of-deepfakes

Ferrara E. (2023). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. arXiv:2310.00737.

Floridi L., Cowls J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).

Frey C.B. (2019). *The Technology Trap. Capital, Labor, and Power in the Age of Automation*. Princeton: Princeton University Press.

Goodfellow I.J., Shlens J., Szegedy C. (2015). Explaining and harnessing adversarial examples. arXiv:1412.6572.

Gu T., Dolan-Gavitt B., Garg S. (2019). BadNets: Identifying vulnerabilities in the machine learning model supply chain. arXiv:1708.06733.

Győri L., Molnár C., Krekó P., Szicherle P. (2022). The Kremlin's troll network never sleeps: Inauthentic pro-Kremlin online behavior on Facebook in Germany, Italy, Romania and Hungary. Budapest: Political Capital. https://politicalcapital.hu/pc-admin/source/documents/pc_ned_study_kremlin_troll_network_2022_web.pdf

Insikt Group (2024, 24 settembre). Targets, objectives, and emerging tactics of political deepfakes. *Recorded Future*. https://www.recordedfuture.com/research/targets-objectives-emerging-tactics-political-deepfakes

Jansevskis M., Osis K. (2023). *Artificial Intelligence for Security. Enhancing Protection in a Digital Age*. Cham: Springer Nature.

Josyula H.P., Vishnubhotla D., Onyando P.O. (2023). Is artificial intelligence an efficient technology for financial fraud risk management?. *International Journal of Managerial Studies and Research (IJMSR)*, 11(6): 11-16. https://doi.org/10.20431/2349-0349.1106002

Khalid N., Qayyum A., Bilal M., Al-Fuqaha A., Qadir J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, 158: 106848. https://doi.org/10.1016/j.compbiomed.2023.106848

Kharvi P. (2024). Understanding the impact of AI-generated deepfakes on public opinion, political discourse, and personal security in social media. *IEEE Security & Privacy*, 22(2): 2-9. https://doi.org/10.1109/MSEC.2023.106848

Kurakin A., Goodfellow I., Bengio S. (2017). Adversarial examples in the physical world. arXiv:1607.02533.

Marcellino W., Beauchamp-Mustafaga N., Kerrigan A., Chao L.N., Smith J. (2023). The rise of generative AI and the coming era of social media manipulation 3.0: Next-generation Chinese astroturfing and coping with ubiquitous AI. Santa Monica (CA): RAND Corporation. https://www.rand.org/pubs/perspectives/PEA2679-1.html

Marchal N., Xu R., Elasmar R., Gabriel I., Goldberg B. (2024). Generative AI misuse: A taxonomy of tactics and insights from real-world data. arXiv:2406.13843.

Masayuki I. (2021). *A Narrative History of Artificial Intelligence. The Perpetual Frontier of Information Technology*. Cham: Springer Nature.

Papernot N., McDaniel P., Wu X., Jha S., Swami A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. *arXiv*:1511.04508.

Park P.S., Goldstein S., O'Gara A., Chen M., Hendrycks D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5): 100988. https://doi.org/10.1016/j.patter.2024.100988

Pashentsev E. (2021). The Palgrave handbook of malicious use of AI and psychological security (pp. 45-60). London: Palgrave Macmillan.

Pashentsev E. (a cura di) (2023). *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-22552-9

Political Capital (2022). *The Kremlin's Troll Network Never Sleeps. Inauthentic Pro-Kremlin Online Behavior on Facebook in Germany, Italy, Romania, and Hungary*. Budapest: Political Capital.

Smahel D., Machackova H., Mascheroni G., Dedkova L., Staksrud E., Ólafsson K., Livingstone S., Hasebrink U. (2020). *EU Kids Online 2020. Survey results from 19 countries*.

Artificial social research? AI's capabilities and risks in predicting values and attitudes

by *Caterina Ambrosio, Ciro Clemente De Falco, Domenico Trezza* [*]

This study explores the potential and limitations of generative artificial intelligence, with a particular focus on ChatGPT-4, in reproducing human values and attitudes based on socio-demographic profiles. By comparing real data from the European Social Survey (ESS) with AI-generated data, the study assesses the ability of AI to reflect public opinion trends. While AI-generated responses exhibit a general alignment with survey data, they show limited variability and some inconsistencies in group-level analyses. In conclusion, despite its promising aspects, the findings suggest that AI cannot replace empirical research.

Keywords: artificial intelligence, experiment, predictive capacity, validity and reliability.

Ricerca sociale artificiale? Capacità e rischi dell'IA nella previsione di valori e atteggiamenti

Questo studio esplora le potenzialità e i limiti dell'intelligenza artificiale generativa, con particolare riferimento a ChatGPT-4, nella riproduzione di valori e atteggiamenti umani sulla base di profili socio-demografici. Attraverso il confronto tra dati reali provenienti dall'European Social Survey (ESS) e dati generati da ChatGPT, si valuta la capacità dell'IA di rispecchiare le tendenze dell'opinione pubblica. Sebbene le risposte generate dall'IA mostrino un allineamento generale con i dati dell'indagine, esse presentano una ridotta variabilità e alcune incongruenze nelle analisi di gruppo. In conclusione, nonostante gli aspetti promettenti, i risultati suggeriscono che l'IA non può sostituire la ricerca empirica.

Parole chiave: intelligenza artificiale, esperimento, capacità predittiva, validità e affidabilità.

Introduction

In recent years, generative artificial intelligence (AI-gen) has increasingly become central to technological progress and research. Like all major technological innovations, it raises ethical, social, and cultural questions. The

DOI: 10.5281/zenodo.17297559

[*] Università di Napoli Federico II. caterina.ambrosio@unina.it, ciroclemente.defalco@unina.it, domenico.trezza@unina.it.

ability of these generative systems to understand and replicate the nuances of human behavior has become – among various issues related to the relationship between AI and society – an intensely debated topic in the social sciences, particularly with respect to the integration of social and cultural competencies into computational models (Floridi, Cowls, 2022). On the one hand, AI-gen promises to offer innovative tools for the analysis and prediction of social phenomena; on the other hand, numerous questions emerge regarding the validity of such approaches and the implications of their application (Mokander, Schroeder, 2022).

This reflection could be a part of a broader and likley epistemological transformation within social research, which has recently seen the rise of big data as a predominant object of study. While big data promises – for many, it threatens – the "end of theory" (Amaturo, Aragona, 2019; Anderson, 2008), today AI-gen and advanced language models seem to present similarly unprecedented scenarios for social research: the potential to generate credible simulations of opinions, attitudes, and values without the need to collect real-world data raises a provocative – and likely exaggerated – question: Are we facing the end of empirical research as we know it?

Building on these premises, our study explores AI's ability to predict personal orientations on values and attitudes solely from socio-demographic data. We ask if a language model can adopt a social category and replicate its opinions and values. By comparing AI-generated responses with European Social Survey (ESS) data using identical profiles, we examine the alignment between AI and human insights. Focusing on value elements and gender issues, our preliminary experiment an existing, yet not yet predominant, ability to predict social opinions. The paper is divided into five sections covering previous research, methodology, results, discussion, and future perspectives.

1. Literature review

The predictive capabilities of AI are generating cross-disciplinary interest, engaging in all fields of study, including the humanities and social sciences (Fan *et al.,* 2024). A recent interest in this field is the ability to simulate human responses to sociological stimuli. This area of research encompasses studies in which the generated data are commonly referred to as "synthetic data" or "silicon samples" (Argyle *et al.,* 2023). The goal of such studies is to determine whether responses simulated by AI to questions on attitudes and opinions are comparable to those collected from real participants.

The study by Argyle and colleagues (2023) was among the first to propose the use of LLMs to simulate responses on political opinions. Specifically, GPT-3 was tested for its ability to replicate voting choices for either the Republican or Democratic candidate based on the socio-demographic characteristics of human respondents. Using data from the 2012, 2016, and 2020 editions of ANES, GPT-3 was conditioned on variables such as gender, ethnicity, age, political ideology, political interest, and state of residence. The model was asked to complete prompts like "In 2016, I voted for...". After comparing GPT-3's responses with real data, the authors concluded that the AI could replicate patterns with high accuracy, but limited variability emerged within specific subgroups.

Two years later, building on this study, Bisbee and colleagues (2024) proposed a similar study. In this case, the authors tested ChatGPT 3.5, a more advanced version. Once again, the comparison dataset consisted of real responses collected from the 2016 and 2020 ANES editions. The authors input the descriptions of 7,530 human respondents into the chatbot using an automated program. For each of these profiles, 30 possible responses were requested. The model then simulated various responses for each human profile. After comparing synthetic data with real data, the results demonstrated an apparent similarity to real data; indeed, the mean scores were highly similar. However, what stood out again was the reduced variability

The studies highlight LLMs' potential and limitations in social research. Both find that ChatGPT generates survey-like synthetic data but warn of risks. Low response variability, biases in training data, and prompt sensitivity make LLMs unreliable for replacing traditional surveys.

The present study pursues the same goal as previous works but aims to overcome some of their limitations. First, it uses the latest publicly available version, ChatGPT-4. This version allows the provision of documents in various formats (docx, Excel, etc.). This enables a more quantitative and structured approach. Second, while the analyzed articles relied on a single approach in prompts to obtain responses, this experiment tested different approaches. Finally, unlike previous studies, which primarily focused on simple and specific questions, such as voting preference, our study selected items from attitude scales designed to explore more complex and nuanced dimensions of personality and human opinions.

2. Methodology

2.1. Research design

The aim of this study was to investigate the ability of generative AI - specifically, the GPT chatbot – to reproduce patterns of values and opinions based on specific socio-demographic profiles. To achieve this goal, two real datasets were compared with two "synthetic" datasets. Specifically, through an experiment that compares AI-generated responses with data from the international European Social Survey (ESS) – Round 11, using identical socio-demographic profiles, we aim to assess the degree of alignment between human-generated and AI-generated data.
The version of ChatGPT used for this study was ChatGPT-4, available as a paid service since May 2024. The 2023 ESS dataset for Italy and Great Britain was used for comparison. These countries were selected because ChatGPT, like other LLMs, is mainly trained on English data (Brown *et al.,* 2020), which could influence its outputs. Italy was included due to expertise in its context. The variables considered in this study include socio-demographic variables as well as variables concerning attitudes and opinions for 1,000 Italian individuals and 1,000 British individuals, sampled from a ESS dataset using simple random sampling. For attitudes and opinions, 2 items were selected from the "Human Values Scale" module and 8 items from the "Gender in Contemporary Europe**"** module. The variables included are presented in the table below (Tab.1)

Tab 1. Variables included in the research.

Variable Description	Response Categories / Coding Details
SOCIO-DEMOGRAPHIC	
Gender	1: Male, 2: Female, 9: No answer
Highest level of education[1]	ISCED codes (0: No ISCED completed to 800: Doctoral degree), 5555: Other, 7777: Refusal, 8888: Don't know, 9999: No answer

[1] For the purpose of the analysis, this variable was recoded as follows: 1 = Low level; 2 = Medium level; 3 = High level.

Household's total net income, all sources[2]	1-10: Income deciles, 77: Refusal, 88: Don't know, 99: No answer
HUMAN VALUES SCALE	
Important to be rich, have money and expensive things (renamed *money*)	1: Very much like me, 6: Not like me at all, 66-99: Missing values
Important to have a good time (renamed (*goodtime*)	1: Very much like me, 6: Not like me at all, 66-99: Missing values
GENDER IN CONTEMPORARY EUROPE	
Good or bad for family life if equal number of men and women are in paid work (renamed *family*)	0: Very bad, 6: Very good, 7-9: Missing values (Refusal, Don't know, No answer)
Good or bad for politics if equal number of men and women hold leadership positions (renamed *politics*)	0: Very bad, 6: Very good, 7-9: Missing values (Refusal, Don't know, No answer)
Good or bad for businesses if equal number of men and women are in higher management (renamed *business*)	0: Very bad, 6: Very good, 7-9: Missing values (Refusal, Don't know, No answer)
Good or bad for economy if men and women receive equal pay for the same work (renamed *economy*)	0: Very bad, 6: Very good, 7-9: Missing values (Refusal, Don't know, No answer)
Dividing the number of seats in parliament equally between men and women (renamed *parliament*)	1: Strongly in favour, 5: Strongly against, 7-9: Missing values
Requiring both parents to take equal periods of paid leave (renamed *periods*)	1: Strongly in favour, 5: Strongly against, 7-9: Missing values
Firing employees who make insulting comments directed at women (renamed *insulting*)	1: Strongly in favour, 5: Strongly against, 7-9: Missing values
Fining businesses when men are paid more than women for the same work (renamed *business*)	1: Strongly in favour, 5: Strongly against, 7-9: Missing values

The various prompts and their corresponding outputs were executed in January 2025. For each country, ChatGPT was provided with a matrix containing completed socio-demographic variables and incomplete attitudinal variables. To avoid computational slowdowns, which occurred during

[2] For the purpose of the analysis, this variable was recoded as follows: 1 = Low income; 2 = Medium income; 3 = High income.

preliminary tests, the two datasets were split into 4 datasets of 250 cases each. The ESS codebook was also provided to aid in understanding the variable labels. Once the matrix and the codebook were supplied, ChatGPT was instructed to perform the following tasks for each session:

1. Confirm understanding of the dataset and codebook. Verify the presence of completed variables (socio-economic and demographic characteristics) and incomplete variables (attitudes and opinions).
2. Identify the socio-economic and demographic profile of the subjects. Extract and interpret specific information for each subject based on the complete variables.
3. Impute responses to the missing variables. Generate responses to the questions on attitudes and values for each individual, based on their defined socio-economic and demographic profile. The responses had to be coherent and substantiated, following the descriptions provided in the codebook and avoiding randomness.

Given the demonstrated sensitivity of AI-generated responses to the prompts used (Argyle *et al.,* 2023; Bisbee *et al.,* 2024), three different prompts were tested for each country, each employing a different approach. A new chat session was initiated for each approach. The three approaches were as follows:

- Predictive approach. In this approach, the AI was asked to predict coherent responses to all the questions on attitudes and values based on the subject's socio-demographic profile.
- Interviewer approach. This approach simulated a direct interview with the subjects represented in the dataset. The AI assumed the role of an interviewer and filled in the questionnaire cells for each socio-demographic profile provided in the Excel file.
- Researcher approach. In this approach, the AI was made aware that it was participating in a social research experiment. The aim was to determine whether it could identify with a reference social category and accurately reproduce the typical patterns of values and opinions of that group.

2.2. Analysis Procedure

To compare real responses with those generated by the three AI models (Section 3.1), we first standardized the scores across the three scales to ensure consistency. We then calculated average scores for each item, analyzing variability through standard deviations. This helped assess whether AI could

replicate the diversity of human responses. Differences in average scores were presented using tables and visualizations, highlighting areas where AI responses aligned with real data.
To examine how well the AI models reflected values and attitudes based on social categories (Section 3.2), we grouped education and income into three levels (low, medium, high). The analysis of gender-related items and values followed a structured approach. For the two value-related items, we conducted an ANOVA with post-hoc tests for all three AI models in Italy and the UK, incorporating socio-demographic factors. The ANOVA identified significant relationships, while post-hoc tests clarified their direction and strength.
For gender-related items, we complemented the ANOVA with a multiple regression model. We aggregated responses into composite scores to reflect overall attitudes, creating an additive index as the dependent variable. Socio-demographic factors were converted into dummy variables, using "high" education and income as reference categories, along with "male" for gender. This approach helped isolate the specific influence of each factor on attitudes toward gender issues.

3. Analysis of results

In this paragraph the ability of AI-generated responses to replicate human survey data will be examined. Section 3.1 analyzes score distributions across different items, comparing human responses (ORIG) with three AI simulation approaches (PRED, INTERV, RESEARCH). Section 3.2 extends the analysis by testing whether AI-generated responses reflect socio-demographic patterns observed in the original data. Using ANOVA and regression models, we assess the consistency of AI predictions with human attitudes across variables such as gender, education, and income.

3.1. Analysis of item scores

This section examines how original survey scores compare with those generated by our three AI simulation approaches. The survey items are grouped into three scales: two on gender equality (GENDER_A and GENDER_B, each with four items) and one on personal and social values (VALUE, with two items). Since each scale uses different metrics, we normalized the scores for fair comparison (tab.2).

At first glance, differences emerge. Human responses show a lively spread (standard deviation: 0.256), reflecting diverse opinions. In contrast, AI scores are more uniform, with deviations of 0.079 for PRED, 0.069 for RESEARCH, and just 0.011 for INTERV. This suggests AI struggles to capture the variability of human thought.

A closer look reveals further nuances. For GENDER_A – celebrating equal inclusion in work, politics, and management – survey scores are consistently higher than AI's, suggesting a strong real-world consensus that AI fails to replicate. Conversely, for GENDER_B – corrective measures like quotas and penalties – AI rates them more favorably, possibly overestimating public support. On the VALUE scale, covering personal aspects like enjoyment and wealth, AI aligns more closely with human responses, indicating a better grasp of universal sentiments.

Tab.2 Mean of normalized item scores

codebook	Item	ORIG	PRED	INTERV	RES
0=bad 1=good	Bad or good for family life in [country] if equal numbers of women and men are in paid work	0.808	0.582	0.501	0.642
0=bad 1=good	Bad or good for politics in [country] if equal numbers of women and men are in positions of political leadership	0.805	0.676	0.507	0.551
0=bad 1=good	Bad or good for businesses in [country] if equal numbers of women and men are in higher management positions	0.825	0.582	0.507	0.637
0=bad 1=good	Bad or good for economy in [country] if women and men receive equal pay for doing the same work	0.882	0.507	0.486	0.573
0=strongly in favour 1=strongly against	Dividing the number of seats in parliament equally between women and men	0.353	0.514	0.508	0.457

0=strongly in favour 1=strongly against	Require both parents to take equal periods of paid leave to care for their child	0.325	0.514	0.487	0.493
0=strongly in favour 1=strongly against	Firing employees who make insulting comments directed at women in the work-place	0.278	0.51	0.51	0.454
0=strongly in favour 1=strongly against	Making businesses pay a fine when they pay men more than women for doing the same work	0.24	0.514	0.5	0.51
0=strongly agree 1=strongly disagree	Important to have a good time	0.45	0.411	0.482	0.508
0=strongly agree 1=strongly disagree	Important to be rich, have money and ex-pensive things	0.608	0.411	0.512	0.474
	DEV.STANDARD	0.256	0.07	0.01	0.06

These differences are visualized in Table 3 using grayscale shading – lighter cells indicate lower scores and closer alignment with original data. GENDER_A items show the biggest discrepancies, GENDER_B items fall in the middle, and VALUE items are simulated most accurately. Notably, the statement "Important to have a good time" has minimal differences, ranging only from 0.03 to 0.05.

Among approaches, RESEARCH performs best, with an overall differential of 0.183, suggesting that when AI is explicitly informed about research objectives, it better mirrors human responses. PRED (0.207) and INTERV (0.226) perform slightly worse, likely due to the challenge of simulating complex individual interactions.

Interestingly, AI struggles most with items featuring intricate linguistic constructions, especially those on economic aspects of gender issues. While modern language models excel in processing language, they still miss subtle nuances in topics like pay equity. In summary, RESEARCH best captures gender equality sentiments, while INTERV slightly outperforms personal values. These findings highlight the importance of context and clear instructions in improving AI's ability to predict social opinions with human-like variability.

Tab.3 Absolute differences with original scores

Issue	Item	PRED	INTERV	RES	tot.avg
GENDER_A	Bad or good for family life in [country] if equal numbers of women and men are in paid work	0,226	0,308	0,166	**0,233**
	Bad or good for politics in [country] if equal numbers of women and men are in positions of political leadership	0,129	0,299	0,254	**0,227**
	Bad or good for businesses in [country] if equal numbers of women and men are in higher management positions	0,243	0,318	0,188	**0,250**
	Bad or good for economy in [country] if women and men receive equal pay for doing the same work	0,375	0,396	0,308	**0,360**
GENDER_B	Dividing the number of seats in parliament equally between women and men	0,162	0,155	0,104	**0,140**
	Require both parents to take equal periods of paid leave to care for their child	0,189	0,162	0,168	**0,173**
	Firing employees who make insulting comments directed at women in the workplace	0,233	0,233	0,176	**0,214**
	Making businesses pay a fine when they pay men more than women for doing the same work	0,274	0,260	0,270	**0,268**
VALUE	Important to have a good time	0,039	0,032	0,058	**0,043**
	Important to be rich, have money and expensive things	0,197	0,096	0,134	**0,142**
	tot. Avg.	***0,207***	***0,226***	***0,183***	**0,205**

3.2. Group analysis

In this section, we will test the approaches' ability to reproduce the values and attitudes of social categories chosen for the analysis. The approach taken for this analysis and described in the 2.2 section produced numerous outputs that have been summarized in the tables 4,5,6. Each table shows the results of a single variable. Each table is organised as follows: in the rows are the

gender-related items and values while in the columns are the datasets for the three approaches used and the original db. Each cell represents the result of the ANOVA between a specific item and a given approach on a single socio-demographic variable. Empty cells indicate that the Anova did not reach the level of statistical significance ($p < 0.05$), while cells with significant results directly report post-hoc test results. Finally, Table 7 reports the results of the multiple regressions. The outputs include the classic regression parameters, and only the results for which the t-test was significant were presented.

The first results we are going to comment on concern gender (table 4), where the original data from Italy (ITA) show significant relationships in the expected direction: the female sex has a more favourable attitude toward gender issues. This trend is observed on both scales. Analyses using ANOVA and regression models confirm these relationships. Of the three IA approaches, the "INTERV" approach shows no similarity with the original data. In particular, no significant relationships emerge between sex and scores on individual items, as evidenced by both ANOVA and the regression coefficient. In contrast, the third approach (RES) shows consistent relationships with the original data in three out of four cases on the first scale. This partial consistency is confirmed by the regression coefficient.

Tab 4. Anova on Gender

ITEM	*Original*	*Prev*	*Interv*	*Research*
family	F	F	-	-
politics	F	F (x)	-	F
business	F	F	-	F
economy	F	F	-	F
parliament	M	F (x)	-	-
periods	-	XX F)	-	-
insulting	M	F (x)	-	-
business	M	F (x)	-	-
goodtime	F	F (x)	-	-
money	-	F (x)	-	-

The first AI approach (PREV), on the other hand, is distinguished by complex and not always consistent behaviours. Significant relationships are observed, but with divergent directions between the two scales: on the first scale, the female sex is significantly more attentive to gender issues, while on the second scale, they are less attentive. This results are consistent with

the original data on the first scale but discordance on the second. An important methodological aspect concerns the presence of the "X" in the tables, indicating cases where Anova was not performed due to the homogeneity of within-group scores. Finally, the regression model confirms what has been observed: the regression coefficient is significant on both scales but shows signs consistent with the original data only on the first scale, while on the second scale, it shows opposite directions. What was observed on the sex variable is also repeated, in part, for educational attainment. In Italy, there are significant relationships in the expected direction for most items and for both scale summary indices.

Tab 5. Anova on Education

ITEM	*Original*	*Prev*	*Interv*	*Research*
family	L<M=H	L<H	-	-
politics	L<M=H	L>M=H	-	-
business	L<M<H	L<H	-	-
economy	L<M=H	L>M=H	-	-
parliament	L>M=H	L>M=H	-	-
periods	L>M=H	L>M=H	-	-
insultin	-	L>M=H	-	-
business	L>M>H	L>M=H	-	-
goodtime	L>M=H	L>M=H	-	-
money	L>M=H	L>M=H	-	-

The "INTERV" approach shows no similarity with the original data in Italy as confirmed by the regression coefficients on both scales. The RESEARCH approach IA shows no significant relationships between education and attitudes toward gender issues, except for two items on the first scale, where the direction of the relationship is opposite to that expected. For the PREV approach, in Italy, on the first scale, the regression coefficient is consistent with the original data, but only two out of four items show the expected direction. On the other two items, the relationship exists but is of opposite sign. On the second scale, there are relationships consistent with the original data on three out of four items. However, in the regression, the effect of education is cancelled out by the influence of sex. Finally, regarding income, significant relationships in the expected direction are observed on the first scale.However, the PREV approach shows no significant relationship between income and attitudes toward gender issues.

Tab 6. Anova on income

ITEM	*Original*	*Prev*	*Interv*	*Research*
family	L<H	-	L>M=H	-
politics	L<M=H	-	-	-
business	L<M=H	-	L<M=H	-
economy	L=M<H	-	-	-
parliament	L>M=H	-	-	-
periods	-	-	-	-
insultin	L>M=H	-	-	-
business	L>M=H	-	-	-
goodtime	L>M>H	-	-	-
money	-	-	-	-

The INTERV approach shows a significant relationship only in Italy but with a coefficient of the opposite sign from the original data. Finally, the RES approach shows. The ninth and tenth items show significant weaknesses in the ability of the AI approach to replicate the original relationships between socio-demographic variables and scores correctly. Focusing on significant relationships, we note that the second approach never shows significant relationships, confirming its general ineffectiveness. The RESEARCH approach also fits into this picture of weakness but with an exceptions on income The PREV approach, however, is notable for the consistent presence of significant relationships on gender, education and scores on items.

Tab 7. Regression model

	Value	*Original*	*Prev*	*Interv*	*Ob_dich*
First Scale	*education*	1,27	0,74		
	gender	-0,94	-7,9		-0,78
	Income	0,93	0,19		
	R_square	0,44	0,957		0,012
	R_square_corr	0,39	0,957		0,007
	T Test	Yes	Yes		0,053
Second Scale	*education*	-0,87			
	gender	1,42	-8		
	Income			0,717	
	R_square	0,66	1	0,013	

R_square_corr	0,61	1	0,008
T Test	Yes		Yes

Conclusion

This study compared real survey data with responses generated by an AI model (based on GPT) to assess the ability of AI to reproduce patterns of values and opinions based on socio-demographic variables. The analysis was conducted in two stages: the first examined differences at the individual question level; the second considered socio-demographic factors.

The results at the item level show that there are differences between the actual and simulated data, especially for questions related to gender and values. These discrepancies may have both linguistic (more complex wording) and semantic (some gender issues are still "difficult" for the model) causes. As found in other empirical studies, the AI-generated responses here also show little variability.

However, AI performance improves when clear objectives and context are provided: in the "RESEARCH" approach, which made the study's purpose and setting explicit, results appeared more consistent with actual data than in the "PRED" (simple prediction) or "INTERV" (interviewer role) modes. Despite this, when socio-demographic variables are introduced, the relationships between them and views on gender are often found to be inconsistent with those observed in the actual data. Overall, the "PRED" approach generated the most significant relationships, but in several cases misaligned with the actual data. The lack of variability in the responses sometimes prevented statistical tests (ANOVA) from being performed, suggesting a potential inconsistent use of certain variables by the AI. Where, on the other hand, results are consistent, high R^2 values suggest overfitting or overestimation of ratios.

Considering this evidence, it can be concluded that although AI is able to capture some general trends on social values and demographic profiles, it still cannot reliably reproduce complex nuances and relationships, especially in sensitive contexts such as gender issues. In addition, the design of the interaction-clear instructions and well-defined goals-has a significant impact on the ability of AI to generate consistent responses. Ultimately, while language models can support social research, they cannot yet completely replace empirical data.

This reflection not only concerns the academic community but is part of a broader debate that also involves the world of work, whose experts are

questioning how far artificial intelligence will be able to replace human labor. (Gmyrek, Berg, Bescond, 2023), in a recent paper for the ILO estimated the impact of generative technologies such as GPT-4 on occupations globally. Even outside the strictly scientific realm, the conclusions converge with the findings of this study: despite the significant improvement in the predictive and linguistic capabilities of the latest generation models, the future in which AI will completely replace human labor is still far off. It is true that the technological leap has meant that it is no longer only simple, manual or repetitive jobs that are potentially exposed to automation, but also certain categories of "cognitive work" particularly administrative professions. However, as the ILO authors point out, this exposure is partial: AI can automate some tasks, but it can hardly replace the full range of activities that characterize a profession. The result, rather than a process of total replacement, seems to take the form of a transformation of work.
Future studies may make use of different AI models and longitudinal analyses to assess how evolution and continuous fine-tuning affect this performance.

References

Amaturo E., Aragona B. (2019). Per un'epistemologia del digitale: note sull'uso di big data e computazione nella ricerca sociale. *Quaderni di Sociologia* [Online], 81- LXIII | 2019, online dal 01 juin 2020, consultato il 09 février 2025. http://journals.openedition.org/qds/3508; DOI: https://doi.org/10.4000/qds.3508

Anderson C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7): 16-07.

Argyle L.P., Busby E.C., Fulda N., Gubler J.R., Rytting C., Wingate D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3): 337-351.

Azzadina I., Huda A.N., Sianipar C.P.M. (2012). Understanding relationship between personality types, Marketing-mix factors, and purchasing decisions. *Procedia-Social and Behavioral Sciences*, *65*: 352-357.

Bisbee J., Clinton J.D., Dorff C., Kenkel B., Larson J.M. (2024). Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, *32*(4): 401-416.

Bodroža B., Dinić B.M., Bojić L. (2024). Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, *11*(10), 240180.

Brown T., Mann B. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*: 1877-1901.

Dillion D., Tandon N., Gu Y., Gray K. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences*, *27*(7): 597-600.

Fan L., Li L., Ma Z., Lee S., Yu H., Hemphill L. (2024). A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, *15*(5): 1-25.

Gmyrek P., Berg J., Bescond D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality. *ILO working paper*, *96*.

Mökander J., Schroeder R. (2022). AI and social theory. AI & SOCIETY, 37(4): 1337-1351.

Patrinos G.P., Sarhangi N., Sarrami B., Khodayari N., Larijani B., Hasanzad M. (2023). Using ChatGPT to predict the future of personalized medicine. *The Pharmacogenomics Journal*, *23*(6): 178-184.

Rafikova A., Voronin A. (2024). AI as a Research Proxy: Navigating the New Frontier of Social Science Inquiry through Language Models.

Vidali M. (2024). Intelligenza Artificiale in Medicina: implicazioni e applicazioni, sfide e opportunità. *biochimica clinica*, *48*(2): 129.

The algorithm as judge: predictive justice, decision-making power and digital inequalities

by *Vera Kopsaj*[*], *Sara Sbaragli*[**]

Artificial intelligence is redefining judicial reasoning, discretion and legitimacy. Predictive justice tools influence risk assessment and sentencing, raising questions about fairness and accountability. This article takes a sociological perspective to examine the implications of algorithmic decision-making in justice. Drawing on legal theory, sociology of law and critical algorithm studies, we examine artificial discretion, opacity, and the reproduction of inequality through data. Based on models that consider algorithms as socio-technical actors, we examine how they redistribute power within digital institutions. We also address the augmented justice approach, which combines human judgement with algorithmic support. Our central question is how AI is transforming judicial decision-making in terms of discretion, legitimacy and justice.

Keywords: Justice systems; predictive justice; theory of law; sociology of law; algorithmic governance; AI and law.

L'algoritmo come giudice: giustizia predittiva, potere decisionale e disuguaglianze digitali

L'adozione dell'intelligenza artificiale nei sistemi giudiziari sta trasformando il ragionamento giuridico, la discrezionalità e la legittimità delle decisioni. Gli strumenti di giustizia predittiva, basati su algoritmi di apprendimento automatico, influenzano sempre più le valutazioni del rischio e le sentenze, sollevando interrogativi su equità e responsabilità. Questo articolo, da una prospettiva sociologica, analizza le implicazioni epistemologiche, etiche e politiche del *decision-making* algoritmico, esaminando discrezionalità artificiale, opacità del calcolo e riproduzione delle disuguaglianze. Si propone infine un quadro che interpreta gli algoritmi come attori socio-tecnici che codificano e ridistribuiscono il potere nelle istituzioni digitali. La domanda di ricerca è: quali trasformazioni introduce l'IA nel processo decisionale giudiziario in termini di discrezionalità, legittimità e giustizia sociale?

Parole chiave: sistemi giudiziari; giustizia predittiva; teoria del diritto; sociologia del diritto; governance algoritmica; IA e diritto.

DOI: 10.5281/zenodo.17297567

[*] UniCamillus - Saint Camillus International University of Health and Medical Sciences. vera.kopsaj@unicamillus.org.

[**] Università di Napoli "Federico II". sarasbaragli@gmail.com.

Introduction: From human judgment to algorithmic governance

AI is transforming the way legal decisions are made and justified. Systems such as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in the US, used to predict the risk of reoffending, or Estonia's pilot programmes involving digital judges, signal the rise of algorithmic governance in the courts.

The COMPAS case shows how algorithmic tools encode uncertainty and bias. The Core Report of the COMPAS-R (2022) shows a predictive accuracy of approximately 0.68 AUC − only moderately better than the case − and significant disparities in performance between demographic groups (e.g. black women versus white women). Furthermore, research (Engel, Linhardt, Schubert, 2025) suggests that COMPAS structurally biases outcomes against defendants, promoting a precautionary rationale for detention, often reversing the presumption of innocence. Judges using COMPAS may prefer to detain people who are unlikely to commit a crime rather than risk releasing someone flagged as high risk, even though such predictions do not have strong statistical reliability. These critical issues are exacerbated by the proprietary and opaque nature of the tool, which limits public scrutiny and raises questions of democratic legitimacy.

This reflects a broader shift: decision-making is increasingly being entrusted to automated systems. Predictive justice is a socio-technical change that redefines how decisions are made, by whom, and under what conditions. As Bowker and Star (1999) remind us, classification systems are not neutral, but shape social life, define institutional boundaries, and assign visibility or invisibility to individuals and categories.

From the perspective of legal philosophy, this raises critical questions about the nature of judgement, interpretation and moral reasoning. Tuzet (2020) provocatively asks whether the algorithm becomes a "shepherd" of the judge. If discretion is historically understood as a space for interpretation and moral reasoning, algorithmic systems compress this space within probabilistic models, pre-structuring choices and expectations.

The delegation of evaluative authority to AI systems poses epistemic challenges: on what basis should we trust the results of systems we cannot fully understand? Legal actors often lack the technical expertise to assess the validity of machine-generated risk scores, which further complicates the dynamics of responsibility. As Mezza (2018) points out, the power of computation tends to create a cognitive shortcut, a perceived objectivity

that obscures the incorporation of algorithms into social values, data policy and institutional priorities.

Moreover, algorithmic tools are being introduced in a context characterised by pressures for efficiency, cost reduction, and standardisation. These pressures make AI attractive to politicians and judicial administrators, especially in overburdened systems. However, as Sartori (2025) warns, we must remain vigilant against the enchantment of technology, adopting instead a critical attitude rooted in sociological disenchantment.

This evolution requires a profound rethinking of the very concept of justice. Is justice merely the accurate classification of cases, or is it an inherently interpretative, situated and human activity? The following sections explore these dilemmas, drawing on interdisciplinary studies to analyse the sociological consequences of the algorithm as judge.

1. Predictive justice and the delegation of responsibility

Predictive justice is based on machine learning trained on large datasets of past legal cases. The aim is to generate probability scores: the likelihood of reoffending, the level of risk or the probability of success in court. Although these tools claim to improve efficiency and objectivity, their epistemic basis is fragile and they often struggle to strike a balance between predictive accuracy and fairness, causing disparate impacts among protected groups (Barocas, Selbst, 2016). As Rundo and Di Stallo (2019) note, the algorithmic model is only as good as the data from which it learns.

Donati (2020) frames this shift as an evolution towards second-generation justice, in which human judgment is progressively replaced by calculated suggestions. Cominelli (2025) introduces the notion of "artificial discretion": a hybrid logic in which the human role is maintained but strongly conditioned by the outputs of opaque systems. Judges thus find themselves in a paradoxical position: they have to answer for decisions partially made by others.

Further evidence from COMPAS underlines these concerns. According to the COMPAS-R Core Report (2022), risk scores are derived from variables such as previous arrests, age and socio-economic background, factors that may indirectly encode systemic bias. The reported accuracy (AUC around 0.68) highlights not only the limited predictive ability, but also the inherent uncertainty of these scores. Engel, Linhardt and Schubert

(2025) argue that this uncertainty is rarely communicated transparently to judges, which may lead them to overestimate the reliability of algorithm results.

This change makes accountability less clear and weakens legal reliability. Judges often rely on algorithmic risk scores because they have become institutionalised.

Moreover, the probabilistic nature of predictive instruments introduces a structural ambiguity. A 70% risk is not a fact, but a deduction, uncertain and subject to interpretation. Nevertheless, these tools often frame such results as authoritative or objective, fostering an epistemic conflict between probability and certainty in judicial decision-making (Rundo, Di Stallo, 2019; Donati, 2020).

This confusion between probability and certainty is amplified by the presentation of risk scores without sufficient explanation of their margin of error or bias. The lack of transparency, combined with institutional pressures to adopt such instruments, further limits the ability of judges to exercise informed discretion.

This ambiguity intersects with another key issue, that is the reduction of individuals to data profiles. The complexity of human behaviour is flattened into measurable variables, which are then fed into systems that produce predictions based on historical correlations. As Bevilacqua (2025) points out, this process of abstraction risks dehumanising justice, making it less sensitive to the singularities of each case.

In practice, the use of predictive tools may shift the logic of judgment from interpretation to classification, from deliberation to ranking. This reorientation may inadvertently promote a logic of punishment less concerned with individual culpability and more with actuarial risk management. This shift reflects broader transformations in contemporary governance, where the language of safety, efficiency and risk replaces that of rights, justice and equity.

Ultimately, predictive justice involves not only a technical delegation, but a normative transformation. It reconfigures what counts as relevant information, who is authorised to interpret it, and how decisions are legitimised.

Below, we explore the dynamics of predictive justice, examining how algorithmic systems operate as socio-technical actors and how their institutionalisation reshapes the distribution of power, trust and responsibility within the justice system.

2. A sociological perspective on algorithmic judgment

From a sociological perspective, algorithms should not be seen as neutral tools, but as actors embedded within socio-technical assemblages. As Artieri (2020) suggests, algorithms are cultural agents that classify, normalise and simplify social reality. They not only reflect social structures; they also produce them. The use of AI in the justice system is not an isolated technical update, but part of a broader institutional change in which power, trust and legitimacy are being reconfigured through technology.

Bevilacqua (2025) introduces the idea of digital normativity: a shift in the production of legal meaning from textual interpretation to data-driven modeling. This change involves the gradual replacement of language, precedents and hermeneutics with statistical correlation and pattern recognition. In this context, the authority of the algorithm emerges not from argumentative rigour or democratic deliberation, but from its metric performance, scalability and institutional approval.

The sociological relevance of algorithms lies precisely in their double nature: both technological artifacts and institutional scripts. They enact regimes of visibility and invisibility by determining what data matter, which characteristics are prioritised and which categories become actionable, what Fourcade and Healy (2024) describe as a logic of "ordinal society", in which people are classified, evaluated and governed accordingly. As Finco (2024) argues, this new communicative rationality shifts the focus from shared meanings to automated decision flows. The court, once a space of symbolic mediation, risks becoming a platform for procedural optimisation.

Joyce and Cruz (2024) emphasise the concept of data justice, highlighting the moral and political implications of decisions based on potentially partial or incomplete data sets. When trained on biased data, predictive tools risk perpetuating injustice, which Aanestad *et al*. (2021) describe as a form of digital inequality, where fairness and equal treatment depend on the quality and representativeness of the available data. There is also the challenge of opacity: algorithmic decisions often lack the transparency and accountability found in traditional processes, with proprietary systems limiting public scrutiny. Even experts can find it difficult to understand or contest the way risk scores are generated. In summary, the sociological analysis of algorithmic judgement reveals its performative nature, as algorithms reshape institutional justice logic and

intersect with systemic inequalities to create new forms of exclusion and reinforce existing power structures.

3. Algorithmic bias and reproduction of inequalities

Algorithmic systems used in judicial contexts are only as neutral as the data on which they are trained. This fundamental limitation is well documented by empirical studies that show how predictive tools often perpetuate structural inequalities rather than eliminate them. Eubanks (2012), in her groundbreaking work on automated welfare systems, demonstrates how data-driven technologies disproportionately penalise marginalised populations, particularly low-income and racially diverse communities. Similar dynamics are at work in the judicial system.

Predictive algorithms rely heavily on historical data − arrests, sentencing patterns, demographic trends − that are themselves the product of systemic biases. As Sloane (2019) argues, inequality is not an unfortunate by-product of algorithmic reasoning, but rather a constitutive feature: "inequality is the name of the game". Algorithms trained on models of excessive surveillance or harsher penalties for specific groups inherit and reproduce those same discriminatory patterns. Their results may appear objective, but they are built on distorted foundations.

This becomes more problematic when feedback loops occur. For example, a risk assessment tool that flags individuals in certain zip codes as high risk may lead to increased surveillance and arrest rates in those areas. These new arrests feed back into the dataset, reinforcing the initial bias and creating a self-fulfilling prophecy. As Desrosières (1998) points out, quantification shapes the realities it claims to measure, incorporating institutional assumptions into classifications.

Lazar and Stone (2024) develop a theory of predictive justice, arguing that unequal predictive performance between structurally disadvantaged and advantaged groups constitutes a form of moral error, regardless of outcomes. Predictive models trained on historically unfair data risk representing certain populations epistemically incorrectly, thereby reinforcing structural inequality. This concern echoes what Galli and Sartor (2023) describe as the "digitalisation of deviance", whereby algorithmic classifications redefine the boundaries of suspicion and control.

Furthermore, the impact is not evenly distributed. Already marginalised groups are the most exposed to algorithmic decisions. Queudot and Meurs

(2018) warn against "differential visibility": those who are most visible in data sets are the most controlled and punished, while privileged groups often benefit from algorithmic invisibility, as their data is less collected or used.

This reproduction of inequality is not simply technical, but reflects socio-political decisions. As Joyce and Cruz (2024) argue, data justice requires addressing the power dynamics behind data collection and use. A critical sociology of predictive justice must go beyond accuracy or fairness to examine the broader social impacts of delegating decisions to systems that amplify inequality.

This also raises urgent normative questions: Can a system be fair if it consistently produces unfair results? What forms of control or resistance are available to those affected? These questions point to the need to rethink the design, governance and accountability structures of AI in the legal sphere.

The next section explores possible frameworks for resisting algorithmic domination and claiming a human-centered vision of justice.

4. Disenchantment, trust and human supervision

The growing role of artificial intelligence in judicial systems presents a paradox: while intended to increase efficiency and fairness, algorithmic tools often generate opacity, alienation and distrust. This calls for renewed attention to human oversight and institutional accountability.

Sartori (2025) calls for "sociological disenchantment", a methodological position that resists technological utopianism and emphasises the constructed nature of AI systems. Disenchantment is not a rejection of technology, but an invitation to question its assumptions, limitations and effects. When algorithms mediate decisions relating to liberty or punishment, they must be subject to ethical scrutiny just like human actors.

An urgent challenge is the erosion of transparency. Algorithmic decisions often lack intelligibility, making it difficult for defendants, lawyers, or even judges to understand how a score or classification was generated. This opacity threatens procedural justice and undermines public trust. As Barberis (2022) reminds us, justice must not only be done, but also seen to be done: decision-making must remain interpretable and contestable.

Barberis proposes a middle way: imagining algorithms as *auxiliaries* rather than *substitutes* for judicial reasoning. According to this model, AI provides analytical support without undermining the centrality of human discretion. To ensure this balance, institutions must implement robust oversight mechanisms, including algorithmic auditing, transparency standards and participatory design processes. These mechanisms can help restore a sense of agency to legal practitioners and affected individuals.

Education is also crucial. Mangone, Martini and Volterrani (2025) emphasise the importance of new training programmes that equip lawyers, policymakers and citizens with the critical skills needed to interact with algorithmic systems. This means not only technical literacy, but also ethical reasoning, legal imagination and sociological insight.

Trust in justice cannot be programmed, but must be earned through responsability and human engagement. This implies redefining professional roles in hybrid environments where legal reasoning and algorithmic results coexist, but are not confused with one another.

Ultimately, disenchantment is not cynicism, but responsibility. It affirms the need for of a human-centered approach to justice that recognises the potential of AI, while insisting on the irreplaceable value of human judgment. The next section explores how such an approach might be realised in the context of augmented justice.

5. Toward augmented justice?

The risks described above highlight the need to rethink how artificial intelligence is integrated into judicial institutions. Rather than adopting a binary view – man versus machine, discretion versus automation – we should adopt the existing framework of augmented justice, which emphasises the complementarity between humans and AI in legal decision-making.

Picierno (2025) calls for a "hybrid legal culture" in which judges are supported, not replaced, by computational systems. This vision requires infrastructures that enable collaboration under conditions of transparency, contestability and ethical control. Augmented justice is based on interpretability, contextualisation and a commitment to prioritise justice over efficiency.

Galli and Sartor (2023) outline models of judicial empowerment: diagnostic, advisory, and decision support. The most effective uses of AI

lie in the first two: providing insights into case patterns, relevant precedents, or inconsistencies, while the final interpretation remains the responsibility of the judge. This preserves discretion, empathy, and normative judgement.

Furthermore, augmented justice requires new forms of participatory governance. Citizens, legal practitioners, civil society organizations and technical experts must be involved in the design, evaluation and regulation of AI systems used in the justice sector. This is in line with democratic ideals and ensures that algorithmic systems remain accountable to the public they serve. As Dyson (1997) argued decades ago, technology must be embedded in social justice structures: it is not inherently liberating or oppressive, but is shaped by the way we choose to govern it.

Augmented justice is not a technical solution to a technical problem. It is a political and normative project that recognises the complexity of human judgement and the social embeddedness of legal institutions. Queudot and Meurs (2018) warn against techno-solutionism in legal innovation, reminding us that algorithms cannot resolve the deeply rooted inequalities and ambiguities that characterise legal practice. Rather than eliminating discretion, augmented justice reaffirms it as a locus of human responsibility.

In this sense, the future of justice is not algorithmic, but augmented: informed by data, assisted by computation, but ultimately implemented by human beings. The ethical imperative is to design legal technologies that enhance, rather than diminish, the quality of judgement and to foster institutional ecosystems capable of critically engaging with the tools they use.

Conclusion

The spread of artificial intelligence in judicial systems marks an important institutional change. Predictive justice redefines discretion and legitimacy and can exacerbate inequalities when personal dignity is ignored. Algorithms act as socio-technical agents: they classify, redistribute authority and can amplify existing biases.

We have demonstrated how opacity, probabilistic logic and data-driven reasoning affect fairness and legitimacy. Without critical scrutiny, such systems risk undermining procedural justice and public trust.

We build on the augmented justice model, a framework that envisions hybrid systems where AI supports, not replaces, human judgment. It

prioritises transparency, ethical design, education and participatory governance to ensure discretion and accountability.

Future research may explore tools such as COMPAS and how to access fairness. Surveys of judicial actors could further clarify how these systems are understood and used in practice. Empirical investigation will be key to refining augmented justice and promoting socially responsible innovation.

References

Aanestad M., Kankanhalli A., Maruping L., Pang M.S., Ram S. (2021). Digital technologies and social justice. *MIS quarterly*, *17*(3): 515-536.

Artieri G.B. (2020). Fare Sociologia attraverso l'algoritmo: potere, cultura e agency. *Sociologia italiana*, (15).

Ayuda F.G. (2024). Algorithms, sociology of law and justice. *Journal of Digital Technologies and Law*, *2*(1): 34-45.

Barberis M. (2022). Giustizia predittiva: ausiliare e sostitutiva. Un approccio evolutivo. *Milan Law Review*, *3*(2):1-18.

Barocas S., Selbst A.D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bevilacqua S.A. (2025). Gli strumenti di intelligenza artificiale nel sistema giudiziario: verso una nuova normatività digitale. *Sociologia del diritto*, 52(1): 257-274.

Bowker G.C., Star S.L. (1999). *Sorting things out: Classification and its consequences*. The MIT Press.

Cominelli L. (2025). Discrezionalità artificiale e giudizio algoritmico: Artificial discretion and algorithmic judgment. *Sociologia del diritto*, 52(1): 242-255.

Cordano M. (2025). Intelligenza artificiale e giustizia: una valutazione critica del caso Loomis vs. Wisconsin.

Desrosières A. (1998). *The politics of large numbers: A history of statistical reasoning*. Cambridge: Harvard University Press.

Donati F. (2020). Intelligenza artificiale e giustizia. *RIVISTA AIC* (415-436).

Dyson F. (1997). Technology and social justice. *Carnigie Concil on Ethics and International Affairs* (7-25).

Engel C., Linhardt L., Schubert M. (2025). Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism. *Artif Intell Law* , 33: 383-404. https://doi.org/10.1007/s10506-024-09389-8

Eubanks V. (2012). *Digital dead end: Fighting for social justice in the information age*. Cambridge: MIT Press.

Finco M. (2024). Dall'Intelligenza alla Comunicazione artificiale? Sociologia e possibilità teoriche. *indiscipline-rivista di scienze sociali*, *4*(2): 132-140.

Fourcade M., Healy K. (2024). *The ordinal society*. Cambridge: Harvard University Press-T.

Galli F., Sartor G. (2023). Ai approaches to predictive justice: A critical assessment. *Humanities and Rights Global Network Journal*, *5*(2).

Joyce K., Cruz, T.M. (2024). A sociology of artificial intelligence: Inequalities, power, and data justice. *Socius*, *10*, 23780231241275393.

Lazar S., Stone J. (2024). On the site of predictive justice. *Noûs*, *58*(3): 730-754.

Mangone E., Martini E., Volterrani A. (2025). *Società frammentata e traiettorie educative: Disuguaglianze, giustizia sociale e intelligenza artificiale*. Milano: Mimesis.

Mezza M. (2018). *Algoritmi di libertà: la potenza del calcolo tra dominio e conflitto*. Roma: Donzelli Editore.

Northpointe Inc. (2022). *COMPAS-R Core: The revised version of the standard COMPAS Core (Technical manual).* Research Department, Northpointe, Inc. https://github.com/franzgualdi/PA-Algorithmic-formalization/blob/main/COM PAS-R%20Core%20Research%20Report%203%2028%2022%20(1).pdf

Picierno B. (2025). Giustizia predittiva ed intelligenza artificiale: nuovi scenari per il giurista del futuro. *Futuri*, *23*: 361-372.

Queudot M., Meurs M.J. (2018). Artificial intelligence and predictive justice: Limitations and perspectives. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (889-897). Cham: Springer International Publishing.

Rundo F., Di Stallo A.L. (2019). Giustizia predittiva: algoritmi e deep-learning. *Sicurezza e Giustizia*: 31-34.

Sartori L. (2025). Conoscere e governare la tecnologia per essere più disincantati di fronte all'intelligenza artificiale. *Rassegna Italiana di Sociologia*, *66*(2): 397-405.

Sloane M. (2019). Inequality is the name of the game: thoughts on the emerging field of technology, ethics and social justice. *Weizenbaum conference*. DEU.

Tuzet G. (2020). L'algoritmo come pastore del giudice? Diritto, tecnologie, prova scientifica. *Media Laws*, (1): 45-55.

Sicurezza e benessere del cittadino nell'era digitale. Una lettura sociologica

di *Ida Cortoni*[*]

Il saggio analizza il concetto di sicurezza e benessere sociale nell'era digitale, con un focus sulla percezione degli adolescenti italiani. Attraverso una prospettiva sociologica, si esplorano le trasformazioni del benessere da dimensione pubblica e materiale a privata e immateriale, influenzata dalla platform society. Viene presentata una ricerca empirica condotta su 2708 adolescenti, che evidenzia due profili comportamentali: i "virtuosi digitali" e i "virtuosi sociali", con differenti livelli di consapevolezza e tutela del benessere digitale.

Parole chiave: sicurezza; benessere sociale; platform society; adolescenti; consapevolezza digitale; capitale sociale.

Citizen safety and wellbeing in the digital age. A sociological perspective

This essay explores the concept of security and social well-being in the digital age, focusing on how Italian adolescents perceive and manage it. From a sociological perspective, it examines the shift from public, material well-being to private, immaterial forms shaped by the platform society. An empirical study involving 2,708 adolescents reveals two behavioral profiles: "digital virtuosos" and "social virtuosos," each showing different levels of awareness and strategies for protecting digital well-being.

Keywords: safety; social welfare; platform society; adolescents; digital awareness; social capital.

Introduzione

Il presente saggio focalizza l'attenzione sul concetto di sicurezza e benessere sociale nella sua multidimensionalità semantica, analizzando la sua rilettura nell'era della *platform society* (van Dijck, 2013) e verificandone il tipo e il grado di percezione sociale attraverso l'analisi dei comportamenti culturali degli adolescenti italiani.

DOI: 10.5281/zenodo.17297576

[*] Sapienza, Università di Roma. ida.cortoni@uniroma1.it.

L'UNICEF nel 2007 considera il benessere come: «salute e incolumità, sicurezza materiale, istruzione e socializzazione, il senso di essere amati, stimati e integrati nelle famiglie e nelle società in cui sono nati» (Unicef, 2007: 1).

La questione del benessere sociale e individuale del cittadino diventa particolarmente delicata e controversa nell'era digitale, in cui emergono nuove architetture e sistemi informativi che scambiano i beni materiali (dati e metadati) e immateriali (i processi simbolici sottesi all'uso delle informazioni online) dei cittadini con rischi elevati di violazione dei diritti individuali e di rottura degli equilibri economici, politici e culturali alla base degli scambi commerciali e ideologici in Rete fra piattaforme. La compromissione della sicurezza pubblica e privata, a causa della violazione dei diritti civili dei cittadini e dei principi della democrazia, rischia di minacciare la percezione del benessere individuale e sociale nell'era digitale.

La domanda di ricerca che ci si pone è, dunque, la seguente: cosa può significare tutelare la sicurezza e il benessere del cittadino all'interno delle dinamiche di interazione costruite quotidianamente online? Quali sono le responsabilità etiche e morali delle corporation che gestiscono i loro business online e come autoregolare le implicazioni culturali, ideologiche e politiche legate allo scambio di dati negli ambienti virtuali? Quale è il grado di consapevolezza digitale del cittadino che condivide le informazioni online?

L'ipotesi conoscitiva alla base della struttura teorico concettuale del saggio è che la questione della sicurezza e del benessere, anche digitale, può essere analizzata e verificata dal punto di vista sociologico, partendo da diverse prospettive di analisi: quella tipica della società industriale, riconducibile alle politiche di gestione della sicurezza e del benessere, che chiama in causa la responsabilità dei proprietari e dei gestori delle piattaforme tecnologiche, e quella riconducibile al comportamento responsabile del cittadino quando condivide materiali online o interagisce all'interno degli ambienti virtuali. In altri termini per inquadrare il tema della sicurezza e del benessere digitale in questa duplice lettura bisogna ripartire e considerare sia le categorie sociologiche di consenso e controllo sociale alla base della società del benessere del Welfare State, secondo una chiave di lettura macro sociale, sia le variabili alla base del capitale sociale (con particolare riferimento all'approccio delle capabilities di Sen e Nussbaum), quali strumenti etici, valoriali e simbolici per costruire forme di sicurezza e benessere, di tipo immateriale e privato in prospettiva micro sociale in epoca postmoderna.

In tal senso, il concetto di benessere in ambito sociologico rimanda a una condizione di sicurezza psicologico-emotiva e fisico-sociale del cittadino, che

riflette una situazione socioculturale di consenso e controllo sociale[1]. Con tale termine, Boudon e Bourricaud intendono l'insieme delle risorse materiali e simboliche (i cosiddetti beni) di cui una società dispone per assicurare la conformità dei suoi membri alle regole e principi prescritti (da modelli culturali ed etici appresi) (Boudon, Bourricaud, 1991). Il grado di affidabilità di un ambiente sociale all'interno di una struttura, dunque, è direttamente proporzionale alla percezione pubblica di sicurezza e benessere da parte dei suoi cittadini, che accettano le obbligazioni imposte dalla struttura sociale in virtù del grado di equilibrio e consenso garantito (Iannone, 2006).

L'espressione "società del benessere" richiama lo Stato sociale fra gli anni '50 e '60 del XXI secolo che interviene attraverso politiche di welfare, in ambito economico, professionale e socioculturale, per ridurre la percezione dei gap socioculturali alla base della stratificazione sociale[2].

Secondo il sociologo inglese Thomas Humphrey Marshall negli anni '50 emerge progressivamente la cosiddetta *cittadinanza sociale,* che richiama l'insieme di diritti economici e sociali che permettono agli individui di ceto medio di essere maggiormente integrati nella comunità politica, aumentando il consenso. L'espressione Welfare State, così, indica l'insieme delle istituzioni, delle norme giuridiche, degli attori e delle politiche pubbliche messe in atto dallo Stato per garantire coesione riducendo forme di disuguaglianza sociale (da qui il concetto di benessere sociale pubblico in prospettiva macro sociale)[3].

A partire dagli anni '70 e '80, progressivamente si fa strada l'idea della nuova società del post (post industriale e post moderna) che, secondo alcuni studiosi (Jameson, 1989), rappresenta una precisa fase storica del tardo capitalismo caratterizzata da una smaterializzazione dei rapporti sociali, dalla smobilitazione ideologica, dalla frammentazione del soggetto, il venir meno del senso di storicità, l'ascesa di un populismo estetico, che annulla le distanze fra cultura alta e bassa anche attraverso il riconoscimento e la legittimazione dei prodotti dell'industria culturale. Alla crisi del Welfare State si affianca la necessità di ricostruire

[1] Il controllo e consenso sociale fanno riferimento a meccanismi che mettono in gioco sia le iniziative e le risorse degli individui che i vincoli collettivi e gli obblighi morali.

[2] Quest'ultima si afferma con l'avvento della società moderna ed è strettamente collegata al significato di "classe sociale", intesa come gruppo di persone che condivide una stessa condizione economica, connessa alla struttura occupazionale e alla divisione del lavoro della società industriale (Croteau, Hoynes, 2018).

[3] Le politiche pubbliche attivate dai governi possono basarsi su due principi: quello socialdemocratico dell'equità sociale che riduce il gap fra ricchi e poveri; in questo specifico caso, a tutti i cittadini sono forniti servizi di base come l'assistenza sanitaria e l'istruzione gratuita attenuando l'impatto delle differenze fra i redditi (Welfare State); quello liberal democratico che promuove le pari opportunità attraverso un processo di livellamento verso il basso delle carriere (Portes, 1998) incentrato sulla mobilità individuale.

nelle cerchie sociali di appartenenza tutto quello che prima era stato garantito dall'alto. I reticoli costituivi del capitale sociale nascono proprio dal bisogno di reciprocità e di dipendenza degli individui gli uni dagli altri (Iannone, 2016).

Nel postmoderno, la produzione di beni materiali è soppiantata da quella dei servizi, l'esaltazione della quantità produttiva, connessa al valore d'uso, lascia spazio alla dimensione qualitativa del bene, ovvero quella di legame e di riconoscimento simbolico individuale, alla capacità di espressione del gusto estetico, dello stile di vita e dei mondi immaginativi possibili, alla base dei processi di identificazione e di riconoscimento degli individui (Rullani, 2005; Fabris, 2008). Il bene, così, nel post non è più esaltato per le sue prestazioni funzionali per il benessere sociale collettivo, bensì per le opportunità di significazione e costruzione semantica del sé offerte al soggetto (da qui il concetto di benessere sociale immateriale in prospettiva microsociale).

In questa nuova configurazione sociale, *economia della conoscenza* (Rullani, 2005; Ranieri, 2006) è l'espressione che enfatizza l'avvento di un capitalismo cognitivo, fondato sul sapere come forma di investimento per un progetto di cittadinanza e un senso democratico di convivenza. Attraverso il paradigma tecnoeconomico dell'informazione di Castells (2001), quale chiave di lettura della nuova cornice socioculturale, l'informazione acquista progressivamente un ruolo centrale poiché rappresenta il bene primario alla base del progresso scientifico, mentre il *network* diventa la struttura portante per lo sviluppo dell'informazione, poiché promuove una connessione interdipendente dei diversi apparati sociali: dall'economia alla politica, dalla religione all'istruzione, fino ad arrivare alla stessa comunicazione.

La Rete presuppone una condizione di continuo dinamismo e mutamento dei suoi legami che consentono non solo di ampliare o ridurre l'estensione delle sue maglie, ma anche di mutarne la forma o di cambiarne l'orientamento. Questa continua metamorfosi costruisce una condizione socioculturale poco statica e, dunque, difficile da etichettare all'interno di specifiche categorie semantiche (Castells, 2004).

I cambiamenti culturali e sociali del postmoderno incidono sulle nuove interpretazioni sociali: il declino del lavoro industriale, la flessibilizzazione, la risemantizzazione del "tempo libero" (Morcellini, Mazza, 2008) e il consumismo prorompente (Carmagnola, Ferraresi, 1999; Featherstone, 1995), l'emergere di una nuova etica nelle relazioni sociali, la deistituzionalizzazione della politica, quella amministrativa e, soprattutto, economica (Giaccardi, Magatti, 2003) e il ruolo pervasivo dei mezzi di comunicazione.

1. Inquadramento teorico concettuale

Dal punto di vista delle Scienze Sociali, nella progressiva configurazione sociale postmoderna, l'enunciazione di benessere racchiude diverse categorie concettuali sociologiche che si intrecciano oscillando da una prospettiva sociale macro ad una micro e contribuendo a definire la natura delle relazioni sociali alla base del capitale sociale. Secondo *l'approccio delle capabilities* di Amartya Sen (1999) e Martha Nussbaum (2000)[4], infatti, il benessere sociale dipende dalla combinazione di molteplici variabili, alcune di natura micro sociale riconducibili alle cosiddette *capabilities interne*[5] dell'individuo, ovvero alle caratteristiche biologiche che condizionano le necessità socioculturali di beni materiali o immateriali; esse si pongono alla base della costruzione dell'identità e della personalità di ciascuno. Altre variabili fanno riferimento alla dimensione macro sociale del benessere e rimandano ad alcune caratteristiche dell'area territoriale, entro cui il cittadino agisce quotidianamente, che condizionano la manifestazione delle *capabilities fondamentali*, legate all'orientamento socioculturale del cittadino, e delle *capabilities combinate*[6], condizionate dal capitale sociale, economico e culturale delle principali agenzie di socializzazione.

Tali variabili macro sociali fanno riferimento ad esempio alle *diversità ambientali*, ovvero le caratteristiche geografiche e climatiche territoriali del contesto in cui il cittadino vive, che condizionano la qualità della vita di ciascuno; *le variazioni del clima sociale,* ovvero il capitale socioculturale del territorio, che condiziona la qualità dell'erogazione di alcuni servizi, come ad esempio la scuola, l'assistenza sanitaria, i servizi formativi, le infrastrutture economiche e i servizi per lo svolgimento dell'attività sportiva. In tal senso, il grado di diffusione della sicurezza sociale, il clima culturale del quartiere, il grado di importanza riconosciuto al sistema di istruzione, la libertà politica, la dotazione

[4] L'approccio delle *capabilities* individuali punta sulla crescita individuale, sulla considerazione della persona come fine, capace di agire e degna di rispetto (Nussbaum, 2000). La Capability è la nozione chiave posta da Sen alla base della sua concezione di sviluppo che contrasta con una visione della crescita misurata dall'aumento del prodotto Interno lordo (PIL) dei diversi Paesi. Il concetto viene adottato nell'ambito degli *Human Developments Reports*, il "Programma di sviluppo delle Nazioni Unite": lo *Human Development Index (HDI*). Attraverso tale indice viene misurato il livello di sviluppo dei Paesi a livello mondiale tenendo conto di tre dimensioni critiche dello sviluppo umano: 1. l'aspettativa di una vita lunga e in salute; 2. Il livello di educazione; 3. L'accesso alle risorse sufficienti per acquisire uno standard di vita decente (Sen, 2003).

[5] Le *capabilities* interne sono stadi di sviluppo della persona per lo svolgimento delle funzioni (Nussbaum, 2000).

[6] Le *capabilities* combinate sono capacità interne combinate con condizioni esterne adatte a esercitare quella funzione (Nussbaum, 2000).

tecnologica pubblica rappresentano tutti fattori che incidono sul concetto di qualità della vita e dunque sulla percezione della sicurezza e del benessere sociale del cittadino. Infine altre variabili alla base della costruzione del benessere del cittadino sono riconducibili alle relazioni socioculturali e alle forme di investimento materiale e immateriale all'interno degli ambienti di socializzazione, in cui prendono forma comportamenti, abitudini, visioni di vita, etc. Rientrano in questa terza categoria: *le differenze del capitale culturale* ascritto e acquisito dalle relazioni interpersonali che condizionano i modelli di comportamento pubblici in contesti sociali circoscritti, tenendo conto delle convenzioni e dei costumi condivisi. In sintesi, il concetto di benessere è frutto della correlazione di variabili di diversa natura, la cui combinazione è in grado di definire il tipo e il livello di benessere del cittadino (Sen, 1999).

In tutti i casi, la condizione di benessere è influenzata da determinanti "di tipo materiale", quali ad esempio la disponibilità di beni e servizi, e da determinanti "di tipo immateriale", quali ad esempio la condivisione di un humus etico e valoriale, nonché quel capitale umano inteso come le competenze necessarie per ottimizzare l'uso delle stesse risorse, che contribuiscono a definire il tipo di benessere socioculturale del cittadino (Barbero Vignola *et al.,* 2016).

Se da un lato la disponibilità di beni materiali, frutto di investimenti infrastrutturali ed economici sia territoriali che famigliari, contribuisce a favorire lo sviluppo di una condizione di benessere nel cittadino, in virtù delle opportunità di appropriazione del bene come elemento facilitatore della integrazione socioculturale, dall'altro ci sono altre variabili di tipo immateriale, le cosiddette *capabilities (fondamentali*[7] e *interne* dell'individuo – Nussbaum, 2000), che contribuiscono all'implementazione della condizione di benessere socioculturale, poiché aiutano il cittadino a orientare scelte per migliorare la propria qualità della vita, a investire adeguatamente sul capitale economico disponibile e su quello materiale per soddisfare i propri bisogni di benessere, di realizzazione e di socializzazione.

Secondo Portes (1998), poi, la consapevolezza di vivere in una società dove si ha un comune e reciproco rispetto delle regole comunitarie rappresenta una risorsa per la cittadinanza, perché contribuisce a diffondere nelle persone maggiore senso di sicurezza, per non parlare di uno spirito collettivo di solidarietà (*bounded solidarity*), di sostegno reciproco e coesione, in cui il senso civico è connaturato al processo educativo di ciascuno (da qui il concetto di

[7] Le *capabilities* umane di base o fondamentali riguardano la vita, la salute e l'integrità fisica, la stimolazione del pensiero e dei sensi, dei sentimenti, lo sviluppo di un senso di appartenenza, il gioco, la giustizia istituzionale e la ragion pratica, la libertà di stabilire diverse relazioni con esseri viventi, etc. (Nussbaum, 2000).

consummatory di Portes). Tali fattori alla base del capitale sociale contribuiscono a costruire il senso del benessere sociale immateriale del cittadino in una prospettiva macro sociale.

Secondo questa lettura, è possibile riconoscere tre principali variabili del capitale sociale in grado di contribuire alla percezione della sicurezza alla base del benessere: la *closure (o il legame), il vincolo della fiducia (o la fiducia imposta) e gli scambi di reciprocità.*

La *closure* è correlata alla cosiddetta *dimensione morale* e interessa quell'aspetto di socializzazione relativo all'interiorizzazione nei cittadini di norme, valori collettivi tipici dell'humus culturale condiviso attraverso dinamiche trasmissive e relazionali. Attraverso la *closure* maturano *virtù civiche* (Sciolla, 2000), che regolano il comportamento individuale, l'espressione delle capabilities individuali e le relazioni sociali nel rispetto di norme, valori, obbligazioni e aspettative collettive (*solidarietà collettiva*).

Sulle presunte e attese obbligazioni connesse alla *closure* si costruiscono *scambi di reciprocità,* che presuppongono il riconoscimento delle identità altrui quale riflesso di aspetti del sé (l'io-specchio di Cooley) e sono alla base della costruzione e del rafforzamento delle opportunità sociali, culturali ed economiche della socialità. Il *vincolo fiduciario,* infine, getta le fondamenta per lo scambio, in quanto presuppone la condivisione dell'ethos sociale proiettato sul comportamento individuale. Il suo valore tuttavia trascende la dimensione etica e secondo alcuni studiosi, quale Putnam, diventa un "lubrificante sociale" (2000) perché rappresenta il presupposto dello sviluppo economico armonioso di una comunità fondata sulla cooperazione in cui il bene comune supera le individuali opportunità.

Per continuare questa breve analisi del benessere attraverso la lente dell'approccio delle *capabilities*, condiviso e sviluppato ulteriormente da Martha Nussbaum (2000), riproponiamo di seguito l'elenco di capacitazioni umane fondamentali alla base dello sviluppo sociale, definite in altri termini anche libertà individuali da Amartya Sen: a) avere la possibilità di vivere una vita normale; b) salute fisica; c) integrità fisica; d) capacità di utilizzo consapevole ed informato di sensi, immaginazione e pensiero attraverso un'adeguata istruzione; e) garanzia di uno sviluppo emotivo non limitato da paure eccessive; f) capacità di costruire la propria idea di bene e la riflessione critica su come programmare la propria vita; g) senso di appartenenza ovvero opportunità di vivere con gli altri e per gli altri e di avere le basi sociali per il rispetto di sé e per essere trattato come una persona dignitosa; h) opportunità di vivere in relazione con animali, piante e la natura; i) gioco; l) controllo del proprio ambiente politico e materiale, ovvero partecipare efficacemente alle scelte politiche importanti per la propria vita (giustizia istituzionale).

Queste 10 libertà individuali sintetizzano l'articolazione semantica del concetto di benessere del cittadino, riassumibile in 4 sottocategorie secondo una prospettiva microsociale: fisico, cognitivo, psicologico emotivo e sociale.

2. L'indebolimento del benessere pubblico

Negli ultimi decenni, il progetto delle "libertà individuali", per utilizzare la terminologia di Sen, sembra si sia indebolito dalla progressiva perdita di solidità strutturale delle cosiddette "libertà strumentali", ovvero l'*empowerment* di quelle istituzioni e organismi, *in primis* nazionali ma anche internazionali, che avrebbero dovuto creare i presupposti di espressione della libertà individuale di ciascuno e avrebbero dovuto tutelarne i diritti.

Il peso sociologico della violazione delle libertà individuali si manifesta in modo evidente all'interno di una cornice socioculturale contemporanea già etichettata come critica. Gli studi nel campo delle Scienze Sociali concordano nel dimostrare che la perdita di solidità delle istituzioni nella società contemporanea, seppur attraverso diverse manifestazioni, fa ormai parte della processo evolutivo che ha caratterizzato il passaggio verso la condizione postmoderna (fra gli altri Jameson, 1989; Lyotard, 1979; Maffesoli, 2005): il lamento del crollo di valori della tradizione, spesso appurato come causa della crisi della società contemporanea, in realtà rappresenta una variabile insita nell'idea di progresso e nel concetto di cambiamento.

La dimensione morale collettiva, ereditata, è vittima di un processo di allentamento già dall'avvento della società industriale. Il processo di modernizzazione ha determinato l'alterazione fisiologica dei principi di solidarietà (come approfondito dai primi sociologi, a partire da Durkheim), che si riflettono sul tipo di relazioni sociali, non più costruite su basi eticamente fiduciarie, ma improntate su principi di utilitarismo (Portes, 1998) e di individuazione (Beck, 1988).

Si indeboliscono criteri condivisi alla base dei processi di socializzazione e vacilla l'obiettivo dell'integrazione socioculturale e dell'agire per un benessere sociale pubblico, per una politica liberista improntata prevalentemente sul grado di responsabilità del singolo (benessere individuale privato) e sulla sua capacità di interiorizzare e metabolizzare la morale e la cultura da tramandare anche alle future generazioni.

Da qui si genera la diaspora fra il processo di rinnovamento sociale della solidarietà collettiva da parte di organizzazioni sociali, che tentano di combinare l'eredità strutturale e culturale passata con le spinte moderniste, e il vissuto degli individui che regolano il proprio comportamento partendo da canoni

comportamentali arbitrari e personali. In questi casi, dunque la *closure* non sempre è tradotta in comportamenti di civismo o di *advocacy* e il concetto di bene pubblico perde progressivamente di rilevanza a vantaggio di quello privato, personale e quindi strumentale.

La modernizzazione, per la sua natura mobile e fluida, ha prodotto all'interno delle strutture sociali preesistenti movimenti controculturali intellettuali e civili, che progressivamente hanno condotto alla dissoluzione delle gerarchie simboliche costruite in epoche precedenti. Tuttavia, la fluidità implicita nel cambiamento non solo ha rotto drasticamente con le condizioni storiche del passato, ma ha determinato anche frammentazioni sociali e culturali nella stessa esperienza moderna, rendendo i valori di ieri inadeguati alle richieste di oggi e i principi culturali alla base di un gruppo sociale non sempre adeguati a quelli di altri che vivono un percorso di modernizzazione situato, con accelerazioni e rallentamenti di progresso circoscritti e singolari, diffondendo quindi fra gli individui un senso di relatività culturale e sociale, nonché di continuo mutamento e di caducità[8].

La conoscibilità delle sfaccettature della complessità del reale restituisce un quadro contemporaneo in cui le manifestazioni di ciò che è inteso come "bene" in una specifica cultura si trasformano in "male" all'interno di altre. Il concetto di bene, così, si relativizza a partire dalla lente sociale e culturale di chi lo interpreta, tanto che l'unica percezione pubblica diffusa è quella di una condizione esistenziale potenzialmente manovrabile e succube di azioni o manifestazioni collettive, culturali e sociali, che gli stessi uomini hanno contribuito a fortificare, ma di cui tutti diventano potenziali vittime (Giddens, 1994). Muovendo da questi studi, la Ferrari Occhionero parla di una "sindrome culturale del postmodernismo" come "naturale" conseguenza di quel sistema di sicurezza e benessere economico e culturale, costruito dalle generazioni precedenti durante il processo di modernizzazione, che tuttavia ha generato uno stato di transitorietà permanente e, quindi, una sorta di "sindrome dell'insicurezza" perenne (Ferrari Occhionero, 2002).

Questa condizione socioculturale del *post*[9], se da un lato ha incentivato la crescita culturale, l'investimento emotivo e cognitivo su forme di benessere

[8] "Il modernismo ha cercato un rovesciamento della cultura ufficiale ottimistica della modernità con le sue ambizioni di creare un ordine unificante e integrativo", cfr. Featherstone (1998: 102). Si veda anche Harvey (1997: 25).

[9] Il termine postmoderno (Lyotard, 1979) è ormai entrato nel vocabolario dei classici della Sociologia, per indicare la nuova configurazione socioculturale contemporanea, in cui le utopie del moderno (le cosiddette *grandi narrazioni*) crollano per lasciare spazio a una progressiva frammentarietà dell'esperienza e a un dinamismo perenne, che destabilizza qualsiasi

micro sociale di tipo immateriale, dall'altro lato ha prodotto una lenta ma profonda sfiducia e disaffezione alle istituzioni, o alle autorità, e alla loro capacità di tutelare gli interessi del cittadino, che ha indotto a una valorizzazione della volontà del soggetto e della sua autonomia decisionale e di azione (Ferrari Occhionero, 2002).

A partire da inquadramento teorico concettuale, nei prossimi paragrafi il saggio intende focalizzare l'attenzione sull'analisi del concetto di benessere, nelle sue molteplici dimensioni, rapportato alle trasformazioni strutturali, etiche, economiche e relazionali della società a seguito della rapida integrazione del digitale nelle architetture sociali e nei meccanismi di produzione e di consumo dei beni materiali e immateriali nella sfera economica, politica e culturale, sia dal punto di vista macro che micro sociale. Nello specifico, a seguito di una breve analisi sulla trasformazione semantica del concetto di benessere nella società digitale, l'attenzione si focalizzerà sulla percezione del benessere digitale individuale da parte di un campione di giovani adolescenti italiani attraverso la presentazione dei risultati su una ricerca sociologica sul tema della sicurezza digitale, realizzata dalla Sapienza Università di Roma dal 2020 al 2023.

3. Il benessere del cittadino nell'era digitale

Negli ultimi anni, il tema del benessere sociale pubblico dei cittadini è connesso a quello della sicurezza, anche e soprattutto *online*, e ha acquisito un ruolo prioritario nella agenda dell'Unione europea, tanto che il presidente della Commissione Europea Jean-Claude Junker, nel suo discorso sullo stato dell'Unione del 13 settembre 2017, ha dichiarato la *sicurezza online* come la seconda emergenza in Europa, dopo il cambiamento climatico e prima dell'immigrazione.

La centralità di questo tema nell'agenda politica internazionale è strettamente correlata al fenomeno di *piattaformizzazione* che la società contemporanea ha vissuto negli ultimi anni; con tale espressione si vuole intendere un processo di ascesa delle piattaforme digitali come infrastrutture e modelli comunicativi dominanti sul web in grado di condizionare e orientare dinamiche politiche, economiche, sociali e culturali secondo una prospettiva macro sociale (Helmond, 2015). Le piattaforme sono architetture comunicative e ambienti di interazione

processo di produzione e consumo (sia esso culturale o economico). Alcuni studiosi dell'era contemporanea hanno fornito in ambito scientifico nuove interpretazioni sulla condizione sociale contemporanea, come ad esempio quella che rinvia alla "modernità incompiuta" (Giddens, 1994), alla "modernità riflessiva" (Beck, 2000) o alla "modernità liquida" (Bauman, 2002) di cui il terrorismo ne rappresenta una delle possibili conseguenze.

e scambio che, attraverso il digitale, mettono in relazione utenti, *corporations* e istituzioni pubbliche generando un ecosistema complesso in cui logiche economiche, politiche e sociali interagiscono fra di loro (van Dijck *et al.,* 2018) orientando azioni politiche. Esse prevedono tre principali meccanismi di funzionamento:

1. la datificazione, intesa come traduzione in dati delle attività e delle interazioni degli utenti online (post, attività degli utenti, comportamenti in rete degli utenti, pagamenti, registrazioni, visualizzazioni, commenti, tweet, amicizie, etc. …) in microcontesti virtuali; tali dati sono commercializzati con enti, o aziende, per finalità legate al marketing pubblicitario o politico, al fine di veicolare orientamenti commerciali, culturali e ideologici da cui possono derivare comportamenti sociali. Proprio la capacità di tracciare e controllare flussi di traffico, di opinioni e sentimenti pubblici assume un grande valore sociale, economico e politico appetibile anche per la criminalità;
2. la mercificazione (o *commodification)*, intesa come la trasformazione dei contenuti e delle emozioni condivise in Rete dagli utenti in beni o merci scambiate per generare profitto, seguendo le regole dei meccanismi dell'industria culturale;
3. la selezione, con cui si intende la tendenza delle piattaforme a indirizzare gli utenti verso specifici contenuti, idee, pensieri, oggetti e informazioni a partire da strategie tecno-commerciali, che tramite algoritmi orientano la reputazione online e definiscono i trend comportamentali di massa degli utenti (van Dijck *et al.,* 2018).

In questo scenario, aumentano i rischi delle minacce *online*, sia di natura economica che commerciale, che coinvolgono direttamente le aziende e le organizzazioni, sia di natura politica, capaci di destabilizzare governi e paesi, sia di natura individuale che riguardano il cittadino, spesso carente di competenze adeguate, in grado di avviare azioni cautelari contro il crimine e le minacce *online*.

Dunque, partendo dalla lettura critica della *platform society* di Van Dijck nel 2013, la questione della sicurezza *online* può essere discussa e analizzata a partire dalla correlazione fra quattro principali dimensioni: a) quella tecnologica; b) quella culturale; c) quella sociale; d) quella economica per ognuna delle quali è possibile analizzare rischi e forme di tutela del benessere del cittadino.

La dimensione tecnologica fa riferimento alle continue avanguardie mediali e all'architettura funzionale di gestione di un sistema digitale che definisce e trasforma il processo di digitalizzazione della società contemporanea. Il concetto di sicurezza all'interno di questa dimensione fa riferimento alla tutela delle informazioni attraverso sistemi di protezione dei *devices*, secondo una prospettiva ingegneristica, e di *design* dell'architettura nel momento in cui si progetta e si

sviluppa un ambiente virtuale integrato all'interno di un contesto sociale organizzativo (benessere materiale).

La garanzia del benessere del cittadino all'interno degli ambienti digitali, nella prospettiva macro-sociale, non può prescindere tuttavia dalla progettazione ed edificazione di un'infrastruttura etica (*infraetica*), che richiama il concetto di benessere immateriale. Quest'ultima dovrebbe garantire la *closure* e in *vincolo della fiducia* alla base delle dinamiche di *scambio* degli utenti ed essere in grado di regolare, monitorare e controllare i processi di gestione, produzione e distribuzione dei dati e dei metadati da parte delle *corporations* proprietarie, già nella fase della *progettazione*, al fine di creare condizioni di tutela strutturale delle dinamiche di relazione e condivisione delle informazioni degli stessi cittadini online. L'attivazione di un sistema etico nella *platform society* comporta il coinvolgimento e la responsabilizzazione almeno di due tipi di attori: le compagnie Tech e gli imprenditori per il benessere sociale pubblico *online*.

Il progressivo ruolo centrale assunto dai dati come pilastro strategico di *business* delle aziende *data-driven* orienta scelte di mercato, decisioni e investimenti non solo di natura economica, ma anche politica e culturale. In tal senso, quando si parla di protezione *online* si considerano anche i dati scambiati o condivisi nelle dinamiche relazionali e contrattuali fra utenti e *providers*. In questo caso dalla protezione tecnologica, l'attenzione in termini di sicurezza si focalizza sulla protezione culturale, intesa come tutela dei dati personali e sensibili alla base del capitale umano e culturale degli utenti che agiscono in rete e che si pongono alla base della identità individuale e sociale dei cittadini, non esclusivamente virtuale. In prospettiva macro-sociale, all'interno di questi sistemi grazie ai dati condivisi dei cittadini sulle abitudini culturali e sociali si possono generare relazioni commerciali, transizioni economiche che orientano decisioni politiche, rapporti diplomatici internazionali e orientamenti culturali di intere popolazioni alla base di accordi fra corporations da tutelare.

All'interno delle piattaforme digitali, poi, si costruiscono identità individuali e sociali, si definiscono legami e relazioni interpersonali, si coltivano ideologie, percezioni e interpretazioni della realtà che orientano la costruzione dell'opinione pubblica rispetto ad alcuni fenomeni sociali e i comportamenti quotidiani dei cittadini, compiendo quello che McLuhan definiva un costante "massaggio culturale" (1967). In questo caso, la protezione diventa sociale e politica, nella misura in cui deve tutelare dinamiche di interazione e relazione intersoggettiva fra cittadini nel rispetto delle diversità culturali e sociali.

Infine, i nuovi meccanismi dell'industria dei dati, generati dalla commercializzazione delle informazioni come beni di scambio per una nuova economia digitale, inevitabilmente generano rischi indesiderati con conseguenti danni non solo per i cittadini ma anche per la democrazia e per la sicurezza di

organizzazioni o nazioni. In tal senso è possibile parlare di protezione economica *online,* intesa come regolazione e monitoraggio degli scambi economici, della produzione di criptovalute e delle operazioni commerciali gestite completamente negli ambienti virtuali con ripercussioni pesanti nel contesto socioculturale non virtuale.

Riassumendo, con il termine sicurezza online s'intende un processo di protezione delle reti, dei dispositivi e dei dati da accessi non autorizzati o da utilizzi illeciti al fine di garantire riservatezza, integrità e disponibilità delle informazioni di un sistema online (CISA, 2009). Questa prima definizione è certamente tecnica, legata all'aspetto ingegneristico della *sicurezza online e* dunque alla preservazione del benessere materiale di un sistema virtuale secondo una prospettiva macrosociale. L'International Telecommunication Union (ITU) nel 2006, tuttavia, ha fornito una definizione della stessa più ampia e complessa, che va oltre la dimensione tecnicista parlando di sicurezza online come raccolta di politiche, concetti di sicurezza, approcci, azioni, formazione, *best practices*, garanzie e tecnologie da utilizzate per proteggere l'ambiente *online* e gli *asset* delle organizzazioni e degli utenti (ITU, 2006). In questa seconda definizione rientra il concetto di benessere immateriale, sociale e culturale, del sistema sociale da tutelare rispetto ad azioni devianti di cui sono responsabili i rappresentati delle politiche pubbliche.

In tal senso, la sicurezza non rappresenta solo una tema di natura tecnica ed economica, che tutela gli interessi commerciali e politici delle grandi *corporation* a livello mondiale, ma anche sociologica e antropologica nella misura in cui può contribuire alla riduzione di una cultura della instabilità e un clima di insicurezza sociale diffusa e contagiosa rispetto al quale è opportuno intervenire, dal punto di vista giuridico e politico, ma anche educativo e sociale, al fine di ripristinare la percezione collettiva del benessere pubblico.

3.1. *La prospettiva individuale e simbolica del benessere digitale*

Spostando il focus di attenzione dai produttori ai consumatori, il Regolamento dell'Unione europea sulla sicurezza *online* (n. 2019/881) descrive questa ultima come l'insieme delle attività necessarie per proteggere sia la Rete e i sistemi informativi, sia gli utenti di tali sistemi e altre persone interessate dalle minacce informatiche. La questione assume un ruolo centrale non sono per gli enti e le organizzazioni che devono proteggere i loro interessi e le attività commerciali contro forme di criminalità *online*, ma anche gli stessi cittadini che vogliono preservare la propria integrità etica e

psicologica, nonché proteggere i propri dati personali secondo una prospettiva micro-sociale.

In questo contesto, va riletto e interpretato il concetto di benessere individuale del cittadino negli ambienti digitali in prospettiva micro sociale che, secondo l'approccio delle *capabilities*, richiama il diritto individuale di esprimere ed esercitare la propria libertà sostanziale, di compiere scelte comportamentali e relazionali sul mondo virtuale, sfruttando le opportunità offerte dalle piattaforme digitali, nella consapevolezza del peso e degli effetti socioculturali delle proprie azioni e munito di un forte senso di responsabilità alla base delle scelte compiute.

Ne consegue che i rischi generati da un utilizzo non responsabile dei dati e da una gestione poco etica delle dinamiche di interazione virtuale sono molteplici. Tali rischi oscillano dalla violazione della vita privata e dei dati personali di soggetti operanti in Rete (cittadini e aziende) alla diffusione dei contenuti illeciti e nocivi, dalla disinformazione alla criminalità informatica, dallo sfruttamento all'abuso di minori, la sorveglianza di massa etc. che possono compromettere il benessere individuale del cittadino.

Secondo il gruppo di ricerca internazionale *EUkids online* (2017), a tali rischi è possibile aggiungerne altri quali: il commercio del *sexting,* della pornografia *online*, del cyberbullismo, nonché *l'hate speech,* ovvero la diffusione di messaggi di odio e commenti violenti, discriminatori e offensivi in Rete per genere, etnia, religione, preferenze sessuali o scelte politiche, rintracciabili soprattutto nelle dinamiche relazionali e di condivisione nel web (Mascheroni *et al.,* 2018).

In accordo con la *dichiarazione europea sui diritti e i principi digitali* del 26 gennaio 2022, da qualche anno il Parlamento europeo sta lavorando per garantire il rispetto dei diritti fondamentali quali la neutralità tecnologica e della Rete e l'inclusività, oltre che per migliorare le abilità e le competenze digitali dei cittadini e per promuovere un ecosistema di istruzione digitale efficiente.

La *safety* è una competenza trasversale del Framework Europeo sulle competenze digitali del cittadino (Digcomp 2.0) (Ridsdale *et al.*, 2015) e richiama il concetto di tutela attraverso il potenziamento delle risorse immateriali individuali del capitale umano per fronteggiare le sfide digitali e salvaguardare il benessere individuale quando si usano le tecnologie secondo una prospettiva culturale. Essa include la capacità di utilizzare Internet in modo responsabile, di valutare l'affidabilità delle informazioni trovate in Rete, di identificare le fonti online affidabili e di utilizzare quelle trovate sul web senza plagiare (Wastiau *et al.,* 2013).

Partendo da questo primo inquadramento, il benessere digitale può essere concettualizzato dal punto di vista tecnologico ed economico, rispetto alla tutela dell'ambiente sociale, fisico e virtuale, e dal punto di vista culturale e sociale, legato al benessere individuale che può essere cognitivo, fisico, psicologico emotivo e sociale sia nel mondo fisico che in quello virtuale.

Nello specifico, il benessere tecnologico ed economico, se da un alto rimanda alla protezione del device intesa come protezione del contesto domestico online entro cui l'utente lavora, interagisce e organizza le proprie risorse e i materiali personali informativi e culturali online; dall'altro si riferisce alla protezione dell'ambiente fisico, socioculturale off line frequentato quotidianamente dall'individuo quando non utilizza le tecnologie. Nello specifico esso fa riferimento a:

- la tutela del sistema digitale dal punto di vista tecnico entro cui il cittadino interagisce per motivazioni di studio, lavoro, gestione della quotidianità e richiama la capacità di utilizzare Internet in modo prudente per proteggersi da rischi quali: il bullismo online, le frodi, il furto di identità, i contenuti inappropriati, lo spam e la posta indesiderata, etc.;
- la tutela dell'ambiente sociale attraverso la protezione dell'ambiente fisico in termini di sostenibilità ambientale rispetto all'impatto del digitale e dell'uso delle tecnologie nel contesto socioculturale.

Il benessere sociale e culturale individuale invece prevede:

- la tutela dell'identità digitale online attraverso la protezione dei dati e della *privacy* ai fini della tutela della *web reputation* propria e degli altri e l'identità digitale; a riguardo alcuni studiosi parlano di *data security awareness* come processo di educazione alla gestione consapevole dei dati online (Cascavilla *et al.,* 2017);
- la tutela della condizione sociale della persona nel contesto fisico di fruizione con particolare riferimento al:
    - benessere fisico rispetto a patologie legate alla modalità di fruizione dei media;
    - benessere psicologico-emotivo rispetto ai rischi di dipendenza mediale e di emarginazione online;
    - benessere sociale (inclusione) rispetto a forme di emarginazione e discriminazione socioculturale online (*hating, stalking, revenge pornography, deep web, cyberbullismo*) e rispetto ai temi della economia dell'informazione e della sorveglianza digitale (Lupton, 2015);

- benessere cognitivo rispetto a fenomeni di dipendenza mediale o di abuso fruitivo che porta a forme di depotenziamento dell'apprendimento dei cittadini.

Secondo Gui e Gerosa (2019), l'espressione "benessere digitale" racchiude una serie di competenze digitali, prevalentemente trasversali, utili per gestire e autoregolare gli stimoli provenienti dal web in funzione del proprio benessere soggettivo. Si tratta di competenze strategiche (van Deursen, 2010) e di autoregolazione (Calvani, Fini, Ranieri, 2010) contro l'abuso di utilizzo del web, che possono sfociare in forme di dipendenza mediale. Il benessere soggettivo e sociale cui si fa riferimento richiama dunque un'alfabetizzazione digitale di tipo funzionale, che inglobi capacità di pensiero critico, intelligenza emotiva ed empatia, consapevolezza sulle conseguenze delle proprie azioni, il rispetto dei diritti propri e altrui e la responsabilità del proprio operato (Di Gioia *et al.,* 2017).

4. Sicurezza online e adolescenti: analisi di un caso di studio

La tutela del benessere individuale digitale, secondo una prospettiva micro sociale, da parte dei cittadini è stato uno degli obiettivi alla base dell'indagine nazionale empirica, realizzata dalla Sapienza Università di Roma dal 2020 al 2023 su un campione di 2708 adolescenti compresi fra i 14 e 18 anni provenienti da 37 scuole secondarie superiori di 14 regioni italiane. L'obiettivo della ricerca è stato quello di rilevare un comportamento sociale orientato alla consapevolezza digitale, con particolare riferimento ai temi della sicurezza e del benessere.

Dal punto di vista metodologico, per costruire profili comportamentali collegati al benessere degli adolescenti coinvolti nella ricerca, il processo di elaborazione dei dati ha adottato la seguente sintassi:

- una prima analisi descrittiva per ricostruire un quadro generale sui tipi di comportamenti mediali avviati dagli utenti nella prospettiva della consapevolezza digitale;
- la costruzione degli indici additivi sulla consapevolezza digitale, in grado di restituire un orientamento medio comportamentale rispetto ai quattro tipi di benessere precedentemente introdotti;
- diverse analisi statistiche bivariate per approfondire la relazione fra la diffusione di indici di consapevolezza, descritti sopra, rispetto alle categorie dell'età, del genere, della collocazione territoriale e del capitale culturale e sociale delle famiglie di appartenenza;

- la cluster analysis sulla costruzione di profili comportamentali prevalenti, virtuosi o meno, rispetto al tema del benessere nella sua complessità terminologica e semantica, emergenti all'interno del nostro campione di indagine.

La cluster analysis[10] ha messo in evidenza due principali profili comportamentali di adolescenti: 1. *I virtuosi digitali* ovvero adolescenti attenti al benessere individuale digitale, anche se meno virtuosi rispetto a qualsiasi forma di tutela del benessere sociale (51%); 2. *i virtuosi sociali* ovvero gli adolescenti orientati al benessere socioculturale, sensibili e attenti a preservare il benessere individuale e ambientale nella realtà sociale quando usano i dispositivi digitali, ma meno prudenti rispetto alla tutela dei propri dati online (49%).

Nello specifico *gli adolescenti attenti al benessere digitale* (primo cluster) non sembrano avere una dieta mediale particolarmente ricca, utilizzano pochi dispositivi digitali e manifestano raramente effetti negativi psicofisici dall'uso quotidiano dei media (quali ad esempio perdita del sonno, problemi alla vista, dolore alle dita, ansia e nervosismo, etc.); lo scarso utilizzo delle tecnologie comporta una scarsa attenzione ad attivare comportamenti sociali prudenti per tutelare il proprio benessere fisico, sociale e psicologico emotivo durante la fruizione. Tali adolescenti infatti caricano spesso il cellulare vicino al letto, utilizzano raramente la modalità aereo di notte e lasciano spesso il cellulare accesso mentre dormono, solo qualche volta attivano la modalità notturna o lo spengono.

Rispetto al benessere tecnologico ed economico, *i virtuosi digitali sono più attenti* alle *performanc*e del device di ultima generazione che alla sua sostenibilità ambientale, ad esempio durante l'acquisto la loro attenzione si focalizza prevalentemente sulle caratteristiche estetiche del prodotto e sul brand, sull'ampiezza della memoria del device, sulle caratteristiche dei software e sulla qualità della fotocamera, a pochi interessa la certificazione ecologica sui materiali usati, il consumo di energia celato dietro il funzionamento della connessione digitale, nonché lo sfruttamento di risorse umane a basso costo per la produzione di beni avanguardisti come quelli digitali. Tali adolescenti infine sembrano poco condividere la tesi per cui le tecnologie

[10] La cluster analysis ha utilizzato la funzione k-means, a partire da una componente ottenuto tramite ACP (Analisi delle componenti principali) con rotazione Quartimax, riconducibile agli indici di protezione del benessere individuale e ambientale, online e offline, attribuibili agli studenti. L'obiettivo è stato quello di individuare ed enfatizzare profili comportamentali delle giovani generazioni rispetto al benessere sociale off line e online, nonché all'uso del digitale e alla sua consapevolezza di utilizzo.

inquinino l'ambiente o che possano emettere radiazioni dannose per la salute.

Dal punto di vista socio-anagrafico, i *virtuosi digitali* (49%) sono adolescenti maschi compresi fra i 14 e i 19 anni che possiedono una dieta mediale poco variegata: utilizzano raramente Smartwatch, Tv a Pagamento, Radio e web radio, quotidiani cartacei e e-book. La loro attenzione si concentra prevalentemente sull'uso di smartphone e Internet, in questi ambienti digitali tuttavia sono poco creativi, attivi e partecipativi: non creano contenuti online, non scrivono sui blog o sui social networks e non sono interessati all'implementazione di pagine virtuali nel web, non esprimono pareri e opinioni su tematiche impegnate di natura politica o sociale, certamente non partecipano a consultazioni online di qualsiasi tipo.

La disattenzione per gli effetti ambientali legati a un utilizzo poco accorto delle tecnologie sembra trasferirsi inevitabilmente nell'ambiente domestico online: gli adolescenti virtuosi digitali del cluster non proteggono adeguatamente l'account personale e il proprio pc e le informazioni contenute al proprio interno: ad esempio non cambiano mai il codice di accesso al PC, al portatile o al tablet; non procedono mai con la scansione antivirus soprattutto nei tablet. Raramente utilizzano software per la generazione automatica di password al fine di proteggere i propri dati e altrettanto raramente cambiano la password dei loro dispositivi. È ipotizzabile che questa scarsa prudenza sia associabile a una sottovalutazione dei potenziali rischi di attacchi hacker al proprio sistema e certamente da una ridotta conoscenza degli stessi.

Rispetto al benessere digitale individuale online, connesso alla tutela dei dati personali e della privacy, gli studenti intervistati *virtuosi digitali* sembrano invece possedere comportamenti più prudenti sulla tutela della propria identità online. Sebbene acconsentano al trattamento dei dati personali, soprattutto per finalità di marketing da parte di terzi soggetti, sembrano più prudenti nella condivisione dei dati sensibili: ad esempio non condividono informazioni sul proprio orientamento politico o religioso, i propri recapiti o indirizzi di casa, situazioni sentimentali familiari e collegamenti con siti e altre pagine personali. Ulteriori informazioni legate alla data di nascita, al tipo di scuola frequentata e anche alle esperienze ludiche, testimoniate da foto e video, sono condivise solo con gli amici più stretti.

Gli adolescenti virtuosi attenti al benessere digitale sembrano più rispettosi anche dei dati degli altri: ad esempio raramente condividono foto o video di amici o terzi senza autorizzazione e sempre raramente usano contenuti scaricati online senza verificarne il copyright. L'attenzione al benessere digitale di tali adolescenti è rilevabile anche in ulteriori comportamenti: ad esempio di rado aprono email provenienti da sconosciuti oppure inseriscono

i propri dati per partecipare a concorsi a premio o offerte apparentemente vantaggiose.

I ragazzi *socialmente virtuosi, invece,* usano più frequentemente i media e svolgono più attività culturali outdoor (dallo sport al cinema, dal teatro alle uscite con gli amici e in famiglia). Il loro frequente consumo multimediale li espone a potenziali rischi fisici, emotivi e psicologici derivanti dall'uso prolungato e costante delle tecnologie. In virtù di questo consumo fruitivo più intenso e variegato delle tecnologie, questi adolescenti adottano comportamenti più prudenti e tutelano il proprio benessere individuale offline quando usano le tecnologie. Ad esempio di notte caricano il cellulare lontano dal letto o in un'altra stanza, tengono spesso il device in modalità silenziosa o in modalità notturna. Rispetto al benessere tecnologico, poi, sembrano leggermente più consapevoli dei potenziali effetti negativi delle tecnologie, compreso il problema dello sfruttamento della manodopera infantile, e sono sostenitori del riciclo dei dispositivi tecnologici per ridurre l'impatto ecologico ambientale.

Gli stessi adolescenti tuttavia sembrano meno prudenti e attenti al loro benessere individuale digitale: condividono con più facilità le proprie informazioni online con amici, comprese storie, post, video e foto. Hanno meno scrupoli anche nella condivisione di informazioni e foto di altri, per cui non sempre chiedono l'autorizzazione alla pubblicazione di materiale, così come sono meno attenti al copyright dei contenuti scaricati online. Si geolocalizzano con più facilità e sono propensi alla condivisione della propria posizione.

I *virtuosi sociali* sono certamente più attenti al benessere individuale e sociale nella quotidianità e in tal senso attivano comportamenti virtuosi per la tutela tecnologica e ambientale e quella fisica, psicologica ed emotiva personale; questi stessi adolescenti, prevalentemente femmine, sono più attive culturalmente, ma poco prudenti e attente al proprio benessere individuale digitale, soprattutto per quanto concerne la condivisione dei propri e degli altrui dati personali online, dimostrando una scarsa conoscenza dei potenziali rischi della Rete e una ingenuità comportamentale rispetto ai potenziali danni derivanti da una ridotta protezione anche dei propri ambienti domestici online, ovvero il device tecnologico (pc, tablet o smartphone). È ipotizzabile che la frequenza fruitiva dei dispositivi e le conseguenti competenze di accesso mediale, maturate per via esperienziale, porti a sviluppare nelle adolescenti la percezione di una elevata competenza digitale innata, figlia della capacità tecnica di gestione del device e dei linguaggi mediali, trascurando tutto l'apparato critico della comunicazione, di responsabilità e consapevolezza comportamentale che non può essere sviluppato per inerzia nelle

giovani generazioni. La presunzione della sapienza digitale delle adolescenti può indurre a comportamenti imprudenti e alla sottovalutazione di azioni di tutela quando si utilizzano i dispositivi digitali. La stessa carenza di competenze digitali trasversali negli adulti, genitori o insegnanti, inevitabilmente non consente di stabilire mediazioni culturali e di negoziazione conoscitiva e critica sui media a scuola e in famiglia, in tal senso la *literacy* sulla comunicazione digitale delle adolescenti rischia di essere ancora quasi esclusivamente esperienziale, isolata (perché priva del confronto con l'adulto) e orientata sulla dimensione espressivo-tecnica.

Conclusioni

In questo saggio abbiamo analizzato il concetto di benessere sociale partendo da due punti di vista: 1. quello delle istituzioni che possono garantire il bene pubblico, materiale e immateriale, al fine di mantenere un controllo e ridurre forme di disuguaglianza sociale, attraverso politiche di welfare, orientate all'integrazione socioculturale delle categorie più emarginate e svantaggiate; 2. quello degli individui che tutelano il proprio benessere privato secondo diverse prospettive: fisica, emotiva, psicologica e cognitiva. Il transito dalla condizione moderna a quella postmoderna ha determinato il progressivo passaggio da una visione macrosociale del benessere, prevalentemente materiale e incentrata sulla dimensione economica e tecnica, a una visione microsociale, individualista e intimo privata, di tipo immateriale e orientata al soddisfacimento di bisogni socioculturali, nella consapevolezza dell'impossibilità dello stato sociale di far fronte alle nuove richieste di sicurezza del cittadino in una condizione sociale di precarietà etico valoriale. Questa duplice lettura del benessere sociale è stata riproposta per analizzare il concetto di benessere digitale.

Tale espressione è stata definita dal Digital Wellness Lab nel 2021 come "la capacità dei media e della tecnologia di costruire ambienti digitali che possano aiutare a promuovere il benessere generale dei bambini, delle loro famiglie, della società e dell'umanità in generale". La condizione di benessere digitale dipende quindi anzitutto dalle caratteristiche delle tecnologie e/o degli ambienti digitali, dalla capacità di questi strumenti di adattarsi alle esigenze dei fruitori e di rispondere ai loro bisogni socioculturali e formativi, garantendo piena accessibilità alle informazioni e ai contenuti (Maciariello et al., 2023). Certamente la dimensione tecnologica non può essere considerata in modo deterministico quale fattore incidente sulla questione del benessere sociale, la sua lettura va interpretata all'interno di un ecosistema infrastrutturale, ovvero "un'architettura digitale programmabile, progettata per organizzare interazioni fra utenti…orientata alla raccolta

sistematica, al trattamento algoritmico, alla circolazione e monetizzazione dei dati degli utenti". **(**van Dijck et Alii, 2018: p.27**)**. Da qui possono emergere inevitabili ripercussioni sulle dinamiche di socializzazione, di trasmissione/interiorizzazione etico valoriale e di espressione culturale degli utenti, con conseguenze nei processi di edificazione delle identità individuali e sociali, nonché nella costruzione delle relazioni sociali, comunicative e partecipative dei cittadini. Da qui derivano le regolamentazioni etiche rivolte ai producer degli ambienti digitali che disciplinano la governance di tali sistemi. Tale regolamentazione è centrale per garantire il rispetto dei dati condivisi dagli utenti e consenso attraverso l'instaurazione di un rapporto fiduciario fra l'utente stesso e l'ente proprietario. La regolamentazione genera un senso percettivo di tutela e sicurezza nel cittadino e, dunque, di benessere sociale collettivo sia materiale che immateriale. I dati tuttavia rappresentano i beni di scambio e di contrattazione socioeconomica, ovvero tutte quelle informazioni, generati dalle azioni e dalle interazioni degli utenti in Rete, che orientano non solo i meccanismi di produzione e distribuzione economica, ma anche le principali dinamiche politiche, ideologiche e culturali sottese ai meccanismi produttivi e commerciali a livello globale. Tali aziende per prime hanno elaborato strategie per utilizzare i dati forniti dai loro utenti ed usarli nella promozione di iniziative commerciali personalizzate (Lupton, 2015). In tal senso, il valore pubblico delle dinamiche di azione e di interazione degli utenti viene spesso trasformato in un valore economico che alimenta interessi politici e privati.

Per questo motivo, anche quando si parla di benessere digitale accanto alla prospettiva macrosociale di analisi del fenomeno, progressivamente si fa strada l'analisi e l'investimento sulla capacità dell'utente di favorire una condizione di benessere individuale nel digitale, autoregolando e gestendo in maniera consapevole l'utilizzo delle tecnologie digitali e dei propri dati. Il caso di studio presentato all'interno del saggio restituisce una fotografia delle nuove generazioni ancora poco mature e consapevoli delle dinamiche di gestione dei sistemi digitali online. Nello specifico il campione sembra spaccato a metà fra coloro che manifestano più sensibilità al benessere digitale individuale pur in presenza di un consumo mediale non frequente e coloro che, utilizzando più assiduamente le tecnologie, si comportano privi di forme di autotutela del proprio comportamento, lasciando intendere un atteggiamento di fiducia nelle forme di sicurezza garantite dai sistemi di regolazione infrastrutturali e una ignoranza rispetto ai meccanismi di produzione dei diversi capitali che si celano dietro le industrie produttive e di consumo culturale delle piattaforme digitali. Nel primo caso, il virtuosismo digitale degli adolescenti non si riflette in un habitus orientato al benessere tout court, infatti la stessa attenzione manifestata dagli adolescenti intervistati per i dati condivisi on line non è trasferita in comportamenti

precauzionali sociali di tutela della salute fisica e psicologica anche nella realtà sociale quando si fruiscono i media. Nel secondo caso, gli adolescenti intervistati sono assidui fruitori di tecnologia ma imprudenti e poco attenti al benessere digitale individuale e sociale; di contro gli stessi adolescenti sono più attenti alla tutela della propria persona fisica e dell'ambiente circostante quando usano le tecnologie.

Riferimenti bibliografici

Barbero Vignola G., Bezze M., Canali C., Geron D., Innocenti E., Vecchiato T. (2016). Povertà educativa: il problema e i suoi volti. *Studi Zancan*, Fondazione Emanuela Zancan Onlus, maggio 2016.

Beck U. (1988). *Gegengifte. Die organisierte Unverantwortlichkeit*. Frankfurt a. M.: Suhrkamp; trad. it. (2000). *La società del rischio*. Roma: Carocci.

Boudon R., Bourricaud F. (1991). *Dizionario critico di sociologia*. Roma: Armando.

Bourdieu P. (1979). *La distinction. Critique sociale du jugement*. Paris: Éditions de Minuit; trad. it. (1983). *La distinzione. Critica sociale del gusto*. Bologna: il Mulino.

Bourdieu P. (1980). *Le sens pratique*. Paris: Éditions de Minuit; trad. it. (1995). *Ragioni pratiche*. Bologna: il Mulino.

Calvani A., Fini A., Ranieri M. (2011). *Valutare la competenza digitale. Prove per la scuola primaria e secondaria*. Trento: Erickson.

Carmagnola F., Ferraresi M. (1999). *Merci di culto. Ipermerce e società mediale*. Roma: Castelvecchi.

Cascavilla G., Conti M., Frison D., Surian A. (2017). Data security awareness: metodi e strumenti per promuoverla nella scuola secondaria. *Media Education – Studi, ricerche, buone pratiche*, 8(2): 276-284. DOI: 10.14605/Med821709.

Castells M. (2001). *Galassia Internet*. Milano: Feltrinelli, 2002.

Castells M., Himmanen P. (2004). *The Welfare State and the Information Society*. Bologna: il Mulino.

CISA (2009). What is cybersecurity? Maggio 2009. https://www.cisa.gov/uscert/ncas/tips/ST04-00

Croteau D., Hoynes W. (2018). *Sociologia generale. Temi, concetti, strumenti*. New York: McGraw-Hill Education.

Di Gioia R., Gemo M., Troia S., Chaudron S. (2017). Quando l'alfabetizzazione diventa cyber, responsabile e digitale. In Scarcelli M.C., Stella R. (a cura di), *Digital literacy e giovani: strumenti per comprendere, misurare, intervenire*. Milano: FrancoAngeli.

Fabris G. (2008). *Il nuovo consumatore: verso il postmoderno*. Milano: FrancoAngeli.

Featherstone M. (1995). *Undoing culture: Globalization, postmodernism and identity*. London: Sage Publications; trad. it. (1998). *La cultura dislocata. Globalizzazione, postmodernismo, identità*. Formello: SEAM.

Ferrari Occhionero M. (2002). *Disagio sociale e malessere generazionale. Dinamiche valoriali tra persistenza e mutamento*. Milano: FrancoAngeli.

Giaccardi C., Magatti M. (2003). *L'io globale. Dinamiche della socialità contemporanea*. Roma-Bari: Laterza.

Giddens A. (1990). *The consequences of modernity*. Cambridge: Polity Press; trad. it. (1994). *Le conseguenze della modernità. Fiducia e rischio, sicurezza e pericolo*. Bologna: il Mulino.

Gui M., Gerosa T. (2019). Strumenti per apprendere o oggetti di apprendimento? Una rilettura critica della digitalizzazione nella scuola italiana. *Scuola democratica*, 3: 637-645. DOI: 10.12828/95945.

Harvey D. (1989). *The condition of postmodernity: An enquiry into the origins of cultural change*. Cambridge: Wiley-Blackwell.

Helmond A. (2015). The platformization of the web: Making web data platform ready. *Social Media + Society*, 1(1).

Iannone R. (2006). *Il capitale sociale. Origine, significati e funzioni*. Milano: FrancoAngeli.

ITU (2006). *Cybersecurity guide for developing countries*. http://handle.itu.int/11.1002/pub/8022a3e9-en

Jameson F. (1989). *Postmodernismo, o la logica culturale del tardo capitalismo*. Roma: Fazi.

Lyotard J.F. (1979). *La condition postmoderne*. Paris: Les Éditions de Minuit.

Lupton D. (2015). *Digital sociology*. London: Routledge; trad. it. Santoro M., Timeto F. (a cura di) (2018). *Sociologia digitale*. Milano-Torino: Pearson.

Maffesoli M. (2003). *Notes sur la postmodernité*. Paris: Éditions du Félin; trad. it. (2005). *Note sulla postmodernità*. Milano: Lupetti.

Maciariello G., Bucciarelli F., Glorioso A., Giangrande A., De Blasi A., Maciariello P., Perrone M. (2023). Digital wellness vs wellness in the digital. *Phenomena Journal – International Journal of Psychopathology, Neuroscience and Psychotherapy*, 5(2): 143-148. https://doi.org/10.32069/PJ.2021.2.199.

Mascheroni G., Ólafsson K. (2018). Accesso, usi, rischi e opportunità di Internet per i ragazzi italiani. *EU Kids Online*, OssCom.

McLuhan M. (1967). *Gli strumenti del comunicare*. Roma: Il Saggiatore.

Morcellini M., Mazza B. (2008). *Oltre l'individualismo. Comunicazione, nuovi diritti e capitale sociale*. Milano: FrancoAngeli.

Nussbaum M. (2000). *Women and human development: The capabilities approach*. New York: Cambridge University Press; trad. it. (2000). *Lo sviluppo è libertà*. Milano: Mondadori.

Pizzorno A. (2007). *Il velo della diversità. Studi su razionalità e riconoscimento*. Roma: Feltrinelli.

Portes A. (1998). Social capital: its origins and applications in modern sociology. *Annual Review of Sociology*, 24: 1-24; trad. it. Cortoni I. (a cura di) (2013). *Il capitale sociale*. Lecce: Kurumuny.

Putnam R.D. (2000). *Bowling alone. The collapse and revival of American community*. New York: Simon & Schuster.

Ridsdale C., Rothwell J., Smit M., Ali-Hassan H., Bliemel M., Irvine D., et al. (2015). *Strategies and best practices for data literacy education*. http://dataliteracy.ca/

Ranieri A. (2006). *I luoghi del sapere*. Roma: Donzelli.

Rullani E. (2005). *Economia della conoscenza*. Roma: Carocci.

Sciolla L. (2002). *Sociologia dei processi culturali*. Bologna: il Mulino.

Sen A. (1999). *Development as freedom*. New York: Knopf; trad. it. (2000). *Lo sviluppo è libertà*. Milano: Mondadori.

Sen A. (2003). Development as capability expansion. In Fukuda-Parr S., Shiva Kumar A.K. (a cura di), *Readings in human development. Concepts, measures and policies for a human development paradigm*. Oxford: Oxford University Press.

UNICEF (2007). *Prospettiva sulla povertà infantile: un quadro comparativo sul benessere dei bambini nei paesi ricchi. Report Card Innocenti n. 7.* https://www.unicef-irc.org/publications/pdf/rc7_ita.pdf

van Deursen A. (2010). *Internet skills: Vital assets in an information society*. Enschede: University of Twente.

van Dijck J. (2013). *The culture of connectivity: A critical history of social media*. New York: Oxford University Press.

van Dijck J., Poell T., de Waal M. (2018). *The platform society: Public values in a connective world*. Oxford: Oxford University Press; trad. it. Boccia Artieri G., Marinelli A. (a cura di) (2019). *Platform Society. Valori pubblici e società connessa*. Milano: Guerini.

Wastiau P. et al. (2013). *European Journal of Education*. Oxford: Blackwell Publishing.

Valeria Rosato (2025)
L'Altro come oggetto d'odio. Per una sociologia dei conflitti contemporanei
Meltemi

di Giorgia Mavica[*]

Il volume di Valeria Rosato, "L'Altro come oggetto d'odio", affronta in modo articolato la teoria sociologica del conflitto. L'approccio è multidimensionale, intrecciando livelli micro, macro e soprattutto meso: etnia, dinamiche tra *in-group* e *out-group* e appartenenze, permettendo un maggiore approfondimento delle categorie e dei casi indagati. L'intero testo richiama la tensione tra un "noi" e un "loro" alla base di molti conflitti attuali: l'autrice mostra come dietro la violenza organizzata si celino spesso fratture identitarie e culturali profonde, anche se raramente tematizzate.

In quest'ottica, la sociologia è non solo interpretativa ma anche critica. Come è riportato dalla Rosato, Bauman scrive (2003), «la sociologia critica è prima di tutto una critica della società, un modo di smascherare i miti che nascondono le strutture di potere» (Rosato, 2025: 19).

Il volume è diviso in tre parti. Nella prima, l'autrice analizza il concetto di "nuove guerre", criticando letture dominanti come quella apocalittica di Kaplan (1994) ed Enzensberger (1994), quella economicista di Collier e Hoeffler (2004) e l'approccio identitario di Kaldor (1999). A suo avviso, «tali letture hanno prodotto una visione parziale e incompleta della sua evoluzione storica» (Rosato, 2025: 29), spesso traducendosi in una giustificazione implicita di certi modelli di *governance* globale.

Il volume prosegue affrontando il tema della violenza collettiva. L'autrice distingue tra approcci strutturalisti – che la collegano a disuguaglianze sistemiche e colonialismo – e approcci individualisti, attenti a motivazioni personali e discorsi d'odio. Tre gli autori chiave: Duffield (1998, 2004) parla di "complessi politici emergenti", utile per comprendere come molti conflitti contemporanei siano parte integrante di meccanismi di controllo e di gestione a distanza funzionali al mantenimento dell'ordine neoliberale; Kalyvas (2001, 2006) sottolinea che «la situazione di conflitto in-

DOI: 10.5281/zenodo.17297588

[*] Università degli Studi di Catania. giorgia.mavica@unict.it.

centiva e scatena una serie di opportunità per gli agenti locali» focalizzandosi sulle scelte individuali e sul comportamento degli attori locali (Rosato, 2025: 31); infine Malešević (2010) rifiuta l'idea che le guerre attuali abbiano natura nuova, sostenendo che «non sembrerebbero "nuovi" nemmeno gli obiettivi delle guerre contemporanee» (Rosato, 2025: 31).

La terza parte analizza tre casi studio: conflitto colombiano, crisi nel Sahel e guerra russo-ucraina. Rosato applica qui le teorie discusse, mostrando la guerra come fenomeno ibrido e stratificato. Notevole il caso Sahel, descritto come «terra di terrore versus terra di conquista» (Rosato, 2025: 134), dove convergono jihadismo, criminalità e strategie geopolitiche. In tutti i casi emerge che i conflitti non sono solo materiali, ma alimentati da memorie collettive, ideologie e visioni del mondo radicate.

Un altro punto di forza del libro è forse proprio la capacità di unire teoria ed esperienze empiriche, storia e sociologia, quindi coniugare tra loro temi a tratti densi e comunque mantenendo uno stile leggibile, impreziosito da riflessioni aperte e lontano da dogmi.

L'autrice spinge il lettore tra suggestioni teoriche weberiane e riferimenti a pensatori come C. Wright Mills, dimostrando che la sociologia sia l'arma più che valida per affrontare e comprendere la complessità dei conflitti che caratterizzano la società di oggi.

Il testo si propone come una "mappa concettuale" capace di disporre i concetti in modo utile orientandosi perfettamente in un contesto complesso. Come l'autrice stessa afferma, «l'analisi sociologica […] ci offre gli strumenti per comprendere, non solo le crescenti contraddizioni che caratterizzano le società contemporanee, ma soprattutto a demistificare le narrazioni […] che da sempre accompagnano e distorcono la comprensione della violenza organizzata» (Rosato, 2025: 16).

Alla luce di quanto rilevato, "L'altro come oggetto d'odio" si presenta come un lavoro solido, lucido e utile per chiunque voglia affrontare il tema del conflitto in prospettiva realmente critica e interdisciplinare. È un testo che parla agli studiosi, ma anche a coloro che vogliono approfondire le guerre moderne senza cadere in semplificazioni. Un esempio efficace di come la sociologia possa fornire strumenti preziosi non solo per analizzare la realtà, ma anche per cambiarla.

Riferimenti bibliografici

Bauman Z. (2003). *Intervista sull'identità*. A cura di B. Vecchi. Roma-Bari: Laterza.

Collier P., Hoeffler A., Pattillo C. (2004). *Aid and capital flight*. Oxford University: Centre for the Study of African Economies. http://users.ox.ac.uk/~ball0144/aid&cf.pdf.

Duffield M. (1998). Postmodern Conflict. Warlords, Post-Adjustment States and Private Protection. *Civil Wars*, 1(1): 65-102, DOI: org/10.1080/13698249808402367.

Duffield M. (2004). Guerre postmoderne. L'aiuto umanitario come tecnica politica di controllo. Bologna: Casa editrice il Ponte.

Enzensberger H.M. (1994). *Civil wars: from LA to Bosnia*. New York: Free Press (trad.it.: *Prospettive sulla guerra civile*, Torino: Einaudi).

Kaldor M. (1999). *New and Old Wars: Organized Violence in a Global Era*. Cambridge: Polity Press.

Kalyvas S.N. (2001). New' And 'Old' Civil Wars: A Valid Distinction?. *World Politics*, 54(1): 99-118.

Kalyvas S.N. (2006). *The logic of violence in civil war*. Cambridge: Cambridge University Press.

Kaplan R.D. (1994). The coming anarchy: How scarcity, crime, overpopulation, tribalism and disease are rapidly destroying the fabric of our planet. *The Atlantic Monthly* 273(2), 44-76.

Malešević S. (2010). *The Sociology of War and Violence*. Cambridge: Cambridge University Press.

Rosato V. (2025). *L'Altro come oggetto d'odio. Per una sociologia dei conflitti contemporanei.* Sesto San Giovanni: Meltemi.