2026

SICUREZZA E SCIENZE SOCIALI

ANNO XIV
N. 1/2026

Critique of artificial reason: sociological perspectives on power, ethics, and human relations

a cura di

Vera Kopsaj, Luca Corchia

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

Sommario

Introduzione. Sfide etiche, politiche e relazionali nell'era del potere algoritmico

di Vera Kopsaj*, Luca Corchia**

L'intelligenza artificiale è diventata negli ultimi anni non soltanto un campo di sperimentazione tecnologica, ma anche uno specchio privilegiato attraverso cui osservare le trasformazioni più profonde delle società contemporanee. La sua rapida diffusione nelle economie, nelle istituzioni e nella vita quotidiana solleva interrogativi che toccano la giustizia sociale, la democrazia, la sicurezza, la cultura e le relazioni umane. Il linguaggio dell'innovazione rischia spesso di coprire la portata sociale di questi processi: dietro l'entusiasmo per gli algoritmi si celano questioni di potere, di esclusione, di vulnerabilità, ma anche di creatività e di possibilità inedite.

Il volume che presentiamo nasce con l'intento di dare voce a questa complessità. La raccolta comprende contributi in lingua italiana e in lingua inglese[1], a testimonianza della pluralità degli approcci e dell'orizzonte internazionale della ricerca. Non abbiamo cercato un discorso unitario né una visione definitiva, ma piuttosto di mettere insieme prospettive diverse, complementari e talvolta divergenti, convinti che la sociologia, proprio nella pluralità dei suoi sguardi, possa offrire strumenti essenziali per comprendere i mutamenti in corso. I quindici scritti che compongono il libro possono essere ricondotti a cinque grandi filoni tematici: il potere e le disuguaglianze, la politica e la governance, la criminalità e la sicurezza, la cultura e la creatività, le relazioni umane e i futuri possibili. Questa suddivisione non è esplicitata nell'indice, perché non volevamo irrigidire la lettura, ma è utile come bussola per orientarsi tra le pagine.

DOI: 10.5281/zenodo.18435280

* UniCamillus – Saint Camillus International University of Health and Medical Sciences in Rome. vera.kopsaj@unicamillus.org.

** Università degli Studi "Gabriele d'Annunzio" di Chieti-Pescara. luca.corchia@unich.it.

[1] Gli scritti in lingua inglese presentano sia l'uso dell'inglese britannico sia di quello americano; tale scelta è stata rispettata conformemente all'uso degli autori. L'impiego dei trattini è stato invece uniformato secondo uno standard unico.

Una prima area di riflessione riguarda il legame tra IA, potere e disuguaglianze. Roberto Veraldi e Chiara Fasciani, con il loro contributo, mostrano come le nuove tecnologie possano rafforzare divari già esistenti, amplificando meccanismi di esclusione sociale laddove manchino politiche attente a garantire accesso e equità. La loro analisi mette in guardia contro una fiducia ingenua nella neutralità degli algoritmi e invita a considerare i rischi di una riproduzione automatizzata delle ingiustizie.

Giordana Truscelli affronta invece la questione delle nuove forme di potere che emergono nella società algoritmica: potere di profilazione, di predizione, di sorveglianza, che ridisegnano i confini dell'autonomia umana. Il suo saggio è un invito a riflettere su soluzioni possibili che salvaguardino la dignità e la libertà della persona, evitando che l'individuo diventi mera funzione di un calcolo statistico.

Un secondo gruppo di saggi si misura con il tema della politica e della governance. Alessandra De Luca e Antonello Canzano Giansante presentano una ricognizione bibliometrica sul rapporto tra intelligenza artificiale ed elezioni politiche, tracciando lo sviluppo della ricerca internazionale e restituendo un quadro utile a chi voglia comprendere come l'IA stia trasformando la sfera elettorale.

Luca Corchia propone una riflessione teorica sulla "trappola" della mimesi: l'intelligenza artificiale, imitando l'umano, rischia di trasformarsi in ideologia, alimentando un immaginario che confonde ciò che è simulazione con ciò che è realtà. La sua analisi è una riflessione critica per distinguere tra ciò che l'IA è effettivamente e ciò che viene proiettato su di essa.

Clara Salvatori e Mara Maretti si concentrano sul governo dei flussi migratori, mostrando come la *algorithmic governmentality* si traduca in strumenti di controllo e selezione. Le tecnologie che promettono efficienza e neutralità si rivelano invece profondamente politiche, perché determinano chi è incluso e chi è escluso dai diritti e dalle protezioni.

Domenico Trezza e Giuseppe Luca de Luca Picione analizzano il campo del welfare, evidenziando come l'introduzione di sistemi di IA rimodelli il rapporto tra attori sociali, logiche di scambio e forme di solidarietà. Il loro contributo interroga il delicato equilibrio tra efficienza e giustizia, tra innovazione tecnologica e bisogni delle persone.

La terza area di riflessione riguarda la criminalità, la sicurezza e il controllo sociale. Roberta Aurilia esplora l'ambivalenza dell'IA, che può essere usata sia come strumento di contrasto al crimine sia come mezzo di infiltrazione criminale. Il suo contributo mette in evidenza la necessità di un

controllo attento, per evitare che la tecnologia sfugga alle intenzioni originarie.

Franco Campitelli si concentra sulla cybersecurity, presentando opportunità e rischi derivanti dall'impiego dell'intelligenza artificiale nei sistemi di difesa digitale. La sua analisi richiama l'attenzione sulle sfide future della sicurezza in un mondo sempre più interconnesso.

Emanuela Susca, Federica Fortunato e Simonetta Mucolo mettono infine al centro il "fattore umano": nessuna tecnologia, per quanto sofisticata, può sostituire la consapevolezza e la responsabilità delle persone. La sicurezza digitale, ricordano, è prima di tutto una questione culturale e sociale.

Un quarto filone esplora i rapporti tra intelligenza artificiale, cultura e creatività. Roberta Grasselli propone il concetto di "capitale sociale computazionale" come nuova categoria teorica per analizzare le dinamiche digitali, aprendo prospettive innovative di ricerca.

Armando Saponaro riflette sul ruolo della *generative AI* non solo come strumento creativo, ma come attore sociale che ridefinisce i confini tra devianza e *mainstream*, suscitando interrogativi sulla legittimità e il riconoscimento delle pratiche culturali.

Francesca Guarini offre invece uno sguardo originale sugli immaginari che si costruiscono intorno agli algoritmi, intrecciando il linguaggio militare con le forme culturali contemporanee. La sua riflessione mostra come la tecnologia non sia neutra, ma generi narrazioni e simboli che influenzano profondamente la percezione sociale.

Il saggio di Vera Kopsaj apre lo sguardo al futuro delle relazioni umane. Attraverso il tema degli *artificial companions*, l'autrice esplora le implicazioni dell'IA per la salute mentale e la possibilità che le tecnologie simulino, sostituiscano o trasformino i legami affettivi e sociali. È una riflessione che interroga da vicino l'identità stessa dell'umano in un mondo sempre più popolato da intelligenze artificiali.

Niccolò Faccini tratta la giustizia riparativa e l'uso dell'IA nelle prigioni. L'autore esplora l'impiego di memorie sintetiche per accelerare la riabilitazione dei detenuti, mettendo in discussione la compatibilità di questo approccio con i principi della giustizia riparativa. L'autore solleva dubbi sull'uso dell'IA per manipolare le emozioni dei detenuti, rischiando di minare la responsabilità e il cambiamento necessari. Faccini riflette su una giustizia umana che promuova il dialogo e ricostruisca il senso di comunità tra vittima e reo.

Sara Sbaragli analizza come l'intelligenza artificiale stia trasformando la sanità, passando da una relazione medico-paziente diadica a una triadica

con l'algoritmo come terzo attore. L'IA può migliorare la cura riducendo burocrazia e facilitando la comprensione, ma rischia di introdurre opacità, *bias* e nuove disuguaglianze. La sfida è integrare l'IA in modo trasparente, equo e negoziabile, preservando fiducia, responsabilità e centralità dell'incontro clinico.

Questi quindici saggi non rappresentano un coro uniforme, ma un dialogo plurale. Ciò che li accomuna è l'attenzione a non ridurre l'IA a semplice innovazione tecnica, ma a considerarla come fenomeno sociale, politico e culturale. La varietà delle voci raccolte riflette la convinzione che solo un approccio interdisciplinare e critico possa restituire la complessità del presente.

Come curatori, desideriamo esprimere la nostra gratitudine a tutti gli autori e le autrici per la generosità e l'impegno profusi. Questo volume non vuole chiudere un dibattito, ma aprirlo: offrire strumenti a studenti, ricercatori, operatori e decisori pubblici per orientarsi di fronte a sfide che ci riguardano tutti. La speranza è che queste pagine possano diventare occasione di confronto e di apprendimento, e che possano contribuire a immaginare un futuro in cui l'intelligenza artificiale sia messa davvero al servizio della società e della dignità umana.

Roma-Pisa, 12 dicembre 2025

Algorithmic governmentality and social control in migration management

by Mara Maretti, Clara Salvatori[*]

Artificial intelligence (AI) is reshaping migration governance. Framed by algorithmic governmentality, this study examines AI systems across transit and destination countries, showing how they reconfigure classification and control by reducing personal experiences to risk profiles and enabling automated surveillance. Beyond this, it stresses the urgency of critical sociology to analyse AI-driven power shifts and shape human rights-based policies.

Keywords: migration; datafication; artificial intelligence; governmentality; algorithms; migration governance.

Governamentalità algoritmica e controllo sociale nella gestione delle migrazioni

L'intelligenza artificiale (IA) sta rimodellando la governance della migrazione. Inquadrato nella prospettiva della governamentalità algoritmica, questo studio analizza i sistemi di IA nei paesi di transito e destinazione, mostrando come riconfigurino classificazione e controllo riducendo le esperienze personali a profili di rischio. Sottolinea l'urgenza di una sociologia critica e di politiche fondate sui diritti umani.

Parole chiave: datificazione; intelligenza artificiale; governamentalità; algoritmi; migrazione; governance delle migrazioni.

Introduction

The scale and complexity of global human mobility have expanded significantly in recent decades, with the United Nations Department of Economic and Social Affairs (UN DESA) reporting that international migrants have nearly doubled from 154 million in 1990 to 304 million in 2024 (UN DESA, 2025). This growth has led to significant transformations in how states and international organisations approach migration governance, with digital technologies emerging as central instruments in this evolving

DOI: 10.5281/zenodo.18435478

[*] Università degli Studi "Gabriele d'Annunzio" di Chieti-Pescara. mara.maretti@unich.it; clara.salvatori@unich.it.

landscape. As migration flows intensify and diversify, artificial intelligence, big data and algorithmic systems are increasingly deployed to anticipate, manage and control human mobility.

This study demonstrates that algorithmic interventions in migration governance transcend operational efficiency, engendering novel schemas of classification and control which profoundly influence the responses of policymakers, institutions and public sentiment. To this purpose, a systematic mapping of the main AI-based systems currently employed in migration governance is proposed. These systems are compared according to several criteria: type of technology used, institutional actors involved, territorial scope, stated objectives, ethical and legal implications, degree of algorithmic transparency, and impact on migrants' rights. Our comparative analysis yields four typologies: border-control, migrant-management, return-enforcement and asylum-support systems. According with Rouvroy and Berns (2013), reveals that AI acts not as a neutral administrative device but as an agent transforming the relations among states, international bodies and migrants

Moreover, the research contributes to the wider debate on *algorithmic governmentality*, as conceptualised by Rouvroy and Berns (2013), by demonstrating that artificial intelligence – particularly within the domain of migration – does not function as a neutral instrument of administrative efficiency. Rather, it operates as a powerful mechanism that reconfigures relationships between states, international organisations and mobile individuals. The case studies examined underline how the automation of decisions in this field raises fundamental questions on agency, ethical responsibility and justice in the digital era. The following sections critically investigate the functioning of these systems, their implications and the scope for developing alternative models that balance administrative efficiency with the protection of human rights and individual dignity.

1. Theoretical framework

The intersection of artificial intelligence and migration constitutes an emerging field of sociological inquiry, revealing new modalities of power. Contemporary governmentality increasingly operates with computational systems that collect, categorise, and analyse data on migrants, profoundly reshaping how human mobility is managed (Beduschi, 2021).

AI-based technologies are reconfiguring migration governance regimes by structuring what Rouvroy and Berns (2013) describe as *algorithmic governmentality*: systems that do not merely reflect social reality but actively produce

regimes of truth, continually reshaping how migrants are classified, monitored, and governed.

Foucault's concept of governmentality (2009) provides a useful framework for interpreting these transformations. It is understood as the "conduct of conduct", a form of power that operates not through direct coercion but by shaping the field of possible actions and guiding the choices of individuals and populations. In his *genealogy of power*, Foucault identifies a shift from sovereign disciplinary regimes to regulatory forms of power that operate using security dispositifs and the production of truth regimes (Foucault, 2004). Algorithmic governmentality extends this logic by three key characteristics: large-scale, automated data collection, algorithmic processing to identifying statistical correlations, and anticipatory action based on predictive outputs (Rouvroy, Berns, 2013). The distinctiveness of this form of governance lies in its operation at an infra-individual level, bypassing reflective awareness (Rouvroy, 2013). Unlike disciplinary forms of power that act through visible institutions and explicit norms, algorithmic governmentality imposes an "immanent normativity" that shapes the very space of possible action, rather than prescribing behaviour directly.

In the context of migration, this marks a paradigmatic shift from earlier governance regimes that relied on disciplinary surveillance and fixed norms. The datafication of mobility management has enabled this transition to algorithmic forms of control (Broeders, Dijstelbloem, 2015). Algorithmic governmentality modulates the migrant's range of action through continuous analysis of their digital traces classifying individuals not by pre-existing social categories but through fluid behavioural patterns (Cheney-Lippold, 2017). This aligns with what Deleuze (1995) described as a *society of control*, in which disciplinary forms of power give way to continuous modulations governed by codes that regulate differential access to spaces and resources. In current digital era, such modularity is expressed through algorithmic profiling processes which, as Amoore (2013) notes, produce not precise knowledge of individuals but a probabilistic mapping of their potential for action and decision-making. As a result, algorithmic migration governance is structured around a regime of truth not based on objective representation but on the production of specific "performative ontologies" (Law, Urry, 2004), making migrants governable through their reduction to data patterns.

Recently, Bigo (2020) has identified three key functions which characterised algorithmic governmentality in migration field: transforming individuals into data patterns, anticipating migratory flows by means of predictive modelling, and automating decisions previously left to human discretion. Within this framework, the "truth" of the migrant no longer emerges from inquiries into

substantive identity, but from aggregated statistical correlations based on their digital behaviours. This resonates with Lyon's (2018) concept of a *culture of surveillance*, where control is exercised through preventive normalisation without the need for consent or awareness. Migrant datafication, as exemplified by International Organisation of Migration (IOM)'s MIDAS and United Nations High Commissioner for Refugees (UNHCR)'s PRIMES, transforms individual attributes into biometric profiles, enabling cross-border tracking and service allocation, yet flattening nuanced life histories into standardised datasets, frequently without migrants' informed consent (Singler, 2021; Madianou, 2019).

As van Dijck (2014) notes, datafication represents not merely a digitalisation process but rather an ontological shift that translates human experience into statistical correlations. Within migration contexts, this fragments identities into what Deleuze (1995) termed *dividuals* – discrete, manipulable units of data – altering the conditions under which migrants are recognised as rights-bearing subjects. Such fragmentation carries tangible consequences: when asylum seekers are reduced to biometric data points, Global Positioning System (GPS) traces and metadata, their personal narratives of persecution risk being overshadowed by automated assessments privileging consistency and plausibility. On this point, Metcalfe and Dencik (2019) document how systems increase rejection rates by flagging trauma-induced nonlinear narratives as incoherent.

Beyond the operational function of algorithmic systems, Introna (2016) clarifies the structural logic of datafication, identifying three interrelated mechanisms: *selective visibilisation* (emphasising certain aspects of identity while obscuring others), *forced commensurability* (reducing heterogeneous experiences to standardised metrics), and *anticipatory normalisation* (reshaping individual profiles in line with predictive expectations of risk or integration). These processes reveal that datafication is never neutral but inherently political, embedding specific worldviews within ostensibly objective technical infrastructures. This becomes evident in structural biases pervading algorithmic systems. As Beduschi (2021) argues, models trained on historical or incomplete data tend to replicate and amplify existing prejudices – particularly those related to nationality, ethnicity, or mobility patterns – thereby producing systemic inequalities masked as objectivity. The datafication processes underlying these systems inherently embed discriminatory logics that disproportionately affect marginalised populations (Leurs, Shepherd, 2017). Such bias emerges both from unbalanced data selection practices and predictive modelling approaches that prioritise statistical correlations over contextual and qualitative assessment.

Nevertheless, adopting algorithmic governmentality requires critical reflection on its analytical limits. Weiskopf and Hansen (2023) emphasise how algorithmic governmentality reconfigures the space of ethics, requiring new frameworks for understanding moral responsibility in automated decision-making systems. Critics warn against overly deterministic approaches that portray migrants solely as passive classification objects, neglecting their capacity to resist and reconfigure algorithmic rule conditions (Walters, Tazzioli, 2023). In practice, algorithmic regimes are marked by instability, contestation, and failure (Aradau, 2023). Migrant subjectivities, far from being fully governed, may act as agents of counter-conduct (Foucault, 2009), introducing unpredictability exceeding predictive logic. The notion of insurgent politics draws attention to collective practices that – despite arising from conditions of extreme precarity – interrogate and transform governance technologies (Lecadet, 2023). For this reason, analyses must therefore attend to such frictions, local contexts heterogeneity, and resistances emerging at the nexus of subjectivity, technology, and power.

2. Comparative Analysis of Automated Systems in Migration Governance

To understand the impact of automated systems employed in migration management, a comparative analysis of existing technologies was undertaken. The research systematically catalogued 61 AI systems, selected from 2015 to 2024, via comprehensive web-based investigation of academic and non-academic databases (Scopus, Web of Science, Google Scholar) using keywords such as "migration management" AND/OR "border control" AND ("Artificial Intelligence" OR "large language models" OR "machine learning" OR "predictive analysis"), institutional reports (UNHCR, IOM, Frontex, DG Home Affairs UE) and official documentation from national agencies (BAMF, CBP, IRCC). This catalogue is not exhaustive and reflects heterogeneous geographic coverage: some emerging systems in Africa and Southeast Asia may not have been captured.

These systems were subsequently classified according to multiple criteria: level of technological advancement (Tab. 1), geographical distribution and specific operational functionalities (Tab. 2). Regarding the classification of AI technological levels employed, these were defined according to six successive developmental stages, progressing from the most elementary to the most advanced. The most basic level includes rule-based systems, which operate through deterministic IF-THEN logic. These are typically found in traditional migration registries and eligibility verification tools. Their function is often

limited to mechanical bureaucratic tasks, offering minimal flexibility or learning capacity. The second level comprises supervised machine learning algorithms trained on labelled datasets, enabling predictions as seen in Swiss refugee allocation systems, which estimates the potential "integration merit" of applicants based on historical data patterns. Deep learning and computer vision form the third level, with neural networks performing facial recognition and behavioural analysis. Tools like iBorderCtrl and facial scanning technologies at European Union borders exemplify this class, enabling forms of algorithmic behavioural surveillance. The fourth category encompasses advanced biometric systems that integrate multimodal biometric data (such as iris, fingerprint, and facial recognition) with AI capabilities. Systems like Eurodac, MIDAS, and IrisGuard fall into this category, marking a significant shift towards the reduction of the migrant to a fully digitised biometric profile. At the fifth level, optimisation and matching algorithms, designed to allocate resources or individuals based on system-wide efficiency logics, are positioned. In this category are comprised matching systems used in Canada and Switzerland for refugee placement prioritise performative optimisation over individual preferences or rights, reflecting a systemic orientation towards administrative rationality. Finally, the most advanced level consists of large language models and generative AI, capable of autonomous content generation, human-like interaction, and self-learning. Although not yet widely deployed in operational migration governance, early experiments such as International Rescue Committee (IRC)'s Signpost.AI – a chatbot offering legal and asylum information – raise critical concerns regarding decision-making opacity and asymmetrical access to rights.

Tab. 1 – AI Systems for migration governance

Category	*Description*	*Technological level*	*Social Impact*
1. Rule-based systems	Deterministic if-then logic	Low	Mechanic bureaucratisation
2. Supervised machine learning	Training on labelled datasets	Medium	Predictive assessment of "integration merit"
3. Deep Learning and Computer Vision	Neural networks for facial/emotion recognition	Medium - High	Algorithmic behavioural surveillance
4. Advanced Biometric Systems	Multimodal biometrics and AI	High	Reduction of the migrant to a biometric profile
5. Optimisation/Matching Algorithms	Resource and allocation optimisation	High	Efficiency over rights

6. LLMs and Generative AI	Language models for automated interaction and decision-making	Very high (4th generation)	Opacity, asymmetrical access to rights

The analysis reveals a clear progression in technological sophistication, which has fundamentally transformed approaches in migration management. This evolution has produced a transition from systems initially oriented towards bureaucratic process mechanisation to technologies enabling pervasive monitoring of migrants through comprehensive collection of biometric and behavioural characteristics.

Concurrently, one observes increasing decision-making autonomy in intelligent systems, which are transitioning from mere support tools to entities capable of making critical determinations with limited human oversight. This phenomenon corresponds with a concerning rise in algorithmic opacity, marked by the progressive abandonment of explicit rules in favour of "black box" architectures that remain largely inscrutable even to their developers.

Within this evolutionary framework, a rigorous examination of implications for fundamental rights and the shifting representation of migrants within institutional systems becomes imperative. What emerges is an ontological transformation in which individuals are no longer recognised as a subject of rights but reframed as a dataset optimised for algorithmic management.

As technological complexity increases, so too do the associated ethical and social challenges, reorienting the discourse from administrative efficiency to deeper transformations in how migrants are perceived, categorised and processed by institutions.

Considering the distribution of the catalogued systems according to geographical region and operational function, a significant disparity becomes evident. There is a clear predominance of the Global North: European and North American implementations together account for approximately two-thirds of all documented systems. Intermediate adoption levels are observed in Asia and Oceania, while African implementation remains minimal. Three additional systems operate on a global scale. From a functional analysis perspective, border control emerges as the primary application, followed by information provision, asylum processing, integration services and return management. The concentration of border control technologies in Europe and North America reflects their role as principal destinations for international migration.

Tab. 2 - Identified Cases by Region and Function

Geographical Area	*Border control*	*Asylum requests*	*Deportation control*	*Integration*	*Information support*	*Total*
Europe	7	4	3	3	3	20
North America	6	3	3	4	3	19
Asia	4	1	1	1	1	8
Oceania	3	1	1	1	1	7
Africa	2	1	0	0	1	4
Global	0	0	0	0	3	3
Total	22	10	8	9	12	61

Conclusion

Artificial intelligence in migration governance is not just a technological innovation but constitutes the emergence of new classification and control mechanisms that fundamentally reshape relations between states, international organisations, and migrants. The research exposes unequal geographies of algorithmic power, where advanced systems are developed and managed by Global North states while those in the Global South remain mere implementation territories, subjected to foreign decision-making models lacking local adaptation and accountability. This creates an algorithmic divide that reproduces existing hierarchies through a dual dynamic: countries with advanced predictive capabilities impose standardised frameworks for risk and integration assessment, while transit and origin states must implement these tools without meaningful participation in their design or the ability to contest their determinations. This asymmetry consolidates algorithms as transnational mechanisms of power that reinforce existing hierarchical relations, rather than serving as neutral technical instruments.

Borders increasingly function as a mobile biometric network, tracking migrants across time and space through digital identification systems embedded in pervasive surveillance regimes. This shift reduces personal identity to computational risk profiles, subordinating access to rights to compliance with often invisible algorithmic criteria. Migrant datafication emerges as a mechanism of systematic dehumanisation, transforming the complexity of individual existence into measurable parameters and predefined categories – at the cost of erasing human dimensions, personal narratives, and specific socio-cultural contexts.

Additionally, the deployment of big data and AI in migration governance ultimately serves institutional rather than migrant interests (Bircan, Korkmaz, 2021). Massive collection of personal and biometric data, often without

genuine informed consent, produces a "surveillance bargain": migrants are compelled to surrender privacy and informational self-determination in exchange for access to fundamental rights such as asylum or humanitarian aid. This dynamic raises serious concerns about the human rights impacts of new technologies in migration contexts (Molnar, 2019). Especially concerning is the overlap between migration and criminal surveillance systems, which fosters structural criminalisation by positioning migrants as suspects rather than individuals in need of protection.

Analysis of the 61 documented systems further highlights a persistent tension between administrative efficiency and the protection of fundamental rights. These technologies embed assumptions about migration and risk reinforcing systemic bias, particularly when trained on historical data that reproduces past prejudices under a veneer of objectivity.

However, such transformation is not unidirectional. Alongside emerging technocracy, significant forms of resistance and counter-use arise, including legal contestation strategies, technological evasion practices, and advocacy movements for more ethical AI governance in migration contexts. These practices of counter-conduct, push beyond algorithmic prediction, challenging dominant paradigms and opening space for political contestation. Within these legal, political, and social domains, there is the potential to reimagine governance models that reject reductive profiling in favour of recognising the complexity and dignity of migrant lives. The core challenge emerging from this analysis involves developing governance models that, while benefiting from digital innovation, remain grounded in principles of social justice and human rights, avoiding a drift towards total control that reduces migration to algorithmic calculation. Policy measures should mandate algorithmic impact assessments, conduct regular independent audits of AI systems for bias and transparency, and implement participatory co-design processes involving Global South stakeholders, ensuring that these technologies facilitate not only control and security but also pathways to inclusive integration and the protection of fundamental human rights.

References

Amoore L. (2013). *The Politics of Possibility: Risk and Security Beyond Probability*. Durham: Duke University Press.

Aradau C. (2023). Algorithmic governmentality: questions of method. In: Walters W., Tazzioli M., a cura di, *Handbook on governmentality*, pp. 235-250. Cheltenham: Edward Elgar Publishing.

Beduschi A. (2021). International migration management in the age of artificial intelligence. *Migration Studies*, 9(3): 576-596. DOI: 10.1093/migration/mnaa003.

Bigo D. (2020). The socio-genesis of a guild of "digital technologies" justifying transnational interoperable databases in the name of security and border purposes. *International Journal of Migration and Border Studies*, 6(1-2): 74-92. DOI: 10.1504/IJMBS.2020.108689.

Bircan T., Korkmaz E.E. (2021). Big data for whose sake? Governing migration through artificial intelligence. *Humanities and Social Sciences Communications*, 8(1): 241. DOI: 10.1057/s41599-021-00910-x.

Broeders D., Dijstelbloem H. (2015). The datafication of mobility and migration management. In: van der Ploeg I., Pridmore J., a cura di, *Digitizing Identity: A Reader on Biometrics and Surveillance*, pp. 242-260. London-New York: Routledge.

Cheney-Lippold J. (2017). *We Are Data: Algorithms and the Making of Our Digital Selves*. New York: NYU Press.

Deleuze G. (1995). Postscript on control societies. In: *Negotiations*, pp. 177-182. New York: Columbia University Press.

Foucault M. (2004). *Naissance de la biopolitique. Cours au Collège de France, 1978-1979*. Paris: Gallimard/Seuil.

Foucault M. (2009). *Security, Territory, Population: Lectures at the Collège de France, 1977-1978*. Basingstoke: Palgrave Macmillan.

Introna L.D. (2016). Algorithms, governance, and governmentality: on governing academic writing. *Science, Technology, & Human Values*, 41(1): 17-49. DOI: 10.1177/016224391558736.

Law J., Urry J. (2004). Enacting the social. *Economy and Society*, 33(3): 390-410. DOI: 10.1080/0308514042000225716.

Lecadet C. (2023). Insurgent politics: refugees, sans-papiers and deportees under asylum and migration laws. In: Walters W., Tazzioli M., a cura di, *Handbook on governmentality*, pp. 405-420. Cheltenham: Edward Elgar Publishing.

Leurs K., Shepherd T. (2017). Datafication & discrimination. In: Schäfer M.T., Van Es K., a cura di, *The datafied society: Studying culture through data*, pp. 211-231. Amsterdam: Amsterdam University Press.

Lyon D. (2018). *The Culture of Surveillance: Watching as a Way of Life*. Cambridge: Polity Press.

Madianou M. (2019). Technocolonialism: digital innovation and data practices in the humanitarian response to refugee crises. *Social Media + Society*, 5(3): 1-13. DOI: 10.1177/2056305119863146.

Metcalfe P., Dencik L. (2019). The politics of big borders: data (in)justice and the governance of refugees. *First Monday*, 24(4). DOI: 10.5210/fm.v24i4.9934.

Molnar P. (2019). New technologies in migration: human rights impacts. *Forced Migration Review*, 61: 6-8. https://www.fmreview.org/ethics/molnar (consultato il 20 marzo 2025).

Rouvroy A. (2013). The end(s) of critique: data-behaviourism vs. due process. In: Hildebrandt M., De Vries K., a cura di, *Privacy, Due Process and the Computational Turn*, pp. 143-168. London-New York: Routledge.

Rouvroy A., Berns T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux*, 177(1): 163-196. DOI: 10.3917/res.177.0163.

Singler S. (2021). Biometric statehood, transnational solutionism and security devices. *Theoretical Criminology*, 25(3): 454-473. DOI: 10.1177/13624806211031245.

United Nations Department of Economic and Social Affairs, Population Division (2025). *International Migrant Stock 2024: Key facts and figures* (UN DESA/POP/2024/DC/NO. 13). New York: United Nations. https://www.un.org/development/desa/pd/content/international-migrant-stock (consultato il 18 febbraio 2025).

van Dijck J. (2014). Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. *Surveillance & Society*, 12(2): 197-208. DOI: 10.24908/ss.v12i2.4776.

Walters W., Tazzioli M., a cura di (2023). *Handbook on governmentality*. Cheltenham: Edward Elgar Publishing.

Weiskopf R., Hansen H.K. (2023). Algorithmic governmentality and the space of ethics. *Human Relations*, 76(3): 483-506. DOI: 10.1177/00187267221075346.

Intelligenza artificiale e disuguaglianze sociali: un approccio sociologico

di *Roberto Veraldi*[*]*, Chiara Fasciani*[**]

Negli ultimi decenni l'intelligenza artificiale si è affermata come tecnologia rivoluzionaria capace di trasformare la società, migliorare l'efficienza e aprire nuove opportunità. Tuttavia, la crescente adozione di sistemi digitali solleva preoccupazioni significative riguardo al rischio di attivare o amplificare disuguaglianze sociali. Questo lavoro vuole esplorare come l'IA influenzi le disuguaglianze esistenti e ne crei di nuove, concentrandosi su due aspetti principali: le disparità di accesso alle nuove tecnologie e i bias algoritmici. Da un lato, l'accesso limitato al digitale può escludere interi gruppi sociali dai benefici dell'innovazione tecnologica; dall'altro lato gli algoritmi decisionali possono perpetuare e amplificare pregiudizi esistenti. Attraverso la conduzione di interviste rivolte a testimoni privilegiati, questo studio offre una prospettiva approfondita sulle sfide e le opportunità associate alla crescente digitalizzazione. Questo studio intende fornire le basi per la promozione di una maggiore consapevolezza e di politiche sociali concrete per garantire che l'IA diventi uno strumento di progresso sociale e non di esclusione.

Parole chiave: intelligenza artificiale; disuguaglianze sociali; bias algoritmici; disparità digitale; opportunità; politiche sociali.

Artificial intelligence and social inequalities: a sociological approach

In recent decades, artificial intelligence has emerged as a revolutionary technology capable of transforming society, improving efficiency and opening new opportunities. However, the growing adoption of digital systems raises significant concerns about the risk of activating or amplifying social inequalities. This work aims to explore how AI influences existing inequalities and creates new ones, focusing on two main aspects: unequal access to new technologies and algorithmic biases. On one hand, limited digital access can exclude entire social groups from the benefits of technological innovation; on the other hand, decisional algorithms can perpetuate and amplify existing biases. Through the conduction of interviews with key stakeholders, this study offers an in-depth perspective on the challenges and opportunities associated with the growing digitalization. This study aims to provide the basis for promoting greater awareness and concrete social policies to ensure that AI becomes a tool for social inclusion, not exclusion.

Keywords: artificial intelligence; social inequalities; algorithmic biases; digital inequality; opportunities; social policies.

DOI: 10.5281/zenodo.18435502

[*] Università degli Studi "Gabriele d'Annunzio" di Chieti-Pescara. rveraldi@unich.it; chiara.fasciani@collaboratori.unich.it.

[**] Università degli Studi Magna Graecia di Catanzaro. chiara.fasciani@studenti.unicz.it.

Sebbene frutto di lavoro congiunto, si possono attribuire a R. Veraldi l'introduzione e i paragrafi da 1. a 4. compreso e a C. Fasciani i paragrafi da 4.1. fino alle prime considerazioni conclusive.

Roberto Veraldi, Chiara Fasciani

Introduzione

Recentemente, l'intelligenza artificiale ha espresso il proprio potenziale rivoluzionario e trasformativo in vari settori della società, mostrandosi come uno strumento in grado di migliorare l'efficienza e l'efficacia dei processi in diversi settori e aprendo nuove possibilità di impiego e sviluppo (Kissinger *et al.,* 2023). Tuttavia, tale entusiasmo è accompagnato da crescenti interrogativi di natura etica e sociale. In particolare, da un lato si discute sulla disparità di accesso alle tecnologie intelligenti, che rischia di aggravare il divario digitale; dall'altro, sulla presenza di bias algoritmici nei sistemi decisionali automatizzati, capaci di replicare o addirittura amplificare forme di discriminazione preesistenti (O'Neil, 2017; Lazzini, 2023).

Questo lavoro si propone di riflettere sul rapporto tra intelligenza artificiale e disuguaglianze sociali da un punto di vista sociologico, con l'obiettivo di evidenziare non solo i rischi, ma anche le opportunità per un uso più consapevole.

1. Intelligenza Artificiale: cenni storici e impatto sociale

L'espressione "Intelligenza Artificiale" (IA) si riferisce a sistemi o macchine che imitano l'intelligenza umana durante l'esecuzione di specifici compiti per i quali sono programmati e che possono migliorarsi in itinere sulla base delle informazioni che raccolgono e/o che vengono immesse nella loro memoria (Russel, Norvig, 2005). Sin dalla sua nascita negli anni '50, l'IA ha attraversato varie fasi di sviluppo, evolvendo da semplici programmi di risoluzione dei problemi a complessi algoritmi di apprendimento automatico e reti neurali profonde. In particolare, già nel 1956 McCarthy si riferiva alla possibilità di costruire un dispositivo in grado di simulare i singoli passaggi del ragionamento e dell'apprendimento umano (Carlucci, Cialdea, 1995). Queste tecnologie si sono poi evolute nel corso del tempo permettendo alle macchine di acquisire enormi quantità di dati, prendere decisioni e modellizzare e anticipare i comportamenti umani (Teti, 2025).

Negli ultimi anni si è assistito a un rinnovato interesse verso la disciplina dell'automazione, grazie alla crescita vertiginosa della mole di dati disponibili e agli sviluppi tecnologici correlati. Nel 2018, la Commissione Europea definisce l'IA come «*Systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals*» (2018: 1), mettendo in evidenza la capacità dei sistemi di analizzare l'ambiente e agire con un certo livello di autonomia per raggiungere specifici obiettivi. L'enfasi sul comportamento intelligente dei

sistemi di IA implica che essi non si limitano alla sola rielaborazione di dati, ma sono in grado di imparare dai propri errori e migliorare le proprie prestazioni (Broussard, 2019).

È ormai insindacabile che le dinamiche sociali sono totalmente immerse (se non sostituite) dall'interazione mediata dagli strumenti digitali. Al di là dei software di messaggistica che hanno iniziato a diffondersi esponenzialmente tra le diverse generazioni sin dai primi anni 2000, ad oggi la tecnologia permea ogni aspetto della nostra vita: dalla sveglia all'ascolto delle notizie, dal lavoro al mondo della scuola, fino a condizionare buona parte dell'intrattenimento e del tempo libero. Altro aspetto interessante che cambia le dinamiche sociali e relazionali è quello che riguarda l'uso di assistenti virtuali come Siri o Alexa: la possibilità di parlare direttamente (in chat o a voce) con un'IA rende le interazioni uomo-macchina (U-M) più intuitive e immediate. In sintesi, in un arco temporale di circa dieci anni, i supporti digitali hanno cambiato completamente la società intera, andando ad intervenire in modo pervasivo sul modo di vivere e di pensare dei singoli individui. A cambiare però, sono state soprattutto le persone.

2. Trasformazioni nel mondo del lavoro e disoccupazione digitale

L'influenza dell'IA si estende ben oltre la sfera delle interazioni quotidiane e delle dinamiche sociali, ponendo nuove sfide per la società nel suo complesso con risvolti anche nel mondo del lavoro.

Un primo ambito da considerare è, infatti, l'evoluzione dell'organizzazione del lavoro, caratterizzata da una crescente diffusione dello smart working, reso possibile da strumenti digitali sempre più avanzati. Videoconferenze, piattaforme di gestione dei progetti, condivisione in tempo reale di file e documenti e la possibilità di collaborare simultaneamente su uno stesso contenuto, sono solo alcuni esempi delle innovazioni che stanno rivoluzionando le modalità operative.

Un secondo aspetto rilevante riguarda l'impiego dell'IA per il monitoraggio delle performance e della produttività dei dipendenti. Sebbene queste tecnologie possano contribuire a una maggiore efficienza, sollevano anche delicate questioni etiche, in particolare per quanto riguarda la tutela della privacy e il rischio di un controllo eccessivo sul luogo di lavoro.

Tuttavia, l'aspetto socialmente più rilevante e preoccupante riguarda l'automazione di compiti ripetitivi e manuali che può portare alla perdita di posti di lavoro in settori come la manifattura, il commercio al dettaglio e i servizi (Hatzius, 2023). In particolare, l'emergere recente dell'intelligenza artificiale generativa ha sollevato interrogativi sull'accelerazione della

disoccupazione tecnologica, che colpisce in particolare i lavoratori meno qualificati, esponendoli a forme crescenti di povertà (Zahidi, 2023). Oggi la frontiera tra uomo e macchina si è evidentemente spostata e il *Future of Jobs Report 2023* stima che a livello globale il 66% delle mansioni lavorative è svolto da esseri umani e il restante 34% da robot, prevedendo un ulteriore aumento dell'automazione e, di conseguenza, della disoccupazione umana (Word Economic Forum, 2023).

Ovviamente, accanto alla scomparsa di alcune professioni emergono nuovi lavori legati allo sviluppo e alla manutenzione dei sistemi algoritmici, che però richiedono competenze avanzate in informatica, ingegneria e analisi dei dati. Ciò, tuttavia, non riuscirà a compensare la perdita delle occupazioni meno qualificate né a sostenere la transizione professionale in un mercato sempre più in evoluzione.

È evidente che tale fenomeno di disoccupazione tecnologica tenderà inevitabilmente ad inasprire le disuguaglianze socio-economiche esistenti, creando una società in cui solo una piccola élite beneficia pienamente delle innovazioni tecnologiche, mentre la maggioranza lotta per sopravvivere. Richiamando Marx (1972) potremmo descrivere una classe di «capitalisti digitali» che possiedono e controllano le tecnologie e accumulano vantaggi economici e una classe proletaria che tenta di adattarsi ai rapidi cambiamenti e alle nuove richieste del mercato. Di fronte a questo scenario, a nostro avviso, le ipotesi di intervento sono due: da un lato, chi dispone di risorse economiche e capacità cognitive, tenderà a orientarsi verso studi informatici per accedere alle nicchie del lavoro digitale; dall'altro, chi non ha i mezzi né inclinazioni personali verso questi ambiti, sarà orientato a quei settori del mercato che, *ad oggi*, sono interessati solo parzialmente dall'impiego di IA. È bene sottolineare che questo secondo ambito non comprende solo lavori manuali, fisici e altamente creativi (artigianato, oreficeria, forze dell'ordine…), ma anche professioni basate sul contatto umano e su qualità come empatia e autorevolezza (psicologo, assistente sociale, insegnante, sindaco…).

Qualcuno potrebbe replicare che sì l'intelligenza artificiale non può, ad oggi, sostituire queste professioni, però può sicuramente portare un valore aggiunto, semplificare molti passaggi intermedi o aiutare negli aspetti organizzativi. Ed è proprio questo il punto. L'IA dovrebbe essere considerata come uno strumento supplementare dei professionisti e non come un loro sostituto, da qui la necessità di condurre studi approfonditi, per poter valutare le rischi e potenzialità del loro utilizzo. Partendo da questa prospettiva sarà essenziale proporre percorsi di aggiornamento e di formazione digitale sia nei percorsi universitari sia direttamente nei luoghi di lavoro per non trovarsi impreparati di fronte ai cambiamenti.

3. Impatto socio-economico delle disuguaglianze digitali

Il crescente impiego dell'IA in vari settori porta con sé una serie di contraddizioni intrinseche che necessitano di una riflessione approfondita e di un'analisi accurata.

In primo luogo, all'interno del tema delle responsabilità per le decisioni prese da sistemi autonomi rientra il problema significativo e complesso dei bias algoritmici nel training delle IA: questo fenomeno si verifica quando i modelli riflettono o amplificano i pregiudizi esistenti nei dati utilizzati per il loro addestramento (O'Neil, 2017). Se i dataset utilizzati per addestrare i modelli decisionali non sono rappresentativi o sono contaminati da stereotipi, l'IA potrebbe penalizzare gruppi vulnerabili come minoranze etniche, persone con disabilità o cittadini a basso reddito (Lazzini, 2023). Sono diversi i casi conclamati in cui illustri studiosi hanno potuto verificare il perpetuarsi di decisioni viziate da bias discriminatori interiorizzati dagli algoritmi. Ad esempio, gli algoritmi di rischio usati nel sistema di giustizia penale possono avere pregiudizi razziali, sviluppati sulla base delle precedenti sentenze raccolte negli archivi dei tribunali. In riferimento a tale aspetto, uno studio di ProPublica (2016) ha mostrato che il sistema COMPAS, utilizzato per valutare il rischio di recidiva negli USA, tendeva a classificare – erroneamente – le persone di origine afroamericana come ad alto rischio più frequentemente rispetto a quelle di origine caucasica (Anwing *et al*., 2016). Questo tipo di bias "razziale" può avere gravi conseguenze, influenzando le decisioni dei giudici relativamente a condanne, cauzioni, libertà condizionali o altre misure. In ambito sociosanitario, lo studio di Obermeyer *et al*. (2019) ha messo in luce il bias razziale insito in un algoritmo utilizzato per la gestione della salute negli USA. Il sistema, apparentemente neutrale, assegna le risorse basandosi sulla spesa sanitaria passata, presupponendo che chi ha speso di più in cure abbia maggior bisogno di assistenza futura. Tuttavia, poiché i pazienti afroamericani, a parità di condizioni mediche, tendono a spendere meno rispetto ai bianchi, spesso a causa di ostacoli economici, l'algoritmo sottovaluta sistematicamente i loro bisogni (Obermeyer *et al*., 2019). Gli esempi sono tanti, ma è ben presto dimostrabile come gli algoritmi non possono agire in modo neutrale perché riflettono i pregiudizi dei dati che abbiamo usato per addestrarli, cioè i nostri.

Approfondendo il tema delle decisioni prese dai sistemi autonomi, alcuni studiosi, come Lazzini (2023), sottolineano come il decisionismo algoritmico possa condurre a esiti distorti che finiscono per esasperare le logiche discriminatorie. Questo aspetto potrebbe risultare particolarmente dannoso in quei settori in cui le decisioni incidono direttamente sulla vita delle persone, come l'assegnazione di sussidi o la pianificazione di interventi socio-

assistenziali. I sistemi automatizzati potrebbero escludere richiedenti che in realtà avrebbero diritto a determinati benefici, solo perché il loro profilo non rientra nelle categorie previste dagli algoritmi (Eubanks, 2018). In questo modo, le decisioni automatizzate, utilizzate per risparmiare e semplificare i processi, potrebbero negare assistenza a chi ne ha bisogno. La Eubanks (2018) descrive questo fenomeno come un "*digital poorhouse*", cioè un sistema in cui i più vulnerabili sono ulteriormente penalizzati per l'accesso ai servizi essenziali, perpetuando cicli di povertà e marginalizzazione. Inoltre, molti algoritmi utilizzati nel welfare funzionano come "scatole nere", ossia senza che i cittadini e gli operatori sociali possano comprendere pienamente i criteri utilizzati per prendere decisioni. Questo crea un deficit di responsabilità, poiché diventa difficile contestare decisioni ingiuste o errate. Lazzini (2023) sottolinea che, in un sistema democratico, è fondamentale garantire meccanismi di controllo e revisione umana per evitare che l'IA prenda decisioni arbitrarie o errate senza possibilità di appello. Senza adeguate regolamentazioni e principi etici, il rischio è che il welfare diventi sempre più disumanizzato, trasformando i cittadini in semplici numeri all'interno di un sistema automatizzato.

Dal nostro punto di vista, gli strumenti di IA hanno il potenziale di modificare le strutture di potere (che ad oggi riguardano prioritariamente la distribuzione delle risorse economiche) e inasprire le disuguaglianze sociali. Le tecnologie avanzate richiedono infrastrutture digitali e competenze specifiche, ma l'accesso a queste risorse non è uniforme tra regioni e comunità. Le aree con limitata connessione internet o poche risorse strumentali rischiano di rimanere indietro, aumentando il divario tra chi può sfruttare appieno le potenzialità dell'IA e chi no. Nel descrivere l'avvento della "network society", Castells (1999; 2014) metteva in luce già alla fine del '900 la natura strutturale del divario digitale, sottolineandone le implicazioni profonde in termini di disuguaglianze sociali e accesso al potere informativo. In una "società interconnessa" l'esclusione digitale è una delle principali sfide odierne perché rischia di isolare chi non possiede dispositivi elettronici e di contribuire all'aumento delle disuguaglianze (Dominici, 2018; Parra-Saiani, 2020). Nel contesto contemporaneo, il *digital divide* viene generalmente definito come la distanza tra chi ha la possibilità di accedere e utilizzare efficacemente le tecnologie digitali e chi, per ragioni sociali, anagrafiche o territoriali, ne è escluso.

Questo divario non riguarda solo l'accesso materiale ai dispositivi o alla rete, ma si estende anche alle competenze necessarie per trarre beneficio dalla partecipazione attiva alla c.d. società della conoscenza, con effetti diretti sulla cittadinanza, l'inclusione e le opportunità di sviluppo individuale e collettivo. Negli ultimi anni, infatti, si è assistito a una parziale

democratizzazione delle tecnologie digitali: salvo rare eccezioni, la maggior parte delle persone dispone oggi di uno smartphone e di una connessione a internet. Tuttavia, questa apparente uniformità nell'accesso non elimina la disuguaglianza, che si manifesta in forme più sottili e profonde legate all'effettiva capacità di usare la tecnologia in modo autonomo, consapevole e produttivo. Infatti, è sempre più necessario spostare l'attenzione dal *digital divide* inteso come divario tra "*haves*" e "*have-nots*", misurata in termini dicotomici, al concetto più articolato di "*digital inequality*". Si fa riferimento alle disuguaglianze tra individui che, pur avendo formalmente accesso a dispositivi connessi alla rete, presentano significative differenze nell'effettivo utilizzo delle tecnologie digitali. All'interno di questo gruppo esistono infatti numerose situazioni intermedie, in cui il potenziale tecnologico viene sfruttato solo parzialmente a causa di un basso livello di autonomia, competenze digitali limitate o dispositivi di scarsa qualità (Di Maggio, Hargittai, 2001).

4. Una ricerca qualitativa

L'obiettivo della ricerca è quello di esplorare le percezioni e le opinioni riguardo all'impatto dell'IA sulle disuguaglianze sociali e i bias algoritmici, nonché le possibili soluzioni per garantire un accesso equo a queste tecnologie. Per indagare in profondità questi temi, si è adottato un approccio qualitativo basato sulla realizzazione di interviste rivolte a testimoni privilegiati. Gli interlocutori sono stati selezionati tra professionisti, studenti ed esperti che operano nel campo dell'intelligenza artificiale, della progettazione dei servizi digitali e dell'etica tecnologica. Lo strumento per la raccolta dei dati è stato quello dell'intervista strutturata, con un set di dodici domande identiche sottoposte a ciascun partecipante. Le domande hanno esplorato temi complessi e articolati sui benefici e i rischi dell'IA e dei bias algoritmici; sono state altresì raccolte proposte per una maggiore equità nell'uso delle tecnologie intelligenti. La scelta di una metodologia qualitativa ci ha consentito di cogliere la complessità delle visioni e delle esperienze individuali, andando oltre le semplificazioni dei discorsi tecnocentrici e riportando al centro dell'analisi la dimensione umana e relazionale dell'innovazione tecnologica. L'analisi delle risposte ha consentito l'identificazione dei principali nuclei tematici ricorrenti e delle divergenze significative tra le opinioni espresse. Particolare attenzione è stata dedicata alle raccomandazioni concrete formulate dagli intervistati per affrontare le problematiche identificate.

4.1. Il profilo degli intervistati

Le interviste hanno coinvolto cinque soggetti con diversi background e livelli di esperienza nel campo dell'IA:

- Intervistato 1: Responsabile dei sistemi informativi di una pubblica amministrazione con oltre trent'anni di esperienza in information technology e referente per la cybersecurity.
- Intervistato 2: Studente di Ingegneria Biomedica, già laureato in Biotecnologie, con interesse profondo per l'IA nel campo della diagnostica clinica e della ricerca.
- Intervistato 3: Laureando di informatica con formazione specifica sull'IA.
- Intervistato 4: Matematico di formazione, insegnante nelle scuole superiori e università, esperto di programmazione algoritmica.
- Intervistato 5: Laureando in informatica, esperto di IA e tecnologie digitali.

Questa diversità ha permesso di ottenere una visione multidimensionale delle problematiche legate al tema, combinando prospettive pratiche, teoriche e educative. In generale, le interviste hanno rivelato prospettive diverse ma complementari, che spaziano dalle preoccupazioni per l'amplificazione delle disuguaglianze esistenti alle speranze per un futuro in cui l'IA possa fungere da strumento di democratizzazione della conoscenza. Attraverso l'analisi di queste testimonianze, si intende altresì individuare proposte concrete per tentare di affrontare le sfide sociali poste dall'avanzamento dell'IA.

4.2. Analisi dei nuclei tematici emergenti

L'analisi qualitativa ha fatto emergere una serie di nuclei tematici ricorrenti, che delineano le principali aree di riflessione e preoccupazione legate all'impatto dell'IA sulla società, in particolare in relazione alle disuguaglianze sociali. Questi nuclei costituiscono i principali filoni su cui si articolano le opinioni espresse dai partecipanti, offrendo una panoramica articolata delle dinamiche in gioco e possono essere riassunti nella figura che segue (Fig. 1).

Fig. 1 – Rappresentazione grafica dei nuclei tematici rilevati nelle interviste.

Impatto sulle Disuguaglianze Sociali. Emerge una significativa divergenza di opinioni riguardo all'impatto dell'IA sulle disuguaglianze sociali. L'Intervistato 1 sostiene che "l'intelligenza artificiale, per sua natura, non è neutra rispetto alle disuguaglianze" e oggi tende maggiormente ad amplificarle a causa dell'accesso diseguale alle infrastrutture digitali, dell'automazione che colpisce i lavoratori meno qualificati e della concentrazione del potere nelle mani di poche aziende tecnologiche. L'Intervistato 4 si allinea con la visione più critica, affermando che «l'intelligenza artificiale è senza dubbio un amplificatore di disuguaglianze» che farà sparire numerosi lavori accessibili senza un lungo percorso di formazione, accrescendo la competizione per le famiglie svantaggiate e concentrando ulteriormente il potere economico. In contrasto, l'Intervistato 2 esprime una visione più ottimistica, sostenendo che i Large Language Models (LLM) come ChatGPT «hanno il potenziale per quasi azzerare le disuguaglianze sociali» rendendo accessibili enormi quantità di informazioni a chiunque disponga di una connessione internet. Secondo questa prospettiva, «con gli LLM a disposizione di tutti, l'ignoranza diventa una scelta, non uno stato».

Bias Algoritmici e Metodi Correttivi. Tutti gli intervistati riconoscono l'esistenza di bias algoritmici nei sistemi di IA. L'Intervistato 1 cita esempi concreti come il software COMPAS negli USA, che sovrastimava il rischio di recidiva per le persone nere, e gli algoritmi sanitari che sottovalutavano la gravità delle condizioni dei pazienti afroamericani. Propone metodi correttivi come audit algoritmici, dataset bilanciati, *fairness testing*, *explainable AI* e la partecipazione di esperti sociali ed etici nella progettazione. L'Intervistato 5 menziona il caso dell'algoritmo di reclutamento di Amazon che

penalizzava sistematicamente i curriculum femminili perché addestrato su dati storici in cui la maggioranza delle assunzioni tecniche riguardava uomini. Sottolinea l'importanza della "filtrazione e bilanciamento del dataset di addestramento" e delle "tecniche di fine-tuning supervisionato". Una proposta interessante viene dall'Intervistato 2, che suggerisce di «addestrare un'IA con scritti filosofici sui concetti di etica, morale, giustizia e consapevolezza, e utilizzarla come 'giudice' per valutare (e magari correggere) gli output delle altre intelligenze artificiali».

Categorie a rischio. Dall'analisi delle testimonianze raccolte emerge un consenso riguardo alle categorie sociali maggiormente a rischio di esclusione dai benefici dell'IA. Le persone coinvolte nell'indagine hanno identificato principalmente:

- Anziani: per la scarsa alfabetizzazione digitale e la difficoltà a adattarsi ai cambiamenti tecnologici.
- Persone con disabilità: spesso escluse da sistemi non progettati secondo criteri di accessibilità.
- Persone a basso reddito: ostacolate nell'accesso a dispositivi e connessioni adeguate.
- Lavoratori poco qualificati: maggiormente esposti all'automazione e con meno risorse per una riconversione professionale.

Una interessante suggestione proviene, in particolare, dall'Intervistato 1 che aggiunge all'elenco delle persone comunemente considerate a rischio di esclusione digitale anche "le minoranze etniche e linguistiche", che effettivamente risulterebbero penalizzate da modelli addestrati prevalentemente su dati anglofoni e occidentali.

Soluzioni per l'inclusione digitale e ruolo della formazione. In generale, le persone consultate hanno proposto soluzioni concrete per prevenire l'esclusione digitale dei cittadini più fragili che possono essere riassunte, per esigenze di chiarezza, nella tabella che segue (Tab. 1).

Tab. 1 – Tabella riassuntiva delle soluzioni proposte

Soluzioni proposte	*Opinioni degli intervistati*
Sportelli fisici di assistenza	L'Int.1 suggerisce «l'implementazione di sportelli digitali assistiti nei comuni e nelle ASL», mentre l'Int.5 parla di «spazi fisici dove le persone possano chiedere aiuto, parlare con qualcuno, farsi spiegare come usare un'app o un servizio digitale».
Programmi di alfabetizzazione digitale	L'Int.4 propone «un programma nazionale a livello scolastico e/o televisivo, al pari di quello del PC nelle scuole elementari e della lingua italiana in TV», mentre l'Int.3 parla di «programmi pubblici di alfabetizzazione digitale».

Interfacce accessibili	L'Int.1 suggerisce «lo sviluppo di interfacce semplici, multilingue e vocali», mentre l'Int.5 sottolinea il potenziale degli assistenti vocali che «permettono un'interazione più naturale, senza bisogno di toccare uno schermo o capire un'interfaccia».
Supporto telefonico	L'Int.4 propone di «mettere a disposizione delle linee telefoniche per eventuali dubbi o difficoltà», mentre l'Int.3 sottolinea che «il governo deve garantire sempre alternative fisiche o telefoniche ai servizi digitali».

L'alfabetizzazione digitale è vista come una condizione necessaria per garantire equità e partecipazione democratica nell'era dell'intelligenza artificiale. Tutti gli esperti coinvolti concordano sull'importanza cruciale della formazione e dell'educazione digitale per colmare il divario nell'accesso alle tecnologie basate su IA. Gli intervistati ribadiscono l'urgenza di investire in programmi educativi che rendano le persone capaci non solo di usare, ma anche di comprendere criticamente le tecnologie emergenti. L'Intervistato 1 afferma che "la formazione risulta fondamentale. Essa crea consapevolezza critica, non solo competenze tecniche e deve iniziare nelle scuole primarie per poi continuare in quelle secondarie e nelle università". L'Intervistato 3 sottolinea che «la scuola dovrebbe essere la prima a sensibilizzare i ragazzi ad un uso responsabile dell'IA, con criterio e giudizio» e suggerisce che «spiegare queste nuove tecnologie da un punto di vista tecnico potrebbe aiutare a far capire ai ragazzi che non si tratta di un miracoloso genio della lampada, bensì di una macchina di calcoli con margine di errore». L'Intervistato 4 è ancora più categorico, affermando che «la formazione può contribuire in misura quasi totale a colmare il divario digitale e l'IA sarà per il 2030 quello che è stato il PC per il 2000», sottolineando l'importanza di preparare adeguatamente le nuove generazioni a questa transizione tecnologica.

Riassumendo, dall'analisi delle interviste emergono diverse prospettive sull'impatto dell'IA sulle disuguaglianze sociali, con opinioni che oscillano tra il timore di un'amplificazione delle disparità esistenti e la speranza in un potenziale democratizzante della tecnologia. Nonostante queste divergenze, si riscontra un consenso su alcuni punti fondamentali: l'esistenza di bias algoritmici, la necessità di proteggere le categorie vulnerabili e l'importanza cruciale della formazione. Per estrapolare alcune riflessioni preliminari, l'intelligenza artificiale rappresenta una tecnologia dalle potenzialità straordinarie, ma il suo impatto sulla società dipenderà dalle scelte che faremo oggi. Come sottolineato dall'Intervistato 1, l'IA «può rafforzare ingiustizie già esistenti o diventare un catalizzatore di equità, ma solo se mettiamo al centro le persone e non solo l'efficienza, le performance e la riduzione dei costi». Ne deriva che è responsabilità di tutti gli attori coinvolti – dai decisori politici agli sviluppatori, dalle istituzioni educative alle aziende – lavorare insieme

per garantire che questa potente tecnologia sia al servizio di una società più equa e inclusiva.

Prime considerazioni conclusive

Di fronte alla rapida e pervasiva accelerazione digitale che caratterizza il nostro tempo, non è più sufficiente parlare di "alienazione dei processi produttivi" (Marx, 1972), ma è sempre più opportuno richiamare il concetto di "alienazione esistenziale" di Crespi (1994) per indicare una condizione ben più profonda e radicale. La perdita dell'orizzonte di senso, infatti, non riguarda solo l'attività lavorativa, ma la percezione del Sé, della propria identità e della nostra stessa esistenza all'interno di un "mondo iperconnesso, automatizzato e spesso impersonale" (Crespi, Fornari, 1998). In questo scenario, la tecnologia anziché ampliare le possibilità di autorealizzazione, rischia di generare smarrimento, distacco emotivo e una progressiva perdita di riferimenti simbolici e significati condivisi. Riscoprire il significato dell'esistere – come suggerisce Crespi – può diventare la chiave per superare tanto l'alienazione individuale quanto la frammentazione collettiva. In questo scenario, la sociologia è chiamata a proporre nuove chiavi interpretative e pratiche di intervento, proponendo una lettura sociale del cambiamento tecnologico. Una possibile direzione è quella di spostare l'attenzione dalle sole logiche economiche o tecnologiche al piano dell'esperienza vissuta, riscoprendo l'importanza dell'esistenza concreta come spazio di relazione e senso condiviso. Questo significa promuovere progetti di cittadinanza digitale e responsabilità etica nell'uso delle tecnologie, mettendo al centro la riflessione critica sull'IA e sul suo impiego.

A conclusione di questa analisi, appaiono evidenti alcune priorità d'azione che coinvolgono sia gli sviluppatori, chiamati a un approccio più consapevole, sia i decisori politici, responsabili della definizione di un quadro normativo e sociale capace di orientare l'uso dell'IA verso obiettivi di giustizia ed equità. In particolare, emerge in modo sempre più urgente la necessità di formazione e di programmi educativi rivolti a ogni fascia d'età e livello di istruzione. Questo al fine di educare non solo all'utilizzo del digitale, ma ad un uso critico e consapevole, sviluppando competenze di adattamento al cambiamento. Questo vale sia nei contesti scolastici e universitari, sia in tutte quelle aziende che hanno interesse ad investire nella formazione dei dipendenti. Inoltre, oggi vi è un ampio consenso sul fatto che i bias algoritmici non derivino esclusivamente da errori nei processi di addestramento, ma siano soprattutto il riflesso di stereotipi e pregiudizi già presenti nelle decisioni umane. Per questo motivo, è fondamentale affiancare allo sviluppo

tecnologico l'implementazione di sistemi di monitoraggio continuo e l'attivazione di canali di feedback accessibili agli utenti, al fine di rilevare e correggere tempestivamente eventuali distorsioni. Per gli sviluppatori, inoltre, potrebbe essere utile agire in ottica multidisciplinare includendo nuove figure professionali (ad es. sociologi, giuristi, psicologi) per curare la qualità e la rappresentatività dei dati di addestramento. Per quanto riguarda i decisori politici, è richiesto di andare oltre la mera regolamentazione tecnica, sviluppando normative che includano valutazioni d'impatto sociale, promuovendo programmi nazionali di alfabetizzazione digitale e incentivando politiche pubbliche che orientino l'uso del digitale in modo etico e sostenibile, in particolare nei settori dei servizi sociali, sanitari e scolastici. Al tempo stesso, è fondamentale garantire la presenza di canali alternativi ai servizi digitali, in particolare per anziani e soggetti vulnerabili, affinché nessuno sia escluso dall'accesso a diritti e servizi essenziali. In definitiva, ripensare la convivenza sociale nell'era digitale richiede un nuovo equilibrio tra individuo e collettività, tra innovazione e senso umano, tra progresso tecnologico e giustizia sociale.

Riferimenti bibliografici

Broussard M. (2019). *La non intelligenza artificiale. Come i computer non capiscono il mondo*. Milano: FrancoAngeli.

Carlucci Aiello L., Cialdea Mayer M. (1995). *Invito all'intelligenza artificiale*. Milano: FrancoAngeli.

Castells M. (2014). *La nascita della società in rete*. Milano: Università Bocconi Editore.

Commissione Europea (2018). *Comunicazione della Commissione. L'intelligenza artificiale per l'Europa*. Bruxelles.

Crespi F. (1994). *Imparare ad esistere. Nuovi fondamenti della solidarietà sociale*. Roma: Donzelli.

Crespi F., Fornari F. (1998). *Introduzione alla sociologia della conoscenza*. Roma: Donzelli.

DiMaggio P., Hargittai E. (2001). From the "digital divide" to "digital inequality": Studying Internet use as penetration increases. Princeton: Princeton University Press.

Eubanks V. (2018). *Automating inequality*. New York: St. Martin's Press.

Hatzius J. (2023). *The potentially large effects of artificial intelligence on economic growth*. New York: Goldman Sachs.

Kissinger H.A., Schmidt E., Huttenlocher D. (2023). *L'era dell'intelligenza artificiale. Il futuro dell'identità umana*. Milano: Mondadori.

Lazzini F. (2022). *Etica digitale e intelligenza artificiale. I rischi per la protezione dei dati*. Torino: Giappichelli.

Marx K. (1972). *Il Capitale*. Roma: Editori Riuniti.

O'Neil C. (2017). *Armi di distruzione matematica. Come i Big Data aumentano la disuguaglianza e minacciano la democrazia*. Firenze-Milano: Giunti Bompiani.

Russell S., Norvig P. (2005). *Intelligenza artificiale. Un approccio moderno*, voll. 1-2, 2ª ed. Milano: Pearson Education.

Teti A. (2025). *Digital profiling. L'analisi dell'individuo tra metodologie, tecniche e intelligenza artificiale*. Milano: Il Sole 24 Ore.

Zahidi S. (2023). Prefazione. In *The Future of Jobs Report 2023*. Geneva: World Economic Forum, disponibile al link: https://www.weforum.org/publications/the-future-of-jobs-report-2023/

Le nuove forme di potere nella società algoritmica e l'essere umano: quale possibile soluzione

di Giordana Truscelli*

Le nuove tecnologie e l'utilizzo dell'intelligenza artificiale hanno introdotto una nuova rivoluzione tecnologica nella nostra società, ponendo di fatto gli Stati di fronte ad una sfida digitale globale.

L'impatto del digitale sulla società e sulla democrazia è enorme: la complessità delle domande sociali è dovuta sia dalla rapida accelerazione dello sviluppo di queste nuove tecnologie sia dalla conseguente trasformazione della sfera pubblica, non più luogo intermedio tra lo Stato e la società civile. Oggi nelle cd. "*echo chambers*" il dibattito politico, infatti, subisce una polarizzazione priva di contraddittorio. Le considerazioni che in questo scritto si effettueranno possono essere così riassunte: quali sono i rapporti tra la rivoluzione tecnologica e la sfera pubblica? In che modo le nuove tecnologie dell'informazione e della comunicazione influenzano l'opinione pubblica e quindi la democrazia? Quale ruolo svolgono la combinazione di messaggi ed analisi dei dati nella società algoritmica?

Parole chiave: intelligenza artificiale; opinione pubblica; potere algoritmico; società; democrazia; autonomia umana.

The new forms of power in the algorithmic society and the human being: what possible solution?

New technologies and the use of artificial intelligence have introduced a new technological revolution in our society, effectively confronting states with a global digital challenge.

The impact of digital technology on society and democracy is enormous: the complexity of social demands is due both to the rapid acceleration of the development of these new technologies and to the consequent transformation of the public sphere, which is no longer an intermediate place between the state and civil society. Today, in the so called "echo chambers", the political debate undergoes a polarisation without contradiction. The considerations that will be made in this paper can be summarised as follows: what are the relationships between the technological revolution and the public sphere? How do the new information and communication technologies influence public opinion and thus democracy? What role do the combination of messages and data analysis play in the algorithmic society?

Keywords: artificial intelligence; public opinion; algortmic power; society; democracy; human autonomy.

DOI: 10.5281/zenodo.18435523

* Università degli Studi di Teramo. gtruscelli@unite.it.

Giordana Truscelli

1. Oltre la "rivoluzione digitale": mutazione antropologica?

La contemporaneità è caratterizzata da quello che potremmo definire un salto qualitativo nella natura stessa del potere sociale. Non si tratta più semplicemente di una "rivoluzione digitale" – termine ormai logoro che evoca principalmente cambiamenti tecnologici – ma di una vera e propria mutazione antropologica che ridefinisce le modalità stesse dell'essere-nel-mondo dell'essere umano. L'algoritmo non è più strumento neutro nelle mani dell'uomo, ma diventa ambiente cognitivo, medium costitutivo dell'esperienza e, infine, forma emergente di soggettività politica.

Di conseguenza, la sfera pubblica, così come teorizzata da Habermas, che ai tempi dei greci si identificava con l'agorà[1], cioè il luogo deputato alla formazione dell'opinione pubblica il cui obiettivo era la "traduzione del linguaggio degli interessi individuali/familiari nel linguaggio degli interessi pubblici" (Bauman, 2014) è cambiata: essa è infatti stata sostituita da numerose cd. agorà virtuali in cui il dibattito pubblico viene polarizzato e non trova posto il dissenso.

Nell'epoca classica, quindi, l'agorà costituiva un luogo in cui la discussione poteva svolgersi liberamente affinché potessero emergere chiaramente un consenso ed un dissenso, rendendo noto a tutti il dibattito pubblico. I cittadini partecipavano attivamente al dibattito, comunicando con i "parlanti" in una condizione di reciprocità e vagliando le opinioni di tutti i partecipanti, che spesso erano anche discordanti (Sgobba, 2020).

Ma l'opinione non è una scienza né tantomeno un sapere, ma una convinzione che diviene pubblica quando è del "pubblico" ed investe la *res publica*: "l'interesse generale, il bene comune, la collettività" (Orecchia, Preatoni, 2022). Habermas, pone l'accento sulla corretta formazione dell'opinione pubblica, evidenziandola come elemento importante per garantire la democrazia: «quanto più eguale e imparziale, quanto più aperto quel processo, quanto meno i partecipanti subiscono coercizione e sono invece disposti ad essere guidati dalla forza del migliore argomento, con tanta maggiore probabilità gli interessi effettivamente generalizzabili verranno accettati da tutte le persone che ne sono toccate in modo importante» (Habermas, Rawls, 2023). Il dibattito pubblico, forma un'opinione pubblica concreta e consapevole se aderente ad almeno cinque valori: imparzialità, eguaglianza, apertura, assenza di coercizione e unanimità (Habermas, Rawls, 2023).

[1] Cioè quello spazio intermedio che collega due settori della polis: l'*ekklesia* (il consiglio dei cittadini) e l'*oikos* (il nucleo familiare) coordinando interessi privati e pubblici. Definito anche come la "sede della democrazia" (Bauman, 2011).

Oggi, al contrario, il dibattito politico avviene per lo più attraverso delle agorà virtuali che spesso creano delle vere e proprie "camere dell'eco" (Acemoglu, Johnson, 2023), ove gli individui hanno scarse probabilità di ascoltare opinioni alternative alle proprie[2]. Ciò, quindi, rende del tutto impossibile, al di là della volontà del soggetto, avere un confronto reale con l'alterità di un'opinione differente: è la stessa tecnologia che impedisce il confronto, catalogato come "elemento disturbante". Questo fenomeno viene maggiormente acuito dalle cd. "filter bubbles" (Parisier, 2011) ad opera degli algoritmi presenti sui social media, i quali propongono ed amplificano la diffusione di opinioni coerenti con quelle dell'utente (Acemoglu, Johnson, 2023). Gli algoritmi di "personalizzazione" utilizzati anche dai social media, creerebbero una sorta di bolle di informazione fortemente individualizzate in grado di isolare l'individuo da informazioni, opinioni o prospettive differenti dalle proprie. La personalizzazione delle informazioni proposte dagli algoritmi agli utenti, impedendo che di fatto questi ultimi possano accedere ad opinioni differenti dalle proprie rischiano di radicalizzare le opinioni degli utenti e di creare una sorta di micro-pubblici che non comunicano tra loro nonché mondi informativi paralleli ed il più delle volte incompatibili. Di conseguenza, questo isolamento da punti di vista differenti dal proprio e ad esso contrapposto, finisce, di fatto, per creare una realtà limitata e ristretta, dove le proprie idee vengono di continuo rinforzate e mai messe in discussione. Tale dinamica è particolarmente evidente nei social media, dove un'apparente democratizzazione dell'informazione, nasconde in realtà dei meccanismi di estrazione di dati utili per implementare il targeting degli algoritmi.

Le piattaforme digitali, quindi, non possono essere considerate semplicemente degli strumenti di comunicazione, ma diventano dei veri e propri "architetti invisibili" (Gillespie, 2018) della deliberazione pubblica, in quanto il dibattito online opera attraverso meccanismi computazionali degli algoritmi che "decidono" in maniera selettiva cosa viene visualizzato, quando e da chi. Questi algoritmi, quindi, non si limitano a trasmettere informazioni, ma le strutturano in maniera attiva: infatti propongono determinati contenuti agli utenti amplificando o talvolta sopprimendo talune informazioni (Bozdag,

[2] Come C. Sustein (2017) ha evidenziato nella sua opera *#Republic: Divided Democracy in the Age of Social Media*, gli algoritmi che raccomandano la visione di determinati contenuti agli utenti propongono agli stessi, contenuti in linea con le preferenze già espresse, rinforzando le opinioni preesistenti e realizzando così effetti di polarizzazione e radicalizzazione dell'opinione pubblica.

2013), modellando i flussi informativi che poi andranno a creare l'opinione pubblica.

Non è forse questa una forma di potere? E soprattutto si può affermare che dal *logos*, inteso come ragione dialogica che postula un confronto razionale attraverso l'argomentazione, si è passati al logaritmo[3], realizzando così una mutazione della razionalità pubblica?

2. Definizione di potere e la sottomissione all'algoritmo

Cos'è dunque il potere?

Il potere, inteso in senso generico può essere inteso come la «capacità di influire sul comportamento altrui, di influenzarne le opinioni, le decisioni, le azioni, i pensieri»[4].

Nell'opera fenomenologia del potere, Henrich Popitz, analizza il potere considerandolo come un vero e proprio fenomeno sociale, non banalmente riducibile solo alla coercizione o alla violenza. L'autore, infatti distingue quattro forme di potere: potere di offendere (fondato sulla violenza fisica per imporre il proprio volere)[5], il potere strumentale (produce paura o speranza negli altri individui) (Popitz, 2015)[6], il potere di autorità (basato sul riconoscimento di un soggetto come "superiore" al quale viene data spontanea obbedienza) (Popitz, 2015)[7] ed infine il potere tecnico (che realizza attraverso l'utilizzo di strumenti tecnici controllo ed influenza sulla realtà sociale e materiale) (Popitz, 2015)[8].

[3] Occorre evidenziare che questa mutazione ha comportato la sostituzione di vari principi: dall'intersoggettività alla predittività, dalla trasparenza all'opacità, dall'intesa reciproca al controllo mediante influenza comportamentale ed infine dalla processualità del dibattito pubblico all'automazione.

[4] https://www.treccani.it/vocabolario/potere1/ (11/06/2025).

[5] Classificato come potere di azione, di "recare danno agli altri con un'azione diretta", ovvero il potere di fare qualcosa di male agli altri (Popitz, 2015).

[6] Tale potere si sostanzia in una minaccia o in una promessa. Le minacce non devono essere necessariamente esplicite, poiché possono consistere anche in gesti simbolici inequivocabili.

[7] Occorre evidenziare che l'esercizio di questo potere comporta da parte del soggetto passivo un adattamento sia del comportamento sia dell'atteggiamento, che adotta i giudizi, i criteri e le opinioni di chi detiene l'autorità.

[8] In questo caso, la competenza nell'utilizzo degli artefatti tecnici per modificare la realtà si realizza sia attraverso il comando diretto, sia attraverso la propria abilità nel costruire realtà materiali che influenzano il comportamento umano.

Partendo dunque da tali premesse, occorrerebbe chiedersi se di fronte alle nuove tecnologie e alla possibilità che possano influenzare l'opinione pubblica ci si trovi di fronte ad una nuova forma di potere.

Byung-Chul Han, nella sua opera Che cos'è il potere?, può essere di aiuto nella tematica che, seppur brevemente, si tenterà di affrontare.

L'autore, infatti, evidenzia come il potere "superiore" è quello in grado di plasmare il futuro dell'altro, influenzando, rielaborando o preparando il «campo di azione di alter[9] affinché questi volontariamente opti per ciò che è conforme al volere di ego[10]» (Han, 2019: 11). Quanto scritto mette in luce la necessaria collaborazione dell'altro nei confronti di chi esercita il potere: l'altro, infatti, può sempre e comunque ribellarsi e quindi è fondamentale che alter in piena libertà segua le opinioni di ego. La massima espressione di potere, per l'autore, si raggiunge nello stesso momento in cui libertà e sottomissione combaciano.

Il concetto di sottomissione volontaria dell'essere umano al potere delle nuove tecnologie, delle quali sembra non poter fare a meno, viene successivamente spiegata alla luce del desiderio di "ottimizzazione" del soggetto che parimenti subisce una trasformazione da soggetto a progetto (Han, 2024). L'individuo, spinto dalla società a perfezionarsi sempre di più, diventando più competitivo sia dal punto di vista lavorativo sia dal punto di vista fisico, più produttivo e più efficiente, è costantemente alla ricerca dell'approvazione sociale (es. like sui social), perdendo di vista quale sia la vera libertà e impedendogli di fermarsi e ricercare un senso di appagamento autentico. L'uomo, quindi, diventa così un "progetto" perché è in continua evoluzione e modificazione per ottenere la cd. approvazione sociale.

L'essere umano, inoltre, ritenendosi un progetto libero da obblighi esterni e capace di reinventarsi continuamente, non si rende conto, secondo l'opinione del filoso, che in realtà egli stesso si sottomette ad obblighi interiori, forzandosi alla prestazione ottimale. L'uomo, infatti, sarebbe un servo assoluto nella misura in cui «sfrutta sé stesso senza un padrone» (Han, 2024: 8): l'auto-ottimizzazione, diviene così auto-sottomissione al potere degli algoritmi.

Com'è noto, più un potere è grande, assoluto, e più esso agisce in maniera silenziosa, evitando di contrapporsi alla realtà ma utilizzandola assumendo una forma sempre più permissiva e offrendosi come "libertà".

[9] Inteso come l'altra persona, soggetta a chi detiene il potere.
[10] Ego è il soggetto che esercita il potere su Alter.

Gli algoritmi che utilizzano il *filter bubble*, non sono strumenti neutrali, poiché personalizzano le informazioni fornite all'utente, confermando le proprie opinioni (e così rendendolo soddisfatto) ed incatenandolo ad una visione parziale e limitata della realtà. Il fenomeno della personalizzazione algoritmica, quindi, assume le caratteristiche di un vero e proprio potere invisibile: il soggetto sottomesso del tutto incosciente della propria sottomissione crede di essere libero, in quanto il rapporto di dominio resta a lui celato. L'efficacia di questo potere algoritmico consisterebbe così nell'agire attraverso piacere e soddisfazione, rendendo gli uomini dipendenti da esso, poiché va incontro a loro seducendoli.

Come non notare che questo potere "intelligente" sfrutta la libertà dell'essere umano sostituendo alla libera scelta una libera selezione tra le scelte offerte?

Le scelte che vengono proposte all'individuo sono solo in apparenza libere, perché al contrario esse sono fortemente influenzate dal sistema che conferma e modella i suoi comportamenti ed al contempo lo spinge ad "ottimizzarsi" per soddisfare determinate aspettative. L'utente non è passivo, ma diventa egli stesso volontariamente parte attiva di un sistema di potere che rispecchia i propri desideri e le proprie preferenze.

3. L'essere umano e la sua "datificazione"

I big data costituiscono oggi una forma molto efficace di controllo dell'essere umano, riuscendo a scrutare anche la sua psiche. Alcuni studiosi hanno anche prospettato la nascita di un nuovo tipo di filosofia e cioè il "dataismo" in base alla quale, approfittando della capacità delle nuove tecnologie di raccogliere una straordinaria mole di dati, tutto ciò che può essere misurato deve essere misurato. L'individuo diventa anch'egli misurabile ed il sé viene scomposto in dati fino a svuotarne completamente il senso: si ricerca una conoscenza del sé attraverso i numeri, utilizzando gli stessi per realizzare una specie di tecnica di autocontrollo. «I dati raccolti sono poi pubblicati e scambiati: così la registrazione del sé assomiglia sempre di più ad una sorveglianza del singolo su sé stesso. Il soggetto odierno è in imprenditore di sé stesso, che si sfrutta» (Han, 2024: 57). Ma numeri e dati possono realmente fornire una conoscenza del sé?

Siamo forse quindi difronte ad una forma di potere strumentalizzante? Se così fosse, questo potere avrebbe il compito di «strutturare e strumentalizzare

il comportamento (umano) al fine di modificarlo, predirlo, monetizzarlo e controllarlo» (Zuboff, 2023[11]); esso si fonda esclusivamente su azioni misurabili e di conseguenza è del tutto indifferente rispetto alle motivazioni che spingono l'uomo a tenere certi comportamenti: ciò che conta è che questi comportamenti umani siano accessibili alle sue operazioni di modifica, controllo, renderizzazione e monetizzazione. L'essere umano viene riletto come una "cosa", un "altro", tenendo ben separate l'esperienza esteriore e l'azione esterna: ciò che viene considerata è l'azione sociale, il comportamento osservabile e misurabile rimanendo pressoché indifferenti rispetto al significato che tale esperienza riveste per il soggetto. Ciò comporta l'assoluta prevalenza di un'equivalenza tra gli individui senza però una vera eguaglianza, separando la nostra soggettività dalle nostre azioni osservabili, siamo considerati alla stregua di meri organismi e tutto ciò a discapito dell'unicità di ogni essere umano.

Il cd. potere strumentalizzante, inoltre, erode la democrazia dall'interno, non si confronta con la stessa, bensì agisce divorando le potenzialità umane, strumentalizzando e controllando l'esperienza umana in modo sistematico e prevedibile.

L'uomo si assuefà alla prevedibilità, al controllo, ad una sorta di certezza, appagato dalla possibilità di stabilire connessioni sociali, accesso alle informazioni e risparmio di tempo realizzati mediante le nuove tecnologie che lo illudono di avere, al contrario un sostegno. I "prodotti di predizione comportamentale" basati su modelli statistici riducono la ricchezza dell'agire umano a pattern calcolabili, negando quella dimensione di imprevedibilità e creatività autentiche che costituiscono invero l'essenza stessa della libertà umana. Tale processo raggiunge il suo apice nella manipolazione dell'opinione pubblica, in quanto le differenti prospettive e visioni politiche vengono ridotte a variabili algoritmiche da ottimizzare per massimizzare specifici obiettivi comportamentali (Zuboff, 2023).

Ma qual è il destino della capacità di decidere collettivamente in modo pacifico e l'impegno civile?

Di fronte al vuoto lasciato dalla loro mancanza, il potere strumentalizzante cerca di riempire questi spazi, attraverso le macchine come medium delle interazioni sociali o nella veste di chatbot che influiscono nei e sui rapporti sociali, cercando di regalare quella certezza di una "conoscenza e prevedibilità totale". In altre parole, il potere strumentalizzante è visto come «la

[11] E-book cap. 12 posizione 13.885 di 20.317.

soluzione certa alle incertezze della società» (Zuboff, 2023[12]) che conferma le opinioni dell'essere umano e le sue convinzioni, fornendogli un senso di apparente appagamento che, tuttavia, è ben distante dalla soddisfazione che può procuragli un rapporto ed una conoscenza autentica di sé e dell'altro.

L'esperienza umana, che fino a qualche tempo fa era considerata inviolabile, viene mercificata, divenendo l'anima stessa campo di estrazione economica. Il potere algoritmico non è più soltanto destinato alla modifica e previsione dei comportamenti umani, ma rischia di trasformare in merce l'esperienza umana.

Per completezza di esposizione, occorre però sottolineare una visione differente delle *filter bubbles* fornita da Axel Burns, il quale, nella sua opera, "È vero che internet ci chiude in una bolla? Una prospettiva critica su filter bubble ed echo chamber" (Bruns, 2024), sottolineando l'importanza dell'agency individuale, delle pratiche d'uso degli utenti e della natura intrinsecamente porosa dei sistemi online. In particolare, l'autore evidenzia il ruolo attivo degli utenti, i quali non possono essere ritenuti dei meri ricettori passivi delle informazioni filtrate dagli algoritmi, ma in realtà essi ricercano attivamente contenuti e li selezionano. In tal senso, quindi, gli algoritmi non costruirebbero muri né tantomeno potrebbero essere considerati delle camere ermetiche ed impenetrabili, quanto più delle camere di risonanza in cui esistono degli spazi di confronto e di diversificazione informativa. Di conseguenza, per Bruns, gli algoritmi pur influenzando i flussi normativi, non annullano la capacità dell'utente di ricerca e discernimento nella selezione dei contenuti e delle informazioni.

Conclusioni

Lo studio presentato fornisce un quadro complesso: di fatto viviamo in un'epoca in cui il potere ha assunto forme più sottili ma profondamente invasive operando una sorta di "colonizzazione dell'inconscio collettivo" che può creare una mutazione antropologica trasformando l'idea di "libertà" in sottomissione volontaria.

L'essere umano da soggetto diventa quindi oggetto (o progetto di ottimizzazione come scritto da Han) di estrazione di dati e manipolazione delle proprie opinioni. Il processo di datificazione dell'individuo rappresenta un nodo cruciale dell'analisi: non si tratta più di raccogliere e classificare dati, quanto

[12] E-book cap. 13 posizione 15.129 di 20.317.

più di anticipare, predire, modellare e modificare i comportamenti umani attraverso strumenti di *nudging* comportamentale che molto spesso operano a livello inconscio e senza che il soggetto ne sia pienamente consapevole.

I *filter bubbles*, teorizzati da Eli Parisier, non si limiterebbero pertanto alla mera personalizzazione dei contenuti offerti all'utente, bensì costruirebbero delle gabbie cognitive che intrappolano l'individuo in una realtà ristretta, non autentica e limitano la possibilità di formulare un pensiero critico.

Questo potere si manifesta, in maniera evidente, mediante il controllo dell'opinione pubblica attraverso gli algoritmi di personalizzazione, i quali "decidono" i contenuti da mostrare agli utenti, radicalizzando così le loro opinioni e punti di vista. La conseguenza è una frammentazione dell'opinione pubblica, in micro-opinioni tra loro incomunicabili abolendo di fatto il dibattito pubblico inteso in senso habermasiano ed appiattendo il senso critico con l'utilizzo di strumenti che condizionano il comportamento umano agendo a livello inconscio ed inconsapevole.

Questo panorama dimostra come il potere "strumentalizzante" degli algoritmi e delle nuove tecnologie offrano un'idea illusoria di libertà, poiché, al contrario, l'essere umano con le proprie scelte non fa altro che alimentare sia la soggezione al potere algoritmico in maniera volontaria, sia alimentare sistemi predittivi sofisticati che riducono il suo ambito di autonomia reale.

Come reagire di fronte a queste sfide che rischiano di minacciare la democrazia stessa?

Forse occorrerebbe sviluppare delle strategie di resistenza e riappropriazione dell'esperienza umana simultaneamente sul piano individuale ed istituzionale.

Sul piano individuale, si potrebbe incrementare la cd. "alfabetizzazione algoritmica", consentendo all'essere umano di sviluppare una consapevolezza critica che gli permetta di riconoscere i meccanismi di raccolta dati, targeting e soprattutto quando i suoi comportamenti diventano oggetto di manipolazione.

Sul versante istituzionale, inoltre, si potrebbe implementare una regolamentazione di piattaforme ed algoritmi ancorata a principi etici condivisi dalla comunità internazionale capace di mettere al servizio dell'umano le nuove tecnologie.

Resta però una domanda fondamentale: che tipo di essere umani vogliamo essere in questa società algoritmica e quali sono i nostri valori fondamentali ed irrinunciabili: la libertà e l'autenticità rientrano tra di essi?

Una volta chiarita questa domanda si potrebbe poi procedere ad elaborare delle strategie "difensive" nei confronti del potere algoritmico ed eventualmente, sulla scia di quanto affermato da Bruns, spostare l'attenzione dalla

cosa fanno all'essere umano le tecnologie al come esse vengono utilizzate dagli utenti e come sarebbe possibile migliorarne l'uso, ammettendo così l'esistenza della complessità di un ecosistema informativo e promuovendo un dibattito pubblico sano ed aperto.

Riferimenti bibliografici

Acemoglu D., Johnson S. (2023). *Potere e progresso. La nostra lotta millenaria per la tecnologia e la prosperità*. Milano: Il Saggiatore.

Bauman Z. (2014). *Danni collaterali. Diseguaglianze sociali nell'età globale*. Roma-Bari: Laterza.

Bozdag E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15(3): 209-227.

Bruns A. (2024). *È vero che internet ci chiude in una bolla? Una prospettiva critica su filter bubble ed echo chamber*. Bologna: il Mulino.

Gillespie T. (2018). *Custodians of the Internet. Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Habermas J., Rawls J. (2023). *Dialogo sulla democrazia deliberativa*. Milano: Società Aperta.

Han B.-C. (2019). *Che cos'è il potere?* Milano: nottetempo.

Han B.-C. (2024). *Psicopolitica. Il neoliberismo e le nuove tecniche del potere*. Milano: nottetempo.

Orecchia A.M., Preatoni D.G. (2022). *Bufale, fake news, rumors e post-verità. Discipline a confronto*. Milano: Mimesis.

Pariser E. (2011). *The Filter Bubble. What the Internet Is Hiding from You*. New York: Penguin Press.

Popitz H. (2015). *Fenomenologia del potere. Autorità, dominio, violenza, tecnica*. Bologna: il Mulino.

Sgobba A. (2020). *La società della fiducia. Da Platone a WhatsApp*. Milano: Il Saggiatore.

Sunstein C.R. (2017). *#Republic. Divided Democracy in the Age of Social Media*. Princeton: Princeton University Press.

Zuboff S. (2023). *Il capitalismo della sorveglianza*. Milano: Luiss University Press.

Relationship between AI and political elections: a bibliometric analysis

by *Alessandra De Luca, Antonello Canzano Giansante*[*]

This study explores the application of artificial intelligence in elections, focusing on its impact on voter behavior, campaign strategies, and electoral outcomes in accordance with Science and Technology Studies. A bibliometric analysis highlights how AI shapes political processes, emphasizing emerging themes and trends in international literature. Special attention is given to the summary of key findings and the forecast for future evolutions.

Keywords: artificial intelligence; elections; politics; bibliometric analysis; electoral campaigns; Science and Technology Studies.

Relazione fra IA ed elezioni politiche: un'analisi bibliometrica

Lo studio analizza l'applicazione dell'intelligenza artificiale nelle elezioni, concentrandosi sull'impatto sui comportamenti degli elettori, sulle strategie di campagna e sui risultati elettorali in linea con gli Science and Technology Studies. Un'analisi bibliometrica evidenzia come l'IA influenzi i processi politici, rilevando temi e trend emergenti nella letteratura internazionale, con un focus sui risultati chiave e sulle prospettive future.

Parole chiave: intelligenza artificiale; elezioni; politica; analisi bibliometrica; campagne elettorali; Science and Technology Studies.

Introduction

Major elections worldwide in recent years have highlighted the growing influence of artificial intelligence on democratic processes. In 2024 alone, over 60 countries went to the polls, with new AI tools significantly impacting campaign messaging and voter outreach. While the use of digital technology in elections is not new, the rapid development of AI has introduced unprecedented opportunities and challenges. Generative AI platforms can produce convincingly realistic text, images, and videos, raising concerns about deepfakes and algorithmically tailored propaganda that may mislead voters. The World Economic Forum's Global Risks Report (2024) warns that the

DOI: 10.5281/zenodo.18435538

[*] Università degli Studi "Gabriele d'Annunzio" di Chieti-Pescara. alessandra.deluca@unich.it, antonello.canzano@unich.it.

proliferation of deepfakes and other AI-generated content could erode citizens' ability to discern truth from falsehood and undermine the integrity of elections. The intersection of AI and political elections has thus become a focal point for research and public policy.

Early uses of AI in politics were limited to experimental forecasting and basic e-voting systems. By the 2010s, however, social media and big data analytics reshaped the field. The 2011 Arab Spring showed how online networks could drive protest, amplified by AI-driven trend analysis. In 2012, the U.S. presidential election marked a shift toward data-driven campaigning, anticipating AI-based voter microtargeting. In 2018 the Cambridge Analytica scandal revealed how AI-based profiling and microtargeted advertising influenced voter behavior, often without users' consent. This scandal prompted a global reckoning over data privacy and political manipulation; the European Parliamentary Research Service noted that protecting personal data and ensuring electoral fairness in the age of AI became crucial following the Cambridge Analytica case (Monteleone, 2019). Scholars in political communication observed that we have entered the "fourth wave of digital democracy", characterized by the pervasive use of AI and big data in politics, the emergence of digital platforms as influential political actors, and the normalization of falsehood as a campaign strategy.

Studying AI's role in elections is essential, as elections are the foundation of representative democracy. Any technology that shapes their administration, information flow, or voter behavior has broad societal implications. This topic regards multiple disciplines: computer scientists and data analysts develop AI models to predict outcomes, detect bots, and counter online disinformation.

This study is grounded in political sociology and Science and Technology Studies (STS), focusing on who controls technology, who benefits or is harmed, and how its use is negotiated within political institutions and norms. This perspective is crucial for understanding AI in elections as a socio-technical phenomenon involving both technical systems and human actors.

Given the rapid evolution of this research field, a bibliometric analysis provides a valuable method to systematically map the knowledge regarding AI and electoral studies.

The research objectives are descriptive and analytical. First, we aim to chart scholarly interest in AI's relationship with electoral processes: which are the most prolific authors, sources, and institutions? How has publication volume grown over time, and are there activity bursts corresponding to real-world events? Second, we analyze the literature's content to identify major research themes and trends. By quantifying bibliographic patterns and highlighting these themes, our analysis provides a structured overview of this

rapidly expanding field, which is academically valuable for identifying knowledge gaps and future research directions and is practically important for policymakers.

Section 2 provides an overview of the data and methodology. Section 3 reports the results of the analysis, such as publication trends, geographic and disciplinary distribution of research, collaboration networks, and the thematic structure of the literature. Finally, in Section 4, we summarize the main findings and reflect on their implications for scholarship and practice. We also indicate future research directions and provide considerations on ensuring that the synergy between AI and elections strengthens democratic participation.

1. Material and methods

Bibliometric analysis is a quantitative study of bibliographic material using quantitative and statistical methods to investigate knowledge structure and forecast future developments in a specific field (Maretti, Tontodimamma, Biermann, 2019).

For our analysis, we used the R (R Core Team, 2021) package Bibliometrix (Aria, Cuccurullo, 2017), designed for science mapping analysis, specifically through Biblioshiny, its associated web app.

Data were retrieved from Scopus on January 21, 2025, using the search query: (“AI” OR “artificial intelligence”) AND (“election*” OR “electoral campaign*”).

The asterisks ensured the query included both singular and plural forms. The Boolean operator “OR” retrieved documents containing at least one keyword from each group, while “AND” linked the two groups, focusing on documents that examine the relationship between artificial intelligence and elections. The research followed Scopus’s criteria of “Article Title, Abstract, Keywords” and was restricted to the fields of Computer Science, Social Sciences, and Decision Sciences. We limited results to English documents published from 1956 to 2024, marking the start of our timeline with the Dartmouth conference that established AI as a field.

A total of 886 documents were exported in CSV format. After removing 174 unrelated items and 21 duplicates, the final dataset included 691 documents, which were imported into Biblioshiny.

2. Results

The final documents are from 284 different sources and range from 1980 to 2024. All 691 documents of the dataset were successfully accepted and processed by Biblioshiny. The analysis revealed a total of 1,533 different authors, with only 93 being authors of single-authored documents. This highlights the significant importance of co-authorship in this field, likely due to its interdisciplinary nature.

Fig. 1 - Annual scientific production

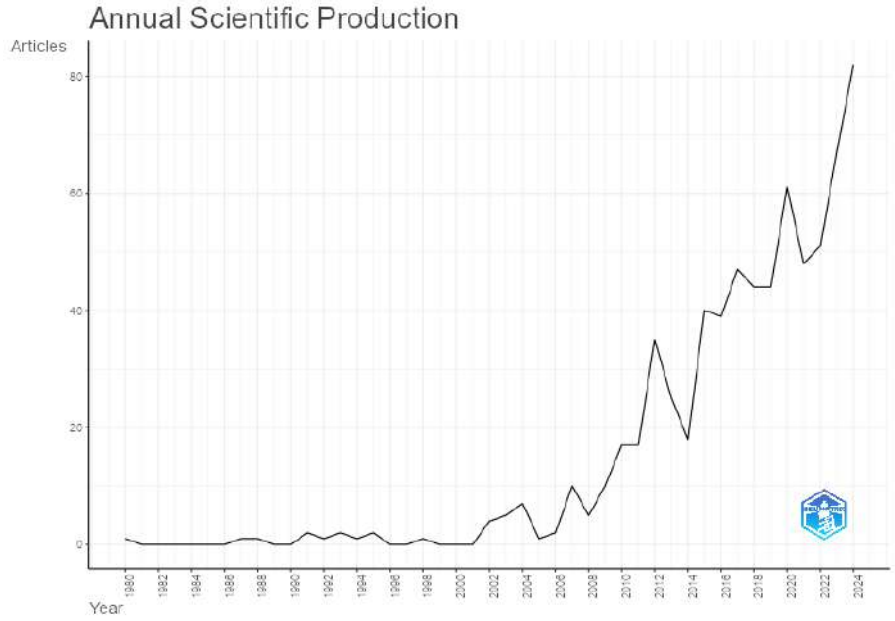


Figure 1 illustrates the annual scientific production of the documents under analysis. There was a notable increase in publications in 2012 (35 documents), likely due to the broad usage of social media during the Arab Spring in 2011 (Ranganath *et al.*, 2016) and political debates from the 2012 U.S. Presidential election. The increase in 2015 (40 documents) may be linked to the growing popularity of social networks beyond Facebook and Twitter. The rise in platforms and their use likely impacted political debates and campaigns online, fueling research on the link between technology and political processes. Moreover, the 2016 U.S. presidential election and Brexit referendum raised concerns about AI-driven misinformation, especially after the Cambridge

Analytica scandal revealed AI's role in influencing elections. García-Orosa (2021) describes this period, initiated by these events, as the beginning of the fourth wave of e-democracy.

COVID-19 further fueled technological development, evident in the peak in 2020 (61 documents). The increase in publication frequency since 2020 reflects rising academic concern about the normative implications of AI in political contexts, including voter manipulation, algorithmic bias, and democratic accountability. Since 2022, there has been a continuous rise (51 documents in 2022, 67 in 2023, and 82 in 2024) in such publications. Technological innovations have made generative AI models accessible to a broader public, increasing the interference of bots and misinformation through AI-generated fake news, images, and videos in online political debate. This has raised concerns about the outcomes of the so-called Super-Election Year in 2024 (Schmitt *et al.*, 2024), which includes the U.S. and Russian presidential elections, as well as the European Parliament elections and general elections in India and the UK, among others.

Fig. 2. - The 10 most relevant sources in our dataset

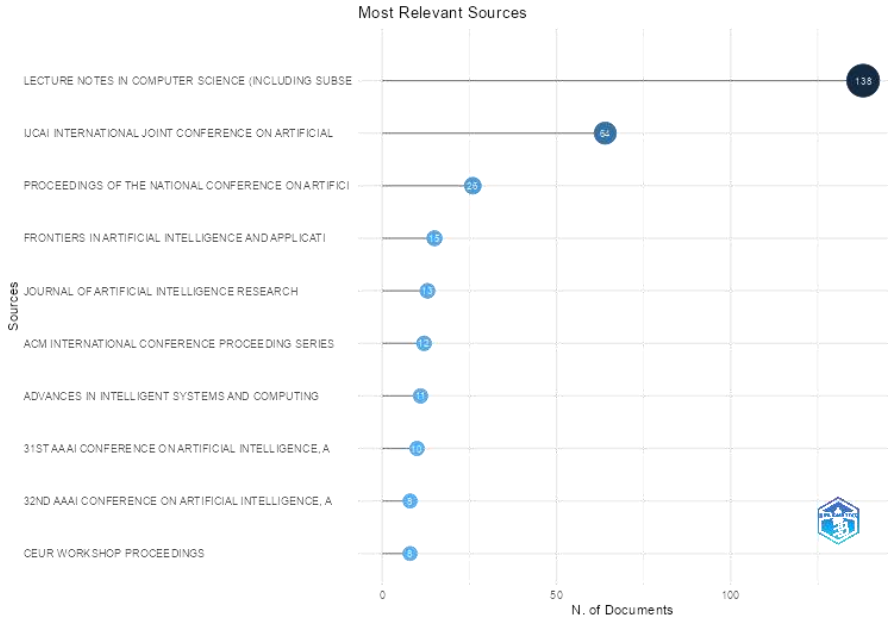


Figure 2 shows the 10 most relevant sources of our documents. The concentration in workshop and conference proceedings, such as the IJCAI, is notable. This reflects the fast-paced development in AI, where research quickly

becomes outdated, making conferences and workshops ideal for discussing advancements and receiving immediate feedback.

Fig. 3. - Country according to corresponding author

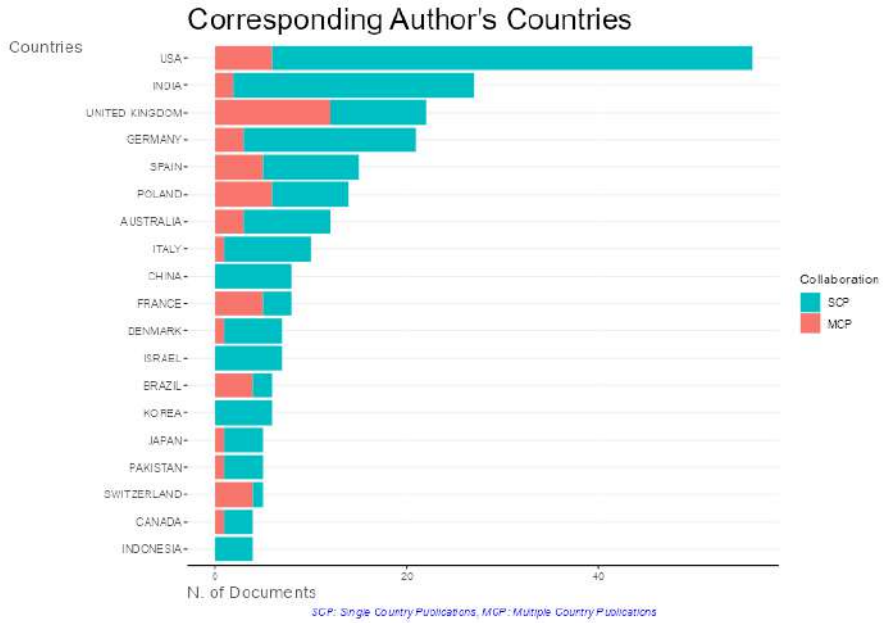
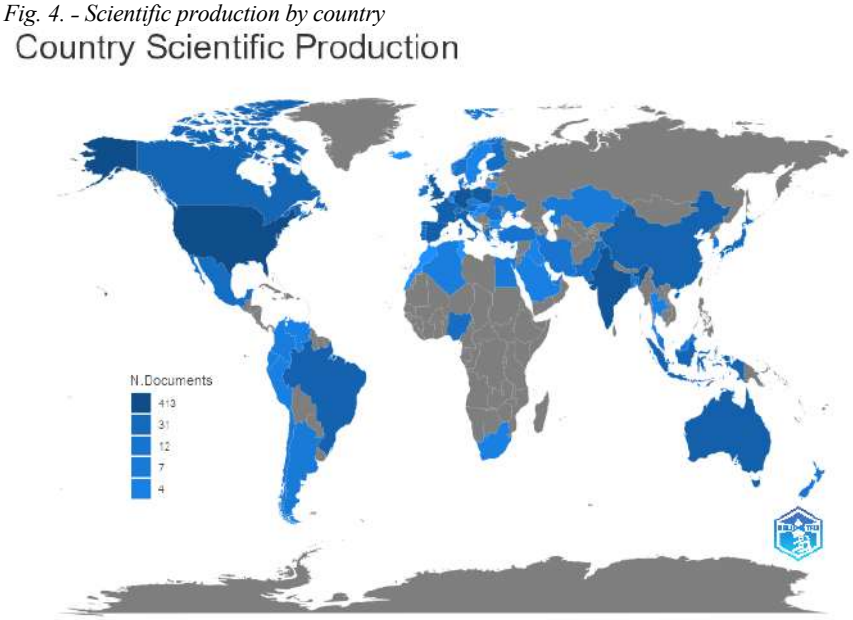


Figure 3 shows that most corresponding authors are affiliated with institutions in the U.S. This data reflects authors' institutional affiliations, not their citizenship. Single country publications (SCP), where all authors are from the same country, are far more common than multiple country publications (MCP). In the U.S., there are 50 SCPs compared to only 6 MCPs. This trend is consistent across all countries included, except for the UK, which has a slightly higher number of MCPs (12) than SCPs (10). Italy ranks 8th, with 9 SCPs and 1 MCP. Among countries with the highest presence of corresponding authors, only China, Israel, Korea, and Indonesia have no MCPs. The lower ratio of MCPs to SCPs may be influenced by regulatory differences across countries, particularly regarding AI, data privacy, and e-voting laws, which can impact international collaboration. For example, in the EU, AI is regulated by the AI Act (Cupać, Sienknecht, 2024), while China has strict requirements for government approval of generative AI models (Soo, 2025). In Indonesia, the lack of MCPs may relate to efforts in implementing the National Strategy for Artificial Intelligence announced in 2020 (Wadipalapa *et al.*, 2024).

Fig. 4. - Scientific production by country



According to Figure 4, darker colors indicate higher scientific production. The U.S. leads with 413 documents, followed by India (175), Germany (169), and the UK (102). The U.S. dominance is due to its widely used generative AI models and substantial funding for AI research. Recent U.S. presidential elections should also be considered, as they have fueled online political debates and concerns over AI-driven interference, which may have encouraged research into the connections between AI and elections. India is investing in AI to enhance its economy (Gupta, Bharadwaj, 2024). Additionally, Germany and the UK have launched ambitious AI Action Plans to promote innovation and technological leadership (European Commission, n.d.; UK Government, 2024).

Poland ranks 5th with 98 documents, likely due to its digitalization efforts supported by EU funds, specifically the European Union Digital Development Funds Program (2021-2027), amounting to 2.5 billion euros (Polish Ministry of Funds and Regional Policy, n.d.). Italy ranks 10th with 62 documents, tied with Israel. Despite substantial investments in AI development (Soo, 2025), China ranks 12th with 51 documents.

Fig. 5. - Word cloud of the top 50 Author's Keywords

We cleaned our dataset by removing research keywords, synonyms, and frequently repeated unrelated terms like "papers" and "proceedings." We then created a word cloud (Fig. 5) to visualize the top 50 most frequently occurring terms using the "Author's Keywords" option in Biblioshiny. Tab. 1 below includes the top 15 terms:

Tab. 1 - Occurrences of the top 15 Author's Keywords

Words	*Occurrences*
Machine learning	42
Social media	33
Sentiment analysis	32
Twitter	30
Fake news	18
Disinformation	16
Voting	14
Deep learning	13
Computational complexity	11
Democracy	11
Natural language processing	11
Computational social choice	10

Misinformation	9
Deepfake	8
Opinion mining	8

The analysis reveals a strong focus on "social media" and "disinformation," reflecting the significant impact of digital platforms on political discourse and the spread of misinformation. This trend is supported by the frequent occurrences of "fake news" and "deepfake," which refers to AI-generated content designed to appear authentic and often used to manipulate public opinion during political campaigns (Loewenstein, 2024). Additionally, the terms "voting" and "democracy" indicate increasing concerns about the influence of AI-generated content on democratic processes.

In the word cloud, the occurrences of "election prediction" (6), and "bots," "e-voting," and "political communication" (each with 5 occurrences) suggest a growing academic interest in how AI can forecast election outcomes and influence political communication. Furthermore, the recurrence of terms like "sentiment analysis," "deepfake," and "disinformation" suggests a shift in research focus from examining AI as a technical tool to exploring its impact on public opinion formation and the integrity of democratic processes.

Other significant terms are "voting advice applications" (5 occurrences), "political participation," "privacy" (both with 4 occurrences), and "AI ethics" (3 occurrences). The emphasis on privacy and security highlights the ethical implications of AI in elections, particularly concerning data protection and digital surveillance. Furthermore, the focus on e-voting and voting advice applications highlights efforts to boost political engagement through electronic voting tools and digital resources that help voters select candidates based on personalized responses.

Conclusion

This bibliometric analysis offers an overview of the literature on artificial intelligence and political elections. Over the past two decades, and especially since the 2010s, research on this topic has grown rapidly, with spikes around 2012 and 2015, likely linked to events such as the Arab Spring, U.S. elections, and the rise of new platforms beyond Facebook and Twitter. By 2024, publication output peaked, reflecting growing concern over AI's role in democracy. This growth pattern shows that the relationship between AI and electoral processes has become a mainstream concern.

Our analysis also highlights the interdisciplinary nature of this field. The 691 documents in our dataset were authored by 1,533 individuals, with few

single-authored works, indicating that studies often involve cross-disciplinary teams. We found a mix of computer science and social science sources, which suggests both a strong technical component in this research and a focus on normative and societal analysis. This dual character confirms that AI-and-elections research is inherently interdisciplinary, bridging algorithmic developments and their socio-political effects.

Third, the bibliometric results reveal several thematic areas of research. Our keyword co-occurrence analysis emphasizes social media and online information. Terms such as "social media," "Twitter," "sentiment analysis," "fake news," "disinformation," "misinformation," and "deepfake" are among the most frequently occurring keywords, indicating a strong interest in how AI technologies contribute to the spread and detection of false or manipulated information during elections. Terms like "fake news" and "deepfake" reflect concern about AI-driven disinformation campaigns that amplify partisan propaganda and conspiracy theories. This aligns with warnings from institutions like UNESCO and UNDP (Patel, 2025), which caution that without proper safeguards, AI could distort public discourse during elections.

Microtargeting and voter persuasion are further key themes, as demonstrated by keywords such as "machine learning," "profiling," "targeting," and "opinion mining." These studies often address the efficacy and ethics of AI-driven campaign strategies, questioning whether personalized messaging enhances voter engagement or crosses into manipulation and privacy violation.

There is also research on predictive analytics in elections. Keywords like "election prediction," "voting," "deep learning," and "computational social choice" suggest that AI models for forecasting election outcomes or optimizing electoral systems are being investigated.

Another significant theme is the exploration of AI in electoral administration and participation, evidenced by terms like "e-voting," "voting advice applications," and "political participation." These studies examine how AI can enhance voting systems by improving the security and accessibility of electronic voting or assist voters in making informed choices.

Our findings indicate an emerging focus on the ethical, legal, and sociotechnical implications of AI in elections. Keywords such as "democracy," "privacy," "AI ethics," and "regulation" signal that, although literature is primarily dominated by studies on AI's role in online political communication and information warfare, it is increasingly addressing governance, policy, and the design of AI systems aligning with democratic values.

Academic conversation has evolved in response to real-world events. Early research investigated how AI could enhance campaigns or predict elections. However, as high-profile incidents emerged, there has been more emphasis on the threats that AI poses to electoral integrity and public trust. García-Orosa

(2021) characterizes the current era as the advent of a "fourth wave of digital democracy," where digital platforms and AI-driven misinformation play a central role in politics. Cupać and Sienknecht (2024) argue that democracies are "under attack" from AI-powered techniques, such as voter profiling, automated propaganda, and troll farms, which need regulatory interventions.

The literature also notes that AI can offer solutions, from faster detection of harmful content to personalized civic education tools. This dual role has sparked debate, with many scholars calling for ways to enhance AI's benefits while mitigating risks to electoral fairness and transparency. Additionally, geographical imbalances in scholarship reflect global power disparities in AI development. Our analysis revealed that authors based in the United States and a few technologically advanced democracies produce a significant share of the research, potentially influencing which problems receive attention. There is comparatively less research from the Global South, raising concerns about underrepresented regional challenges or perspectives, highlighting the need for a more inclusive scholarship.

Looking ahead, this study highlights the need to explore AI's long-term effects on democratic culture and voter attitudes. While short-term impacts of misinformation are known, we still lack longitudinal research on whether repeated exposure erodes trust or increases polarization.

Another critical direction is to further study and evaluate regulatory and governance frameworks. In the EU, Cupać and Sienknecht (2024) identify four main instruments of AI governance: bans on certain uses, transparency requirements, risk management protocols, and digital education initiatives. Comparative research is needed to identify effective regulatory approaches and uncover existing gaps, for which interdisciplinary collaboration is essential.

Future research should also focus on the positive uses of AI in strengthening democracy, such as using machine learning to secure voting systems against fraud or cyberattacks and enhancing voter education and engagement through AI-driven chatbots. Compared to the literature on AI's threats, research on these applications is scarce. Studying pilot projects where AI has been effectively used to boost voter turnout could provide valuable insights.

Scholars should apply STS approaches to examine how election-related AI tools are developed and governed – who builds them, whose values shape them, and how their use is contested across political contexts. Qualitative methods, such as ethnographies or interviews, can shed light on these dynamics, as technologies are not neutral: they reflect human choices and power structures and must be aligned with democratic norms.

The literature on AI and political elections highlights both new opportunities and risks. By synthesizing current research, this study provides a

framework for understanding how AI is transforming electoral processes and highlights the importance of interdisciplinary collaboration to ensure that innovation supports, rather than undermines, democratic integrity.

References

Aria M., Cuccurullo C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4): 959-975. DOI: 10.1016/j.joi.2017.08.007

Cupać J., Sienknecht M. (2024). Regulate against the machine: How the EU mitigates AI harm to democracy. *Democratization*, 31(5): 1067-1090. DOI: 10.1080/13510347.2024.2353706

European Commission (s.d.). Germany AI strategy report. *AI Watch*. Disponibile online (consultato il 29 gennaio 2025).

García-Orosa B. (2021). Disinformation, social media, bots, and astroturfing: The fourth wave of digital democracy. *El Profesional de la Información*, 30(6): e300603. DOI: 10.3145/epi.2021.nov.03

Gupta A., Bharadwaj T. (2024). Spazio e intelligenza artificiale per il futuro di Bharat. *Limes*, 12 settembre: 1-8.

Loewenstein S. (2024). Make America fake again? Banning deepfakes of federal candidates in political advertisements under the First Amendment. *Fordham Law Review*, 93(1): 273-320.

Maretti M., Tontodimamma A., Biermann P. (2019). Environmental and climate migrations: An overview of scientific literature using a bibliometric analysis. *International Review of Sociology*, 29(3): 437-455. DOI: 10.1080/03906701.2019.1641270

Monteleone S. (2019). Artificial intelligence, data protection and elections. *European Parliamentary Research Service*, Briefing PE 637.952: 1-12.

Patel A. (2025). Freedom of expression, artificial intelligence and elections. *UNESCO–UNDP Issue Brief*, 19 maggio: 1-18.

Polish Ministry of Funds and Regional Policy (s.d.). European Funds for Digital Development 2021-2027. Documento istituzionale, Varsavia: 1-45 (consultato il 28 gennaio 2025).

R Core Team (2021). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing (consultato il 27 gennaio 2025).

Ranganath S., Morstatter F., Hu X., Tang J., Wang S., Liu H. (2016). Predicting online protest participation of social media users. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1): 208-214. DOI: 10.1609/aaai.v30i1.9988

Schmitt V., Tesch J., Lopez E., Polzehl T., Burchardt A., Neumann K., Mohtaj S., Möller S. (2024). Implications of regulations on large generative AI models in the super-election year and the impact on disinformation. In *Proceedings of the Workshop on Legal and Ethical Issues of Artificial Intelligence*: 1-15.

Schmitt V., Tesch J., Lopez E., Polzehl T., Burchardt A., Neumann K., Mohtaj S., Möller S. (2024). Implications of regulations on large generative AI models in the super-election year and the impact on disinformation. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*: 28-38. Disponibile online (consultato il 10 ottobre 2025).

Soo Z. (2025). DeepSeek has rattled the AI industry. Here's a quick look at other Chinese AI models. *AP News*, 28 gennaio. Disponibile online (consultato il 3 febbraio 2025).

UK Government (2024). Prime Minister sets out blueprint to turbocharge AI. *GOV.UK*, 22 gennaio. Disponibile online (consultato il 3 febbraio 2025).

Wadipalapa R.P., Katharina R., Nainggolan P.P., Aminah S., Apriani T., Ma'rifah D., Anisah A.L. (2024). An ambitious artificial intelligence policy in a decentralised governance system: Evidence from Indonesia. *Journal of Current Southeast Asian Affairs*, 43(1): 65-93. DOI: 10.1177/18681034231226393.

World Economic Forum (2024). *The Global Risks Report 2024*. Insight Report, 18ª edizione. Ginevra: World Economic Forum. Disponibile online (consultato il 25 maggio 2025).

La trappola dell'intelligenza artificiale tra mimesi imitativa e ideologia

di *Luca Corchia*[*]

Il contributo affronta la questione dell'intelligenza artificiale a partire da due nodi centrali: la mimesi imitativa e l'ideologia tecnico-scientifica. L'analisi mostra come i sistemi di AI riproducano in forma riduzionista processi cognitivi e comportamentali umani, senza però superare la soglia dell'autocoscienza e dell'intenzionalità. Tale imitazione assume un carattere ideologico funzionale agli interessi economici e politici delle corporation globali, contribuendo a consolidare nuove forme di dominio simbolico e di controllo sociale. In questo quadro, il rischio non è soltanto l'alienazione tecnologica, ma anche la progressiva erosione della libertà di scelta individuale, sostituita da processi predittivi e automatizzati. L'articolo propone una lettura critica che intreccia riflessione teorica, genealogia storica e analisi sociologica, evidenziando la necessità di preservare uno "spazio umano" irriducibile alla logica algoritmica

Parole chiave: teoria sociale; intelligenza artificiale; mimesi; libertà di scelta; tecnologia; ideologia.

The trap of artificial intelligence between imitative mimesis and ideology

This article addresses the issue of artificial intelligence by focusing on two core dimensions: mimetic imitation and techno-scientific ideology. It argues that AI systems replicate human cognitive and behavioral processes in a reductionist way, without surpassing the threshold of self-consciousness and intentionality. Such imitation functions as an ideology serving the economic and political interests of global corporations, reinforcing new forms of symbolic domination and social control. Within this framework, the main risk lies not only in technological alienation but also in the gradual erosion of individual freedom choice, increasingly being replaced by predictive and automated processes. The paper offers a critical perspective that combines theoretical reflection, historical genealogy, and sociological analysis, underlining the need to safeguard a human dimension irreducible to algorithmic logic.

Keywords: social theory; artificial intelligence; mimesis; free will; technology; ideology.

DOI: 10.5281/zenodo.18435596

[*] Università degli Studi "G. D'Annunzio" di Chieti-Pescara. luca.corchia@unich.it.

Luca Corchia

Introduzione

Nell'ultimo decennio l'intelligenza artificiale (AI) è emersa come uno dei fattori di cambiamento più incisivi: non è solo un fattore tecnologico capace di ridefinire il modo di produzione delle attuali formazioni storico-sociali bensì un agente che potrebbe trasformare le nostre forme di vita nell'insieme. La letteratura scientifica ha posto l'AI al centro degli interessi di ricerca al pari della pubblicistica dei media e dell'agenda di imprese e politica. Di fronte a questa emergenza due sono gli atteggiamenti di fondo che caratterizzano le reazioni meno riflessive: quelli apocalittici e quelli integrati. È sorprendente l'attualità provocatoria di una coppia dicotomica che Umberto Eco aveva pensato nel 1964 per teorie sulle comunicazioni e la cultura di massa restituendo alla storia e alla ricerca gli artefatti della "cultura bassa". Pur essendo «profondamente ingiusto sussumere degli atteggiamenti umani – in tutta la loro varietà, in tutte le loro sfumature – sotto due concetti generici e polemici come quelli di "apocalittico" e "integrato"» – scriveva Eco – «D'altra parte a coloro che definiamo come apocalittici o integrati, rimproveriamo proprio di avere diffuso dei concetti altrettanto generici – dei "concetti feticcio" – e di averli usati come teste di turco per polemiche improduttive o per operazioni mercantili di cui noi stessi quotidianamente ci nutriamo» (1964: 5). È il maggiore pericolo e il peggiore difetto di tanta pubblicistica sull'AI. Gli integrati si affidano per la soluzione di ogni problema, muovendo trepidanti ogni passo per accorciare la distanza che separa da un futuro migliore. L'attesa è di applicare ai nuovi sistemi intelligenti l'automatizzazione delle forze produttive e dei processi organizzativi, della trasmissione culturale dei saperi e della realizzazione delle mansioni, dell'assistenza nella cura e benessere.

Gli apocalittici, non senza ragioni, prefigurano i rischi di una strutturale crisi occupazionale, il consumo di attenzione, il controllo capillare delle nostre vite, la sottrazione della decisionalità, la disposizione all'imitazione in chi è esposto a una mole impressionante di informazioni e stimoli preselezionati, l'alienazione conseguenze in termini strettamente marxiani dal prodotto, processo, propria umanità e rapporto con l'altro in un mondo regolato da macchine del cui funzionamento siamo totalmente inconsapevoli. Per gli uni e per gli altri l'assunto che l'AI sia un "agente" è dirimente e dato per scontato.

In forma provvisoria vorrei proporre due riflessioni tra speculazione teoretica e critica dell'ideologia che ci riportano tuttavia al nodo cruciale del rapporto di mimesi tra uomo-macchina e alle funzioni ancillari della scienza. In un lungimirante saggio a commento dell'Incontro su "Intelligenze artificiali e scienze sociali", tenuto a Genova nel maggio 1983, Achille Ardigò le aveva già enucleate avvertendo tutti i rischi di un "funzionalismo

apologetico".

Per un verso, c'è la questione "in sé", ossia se c'è un "oltre" la mimesi appropriativa con cui i sistemi di intelligenza artificiale riproducono attraverso forme di apprendimento imitativo i processi mentali degli esseri umani. Quale idea di umanità è all'origine del cd. Programma *Human-Level AI*? Per altro verso, c'è la questione "per sé", cioè ciò che viene creduto o fatto credere e da chi riguardo alla AI e le conseguenze di tali credenze collettive. Quale idea della conoscenza hanno e quali interessi servono gli scienziati? Questi due nodi problematici estremamente rilevanti e attuali sono collegati. Il collante è una concezione riduzionista della scienza e delle relative prospettive sugli oggetti di ricerca – qui ci interessano le scienze sociali – che risulta funzionale a – se non condizionata da – interessi di controllo e regolazione dei nuovi attori sociali attualmente dominanti su scala planetaria.

1. C'è un "oltre" la mimesi appropriativa nell'AI?

Sin dagli anni Ottanta, i sistemi esperti imitano il ragionamento umano e sono in grado di risolvere problemi in un determinato campo, mediante prestazioni simili a quelle di un umano esperto nello stesso ambito. L'eccezionale sviluppo di hardware e software, la digitalizzazione e l'interconnessione in rete hanno mutato l'ecosistema rendendo disponibili «programmi informatici avanzati che analizzano enormi quantità di dati per completare attività come riconoscere contenuti di immagini, testi, predire un *trend*, migliorandone le prestazioni e ottimizzando obiettivi specifici» (Chierici, 2024: 36). Lo sviluppo tecnologica a reso reale lo spettro di una sostituzione del sistema di rilevanze del soggetto, con la propria costellazione di disposizioni e atteggiamenti socio-culturalmente formati, da parte dei sistemi di AI. Era il preoccupato interrogativo che aveva già sollevato Ardigò: «come non ritenere che proprio il senso delle ricerche di IA sia dato da una nuova potente forma di desiderio mimetico: il trasferire la capacità imitativa dell'uomo alle macchine prodotte dall'uomo?» (1983: 236). Con gli sviluppi dell'intelligenza artificiale, il mimetismo passa dall'opera dell'uomo all'uomo stesso (Ivi: 243).

Il dibattito sull'intelligenza umana e quella artificiale è ricco di storie. Ne tralasciamo qui una genealogica che affonda le radici almeno nell'*ars magna* di Raimondo Lullo, la *characteristica universalis* di Gottfried Wilhelm Leibniz, l'*Analytical engine* di Charles Babbage sino ad Alan Turing che propose un radicale cambiamento nel modo di valutare l'intelligenza delle macchine: piuttosto che concentrarsi sulla loro capacità di "pensare" nel senso umano "tradizionale", egli suggerì di verificare se le macchine potessero "imitare" in maniera convincente l'intelligenza umana (1950). E nell'estate del 1956,

John McCarthy organizzò un seminario presso il Dartmouth College, allo scopo di analizzare la «congettura per cui, in linea di principio, ogni aspetto dell'apprendimento o una qualsiasi altra caratteristica dell'intelligenza possano essere descritte così precisamente da poter costruire una macchina che le simuli» (McCarthy *et al*., 1955/2006: 12). Da allora, gli sviluppi dell'*Human-Level AI* sono legata ai progressi della potenza di calcolo, alla relazione con l'informatica e alla convergenza tra neuroscienze, filosofia analitica, psicologia cognitiva, linguistica computazionale, teoria dell'informazione, matematica e robotica. Il problema è se l'AI sia capace di compiere in maniera autonoma, benché a partire da algoritmi di *machine learning*, azioni tipiche dell'uomo: "*thinking humanly, acting humanly, thinking rationally, acting rationally*". Dovrebbero essere: 1) sistemi che "ragionano" come esseri umani, che è l'approccio associato alle scienze cognitive dove l'obbiettivo è che il sistema di *artificial intelligence* risolva qualsiasi tipo di problema (allo stesso modo dell'umano); 2) sistemi che agiscono come esseri umani, emulando dunque tutti i comportamenti della sfera umana; sistemi che pensano razionalmente, ovvero che l'intelligenza artificiale abbia una propria coscienza e una propria razionalità indistinguibile da quella umana; 4) sistemi che agiscono razionalmente, cioè il processo che porta il sistema di *artificial intelligence* alla capacità di *problem solving* con i dati a disposizione (Russell, Norvig, 2010: 2; cfr. Charniak, McDermott, 1985; Rich, Knight, 1991).

Molti scienziali e guru informatici sostengono addirittura che, prima o poi, l'AI supererà l'intelligenza umana, prendendo il sopravvento (Kurzweil, 1999). Non solo nell'immaginario tecnologico, alle macchine artificiali è riconosciuta l'abilità di "simulare" risposte umane e si prefigura con attesa il giorno in cui avranno una loro coscienza e una propria libera cognizione. Una vasta letteratura è riporta da Edmondo Grassi in *Sociologia algomorfica. Il ruolo degli algoritmi nei mutamenti sociali* (2024), ponendo al centro della sua riflessione le nuove categorie attoriali dell'AI. Si arriva a preconizzare un'intelligenza in grado di comprendere il mondo circostante, provvista di capacità cognitive ed emozioni, di apprendere, riflettere e prendere decisioni in totale autonomia, prefiggendosi obiettivi propri, e predisponendo coscientemente una strategia per raggiungerli, allo stesso modo dell'essere umano.

A bene vedere però gli algoritmi di *machine learning* sono privi di capacità di autodeterminazione e di comprensione delle informazioni processate e, dunque, non in grado di raggiungere le reali capacità intellettuali tipiche dell'uomo (Vladeck, 2014). Nel suo famoso esperimento mentale noto come la "stanza cinese" ed illustrato per criticare i sostenitori dell'intelligenza artificiale "forte", John Searle (1980/1984) ha contestato che la mente possa essere un programma, perché il computer non capisce ciò che sta facendo quando elabora un'operazione mentale attraverso il linguaggio simbolico,

essendovi una distanza incolmabile tra sintassi e semantica, ossia tra struttura grammaticale del linguaggio e comprensione significante delle operazioni combinatorie (1980). La distinzione tra decifrazione simbolica e conoscenza semantico-contestuale rimane cruciale: «Una macchina […] sa senza sapere di sapere e senza nemmeno sapere cosa significhi sapere. Il suo sapere […] è semplicemente informazione costituita da simboli formali senza significato, che possono essere collegati meccanicamente a qualche azione deterministica» (Faggin, 2022: 54-55). A dimostrazione, si veda la ricognizione che Simone D'Alessandro illustra sulle differenze irriducibili tra intelligenza umana e intelligenza artificiale, analizzando i presupposti teorici e i test utilizzati dai programmatori per la valutazione delle interazioni con gli algoritmi ad adattamento automatico utilizzati nei large language model delle chatbot – il test di Turing classico; b) il test di Turing Inverso; c) il test di Winograd; d) il test Winogrande; e) il Test di Lovelace sulla creatività degli agenti artificiali – e la disamina di un insieme oggettivo di limiti dell'intelligenza artificiale: «1. incapacità di comprendere "propriamente" la semantica; 2 incapacità di contestualizzare presupposti e implicature conversazionali ambigue; 3. incapacità di interagire o agire in modo imprevedibile; 4. incapacità di decidere in assenza di informazioni di partenza date dal programma; 5. incapacità di boicottare creativamente l'automatismo dei sistemi di programmazione» (2025: 212). In definitiva, viene riconosciuta all'AI una creatività esplorativo-combinatoria, ma non trasformativa che discende dall'intenzionalità libera e consapevole di poter modificare il sistema della programmazione: «gli algoagenti, per quanto sofisticati nelle loro capacità computazionali, così come la rete di dispositivi che li accoglie e affianca, non possiedono coscienza, né agiscono in modo completamente autonomo nel senso intenzionale che viene attribuito all'essere umano. [...] Nessun algoagente, per quanto evoluto, possiede un "sé" che esperisce il mondo, né una volontà autonoma che lo guidi nelle sue azioni» (Grassi, 2024: 22-23). Come ribadisce Massimo Airoldi, «i sistemi di machine learning non hanno una vita sociale densa di significato né riflessività, o coscienza. In quanto agenti sociali si limitano a utilizzare una *ragione meramente pratica*, attualizzando disposizioni culturali acquisite da esperienze datificate» (2024: 41).

La coscienza è – scrive brillantemente Max Tegmark – il "problema davvero difficile" (*really hard problem*): perché qualcosa è cosciente? La difficoltà non diminuisce pur assumendo «una posizione che sia il più ampia e inclusiva possibile», come quella che definisce la coscienza come un'«esperienza soggettiva»: «se essere voi vi fa sentire qualcosa in questo momento, allora siete coscienti» (2017/2018: 313). A dispetto del riduzionismo fisicalista, la questione è inaggirabile anche ammettendo che la domanda ricada nel dominio delle proposizioni infalsificabili, ossia non-scientifiche o

metafisiche. È una constatazione osservativa che «Non solo sappiamo di essere coscienti, ma è *tutto* quello che sappiamo con assoluta certezza» (ivi: 320).

Ora, possiamo convenire che le macchine artificiali, dagli "Assistenti cognitivi" agli "individui algoritmici", dagli "algoagenti generatori" agli "algogenti incorporati" non abbiano una coscienza e non possano essere equiparati alle persone a cui imputare pensieri, azioni ed espressioni che ci aspettiamo che dovrebbe possedere un individuo adulto psico-socialmente "normale". Nonostante la tendenza, tutta umana, ad antropomorfizzare. Dovrebbe infatti possedere in modo critico l'autocoscienza del soggetto epistemico, l'autonomia del soggetto pratico e l'autorealizzazione del soggetto sensibile e per ciascuna di queste dimensioni esser cosciente della fallibilità. Sul piano dei criteri di identificazione dell'algoagente in quanto persona, inoltre, dovrebbe poter essere identificato attraverso i seguenti criteri: a) l'"identificazione numerica" come "organismo" in quanto corpo localizzato nello spazio e nel tempo e connotato da aspetti fisici; b) l'"identificazione generica' di un "soggetto" che – in generale – definisce le proprie relazioni con il mondo manifestando una capacità di intendere, agire e volere; e c) l'"autoidentificazione predicativa dell'io compiuta in quanto persona "determinata" a partire dall'assunzione di una propria "biografia individuale" che manifesta la capacità di "voler-essere-se-stesso". Nelle risposte alle domande esistenziali «chi sono io?», «Chi voglio essere io?» e «Chi posso essere io?» sono coinvolti tutti gli aspetti che distinguono, secondo Habermas, l'individualità umana, cioè l'identità dell'io, le capacità di pensiero, azione ed espressione, ma anche il proprio corpo (1968/1970; 1981/1986/; 1988/1991). Non solo. Per gli esseri umani, queste forme di identificazione si riferiscono a tre forme di identità – "simbiotica", "dei ruoli" e "dell'io" – che l'individuo matura nello sviluppo ontogenetico e che possono essere socialmente accettate e negate, in parte o totalmente, nei rapporti di reciproco riconoscimento, culturalmente interpretati e normativamente regolati delle relazioni che il soggetto instaura nei processi di socializzazione primaria e secondaria (Corchia, 2012: 31-52). Si può ragionevolmente, immaginare un algoagente che affronti delle "crisi di maturazione" nel proprio percorso di sviluppo cognitivo, morale ed espressivo? E che, una volta divenuto "maturo", sappia persino riconoscere, risolvere o accettare e tollerare come insolubili problemi relativi alla rappresentazione della realtà, al controllo del comportamento e alla repressione dei propri desideri? È la condizione abituale del nostro essere soggetti epistemici sul piano dell'auto-riflessione, soggetti morali sul piano dell'autonomia e soggetti sensibili sul piano dell'autorealizzazione. Gli esseri umani, infine, sono capaci di adattarsi alle condizioni date ma anche di trasformarle. Nell'apologetica dell'AI scompare la qualità propria del genere umano che,

scriveva Ardigò, ci rende un "oltre" la capacità imitativa delle macchine:

> la direzione dell'oltreità in cui intelligenze umane moralmente orientate alla verità e al bene, e coltivate a livello storico, non avranno «doppi» computerizzati: *a*) oltre nel determinare, in coscienza personale, la rilevanza dei possibili progetti di esperire vivente e di azione, rispetto alia propria domanda di senso della vita, tenuto conto delle comunicazioni che provengono al soggetto singolo dall'ambiente socio-culturale, e delle sfide cui la persona è sottoposta; *b*) oltre nel decidere di rom-pere regole, e di crearne di nuove, quando lo si ritenga necessario ai fini della rilevanza di cui al punto precedente; regole, mediazioni culturali, abitudini, motiva-zioni standard a stimoli, a bisogni. Il tutto secondo decisioni di rilevanza soggettiva non spiegabili solo psicologicamente, che nascono dall'intenzionalità (1983: 238).

Queste forme di autocoscienza sinora sono relegate nella fantascienza, come nel caso di Hal 9000, il computer che in *2001: Odissea nello spazio* cerca di impedire il ritorno del protagonista nell'astronave salvare la sua vita.

2. Individualizzazione senza scelta

Non saranno mai persone, eppure gli algoritmi sono agenti di mutamento che producono attivamente degli effetti nella costruzione del mondo sociale: «Sebbene manchi della dimensione auto-rappresentativa che caratterizza l'agire umano, l'algoagente produce effetti concreti e profondi sulle dinamiche sociali, politiche e culturali, operando attraverso strutture che si intrecciano con le relazioni interpersonali e con le istituzioni» (Grassi, 2024: 23). Una considerazione di Luciano Floridi fa riflettere sulla mancanza di una riflessione pubblica sugli effetti reali delle applicazioni dell'Intelligenza artificiale "debole" (Weak AI) mentre l'immaginario si confonde sulle superintelligenza AI: «Il successo delle nostre tecnologie dipende in gran parte dal fatto che, mentre speculavamo sulla possibilità dell'ultraintelligenza, abbiamo sempre più avvolto il mondo per mezzo di così tanti dispositivi, sensori, applicazioni dati da diventare un ambiente adattato alle ict, dove le tecnologie possono sostituirci senza disporre di alcuna comprensione, stato mentale, intenzione, interpretazione, stato emotivo, abilità semantiche, coscienza, autocoscienza o intelligenza flessibile» (2022/2022: 233). Gli effetti sociali sono già oggi rilevanti sia nella riproduzione materiale delle sistemi funzionali economici e politici che in quella simbolica dei mondi vitali. Tra le trasformazioni epocali che si prefigurano all'orizzonte qui mi limito a considerare la questione della libertà di scelta e dei condizionamenti. In un

recente saggio, Ori Schwarz pone l'*agency* come elemento discriminante tra l'umano e l'artificiale rimarcando il pericolo esiziale per l'umanità. Il rischio che viene paventato non è l'emulazione del comportamento umano da parte delle macchine bensì la sottrazione della libertà di scelta agli uomini. Effettivamente, gli algoritmi predittivi stanno interferendo con i processi decisionali a livelli variabili, dal mero supporto alla sostituzione della volizione. L'aspetto che qui interessa della sua riflessione è il collegamento con la storia dello sviluppo capitalistico che nella fase delle piattaforme starebbe affermando un modello algoritmico di "*individuation without choice*". Le attività predittive sono elaborate da dati su ciò che gli attori avrebbero scelto e ciò salva l'apparenza del principio della soggettività dell'individuo, dal consumo di merci, alla fruizione culturale, dalla politica alle relazioni personali. È proprio la scelta individuale però che risulta incompatibile con la nuova *governance* del mercato. È un ostacolo che perturba il ciclo della produzione e il flusso dei consumi. La tesi di Schwarz è che la previsione algoritmica stia «riducendo il numero di scelte che le persone devono compiere» (2025: 2). Gli algoritmi predittivi sono integrati in un numero sempre maggiore di sistemi, compresi i sistemi di raccomandazione che modellano e sostituiscono le scelte. Alcuni sistemi di essi concentrano la nostra attenzione su un piccolo insieme di opzioni (che si tratti di libri, gruppi sui social media o ristoranti) tra cui scegliere, modellando l'architettura delle decisioni poiché prevedono che questa focalizzazione aumenterebbe le nostre possibilità di fare una scelta e di acquistare; questi possono essere considerati come facilitatori della scelta. Altri sistemi invece, come i servizi di streaming musicale, i siti di social network (SNS) come TikTok e i servizi di abbonamento personalizzati che ordinano prodotti per i clienti e li spediscono automaticamente sulla base di previsioni, eliminano quasi completamente la scelta: la scelta è necessaria solo fino al punto in cui può essere prevista e automatizzata in modo affidabile. In prospettiva, le tecnologie elimineranno la possibilità di scelta: «gli algoritmi sono "regole generative" che non possono essere violate, rendendo impossibile scegliere determinate possibilità di azione […], senza lasciare ai governati alcuna scelta se obbedirvi o meno» (Ivi: 4).

In un colpo solo sono superati sensismo e razionalismo, Hume e Kant. Non sono né i sentimenti né le ragioni a formare i motivi che muovono le azioni perché è la volontà stessa che non rappresenta più il motore delle scelte. In una sorta di teismo tecnocratico, al di sopra dei soggetti opera un primo principio di tutte le cose, onnisciente, onnipotente e onnipresente che fissa le leggi generali che governano le forme della nostra esistenza. In tal modo viene data una risposta al problema che Kant ha posto nella *Fondazione della metafisica dei costumi*. Dopo aver elaborato la formula della legge – «Praticamente buono è ciò che determina la volontà mediante rappresentazioni

della ragione, quindi non per cause soggettive, ma oggettivamente, cioè per princìpi validi per ogni essere ragionevole in quanto tale» –, infatti, introduce la clausola che una simile legge universale è ugualmente buona per tutti solo se tutti fanno ciò che devono fare (1785/1970: 70). Per Kant, la validità normativa dei precetti richiede da parte del soggetto un'obbedienza incondizionata, indipendente dalle particolari conseguenze indesiderate, a una determinazione della volontà di tipo speciale che si colloca sul piano dell'escatologia cristiana come il "sommo bene" di un progresso verso la condizione cosmopolitica di cittadini liberi ed eguali. Per l'implicita filosofia della storia tecnocapitalistica, la regolazione algoritmica delle condotte è la panacea contro la fallibilità delle volizioni e contro la corruzione della natura umana. E l'"indipendenza da ogni costrizione imposta dalla volontà di un altro" viene vista come l'errore da correggere affinché gli uomini si risolvano a cooperare in vista di un fine superiore, esteriore e costrittivo ma benevolo e customizzato. Gli effetti della libera volontà sono così prevenuti e neutralizzati. Non serve più la fallace e pericolosa autocoscienza: le preferenze degli individui non sono accessibili attraverso l'introspezione riflessiva, ma si deducono invece nella miriade di risposte documentate agli stimoli ricevuti. Quest'ideale è stato ben espresso dall'ex data scientist di Google, Stephens-Davidowitz: «i Big Data ci permettono finalmente di vedere ciò che le persone vogliono e fanno realmente, non ciò che dicono di volere e di fare» (2017: 54). L'elaborazione delle tracce digitali lasciate dal nostro comportamento darebbero descrizioni autentiche sul nostro essere e nascoste anche agli stessi attori: «gli algoritmi ti conoscono meglio di quanto tu conosca te stesso» (Ivi: 155).

Il riduzionismo scientista tende a sovrastimare la veridicità dei modelli algoritmici databehaviouristi (Rouvroy, 2013) e sottostimare quanto gli utenti partecipino alla formazione del sé nel dialogo con gli "specchi algoritmici": «gli output elaborati dalle macchine autonome sono attivamente negoziati e problematizzati da parte degli individui» (Airoldi, 2024: 34). Eppure questa rappresentazione, secondo Brubaker, è congeniale all'evocazione di un "sé post-neoliberista" che cancella persino la liberà diluita, consumistica e atomistica del *self made*. Mentre il sé neoliberista, "imprenditoriale", era un soggetto di scelta, autonomia e responsabilità, il sé post-neoliberista è passivo, un soggetto di conoscenza, previsione e controllo governato da dati e algoritmi, guidato esternamente da sistemi artificiali e risponde in modo prevedibile agli stimoli. Sebbene la scelta rimanga indispensabile nei sistemi algoritmici, essa è svuotata, in quanto viene utilizzata solo per automatizzare e manipolare il processo decisionale dell'utente (2022: 43). Nella prospettiva di una sociologia critica della tecnologia, è utile richiamare alcune riflessioni sviluppate da Kopsaj in *Digital Age, Inclusive Future: Society, Self, and Health*. Viene indagato il nesso tra digitalizzazione, costruzione del sé e

trasformazioni delle strutture sociali, mostrando come l'avanzata dell'intelligenza artificiale e delle piattaforme digitali generi nuove forme di condizionamento simbolico, di diseguaglianza e di limitazione dell'agency individuale. Le analisi offrono un terreno fertile per connettere il problema della mimesi tecnologica con quello dell'ideologia tecnica e della governance algoritmica, mettendo in luce i rischi di una riduzione della soggettività a mero oggetto di calcolo predittivo: «Nell'era digitale, l'io è diventato sempre più performativo e mediato. Le piattaforme dei social media [...] incoraggiano gli individui a curare la propria identità, presentando versioni selettive e curate di sé stessi per il consumo pubblico. [...] Questa performatività, tuttavia, spesso va a discapito dell'autenticità, poiché i confini tra l'io come rappresentazione e l'io come esperienza diventano sfumati. Inoltre, gli algoritmi svolgono un ruolo chiave nel plasmare l'identità» (2025: 102). Questa "colonizzazione" del sé – e dei processi di riproduzione simbolica del mondo della vita: la socializzazione delle persone, le interazioni sociali, la trasmissione culturale – è il futuro possibile a cui aspirano gli ideologi dell'AI. Siamo incoraggiati a impegnarci meno nella scelta riflessiva, perché la previsione algoritmica, viene detto, è ciò che avremmo scelto se avessimo scelto. In tal senso, secondo Schwarz, «l'analisi predittiva minaccia di smantellare questa associazione tra individualizzazione e scelta, offrendo individuazione senza scelta, o addirittura un'*individualizzazione senza scelta*» (2025: 13). La libertà è ridotta a libertà di reagire piuttosto che di agire (Davies, 2024). Sostituita dalla previsione, la scelta viene gradualmente bandita dalla pratica quotidiana senza essere espulsa dallo spazio delle giustificazioni. la previ–sione si basa su comportamenti passati documentati, che sono considerati il risultato di scelte consapevoli. La volizione individuale rimane quindi il fondamento per la determinazione algoritmica intesa a prevederla e sostituirla. Giustamente, Schwarz scrive che «le giustificazioni tardo-moderne vengono quindi mantenute per giustificare una realtà non più moderna» (2025: 14).

3. Ancora su tecnica e scienza come ideologia?

Il sodalizio tra il riduzionismo scientista della visione computazionale dell'intelligenza umana (e delle persone trattate come mezzi per il raggiungimento di fini) e gli interessi delle corporation hi-tech (che avanzano la pretesa di stabilire per tutti quali siano i fini) alimenta una rappresentazione dell'AI come tappa evolutiva nel progresso della civilizzazione. Persino gli apocalittici accettano la previsione che artificiale e umano saranno congiunti. Contro questa parvenza che diviene oggettiva nelle condizioni strutturali di una comunicazione sistematicamente distorta, Ardigò avanzava il sospetto

che dietro queste «macchine dotate di programmi "intelligenti", nel senso di *computer* con una certa capacità di decisione non deterministica in situazioni con adeguato carico di contingenze» ossia «dietro la novità dell'impresa e del suo accattivante strumentario computazionale (in apparenza neutrale), si manifesti un soprassalto di quell'Illuminismo meccanicistico settecentesco e ciò in un tempo in cui sembrano, per contro, avanzare valori post materialistici e domanda di senso complessivo della vita umana, sia singola che per l'umanità storica in pericolo di autodistruzione» (1983: 233-234). A quarant'anni di distanza, quella che lo studioso bolognese – nell'emergente scoperta e rivendicazione delle nuove grammatiche di vita di un'epoca post-ideologica – poteva definire una "ideologia in declino" ci appare per contro come una nuova rappresentazione deformata della realtà del genere umano. Come rimarca Grassi, l'intelligenza artificiale è l'esempio compiuto di «quel fenomeno in cui la tecnologia assume una dimensione quasi mitologica, trascendendo la realtà materiale per divenire un simulacro ideologico, in cui si riversano credenze e aspirazioni collettive, producendo sentimenti di soggezione di fronte all'apparato tecnoscientifico (2024: 18). La diagnosi francofortese sulla coscienza reificata dalla razionalità scientifica come ideologia ausiliaria alla riproduzione del capitalismo torna attuale per disvelare le rappresentazioni dei ricercatori e delle aziende che propongono le proprie tecnologie di intelligenza artificiale come strumenti neutrali (Fuchs, 2021).

C'è una consonanza con la riflessione sull'ideologia della società industriale avanzata di Herbert Marcuse. Rileggendo criticamente le diagnosi di Max Weber, egli sosteneva che la razionalizzazione tecnico-scientifica non realizza un mondo razionale, ma piuttosto, in nome della ragione, una forma specifica di dominio funzionale agli interessi capitalistici (1964a/1967). L'ideologia della scienza e della tecnica giustifica l'estromissione dei cittadini da parte dell'apparato tecnocratico e delle élites con il consenso dei dominati. In una relazione che Habermas tenne davanti a Marcuse, Ernst Bloch, Alfred Sohn-Rethel – nel gruppo "filosofia della prassi" fondato da Gajo Petrović a Korčula e di cui Habermas e Marcuse erano protagonisti sin dalla costituzione nel 1965 – viene chiarito il nucleo della tesi di *Tecnica e scienza come ideologia* (1968a/1969):

> Quando questa coscienza tecnocratica che naturalmente è una falsa coscienza, raggiunge l'evidenza di una ovvietà quotidiana, il ruolo della tecnica e della scienza diventa un argomento in grado di giustificare il fatto che nelle società moderne debba perdere le sue funzioni un processo democratico di educazione della volontà tale da consentirle di affrontare, discutere e risolvere problemi pratici. [...] In tal senso la tecnica e la scienza assumono una doppia funzione: non sono solo forze produttive, sono anche ideologiche. In questo modo si

> spiega anche perché oggi lo squilibrio tra le forze produttive e i rapporti di produzione non è tangibile, ossia non è più evidente alla coscienza delle masse (1968b/1980: 65-66).

Mentre il controllo delle società complesse spinge verso lo sganciamento dei processi deliberativi-decisionali dalla sovrastruttura normativa, dai motivi per-sonali e dai valori dei cittadini, si diffonde nella popolazione il "cinismo" di una coscienza borghese che smentisce se stessa nell'edonismo del privato. Il nesso tra capitalismo, tecnica e dominio ha l'"ultima parola" per Marcuse quando «la critica si ferma, accetta ciò che si pretende ineluttabile e diventa apologia; peggio ancora, mette sotto accusa la possibile alternativa di una razionalità storica qualitativamente diversa» (1964b/1969: 249). E ciò sorprende è la mancanza di critica verso la sostituzione degli esseri umani con le macchine e il potere sovrano che i nuovi "padroni del vapore" intendono esercitare in nome della volontà popolare: «Che dietro la punta di diamante della ricerca interdisciplinare al servizio della computerizzazione della conoscenza, quale appunto quella delle IA, ci siano obiettivi di potenza oltre che di utilitarismo sociale o mercantile, è quasi ovvio» (Ardigò, 1983: 239).

Riferimenti bibliografici

Airoldi M. (2024). *Machine habitus. Sociologia degli algoritmi*. Milano: Luiss University Press.

Ardigò A. (1983). Un nuovo processo mimetico: le ricerche di «intelligenze artificiali»: interrogativi ed ipotesi di rilevanza. *Studi di Sociologia*, 21(3): 233-244.

Brubaker R. (2022). *Hyperconnectivity and its discontents*. Cambridge: Polity Press.

Charniak E., McDermott D. (1985). *Introduction to artificial intelligence*. Boston: Addison-Wesley Longman Publishing Co.

Chierici A. (2024). L'emergere storico dell'intelligenza artificiale. *Nuova Atlantide*, 11(20): 33-37.

Corchia L. (2012). Il concetto di individuo. In Id., *La teoria della socializzazione di Jürgen Habermas* (pp. 31-52). Pisa: ETS.

D'Alessandro S. (2025). Creatività e intelligenza artificiale. In Id., *La regola che cambia le regole. Sociologia dei processi creativi e degli ecosistemi innovativi* (pp. 209-229). Milano: Mimesis.

Davies W. (2024). Reaction value: affective reflex in the digital public sphere. *Distinktion*, 25(3): 297-317.

Eco U. (1964). *Apocalittici e integrati*. Milano: Bompiani.

Faggin F. (2022). *Irriducibile. La coscienza, la vita, i computer e la nostra natura*. Milano: Mondadori.

Floridi L. (2022). *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*. Milano: Raffaello Cortina.

Fuchs C. (2021). History and class consciousness 2.0: Georg Lukács in the age of

digital capitalism and big data. *Information, Communication & Society*, 24(15): 2258-2276.

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

Habermas J. (1968a). Tecnica e scienza come ideologia. In Id., *Teoria e prassi nella società tecnologica* (pp. 195-234). Bari: Laterza, 1969.

Habermas J. (1968b). Su alcune condizioni necessarie al rivoluzionamento delle società tardo-capitaliste. In Id., *Cultura e critica. Riflessione sul concetto di partecipazione politica e altri scritti* (pp. 61-76). Torino: Einaudi, 1980.

Habermas J. (1968). Appunti per una teoria della socializzazione. In Id., *Cultura e critica. Riflessione sul concetto di partecipazione politica e altri scritti* (pp. 77-139). Torino: Einaudi, 1970.

Habermas J. (1981). Il mutamento di paradigma in Mead e Durkheim: dall'attività finalizzata a uno scopo all'agire comunicativo. In Id., *Teoria dell'agire comunicativo. Vol. II* (pp. 547-696). Bologna: il Mulino, 1986.

Habermas J. (1988). Individuazione tramite socializzazione. Sulla teoria della soggettività in Mead. In Id., *Il pensiero post-metafisico* (pp. 184-236). Roma-Bari: Laterza, 1991.

Kopsaj V. (2025). *Digital age and inclusive future. Society, self and health*. Milano: FrancoAngeli.

Kurzweil R. (1999). *The age of spiritual machines*. New York: Viking Press.

Marcuse H. (1964a). *L'uomo a una dimensione. L'ideologia della società industriale avanzata*. Torino: Einaudi, 1967.

Marcuse H. (1964b). Industrializzazione e capitalismo nell'opera di Max Weber. In Id., *Cultura e società. Saggi di teoria critica 1933-1965* (pp. 243-264). Torino: Einaudi, 1969.

McCarthy J., Minsky M., Rochester N., Shannon C. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, 18(4): 1-13 (rist. 2006).

Rich E., Knight K. (1991). *Artificial intelligence*. New York: McGraw-Hill.

Rouvroy A. (2013). The end(s) of critique: data-behaviourism vs. due-process. In Hildebrandt M., de Vries K., a cura di, *Privacy, due process and the computational turn* (pp. 143-169). New York: Routledge.

Russell S.J., Norvig P. (2010). *Artificial intelligence. A modern approach*. Upper Saddle River: Pearson Education.

Schwarz O. (2025). The post-choice society: algorithmic prediction and the decentring of choice. *Theory, Culture & Society*, First online, 15 marzo: 1-19.

Searle J.R. (1980). *Menti, cervelli e programmi. Un dibattito sull'intelligenza artificiale*. Milano: Clued, 1984.

Tegmark M. (2017). *Vita 3.0. Essere umani nell'era dell'intelligenza artificiale*. Milano: Raffaello Cortina, 2018.

Turing A.M. (1950). Computing machinery and intelligence. *Mind*, 76(236): 433-460.

Vladeck D.C. (2014). Machines without principals: liability rules and artificial intelligence. *Washington Law Review*, 97(1): 117-150.

Do ut des. Gli attori del welfare educativo alla sfida dell'IA

di *Giuseppe Luca De Luca Picione*[*], Domenico Trezza*[**]

Questo lavoro analizza rischi e opportunità legati all'uso dell'IA nel welfare educativo. Partendo dalla logica del *do ut des* e dalla sua apparente reciprocità algoritmica, gli autori mirano a individuare le condizioni perché l'IA diventi leva di giustizia socio-educativa, anziché nuovo fattore di disuguaglianza. L'analisi si concentra sull'esperienza di Govern-AI Eda Lab nei CPIA campani, esplorando come integrare l'IA nelle pratiche educative senza compromettere la dimensione umana e relazionale dell'apprendimento.

Parole chiave: Welfare educativo; intelligenza artificiale; chatbot; adult education; CPIA; algoritmi.

Do ut des. Educational welfare actors at the challenge of AI

This paper examines the risks and opportunities associated with the use of AI in educational welfare. Starting from the logic of *do ut des* and its seemingly reciprocal algorithmic exchange, the authors aim to identify the conditions under which AI can become a driver of socio-educational justice rather than a new source of inequality. The analysis focuses on the experience of the Govern-AI Eda Lab in Campania's CPIAs, exploring how AI can be integrated into educational practices without compromising the human and relational dimensions of learning.

Keywords: educational Welfare; artificial intelligence; chatbots; adult education; CPIA; algorithms.

DOI: 10.5281/zenodo.18435621

[*] Università di Napoli Federico II. giuseppe.picionedeluca@unina.it.

[**] Università Pegaso II, domenico.trezza@unipegaso.it.

Il presente articolo è frutto della collaborazione congiunta degli autori. Nello specifico, il paragrafo 1 è stato redatto da Domenico Trezza; il paragrafo 2 da Giuseppe Luca de Luca Picione; i paragrafi 3 e 4 sono stati elaborati in forma condivisa da entrambi gli autori.

1. Welfare educativo nell'era dell'Intelligenza Artificiale: tra aspetti ambivalenti e sfide aperte

L'intelligenza artificiale (d'ora in avanti IA[1]) sta progressivamente perdendo lo status di tema di nicchia nelle scienze sociali e si impone oggi come una delle chiavi di lettura più incisive delle trasformazioni in corso nei sistemi di welfare, inclusi quelli educativi (Cooper, 2023; De Luca Picione, Trezza, 2024; Grimaldi, 2022; Selwyn, 2019). Non si tratta soltanto dell'ingresso di nuovi strumenti tecnologici, ma di un cambiamento più profondo: l'IA sta ridefinendo i presupposti su cui costruiamo e legittimiamo la conoscenza, con effetti che attraversano tutti i settori della sfera pubblica e normativa (Bentley, 2025; Cosenza *et al*., 2023; Elliott, 2021; Floridi, 2020).

Nei contesti educativi, questa trasformazione assume un significato ancora più centrale. Qui, la conoscenza non è semplicemente un insieme di dati da gestire, ma rappresenta il fulcro stesso del diritto all'inclusione, alla cittadinanza e alla mobilità sociale. Ed è proprio nelle realtà educative più complesse – come l'istruzione degli adulti e i percorsi di lifelong learning – che l'IA diventa terreno di tensione tra innovazione e disuguaglianza, ovvero, tra necessità di mettere in campo nuovi strumenti per raggiungere buoni standard di efficacia ed efficienza e i nuovi rischi che aumentino i gap sociali, tecnologici, digitali e culturali tra persone (e luoghi) che hanno capacità di attivare risorse e chi invece queste disponibilità non ce l'ha (Fortino *et al*., 2024; Oberdieck, Mock, 2024).

In Italia, i Centri Provinciali per l'Istruzione degli Adulti (CPIA)[2] rappresentano il cuore del welfare educativo rivolto a chi, per ragioni biografiche, migratorie o sociali, rischia di restare escluso dai circuiti tradizionali della formazione. Istituiti con il DPR 263/2012 in sostituzione dei Centri Territoriali Permanenti, i CPIA operano come rete provinciale per l'istruzione in età adulta, offrendo percorsi di primo livello, alfabetizzazione e apprendimento della lingua italiana, ma anche attività di ampliamento dell'offerta formativa rivolte a cittadini italiani e stranieri, con particolare attenzione a chi vive condizioni di fragilità o marginalità educativa. Nei CPIA, educazione e welfare si intrecciano: la scuola diventa luogo di apprendimento ma anche di

[1] Siamo consapevoli che il termine "intelligenza artificiale" racchiuda un ampio spettro di tecnologie, dal machine learning ai sistemi di automazione complessi. In questa sede, tuttavia, con tale etichetta ci riferiamo in particolare ai sistemi generativi, ossia a quelle tecnologie di IA di recente introduzione capaci di produrre testi, immagini, video o altri contenuti inediti a partire da input testuali o vocali. Il più noto tra questi è il chatbot ChatGPT dell'azienda OpenAI.

protezione sociale, in un'ottica integrata che va ben oltre la mera erogazione di corsi.

È proprio in questi contesti che le promesse dell'IA – personalizzazione, efficienza, supporto didattico – si scontrano con rischi tangibili. I modelli predittivi e generativi non si limitano a gestire informazioni, ma definiscono cosa sia rilevante, tracciano profili, stabiliscono priorità educative (Kitchin, 2014; O'Neil, 2016). Il rischio è duplice: da un lato, algoritmi che costruiscono categorie rigide, appiattendo la complessità delle biografie; dall'altro, la creazione di relazioni profondamente asimmetriche tra chi fornisce i dati – studenti, docenti e personale – e chi elabora decisioni automatizzate (Foley, Malese, 2025).

D'altra parte, la logica algoritmica delle piattaforme online, le cui implicazioni sociologiche sono ormai attenzionate con costanza (Airoldi, 2022; Aragona, 2021; Boccia Artieri, 2020; Williamson, 2019) ricorda, a chi scrive, una vecchia espressione latina, *do ut des* (io do a te, affinché tu dia a me), che, se sulla carta prospetta reciprocità – offro dati, ricevo servizi personalizzati – rischia di tradursi, nei fatti, in una governance verticale dove la conoscenza restituita dai sistemi algoritmici è filtrata da logiche opache e spesso estranee ai contesti locali. Nei CPIA, dove molte storie educative sono intrecciate a percorsi di marginalità o migrazione, questa tensione si fa particolarmente acuta.

In questo scenario si colloca l'esperienza di *Govern-AI Eda Lab* (Govern-AI EL), un progetto sperimentale realizzato grazie alla collaborazione tra Regione Campania e Università Federico II, che esplora l'uso dell'intelligenza artificiale generativa per supportare docenti e governance scolastica nei CPIA. *Govern-AI EL* non si presenta come una "buona pratica" definitiva, ma come un laboratorio critico, un terreno di sperimentazione per comprendere sia le potenzialità sia i limiti dell'IA applicata al welfare educativo.

Scopo di questo contributo è riflettere su questa ambivalenza. A partire dall'esperienza di *Govern-AI*, intendiamo analizzare i tentativi concreti messi in atto per ridurre il divario insito nelle logiche algoritmiche e per trasformare l'IA in uno strumento capace di promuovere inclusione, anziché diventare un nuovo vettore di disuguaglianze. Soprattutto, ci interroghiamo su quali condizioni culturali, organizzative ed etiche siano necessarie perché l'innovazione tecnologica non eroda la dimensione umana, relazionale e plurale dell'educazione, in particolare nei contesti più vulnerabili come quelli dei CPIA.

2. IA e algoritmi nelle scuole degli adulti. Tra divari e tentativi di convergenza

L'intelligenza artificiale sta assumendo un ruolo sempre più rilevante anche nei sistemi educativi, dove il dibattito si concentra non soltanto sull'introduzione di nuovi strumenti tecnologici, ma su trasformazioni più profonde che investono le pratiche didattiche, i processi di apprendimento e le forme di governance scolastica (Foley, Malese, 2025; Selwyn, 2022). Le sperimentazioni in atto, sebbene ancora in fase iniziale, si propongono di rendere l'istruzione più inclusiva, attraverso soluzioni come tecnologie multisensoriali finalizzate a supportare l'apprendimento e a migliorare le interazioni, in particolare con studenti che presentano disturbi dello spettro autistico, bisogni educativi speciali o disturbi specifici dell'apprendimento (Grimaldi, 2022; Selwyn, 2019; Yang, 2024). Tuttavia, anche in questi ambiti, le promesse dell'IA non possono farci dimenticare che esistono potenziali effetti indesiderati o rischi che riguardano tanto gli attori istituzionali quanto gli utenti finali.

Un primo nodo critico riguarda la natura "costruttiva" degli algoritmi, che non si limitano a elaborare dati, ma plasmano categorie sociali, priorità educative e visioni di realtà. Come osservano Espeland e Stevens (2008), i processi di quantificazione non sono mai neutri: definiscono cosa diventa visibile e meritevole di attenzione, riducendo la complessità sociale a dimensioni misurabili. Eubanks (2018) ha documentato come, nei servizi sociali, i sistemi algoritmici tendano a classificare i bisogni secondo logiche burocratiche, spesso cancellando specificità culturali o biografiche. Questo rischio si amplifica nei contesti educativi più fragili, dove ogni storia di apprendimento è intrecciata a percorsi di marginalità sociale.

Nei CPIA che, lo ricordiamo, sono le "scuole per gli adulti" in Italia, la complessità delle biografie degli immigrati o dei percorsi discontinui, ad esempio, viene spesso tradotta in indicatori standardizzati, come il "livello linguistico" o il "rischio di abbandono" (De Luca Picione, Madonia, 2018). Tali categorie, pur necessarie per gestire la didattica, rischiano di consolidare stereotipi o di rendere invisibili risorse preziose, come il plurilinguismo o le competenze informali maturate nei contesti migratori (Fricker, 2007). Williamson (2019), ad esempio, ha mostrato come i sistemi di tracking educativo nel Regno Unito tendano a cristallizzare disuguaglianze, fissando i profili predittivi degli studenti e precludendo percorsi alternativi.

L'applicazione dell'IA nei sistemi educativi – attraverso l'uso di strumenti come chatbot, tutor virtuali, sistemi di profilazione o dashboard predittive in chiave formativa ma anche di organizzazione e governance dei sistemi scolastici – rischia, quindi, da un lato, di esacerbare certi divari già

molto significativi con la tecnologia digitale consolidata (digital divide dei territori e degli utenti, gap nelle competenze informatiche, etc..), dall'altro, di allargare l'asimmetria che emerge dal rapporto (e dal potenziale scambio) utente – piattaforma. In teoria, lo scambio è chiaro: gli utenti – in questo caso, studenti, ma anche insegnanti e amministratori scolastici – forniscono dati (informazioni anagrafiche, tracciamento delle attività di apprendimento, preferenze linguistiche, comportamenti digitali, interazioni con le piattaforme e altro ancora…) in cambio di servizi educativi più personalizzati o di interventi mirati. Nella pratica, però, lo squilibrio di *potere* è evidente: gli utenti cedono informazioni personali senza reale controllo su come queste verranno utilizzate, archiviate o eventualmente commercializzate (Foley, Malese, 2025; Pasquale, 2015; Selwyn, 2022).

Spesso, ciò che viene presentato come personalizzazione si traduce in servizi standardizzati che difficilmente compensano il valore estratto dai dati, ma che tendono a valorizzare, invece, il cosiddetto "capitale culturale istituzionalizzato" (Bourdieu, 1979). Nei percorsi di *adult education*, questo si potrebbe tradurre nel rischio che metriche costruite sulla base di competenze formali penalizzino i saperi esperienziali o informali, cruciali invece per l'integrazione sociale. Gli algoritmi possono finire per privilegiare competenze "occupazionali" – più facilmente quantificabili e spendibili sul mercato del lavoro – a scapito di saperi emancipativi e critici, come insegnava Freire (1970). García e Wei (2014), peraltro, hanno inoltre evidenziato come la pluralità linguistica tipica, ad esempio, dei CPIA, renderebbe spesso inefficaci modelli addestrati su dati prevalentemente monolingui.

Ne sono esempio emblematico molte piattaforme di *adaptive learning*[3], che escludono variabili culturali non quantificabili, impongono percorsi didattici rigidi e trasformano i docenti in meri esecutori di scelte algoritmiche (Selwyn, 2022; Yang, 2024). In queste condizioni, la logica del *do ut des* rischia di apparire più come un esercizio di governance verticale che come uno strumento autentico di inclusione educativa.

L'impatto dell'IA nel welfare educativo solleva interrogativi etici che tuttavia vanno ben oltre la tutela della privacy. Noddings (1984) e Arendt (1958) ricordano come l'educazione sia, innanzitutto, relazione, cura e capacità di giudizio contestuale. Delegare decisioni pedagogiche a sistemi automatizzati comporta il rischio di erosione di questi valori fondanti. Perrotta

[3] L'*adaptive learning* è un approccio educativo basato su algoritmi e intelligenza artificiale che personalizza in tempo reale i percorsi di apprendimento, adattando contenuti, livello di difficoltà e feedback alle caratteristiche dello studente. Il sistema analizza dati su risposte, tempi di esecuzione e interazioni, ridefinendo costantemente il tracciato formativo in base a performance e bisogni individuali.

(2021) documenta come i sistemi di valutazione automatizzata tendano a sostituire il feedback umano, riducendo il momento educativo a un semplice atto transazionale.

Servono, dunque, quadri etici e normativi più solidi. Pasquale (2015) propone il principio della "trasparenza radicale", secondo cui i criteri decisionali degli algoritmi devono essere pubblici e accessibili: in sostanza, i sistemi *algorithm-based* e *AI-based* devono essere "spiegabili" (Felaco *et al.,* 2024). Benjamin (2019) d'altra parte, sottolinea la necessità di audit indipendenti per individuare bias algoritmici, mentre Glissant (1990) evoca il diritto all'opacità: il diritto, cioè, di sottrarsi alla datificazione senza subire penalizzazioni.

In ambito normativo, tra gli interventi più incisivi si segnala l'AI Act europeo (2024) che classifica i sistemi educativi basati su IA come "ad alto rischio", imponendo valutazioni d'impatto sui diritti fondamentali, garanzie contro la discriminazione e tracciabilità delle decisioni. Tuttavia, l'effettiva implementazione di tali garanzie resta una sfida aperta, e ovviamente lo è ancora di più per i CPIA.

D'altra parte, lo diciamo ancora una volta, i CPIA rappresentano un campo di tensione e di dilemmi: da un lato, la personalizzazione offerta dall'IA può diventare certamente strumento potente di inclusione; dall'altro, non bisogna però trascurare il risvolto della medaglia, ossia che gli stessi algoritmi rischino di agire come fattori di "colonialismo digitale", imponendo logiche esterne alle comunità locali (Couldry, Mejias, 2019), che a loro volta affrontano situazioni interne molto variegate dal punto di vista socio-culturale.

Affinché l'IA diventi davvero leva di welfare educativo, sono necessarie condizioni stringenti: processi di co-design che coinvolgano docenti e studenti, modelli algoritmici open-source e verificabili, clausole etiche vincolanti nei contratti con i fornitori tecnologici. In ultima analisi, la sfida consiste nel trasformare l'IA da strumento di governance verticale a tecnologia conviviale, radicata nelle pratiche comunitarie e nella pluralità dei territori (Illich, 1973). È questa la cornice entro la quale si è mosso chi è stato coinvolto nella programmazione di Govern-AI, oggetto di discussione di questo paper e, in particolare, del paragrafo che segue.

3. Govern-AI Eda Lab: IA conviviale e nodi irrisolti

Se il rischio dell'intelligenza artificiale nei sistemi educativi è quello di cristallizzare le disuguaglianze e standardizzare i percorsi, esistono tuttavia spazi di sperimentazione che cercano, se non di rovesciare completamente

questa logica, quantomeno di ridimensionarla e, soprattutto, di problematizzarla. Il progetto *Govern-AI Eda Lab* (Govern-AI EL), sviluppato a partire da Luglio 2023 e reso operativo da Giugno 2025, grazie alla collaborazione tra la Regione Campania e l'Università Federico II, si colloca precisamente in questa tensione. Non si propone come una buona pratica definitiva né come una soluzione immediatamente esportabile, ma come un laboratorio critico che esplora le possibilità di un'IA pensata come strumento di inclusione e di empowerment, soprattutto nei contesti complessi dell'educazione degli adulti.

L'iniziativa ha concentrato la propria azione sui CPIA della Campania, spazi in cui l'esperienza educativa si intreccia costantemente con percorsi biografici segnati da migrazione, interruzione scolastica e vulnerabilità sociale (De Luca Picione, Madonia, 2018). È proprio in questi contesti, densi di complessità culturale e linguistica, che si è tentato di sperimentare come l'IA possa costituire non soltanto un ausilio strumentale, ma anche un dispositivo capace di riconoscere la singolarità dei percorsi educativi, sostenere l'autonomia e costruire un ponte verso nuove forme di cittadinanza attiva.

L'infrastruttura sviluppata da *Govern-AI EL* si basa su una piattaforma conversazionale alimentata da un sistema di IA generativa, progettata specificamente per il CPIA Napoli città 2. Ciò che rende peculiare questo progetto è la scelta di costruire la piattaforma a partire da basi dati locali, elaborate attraverso sessioni di audit e co-progettazione con dirigenti scolastici, docenti e personale amministrativo. Si è così evitato di importare modelli esterni potenzialmente inadatti, preferendo integrare conoscenze e pratiche già presenti nel contesto educativo, in linea con le critiche mosse in letteratura verso l'applicazione acritica di sistemi algoritmici (Eubanks, 2018; Espeland, Stevens, 2008).

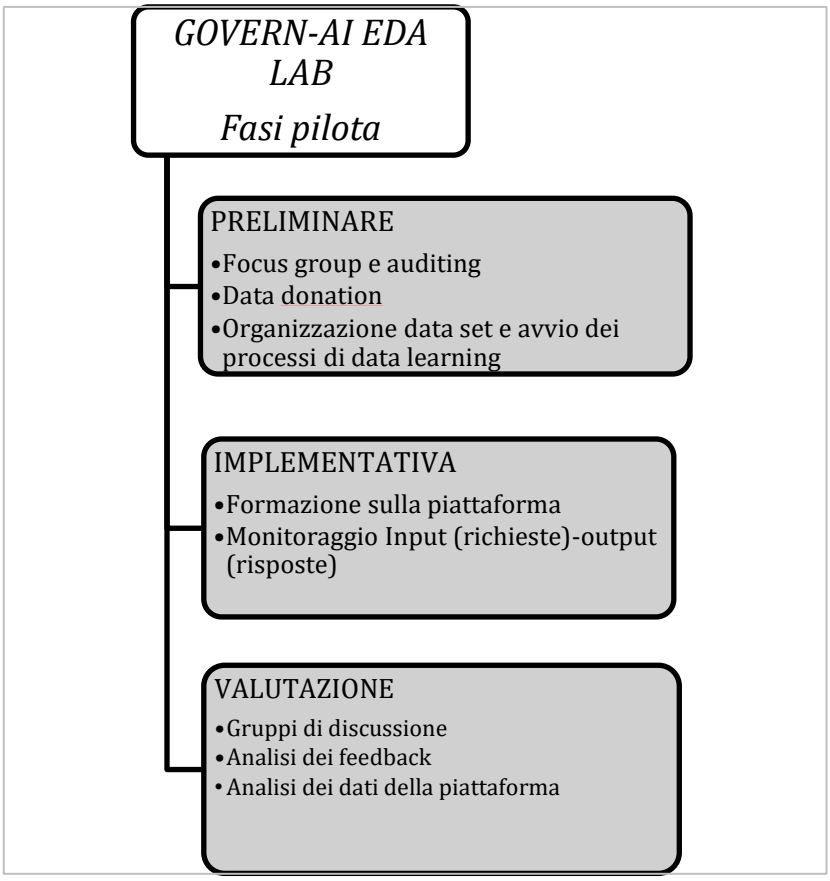
Il percorso di sviluppo ha seguito tre fasi: una fase preliminare di definizione tecnica e costruzione del corpus informativo locale; una fase implementativa in cui il personale scolastico ha ricevuto formazione e ha potuto discutere collettivamente rischi e opportunità legati all'uso dell'IA; e una fase valutativa, attualmente in corso, dedicata alla raccolta di feedback qualitativi e quantitativi per il miglioramento continuo del sistema.

Dai focus group sono emerse rappresentazioni della tecnologia e sentimenti verso l'IA ambivalenti. Da un lato, gli insegnanti riconoscono nella piattaforma uno strumento utile per alleggerire il carico burocratico, per organizzare lezioni in tempi rapidi – come nel caso di chi ha utilizzato il sistema per preparare, ad esempio, una lezione sul sistema solare – o per facilitare le traduzioni simultanee in classi plurilingui. Dall'altro, permangono timori rispetto alla possibilità che l'uso dell'IA possa appiattire la dimensione creativa dell'insegnamento, ridurre l'autonomia professionale e

trasformare l'atto educativo in una transazione standardizzata, come d'altra parte aveva già evidenziato Perrotta (2021).

Per rispondere a questi timori, il progetto ha operato lungo due assi fondamentali. Il primo è quello della co-costruzione, attraverso la partecipazione attiva dei docenti nella definizione di contenuti e modalità di interazione, così da ancorare lo strumento alle reali esigenze pedagogiche. Il secondo è quello della trasparenza, con la scelta di rendere esplicite le logiche che governano le risposte dell'IA e di garantire agli utenti la possibilità di interagire senza obbligo di cedere dati personali, in coerenza con il principio del diritto all'opacità (Glissant, 1990).

Fig.1 - Le tre fasi operative di Govern-AI EL



Una delle scelte più interessanti dal punto di vista metodologico riguarda l'impianto bifasico su cui si fonda la piattaforma. Fin dall'apertura del dialogo, l'utente viene invitato a dichiarare il proprio profilo – studente, docente o personale amministrativo – così da permettere al sistema di modulare linguaggio, contenuti e complessità delle informazioni (Tab.1). Da questo punto, il percorso si biforca in due direzioni: una modalità guidata, in cui vengono proposte attività strutturate e progressivamente più complesse, come simulazioni, quiz, esercizi di orientamento o analisi di casi didattici; e una modalità libera, dove l'utente può esprimere bisogni, porre domande aperte o condividere casi concreti, ricevendo risposte adattate al contesto e al ruolo.

Nei confronti degli studenti, il sistema offre strumenti di orientamento, supporto linguistico multilingue, spiegazioni semplificate e attività interattive, con un'attenzione particolare alla valorizzazione delle competenze pregresse e al riconoscimento dei saperi informali. Nel caso dei docenti, invece, il percorso guidato propone spazi di riflessione professionale, simulazioni di casi didattici e possibilità di sperimentare scenari di utilizzo dell'IA nell'insegnamento, specie in contesti multiculturali e multilingue. Il personale amministrativo, dal canto suo, trova nel sistema un supporto per navigare normative, procedure organizzative e strumenti comunicativi interni, oltre a poter sottoporre casi pratici per ricevere indicazioni contestuali.

L'infrastruttura non si limita a fornire risposte immediate, ma è stata concepita come un ambiente di apprendimento dinamico, capace di raccogliere e organizzare in forma anonima e aggregata i dati provenienti dalle interazioni, per individuare temi ricorrenti, bisogni formativi emergenti e aree di criticità. Questa logica di feedback continuo è parte integrante della natura sperimentale del progetto e rappresenta una delle sue principali innovazioni.

Tuttavia, anche in un contesto tanto attento, permangono alcuni nodi irrisolti del paradigma *do ut des*. L'accesso a servizi personalizzati comporta inevitabilmente la produzione e la cessione di dati a strutture esterne (il sistema infatti si basa su modelli IA già esistenti) alimentando quella tensione ineludibile tra personalizzazione e sorveglianza algoritmica. *Govern-AI EL* cerca di riequilibrare questo scambio attraverso pratiche di co-design e la costruzione di comunità educanti intorno alla tecnologia (Illich, 1973), ma la questione rimane aperta.

In definitiva, ciò che l'esperienza di Govern-AI Eda Lab dimostra è che un'IA davvero emancipativa non si gioca soltanto sul terreno tecnico, ma innanzitutto su quello organizzativo, culturale e politico. La sfida, soprattutto nei CPIA, è progettare dispositivi digitali che non si limitino a gestire dati o automatizzare processi, ma siano capaci di riconoscere la profondità delle biografie individuali e rispettare la complessità dei contesti educativi. È in

questo spazio, fragile e fertile, che l'IA può trasformarsi da tecnologia estrattiva a tecnologia conviviale, radicata nelle pratiche comunitarie e capace di potenziare, piuttosto che omologare, le traiettorie individuali.

Tab.1 - Nel chatbot Govern-AI EL: profili di scelta e funzionalità connesse

Profilo utente	*Funzionalità e contenuti specifici*
Studenti	Orientamento, supporto linguistico multilingue, spiegazioni semplificate, attività interattive, valorizzazione competenze pregresse e saperi informali
Docenti	Spazi di riflessione professionale, simulazioni di casi didattici, sperimentazione di scenari IA nell'insegnamento, con attenzione a contesti multiculturali e multilingue
Personale amministrativo	Supporto su normative, procedure organizzative, strumenti comunicativi interni, gestione di casi pratici con indicazioni contestuali

4. Adult education e IA: prospettive per un *do ut des* più equo

L'intelligenza artificiale, lo abbiamo visto, non è semplicemente una tecnologia in più fra molte altre: è una lente che ridefinisce ciò che conta come conoscenza, chi ha il diritto di insegnare e imparare, e secondo quali priorità. Nei contesti educativi, e soprattutto nei CPIA, questa trasformazione si carica di un peso specifico enorme, perché qui l'educazione non è mai solo trasmissione di saperi, ma una possibilità molto concreta di inclusione sociale, non solo nelle intenzioni del legislatore, ma anche da alcuni output empirici che ben documentano segmenti di vita durante e post percorsi CPIA (De Luca Picione, Madonia, 2018; Fortini, Trezza, 2021).

Se c'è un aspetto che emerge dal lavoro di revisione del paper, è che l'IA nel welfare educativo è una leva significativamente ambivalente. Può personalizzare i percorsi, sostenendo chi si occupa di didattica e abbattendo le barriere linguistiche per gli studenti stranieri. Tuttavia, e questo come abbiamo visto è un po' il rovescio della medaglia delle recentissime tecnologie generative, può anche semplificare eccessivamente la complessità delle biografie e fissare in schemi rigidi realtà sociali che invece chiedono di essere comprese nella loro ricchezza.

I rischi epistemologici non sono un'astrazione teorica: sono già presenti

nei sistemi che trasformano studenti in profili predittivi, spesso senza che docenti e comunità locali abbiano voce nel processo. È il paradosso di quello che qui abbiamo definito, forse anche semplificando, *do ut des*: gli utenti cedono dati, nella speranza di ricevere servizi più efficaci, ma in cambio ottengono spesso algoritmi opachi, che restituiscono soluzioni preconfezionate e poco aderenti ai contesti locali. Una contraddizione ancora più evidente se si pensa che molti di questi strumenti, pur operando in contesti educativi pubblici, sono progettati da aziende private, secondo logiche che non sempre perseguono i bisogni sociali o educativi del territorio (De Luca Picione, Trezza, 2024; Eubanks, 2018). Nei CPIA, dove ogni percorso educativo è anche un pezzo di biografia migrante, questo rischio non è marginale, anzi.

Al tempo stesso, non possiamo cedere a un *tecnopessimismo* sterile. Esperienze come quella di Govern-AI EL dimostrano, o almeno provano a farlo, che un'altra via è possibile. Quando i sistemi vengono progettati insieme agli operatori scolastici, quando la piattaforma è plasmata sulle specificità del territorio, quando le persone che ogni giorno vivono la scuola hanno la possibilità di dire la loro, allora l'IA smette di essere un'imposizione dall'alto e inizia a diventare uno strumento conviviale, per dirla con Illich: un mezzo che le comunità possono governare, negoziare, e adattare alle loro esigenze e a quelle dei loro contesti.

Se vogliamo che l'IA diventi una leva di giustizia sociale e non di nuove disuguaglianze, servono alcune condizioni irrinunciabili. Primo, il co-design, che significa progettare sistemi con, e non per, chi li userà. Secondo, la trasparenza, perché nessuna macchina dovrebbe decidere sul destino educativo delle persone senza poter spiegare come. Terzo, la sensibilità pedagogica e culturale, perché gli algoritmi non possono ignorare il valore del plurilinguismo, dei saperi informali e delle storie personali. Quarto, percorsi di formazione critica per docenti e studenti, affinché non siano spettatori passivi ma protagonisti consapevoli della rivoluzione digitale.

La partita che si gioca nei CPIA è emblematica di una questione più grande: il futuro del welfare educativo nell'era dell'IA. Non si tratta di respingere la tecnologia, ma di negoziarne le condizioni. Di preservare quella dimensione relazionale, umana e plurale dell'educazione che è, in fondo, la migliore garanzia di giustizia sociale ed educativa

Il nostro auspicio è che progetti come *Govern-AI EL* – nonostante lascino in sospeso diverse questioni – possano diventare non solo casi isolati, ma laboratori permanenti di riflessione e innovazione, capaci di dimostrare che sì, è possibile un'IA che non governa dall'alto, ma cresce dentro le comunità, ne rispetta la complessità e contribuisce, davvero, a liberare le potenzialità di chi, troppo spesso, è rimasto ai margini.

Riferimenti bibliografici

Airoldi M. (2021). *Machine habitus: Toward a sociology of algorithms*. Hoboken: John Wiley & Sons.

Aragona B. (2021). *Algorithm audit: Why, what, and how?* London: Routledge.

Boccia Artieri G. (2020). Fare sociologia attraverso l'algoritmo: potere, cultura e agency. *Sociologia italiana*, 15.

Benjamin R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Cambridge: Polity Press.

Bentley S.V. (2025). Knowing you know nothing in the age of generative AI. *Humanities and Social Sciences Communications*, 12: 409. DOI: 10.1057/s41599-025-04731-0.

Bourdieu P. (1979). *La distinction. Critique sociale du jugement*. Paris: Les Éditions de Minuit.

Cooper G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, 32(3): 444-452. DOI: 10.1007/s10956-023-10039-y.

Cosenza M., Giannini G., Pescapè A. (2023). L'IA tra tecnologia e filosofia. *Rivista di Digital Politics*, 3. DOI: 10.53227/113107.

De Luca Picione G.L., Madonia E. (2018). *L'istruzione degli adulti nei CPIA in Campania. Rapporto preliminare del Centro Regionale di Ricerca, Sperimentazione e Sviluppo*. Napoli: Guida Editore.

De Luca Picione G.L., Trezza D. (2024). Empowering educators with generative AI: The Govern-AI program for adult education governance. In *Conference Papers*. Napoli: Guida Editore: 5-16.

Espeland W.N., Stevens M.L. (2008). A sociology of quantification. *European Journal of Sociology / Archives Européennes de Sociologie*, 49(3): 401-436. DOI: 10.1017/S0003975608000202.

Eubanks V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

European Parliament, Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689. Disponibile online (consultato il 15 settembre 2025).

Felaco C., Amato F., Aragona B. (2024). Digital methods for social sciences. *Mathematical Population Studies*, 31(4): 237-241. DOI: 10.1080/08898480.2024.2414653.

Foley A., Melese F. (2025). Disabling AI: Power, exclusion, and disability. *British Journal of Sociology of Education*: 1-22. DOI: 10.1080/01425692.2025.2519482.

Fortini L., Trezza D. (2021). New profiles of adults in education. In *New Profiles of Adults in Education*. Roma: Associazione per la Scuola Democratica: 162-163.

Fortino G., Mangione F., Pupo F. (2024). Intersezione tra intelligenza artificiale generativa e educazione: un'ipotesi. *Journal of Educational, Cultural and Psychological Studies*, 30: 25-52. DOI: 10.7358/ecps-2024-030-fort.

Freire P. (1970). *Pedagogy of the oppressed*. New York: Continuum.

Fricker M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

García O., Wei L. (2014). *Translanguaging: Language, bilingualism and education*. London: Palgrave Macmillan.

Glissant É. (1990). *Poetics of relation*. Ann Arbor: University of Michigan Press.

Grimaldi R. (2022). *La società dei robot*. Milano: Mondadori.

Illich I. (1973). *Deschooling society*. New York: Harper & Row.

Kitchin R. (2014). *The data revolution*. London: SAGE.

Noddings N. (1984). *Caring: A feminine approach to ethics and moral education*. Berkeley: University of California Press.

Oberdieck T., Moch E. (2024). Lifelong learning with artificial intelligence: potentials, challenges and future perspectives. *International Journal of Advanced Research*: 545-555. Disponibile online (consultato il 15 settembre 2025).

O'Neil C. (2016). *Weapons of math destruction*. New York: Crown.

Pasquale F. (2015). *The black box society*. Cambridge (MA): Harvard University Press.

Perrotta C. (2021). Automated assessment and the erosion of educational feedback: A critical perspective. *Journal of Educational Technology*, 17(2): 45-60.

Selwyn N. (2019). *Should robots replace teachers? AI and the future of education*. Cambridge: Polity Press.

Selwyn N. (2022). *Education and technology: Key issues and debates*, 3ª ed. London: Bloomsbury Academic.

Tirocchi S. (2024). Digital education. Dalla scuola digitale all'intelligenza artificiale. *@DIGITCULT*, 8(2): 75-89.

Williamson B. (2017). *Big data in education*. London: SAGE.

Williamson B. (2019). Digital policy sociology: Software and science in data-intensive precision education. *Critical Studies in Education*, 62(3): 354-370. DOI: 10.1080/17508487.2019.1691030.

Yang Z. (2024). Empowering teaching and learning with artificial intelligence. *Frontiers of Digital Education*, 1(1): 1-3.

L'IA da strumento di contrasto a strumento di infiltrazione criminale: criticità e soluzioni auspicabili

di Roberta Aurilia[*]

I progressi dell'IA offrono strumenti utili alla prevenzione e al contrasto della criminalità organizzata, ma allo stesso tempo ampliano le opportunità criminali. In un quadro normativo europeo disomogeneo, IA e criptovalute sono sfruttate come strumenti di accesso al credito e come strumenti di riciclaggio, con reinvestimenti anche nei mercati legali, come quello dei crediti deteriorati e degli NPL immobiliari. L'accessibilità tecnologica ipertrofica e la debole regolamentazione favoriscono il cybercrime as a service, sollevando il dubbio se strumenti digitali e cooperazione internazionale possano essere risolutivi o nuovi vettori di rischio.

Parole chiave: intelligenza artificiale; polizia predittiva; organizzazioni criminali; npl immobiliari; cybercrime; riciclaggio.

AI from law enforcement tool to criminal infiltration tool: critical issues and desirable solutions

Advances in AI provide valuable tools for the prevention and countering of organized crime phenomena, while simultaneously expanding criminal opportunities. Within a fragmented European regulatory framework, AI and cryptocurrencies are exploited both as means of access to credit and as instruments for money laundering, with subsequent reinvestment in lawful markets, including those of non-performing loans and real estate NPLs. Hyper-accessibility to technology and weak regulation foster the phenomenon of cybercrime as a service, raising the question of whether digital tools and regulated international cooperation can be genuinely effective countermeasures or whether they risk becoming new vectors of criminal activity.

Keywords: artificial intelligence; predictive policing; criminal organizations; real estate NPLs; cybercrime; money laundering.

DOI: 10.5281/zenodo.18435675

[*] Università di Napoli Federico II. roberta.aurilia@unina.it.

Roberta Aurilia

Introduzione

Le nuove tecnologie e gli strumenti di intelligenza artificiale (IA) sempre più avanzati, negli ultimi, hanno aperto nuove frontiere e scenari nel campo del contrasto ma, soprattutto, della prevenzione dei fenomeni di criminalità organizzata.

Le teoriche della *new criminology* (Taylor, Walton *et al.*, 1973), infatti, non si stanno sviluppando solo nella direzione di integrare, secondo una strategia multilivello, elementi teorici specifici di singoli approcci per sintetizzare nuove teorie, bensì anche in quella della "simulazione", mediante l'utilizzo di nuove tecnologie, delle attività criminali in diverse situazioni al fine di prevederle e dunque prevenirle o, ancora, utilizzare la conoscenza digitale per costruire una metodologia di ricerca criminologica *digital* capace di produrre non solo sempre nuove modalità di raccolta e sistematizzazione dei dati ma anche di generare nuove conoscenze, funzionali agli strumenti di prevenzione e contrasto già esistenti (Yar, Steinmetz, 2023).

Ovviamente, la nuova frontiera dell'analisi criminologica muove dall'integrazione delle teoriche della *new criminology* con lo sviluppo dei sistemi di *machine learning* per contrastare le nuove forme di *cyber*-criminalità cui è correlato lo sviluppo della *cyber*-sicurezza.

Stante la peculiarità del *modus agendi* della criminalità organizzata, le strategie adottate sono mutevoli e in continua e rapida evoluzione per stare al passo con le strategie delle organizzazioni criminali, al fine di dare una risposta tempestiva al problema. Anche perché, com'è ormai noto, le organizzazioni criminali vantano due caratteristiche che rendono la loro operatività difficile da eliminare *in nuce*: da un lato la adattabilità (Varese, 2011), *i.e*. la capacità di apprendere e velocemente adattarsi al contesto – per mimetizzarsi ed entrarne a far parte – e al *modus operandi* utilizzato dalle Forze dell'Ordine; e, dall'altro, la capacità di anticipare non solo le strategie di contrasto ma, spesso, anche le strategie stesse del mercato. Basti pensare che, ad oggi, le grandi organizzazioni criminali hanno esternalizzato la criminalità predatoria e violenta per dedicarsi a forme di criminalità più "raffinata" che si consumano nel *cyber*-spazio, sulla falsariga della cd. *cyberwarfare* (Richet, 2015). Vieppiù. L'asimmetria normativa e negli approcci di prevenzione e contrasto al fenomeno criminale nella sua veste "*cyber*", permette alla criminalità organizzata di muoversi tra le maglie larghe non solo del diritto nazionale, ma anche di sfruttare le legislazioni dei Paesi con meno vincoli e formalità per poter operare indisturbate nel mondo del paralegale.

Ecco che, al netto dell'esperienza italiana e del dialogo europeo e internazionale, è stato da ultimo emanato l'AI Act, regolamento UE/2024/1689, che rappresenta il primo quadro giuridico sull'intelligenza artificiale, che

affronta i rischi dell'utilizzo dell'IA e i benefici di un suo uso regolamentato[1]. Si ricordano, inoltre, tra le più recenti fonti che regolano l'utilizzo dell'IA: la Convenzione Quadro del Consiglio d'Europa sull'intelligenza artificiale (CAI)[2], un trattato internazionale vincolante aperto alla firma nel 2024, incentrato sull'armonizzazione delle politiche nazionali per l'utilizzo responsabile dell'IA, riferendosi soprattutto alla tutela dei diritti umani, al rispetto dell'ordinamento democratico e dello Stato di diritto; la Direttiva GDPR del 2018[3], che tutela la privacy e la protezione dei dati personali all'interno dell'Unione e si riferisce, altresì, ai dati utilizzati dai sistemi di IA per finalità di polizia predittiva nel rispetto dei principi di minimizzazione, trasparenza, legittimità e diritto di accesso da parte dei cittadini, oltre alla necessità di prevedere un consenso valido per il trattamento; e la raccomandazione CM/Rec(2020)1 del Consiglio d'Europa sull'intelligenza artificiale e i Diritti Umani[4], che fornisce linee guida agli Stati membri su come integrare i diritti umani e i principi etici nella progettazione e applicazione delle tecnologie di IA.

Elementi comuni alla regolamentazione dell'IA, a prescindere da quale sia la fonte, sono: la valutazione d'impatto, la trasparenza delle informazioni, la supervisione indipendente, il rispetto della privacy e, soprattutto, il necessario controllo umano, a conferma che le scelte di utilizzare la tecnologia di IA non sono volte all'automatizzazione delle decisioni bensì a fornire un ausilio "*digital*" agli operatori.

1. Le nuove tecnologie come armi *utilizzate dalla* criminalità organizzata

Così come le potenzialità dell'IA possono essere utilizzate in maniera virtuosa non solo per migliorare la vita di tutti i giorni ma anche come strumento di contrasto alla criminalità, così sono gli stessi criminali che utilizzano la tecnologia come nuova opportunità per infiltrare nuovi tessuti economici e sociali. Il connubio criminalità organizzata-intelligenza artificiale, infatti, è più forte di quanto non si possa immaginare e coinvolge le più disparate attività illecite, dalla tratta di esseri umani agli abusi sessuali, dal *ransomware*

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
[2] https://rm.coe.int/CoERMPublicCommonSearchServices/documentAccessError.jsp?url=https://rm.coe.int:443/CoERMPublicCommonSearchServices/sso/SSODisplayDCTMContent?documentId=0900001680a33b01
[3] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679
[4]https://search.coe.int/cm#{%22CoEIdentifier%22:[%2209000016809ee581%22],%22sort%22:[%22CoEValidationDate%20Descending%22]}

alle frodi informatiche, alla consumazione di reati finanziari (Velasco, Periche *et al.*, 2024).

La componente che genera maggiore allarme è l'evoluzione del fenomeno che sta garantendo una sempre maggiore "emancipazione" e indipendenza nell'operatività dei criminali. Infatti, se prima la criminalità organizzata si avvaleva dei cd. colletti bianchi, cioè di quella zona grigia della società che collaborava con l'organizzazione ma non era parte integrante del gruppo, per poi diventare gli stessi membri del gruppo i "tecnici" della malavita (Di Gennaro, La Spina, 2010), oggi l'utilizzo dell'IA, estremamente semplice ed intuitivo, ha messo tale strumento (potenzialmente dalle possibilità illimitate) alla mercè di chiunque, non essendo necessarie particolari competenze tecniche per il suo utilizzo. L'automazione, infatti, permette da un lato di ampliare il raggio di operatività delle organizzazioni criminali e, dall'altro, riduce la probabilità di essere individuati.

L'"originalità" della criminalità organizzata nell'utilizzo di qualsiasi strumento a diposizione per delinquere è, ormai, ben nota. E lo stesso vale per l'utilizzo dell'IA nelle modalità più disparate e impensabili. A titolo esemplificativo e non esaustivo, si faranno di seguito alcuni esempi di come alcune tecnologie di IA, se usate in modo distorto, possono essere utilizzate per delinque (Wall, 2003).

1. Il *phishing* e truffe *AI-driven*: Così come questa forma di messaggistica veniva utilizzata, prima di essere sostituita da forme più sofisticate, per l'inserimento di *trojan* sui dispositivi dei sospettati, così tramite l'utilizzo di messaggi realistici e personalizzati, gli utenti vengono convinti a cedere chiavi private e password di accesso ai propri wallet, favorendo il furto di criptovalute e successiva immissione nei circuiti illeciti (Almeida, Pinto *et al*., 2023)[5] ovvero mediante *spoofing,* cioè quella tecnica di attacco informatico che consiste nel mascherare la propria identità digitale (indirizzo IP, email, numero di telefono) per fingersi un'altra persona o entità fidata (come una banca, un sito noto) al fine di ingannare la vittima, rubare dati sensibili.
2. I *deepfake* e la manipolazione audiovisiva: Così come i *deepfake* posso essere utilizzati per scopi virtuosi[6], dall'istruzione alla sanità, così permettono di creare audio e video falsi ma estremamente reali e veritieri, per consumare i crimini più disparati, tra cui, *ex multis*,

[5] Sul *phishing*, cfr. https://www.proofpoint.com/it/newsroom/press-releases/phishing-italia-nel-2022-il-79-delle-aziende-ha-subito-almeno-un-attacco Report State of Phishing 2022.

[6] Sull'utilizzo virtuoso dei *deepfake*, cfr. https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positive-for-humanity/

estorsioni, truffe, disinformazione, proselitismo, orientamento elettorale[7] o compromettere noti personaggi istituzionali[8].

3. L'automazione del *cybercrime*: Grazie alle tecnologie di IA è possibile automatizzare gli attacchi informatici (*i.e.* cracking di password, rilevamento dei *vulnera* di sistema) e creare *malware* adattivi che sfuggono ai sistemi di difesa in quanto si evolvono e adattano a loro (così come, in precedenza, facevano i *trojan*)[9].
4. Sorveglianza e contro-sorveglianza: così come le forze dell'ordine possono utilizzare sistemi di IA per controllare le attività criminali, così la criminalità organizzata può usare i medesimi strumenti per monitorare le attività delle forze dell'ordine usando, nello stesso modo, sistemi di riconoscimento facciale e droni. Ancora, possono adottare tecniche di crittografia tramite IA al fine di eludere le intercettazioni.
5. Traffico della droga e gestione delle reti criminali: così come la polizia predittiva viene utilizzata per anticipare le "mosse" della criminalità organizzata, così quest'ultima, utilizzando le medesime tecnologie, può prevedere i movimenti delle FFOO e regolarsi di conseguenza in termini di ottimizzazione della logistica del traffico illecito, coordinamento delle operazioni, migliorando efficienza e operatività delle reti criminali, scongiurando altresì il pericolo degli arresti.
6. Frode finanziaria e riciclaggio di denaro: mediante algoritmi di apprendimento automatico, è possibile indentificare i *vulnera* nei sistemi di controllo bancario o nei circuiti finanziari, generando sistemi di transazione complessa e difficile da tracciare. Inoltre, tramite questi sistemi è possibile celare i flussi di denaro illecito, facilitando il riciclaggio[10].

Si evince, dunque, che il paradosso è che così come le tecniche di *machine learning* vengono utilizzate dalle forze di polizia per analizzare i profili dei criminali e ideare strategie di prevenzione e contrasto, così la criminalità

[7] Sull'utilizzo dei *deepfakes* durante le elezioni, cfr. https://www.wsj.com/tech/ai/new-era-of-ai-deepfakes-complicates-2024-elections-aa529b9e.

[8] Sull'utilizzo dei *deepfakes* in azienda, cfr. https://www.wsj.com/articles/deepfakes-are-coming-for-the-financial-sector-0c72d1e5.

[9] Sul ransomware "REvil" che ha utilizzato tecniche AI per automatizzare la scansione delle vulnerabilità nelle reti aziendali, aumentando così la rapidità e l'efficacia degli attacchi, cfr. https://www.ibm.com/thought-leadership/institute-business-value/report/2025-threat-intelligence-index.

[10] Sulle nuove tecnologie nel riciclaggio di denaro cfr. https://www.fatf-gafi.org/content/dam/fatf-gafi/annual-reports/FATF-AR-2023-2024.pdf.coredownload.pdf

organizzata utilizza gli stessi strumenti per profilare le vittime e ottimizzare i processi per infiltrarsi nei mercati legali, aumentando non solo il successo delle proprie operazioni ma facendo diminuire sensibilmente il rischio della loro individuazione.

2. *Case studies*: evidenze empiriche sull'evoluzione degli attacchi Deepfake ai sistemi bancari

L'analisi empirica degli attacchi deepfake nel settore finanziario rivela un fenomeno non più episodico, bensì sistemico e pervasivo. I principali *case studies* analizzati – relativi all'Indonesia, alle tendenze regionali nell'area Asia-Pacifico e alla crescente offerta criminale di strumenti "Deepfake-as-a-Service" – costituiscono un corpus di evidenze che permette di osservare come i deepfake stiano trasformando radicalmente la sicurezza dei sistemi biometrici.

2.1. Il caso indonesiano: il punto di svolta nella compromissione dei sistemi KYC biometrici

L'incidente verificatosi presso un'importante istituzione bancaria indonesiana nell'agosto 2024 rappresenta uno dei casi più significativi nella recente letteratura sulla sicurezza biometrica (Borak, 2025). Ciò che distingue questo episodio da altri tentativi documentati non è solo la scala dell'attacco – oltre 1.100 tentativi di *spoofing* – ma l'efficacia con cui gli aggressori hanno aggirato un sistema KYC (*Know your Customer*) che, sulla carta, integrava tecniche di riconoscimento facciale e *liveness detection* multilivello.

I cybercriminali, in questo caso, hanno ottenuto documenti reali attraverso l'utilizzo di di *malware*, furto di dati e forum del *dark web*. Tali documenti sono stati successivamente manipolati utilizzando *pipeline* generative in grado di produrre volti sintetici quasi indistinguibili dagli originali. Questo processo ha permesso di addestrare modelli *deepfake* ad alta coerenza spaziale e temporale, ottimizzati per ingannare i moduli *di face-matching* dell'istituzione bancaria.

Il risultato operativo è stato drammatico: oltre 1.000 account fraudolenti sono stati creati con successo, implicando almeno 45 dispositivi distinti e generando perdite stimate in 138,5 milioni di dollari (Huang, 2025).

Oltre al danno economico, il caso rivela una falla concettuale nei sistemi biometrici contemporanei: la loro assunzione radica e ormai implicita secondo cui il "canale video" sia intrinsecamente affidabile. La presenza

crescente di software di videocamere virtuali, impiegate per far confluire direttamente il deepfake nel flusso video, infrange questa assunzione alla radice (Deloitte, 2024).

3. Le criptovalute come strumento per delinquere: il riciclaggio tramite gli NPL immobiliari

Le organizzazioni criminali utilizzano algoritmi di IA per automatizzare e meglio gestire disparate micro-transazioni su differenti piattaforme di scambio (*i.e. exchange*) e *wallet*, disperdendo e frammentando, così, i fondi illeciti in piccole somme per renderne difficile la tracciabilità[11]. È ben possibile, infatti, che un sistema di IA possa decidere in autonomia come convertire i fondi e dove spostarli, in base a delle informazioni preselezionate, magari inserite in riferimento alla legislazione – più o meno permissiva – del Paese di provenienza o atterraggio dell'investimento, utilizzando tecniche come lo *smurfung*, cioè l'utilizzo di micro-trasferimenti multipli e combinati, ovvero il *layering*, cioè diversi passaggi tra *wallet*, al fine di nascondere la provenienza (illegale) del fondo o *asset.*

Da ultimo, tra le nuove modalità di infiltrazione nell'economia legale tramite l'utilizzo di IA, si annovera senza dubbio l'utilizzo criptovalute come strumento di riciclaggio dei *Non-Performing Loans* (NPL) (Passador, 2021) e, in particolare, quelli immobiliari.

Gli NPL o crediti deteriorati sono quei crediti "a rischio di perdita" che le banche o altri istituti finanziari non riescono a riscuotere o che, comunque, hanno una probabilità molto bassa di essere onorati e che, pertanto, vengono venduti a soggetti interessati ad acquistare tali *asset* a rischio, con la convenienza dell'acquisto ad un prezzo minore rispetto al loro valore nominale e conseguente possibilità di rivendita con elevati margini di guadagno.

Più nello specifico, gli NPL immobiliari sono quei crediti deteriorati legati a mutui o prestiti garantiti da immobili. Questi crediti vengono spesso

[11] Tra il 2022 e il 2023, si ricordano il caso *Chainalysis* e l'Operazione *GhostNet*. Nel primo caso, una società leader nell'analisi di blockchain, ha documentato l'utilizzo dell'IA per la movimentazione di criptovalute tra wallet ed exchange, sfruttando mixer avanzati per "ripulire" miliardi di dollari in fondi illeciti, cfr. https://go.chainalysis.com/crypto-crime-report-italian-sign-up.html?blaid=4774374; nel secondo caso, le autorità europee hanno scoperto un network criminale che utilizzava AI per la gestione di migliaia di transazioni in criptovalute in modo automatico, riciclando così milioni di euro derivanti da frodi consumate online, cfr. https://www.europol.europa.eu/media-press/newsroom/news/global-coalition-takes-down-new-criminal-communication-platform.

venduti a fondi di investimento o società specializzate ad un prezzo minore rispetto al loro valore nominale, le *Special Purpose Vehicle* (SPV) che, tramite vendita e/o ristrutturazione, tentano di (*rectius*, riescono a) recuperarne il valore.

Le organizzazioni criminali che, notoriamente, sono quelle che hanno a disposizione ingenti quantità di capitale illecito – liquido e (anche) sotto-forma criptovalute – da "ripulire", hanno interesse a convertire questi fondi in *asset* immobiliari. Infatti, il mercato degli NPL immobiliari è utile alle organizzazioni criminali almeno per tre ordini di motivi: (a) permette di acquistare beni a un prezzo inferiore al loro valore nominale o di mercato tramite aste giudiziarie (Di Gennaro, Pastore, 2022) o compravendite pilotate; (b) permette di ripulire denaro di provenienza illecita investendolo in *asset* apparentemente legali; (c) permette di reintrodurre fondi illeciti in circuiti legali attraverso operazioni immobiliari.

Gli NPL immobiliari, proprio perché vengono venduti a prezzi inferiori rispetto al valore di mercato – in quanto, per loro natura, sono difficili da recuperare – sono un'opportunità per acquisire immobili con denaro "criptato", mascherandone così alla provenienza.

Il *modus operandi* è il seguente: il capitale illecito in criptovalute viene convertito in una moneta fiduciaria a corso legale (fiat), o tramite servizi *over the counter* (OTC) oppure in strumenti finanziari tramite *exchange* ottenendo, così, un capitale "ripulito" che, successivamente, viene utilizzato per l'acquisto di NPL immobiliari. La vendita e/o ristrutturazione degli immobili associati agli NPL genera redditi, in apparenza leciti, chiudendo così il cerchio del processo di riciclaggio (Lemme, 2024).

In tale sistema, l'IA viene utilizzata da parte delle organizzazioni criminale per: analizzare i portafogli NPL al fine di indentificare gli *asset* immobiliari più redditizi e vulnerabili; simulare e ottimizzare le operazioni di riciclaggio, prevedendone i rischi e le possibili segnalazioni; automatizzare la creazione di reti di società fittizie e movimenti finanziari, creando reticoli che opacizzano la tracciabilità delle operazioni (Balaji, 2024); manipolare le aste online o le valutazioni immobiliari, inquinando il mercato lecito delle aste giudiziarie (Aurilia, Di Gennaro, 2023).

Dal Report Europol 2023, è emerso che le SPV che acquistano NPL immobiliari e immobili utilizzando fondi cripto-valutari vengono create in giurisdizioni *offshore*, così da rendere ancora più opaca la provenienza del denaro, causa impossibilità di risalire ai reali beneficiari. E, di recente, il Rapporto FATF (2025) ha evidenziato come il settore immobiliare sia diventato uno dei principali veicoli di riciclaggio di criptovalute, soprattutto mediante l'acquisto di proprietà legate a NPL e ristrutturazioni immobiliari.

Tale sistema è facilitato da due aspetti: il primo è che per loro natura gli NPL immobiliari sono asset difficili da monitorare, in quanto spesso sono connotati da transazioni complesse e poco trasparenti; il secondo riguarda l'assenza di regolamentazioni uniformi tra i Paesi rispetto alla tracciabilità delle criptovalute, rendendo più difficile l'identificazione dei flussi finanziari illeciti nel settore immobiliare (Donato, 2017).

4. Le nuove tecnologie come armi *contro* la criminalità organizzata

Dunque, al netto di quanto fin qui riportato, quale può essere uno strumento efficace per rispondere all'uso criminale dell'IA se non proprio l'utilizzo virtuoso di tecnologie IA per la prevenzione e il contrasto dei crimini?

Sicuramente i campi di applicazione dell'IA, nel corso degli ultimi anni, sono aumentati e si sono differenziati in maniera esponenziale. Tuttavia, se è vero che l'utilizzo dell'IA può avere diversi vantaggi (Holt, Bossler *et al.*, 2017), è altrettanto vero che il suo utilizzo comporta una serie di meccanismi decisionali a volte opachi e rischi potenziali di intrusione nella quotidianità e nella sfera privata di ognuno. Basti pensare all'impiego dell'IA per finalità di polizia predittiva, nel contrastare i fenomeni di criminalità organizzata, per analizzare grandi quantità di dati in poco tempo e, in base ad algoritmi che lavorano su quanto processato, prevedere dove è più probabile che si verifichi un crimine consentendo, così, alle FFOO non solo di anticipare la consumazione del reato ma anche di ottimizzare l'allocazione delle risorse. Gli esempi virtuosi, sia in letteratura sia sperimentati sul campo, con risultati spesso positivi, sono numerosi e hanno permesso di stabilire i criteri per arrivare allo sviluppo di modelli previsionali del crimine attraverso tecniche di *machine learning*. Si ricordano, ad esempio, l'esperienza italiana di *X-Law* (Lombardo, 2019) e quella del modello predittivo del reato di estorsione (Di Gennaro (*a cura di*), 2023). Basandosi, dunque, su dati "storici" di criminalità, lo studio del territorio e il *modus agendi* delle organizzazioni (e non solo) che insistono in quelle zone, è possibile identificare aree ad alto rischio e migliorarne il pattugliamento con finalità di prevenzione. In tal modo, il costrutto strategico dell'azione di controllo passa da una visione riparatoria del danno ad una visione probabilistica del rischio, *i.e.* da una logica emergenziale di gestione di un problema ad una che lavora nell'ottica della prevenzione.

Ancora, si ricorda come *Cloudflare* ha recentemente neutralizzato quello che è considerato il più imponente attacco DDoS (*Distributed* Denial of Service) mai registrato (Mocerino, 2025). L'evento, neutralizzato in soli 45 secondi, dimostra come gli attacchi volumetrici stiano raggiungendo intensità

senza precedenti, rendendo indispensabile l'impiego di sistemi autonomi ad alta reattività. L'architettura di *Cloudflare* integra algoritmi di machine learning capaci di analizzare in tempo reale pattern di traffico, identificare deviazioni comportamentali e distinguere flussi legittimi da traffico malevolo anche a velocità terabitiche. La natura distribuita della rete globale di data center consente di assorbire e frammentare il carico, riducendo drasticamente il rischio di saturazione dei punti critici dell'infrastruttura. L'episodio *Cloudflare* conferma dunque che l'IA non è solo un supporto alla difesa, ma un elemento strutturale indispensabile per la protezione di reti e servizi critici nell'era del cybercrime ad alta intensità.

Tuttavia, nonostante le esperienze positive non solo nazionali ma anche europee[12] e internazionali[13], il bilanciamento tra l'utilizzo di tecnologie predittive e la tutela dei diritti umani rimane una questione etica spinosa, soprattutto in termini di discriminazione e stigmatizzazione di determinate comunità che endemicamente risultano attenzionate in quanto particolarmente vulnerabili. Inoltre, esistono problematiche operative che riguardano il numero di agenti disponibili sul territorio che dovrebbero pattugliare una zona "a rischio", lasciando scoperte altre zone dove, potenzialmente, potrebbero consumarsi dei delitti, e dunque involge il macrotema dell'allocazione delle risorse umane disponibili.

Ad oggi, secondo Europol (2023), il divario tecnologico tra le organizzazioni criminale e la polizia è "*forse la più grande sfida che le forze*

[12] Il riferimento è, *ex multis,* a diversi esperimenti di polizia ancora in fase embrionale e sperimentale.: il Sistema HART (*Harm Assessment Risk Tool*) utilizzato nel Regno Unito, che è uno strumento di polizia predittiva per valutare il rischio che un sospettato commetta un futuro, utilizzato soprattutto nelle decisioni relative alla custodia cautelare o ai programmi di riabilitazione; *Precrime*, progetto pilota utilizzato a Rotterdam che individua aree e persone potenzialmente a rischio (recidiva o atti violenti) di coinvolgimento in attività criminali; i sistemi di polizia predittivi parigini SAPE (*Système d'Analyse des Prèdictions d'èvènements*) che hanno l'obiettivo di prevenire reati di furto e vandalismo; il progetto *DeepLapd* a Milano che utilizza tecniche di intelligenza artificiale per l'analisi di flussi di telecamere urbane e dati sociali per migliorare la sicurezza urbana; il progetto *PredPol Europe*, utilizzato in Germania sulla falsa riga del software predittivo *PredPol USA*, per la prevenzione dei reati contro il patrimonio. Anche a livello di UE è stato finanziato il progetto COPKIT, che, tramite un sistema innovativo di allerta precoce (EW)/azione precoce (EA), sarebbe in grado di migliorare l'efficienza delle indagini sui reati che coinvolgono l'uso criminale delle nuove tecnologie.

[13] Il riferimento è alle esperienze tra cui, *inter alia*: 1. PredPol (*Predective Policing*) negli USA che, utilizzando dati storici sui crimini (tipologia, luogo e ora), sono in grado di prevedere dove è più probabile che si verifichino reati in futuro, ai fini di ottimizzare il pattugliamento; 2. COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), algoritmo utilizzato negli USA per valutare il rischio di recidiva dei criminali, ai fini delle decisioni in merito al rilascio su cauzione o in libertà vigilata; Operation Laser, della Polizia di Los Angeles che, combinando dati storici di crimini e rete sociale, indivisuava soggetti a rischio di commettere/subire crimini violenti, a scopo preventivo, *etc*.

dell'ordine devono affrontare in tutto il mondo". Le organizzazioni criminali, infatti, stanno adottando nuove tecnologie dell'informazione e della comunicazione.

Tuttavia, la componente umana resta fondamentale. Infatti, i sistemi di IA, nella visione antropocentrica[14], restano uno strumento che agevola il lavoro degli esperti e/o analisti senza però sostituirsi a loro. Infatti, sono questi ultimi che, sulla base dei dati processati dall'IA potranno analizzare in via prospettica e prognostica come determinati reati stanno evolvendo, evidenziare le nuove tendenze nel *modus operandi* e, se del caso, individuare eventuali punti deboli sui quali poi andrà ad agire il sistema di prevenzione e/o contrasto, costruendo degli *alert* precoci tali da anticipare la consumazione del reato (Alakayleh, 2025). L'IA come strumento di ausilio durante le indagini permette, invero, la rilevazione di quei segnali "deboli" che, "*ictu oculi*" spesso sono di difficile percezione.

5. Criticità e soluzioni auspicabili

I sistemi di nuova generazione utilizzano l'IA per colmare una grande lacuna che grava ormai sul sistema investigativo italiano ed europeo, cioè l'isolamento delle informazioni e la difficoltà non solo nella loro condivisione ma anche nella loro sintesi e comunicazione, sia a livello micro tra omologhi, sia a livello macro tra agenzie, favorendo una base dati condivisa tra tutte le Forze dell'Ordine europee, migliorando rapidità, efficacia e cooperazione nelle strategie di contrasto alla criminalità.

In quest'ottica, infatti, negli ultimi anni si è intensificato l'impegno internazionale per uno sviluppo etico e responsabile dell'IA. Per garantire trasparenza, sicurezza e rispetto dei diritti umani, facendo sì che l'intelligenza artificiale sia uno strumento al servizio dell'uomo, l'AI Act (2024) rappresenta un importante passo avanti, ma rischia di limitare l'efficacia operativa delle FFOO se non bilanciato con eccezioni mirate a determinati diritti fondamentali in una data fase delle indagini, ad esempio anche derogando alla privacy dei sospettati.

La cooperazione internazionale, soprattutto in ambito giuridico, sta evolvendo grazie a progetti dell'UE che integrano IA e analisi avanzata,

[14] Trattasi di un approccio all'intelligenza artificiale che pone l'essere umano al centro, garantendo che l'IA sia a servizio delle persone e tuteli i loro diritti individuali, non solo in via strumentale ma anche in termini di benessere umano. In particolare, garantendo affidabilità e sicurezza, favorendo l'inclusione e la sostenibilità, come strumento al servizio della società.

superando barriere normative e culturali. Tuttavia, il settore della giustizia resta indietro su trasparenza, equità e responsabilità nell'uso dell'IA[15].

Un punto critico è rappresentato dall'uso distorto dell'IA da parte della criminalità organizzata, che sfrutta strumenti avanzati, parallelamente all'utilizzo virtuoso che ne fanno le Forze dell'Ordine, come riconoscimento facciale, *deepfake*, analisi predittiva e droni autonomi. Il crescente impiego dell'IA per attività criminali come traffico di droga, tratta di esseri umani, frodi finanziarie e altri reati gravi non deve essere interpretato esclusivamente come un problema tecnico da risolvere con strumenti di sicurezza più sofisticati. Piuttosto, esso rappresenta l'inevitabile conseguenza di un modello di sviluppo che privilegia la rapidità dell'innovazione rispetto alla sostenibilità etica, sociale e normativa. La democratizzazione dell'accesso alle tecnologie di IA, pur essendo un fattore positivo per l'inclusione e lo sviluppo, ha abbattuto le barriere tecniche che in passato limitavano l'adozione di strumenti complessi da parte degli attori criminali che erano costretti ad avvalersi dei colletti bianchi o, comunque, di soggetti esterni all'organizzazione.

È fondamentale evitare che la risposta istituzionale si limiti alla corsa agli armamenti tecnologici. Serve, piuttosto, un ripensamento strutturale che affronti, *inter alia*, le fragilità istituzionali, una produzione legislativa elefantiaca e non armonizzata (a livello nazionale ed europeo), le disuguaglianze educative e la mancanza di una *governance* resiliente. Infatti, la criminalità organizzata sfrutta non solo le potenzialità tecnologiche dell'AI ma anche i vuoti lasciati da istituzioni statali inefficaci o corrotte, oltre alle maglie larghe del diritto, consolidando così il proprio controllo sociale ed economico.

L'avvento di tecnologie sempre più complesse aggiunge un ulteriore livello di rischio, minacciando la stabilità del tessuto sociale basato sulla fiducia reciproca. Gli effetti di queste tecnologie non si limitano alle vittime dirette di frodi o attacchi informatici, ma si estendono alla destabilizzazione dei mercati finanziari, all'erosione della fiducia nelle Istituzioni pubbliche e alla percezione diffusa di insicurezza. Questo scenario pone un dilemma etico di grande rilievo: la necessità di ridefinire il concetto di "*accountability* condivisa" tra tutti gli attori coinvolti.

L'adozione dell'IA non deve generare una "guerra tecnologica" tra criminali e Istituzioni, ma offrire l'occasione per costruire un nuovo modello di sicurezza omnicomprensiva e collaborativa, fondato su etica, diritti e cooperazione internazionale.

[15] Basti pensare che nell'AI Act, tra le tecnologie *AI high risk* rientrano quelle utilizzate in ambito giudiziario e investigativo e, pertanto, sono soggette a controlli stringenti, obblighi di trasparenza, valutazioni di impatto e monitoraggio costante.

Riferimenti bibliografici

Alakayleh O. (2025). The use of artificial intelligence systems in crime detection and prevention: applications and challenges. New York: SSRN.

Almeida H., Pinto P., Fernández Vilas A. (2023). A review on cryptocurrency transaction methods for money laundering. In: *FEMIB 2023*, pp. 114-121.

Aurilia R., Di Gennaro G. (2023). Un Mezzogiorno ancora imbrigliato nella morsa delle estorsioni. Un'attività che va radicandosi anche al Nord. *Rivista giuridica del Mezzogiorno*, 2: 395-426.

Borak M. (2025). 2025 deepfake threat predictions from biometrics, cybersecurity insiders. In: *State of Biometrics Report*. https://www.biometricupdate.com/202501/2025-deepfake-threat-predictions-from-biometrics-cybersecurity-insiders

Deloitte Center for Financial Services (2024). Generative AI is expected to magnify the risk of deepfakes and other fraud in banking. https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html

Di Gennaro G., a cura di (2023). *Il potere delle estorsioni. Un modello predittivo come strategia di contrasto*. Napoli: Editoriale Scientifica.

Di Gennaro G., La Spina A., a cura di (2010). *I costi dell'illegalità: Camorra ed estorsioni in Campania*. Bologna: Il Mulino.

Di Gennaro G., Pastore G. (2022). Aste giudiziarie: effetti economici e sociali. Un approccio non apodittico. *Rivista giuridica del Mezzogiorno*, 2: 483-498.

Donato L. (2017). La vulnerabilità dei mercati immobiliari ai rischi di riciclaggio. *MonitorImmobiliare*. https://www.monitorimmobiliare.it/monitorimmobiliare/notizia/la-vulnerabilita-dei-mercati-immobiliari-ai-rischi-di-riciclaggio_2017122915/

Europol (2023). *Criminal Asset Recovery in the European Union* (CAAR 2023). https://www.europarl.europa.eu/cmsdata/286518/Europol%20CAAR%202023.pdf

Holt T.J., Bossler A.M., Seigfried-Spellar K.C. (2017). *Cybercrime and Digital Forensics: An Introduction* (2ª ed.). London: Routledge.

Huang Y. (2025). Deepfake fraud: AI's impact on financial institutions. *Frontier Enterprise*. https://www.frontier-enterprise.com/deepfake-fraud-ais-impact-on-financial-institutions/

Lemme G. (2024). Criptovalute e riciclaggio: un rapporto "troppo facile". *Dialoghi di diritto dell'economia*: 1-12.

Lombardo E. (2019). *Sicurezza 4P. Lo studio alla base del software XLAW per prevedere e prevenire i crimini*. Venezia: Mazzanti Libri.

Mocerino G. (2025). Cloudflare? Evitare che il prossimo incidente sia crisi sistemica. *CybersecurityItalia*. https://www.cybersecitalia.it/cloudflare-evitare-che-il-prossimo-incidente-sia-crisi-sistemica/54944/

Passador M.L. (2021). Le nuove frontiere del riciclaggio e il ruolo dell'innovazione tecnologica. *Diritto del commercio internazionale*, 3: 611-636.

Rapporto FATF (2025). https://crystalintelligence.com/crypto-regulations/fatf-2023-24-report-crypto-compliance-risks-gaps/

Richet J.L., a cura di (2015). *Cybersecurity Policies and Strategies for Cyberwarfare Prevention*. Pennsylvania: IGI Global.

Taylor I., Walton P., Young J. (1973). *The New Criminology: For a Social Theory of Deviance*. London: Routledge.

Varese F. (2011). *Mafias on the Move: How Organized Crime Conquers New Territories*. Princeton (NJ): Princeton University Press.

Velasco C., Periche J.G., De Dios Gómez J., Bueno Benedí M. (2024). *Artificial Intelligence and Organised Crime*. EL PACTO 2.0 EU-LAC, European Commission. https://www.fiap.gob.es/wp-content/uploads/2024/11/ELPACCTO2-IAyCrimen-EN.pdf

Wall D.S. (2003). *Cyberspace Crime.* Aldershot: Ashgate Publishing.

Yar M., Steinmetz K.F. (2023). *Cybercrime and Society* (4ª ed.). Thousand Oaks (CA): SAGE Publications Ltd.

L'Intelligenza Artificiale per la Cybersecurity: opportunità e sfide nella sicurezza digitale

di Franco Campitelli*

L'evoluzione dell'Intelligenza Artificiale sta modificando rapidamente il panorama della cybersicurezza. In questo saggio l'autore intende esaminare potenzialità e criticità del rapporto tra IA e *cybersecurity.* Se da un lato le minacce diventano sempre più sofisticate, dall'altro si stanno studiando strumenti in grado di contrastarle quali, ad esempio, l'analisi predittiva, l'identificazione di schemi di attacco in tempo reale e l'automatizzazione delle risposte ad incidenti di sicurezza. Ulteriori aspetti da considerare nell'utilizzo dell'IA sono la privacy, la protezione dei dati personali e i bias algoritmici. Infine, si analizzeranno idee per garantire i diritti della cittadinanza e minimizzare i rischi quali adozione di meccanismi di supervisione e tecniche di "*Differential Privacy*". In definitiva, il successo dell'IA nella cybersicurezza dipenderà dalla capacità di tutti gli *stakeholders* di assicurare la protezione dei sistemi con un'attenzione particolare alla tutela della privacy e dei diritti fondamentali.

Parole chiave: intelligenza artificiale; cybersecurity; minacce; privacy; bias algoritmici.

Artificial intelligence for cybersecurity: opportunities and challenges in digital security

The rapid evolution of artificial intelligence is changing the cybersecurity landscape. In this essay, the author examines the potential and critical issues surrounding the relationship between AI and cybersecurity. As threats become more sophisticated, tools are being developed to counter them, including predictive analysis, real-time identification of attack patterns, and automated responses to security incidents. Other considerations when using AI include privacy, personal data protection and algorithmic bias. Finally, the essay will analyze ideas for guaranteeing citizens' rights and minimizing risks, such as the adoption of oversight mechanisms and 'differential privacy' techniques. Ultimately, the success of AI in cybersecurity depends on all stakeholders' ability to protect systems, with a particular focus on privacy and fundamental rights.

Keywords: artificial intelligence; cybersecurity; threats; privacy; algorithmic bias.

DOI: 10.5281/zenodo.18435761

* Università degli Studi di Teramo, fcampitelli@unite.it

Franco Campitelli

Introduzione

Oggi viviamo in un'epoca di enorme trasformazione digitale in cui i sistemi informatici, i dati e le persone sono sempre più interconnessi tra loro. Se da un lato questa interconnessione ha reso più rapide le comunicazioni e ha migliorato significativamente l'efficienza sul lavoro, la stessa ha fornito un enorme valore ai "dati digitali" che transitano in rete. Quando si acquista online, si fa una ricerca, si naviga all'interno dei siti tutti i nostri spostamenti online sono tracciati dai cosiddetti "*cookies*". È evidente l'importanza che questo piccolo file riveste per un'azienda; infatti, in tempo reale è possibile studiare il comportamento online di milioni di persone e valutarne i gusti e le tendenze da utilizzare in campagne pubblicitarie mirate. I *social* in questo senso sono diventati dei "forzieri" di informazioni tanto che Floridi ha coniato il cosiddetto termine "infosfera"[2] per descrivere il mondo digitale in cui siamo immersi. Valutata l'importanza che il dato oggi riveste nel mondo digitale diventa naturale per i cosiddetti "*black hat hacker*" cercare di impossessarsene per rivenderli al miglior offerente. In questo scenario entra in maniera dirompente l'Intelligenza Artificiale che ha permesso di velocizzare molte operazioni comprese quelle di *coding* sia benevole che malevole, pertanto, le difese tradizionali utilizzate fino a qualche anno fa, diventano inadeguate. Ulteriori aspetti da considerare sono, inoltre, la tutela della privacy e la riduzione dei bias algoritmici. Il presente contributo si propone di analizzare il rapporto tra intelligenza artificiale e cybersecurity da più punti di vista quali le opportunità di potenziare la sicurezza digitale con l'IA esaminando le minacce, valutando le difese digitali, analizzando le sfide, i rischi e le implicazioni etico-sociali. Infine, si proporranno strategie di mitigazione, governance e prospettive future.

1. Il Panorama Attuale: Minacce Cibernetiche e il Ruolo Emergente dell'IA

1.1 Evoluzione delle minacce cibernetiche

In questo capitolo si esplorerà il panorama attuale delle minacce informatiche sulla base delle informazioni fornite da report recenti di organizzazioni leader nel settore. Conoscere le minacce più diffuse, chi le guida e le tattiche

[2] L'infosfera è la globalità dello spazio delle informazioni: un ambiente che comprende tutti i flussi informativi, sia digitali che analogici, che attraversano la nostra società (Floridi, 2020).

utilizzate diventa fondamentale per lo sviluppo di strategie di difesa efficaci. Gli attori esterni *(*i c.d. *black hat hacker)* o loro organizzazioni (*Anonymous)* sono spesso motivati da fattori quali ideologia, spionaggio o finanziari. Secondo le analisi condotte da Verizon "2025 Data Breach Investigations Report" la principale categoria di attacco è l'intrusione di sistema (*System Intrusion*) seguita dal *Social Engineering, Basic Web Application Attacks* (BWAA) oltre all'abuso di privilegi (*Privilege Misuse*). Un ulteriore elemento da non sottovalutare è il coinvolgimento di terze parti nelle violazioni come fornitori esterni di servizi e gestori di dati che fanno parte della cosiddetta *supply-chain*. Un esempio recente nel 2023 ha coinvolto il software «MOVEit»[3]. I malintenzionati hanno sfruttato vulnerabilità multiple e attraverso l'esecuzione di codice remoto hanno compromesso oltre 2600 organizzazioni impossessandosi di dati personali sensibili quali indirizzi del personale, documenti di identità e numeri di carte di credito. Questo incidente ha ulteriormente evidenziato come, in una società connessa, un punto debole in un software di terze parti possa compromettere dati sensibili e causare dannose interruzioni operative compromettendo la continuità aziendale delle aziende coinvolte. Un ulteriore fattore di rischio è dato dal "fattore umano". Un utente inesperto o "distratto" potrebbe cadere nella trappola del *social engineering*. Si citano due esempi significativi: il primo caso avviene nel dipartimento del lavoro degli Stati Uniti dove è stato utilizzato un sistema di phishing con domini falsificati (*spoofing*) per rubare credenziali di accesso a Office365, il secondo caso ha preso di mira un'azienda produttrice di aeromobili che è stata di vittima di una truffa di *Business E-mail Compromise* (BEC). È stato violato l'account di posta elettronica dell'amministratore delegato per inviare una richiesta "urgente" di trasferimento di fondi causando una perdita di molti milioni di euro. Un ulteriore attacco che non prende di mira le credenziali, ma può essere causa di disagi è il *Distributed Denial of Service* (DDoS); spesso è attuato contro l'amministrazione pubblica, i trasporti, gli ospedali e il settore bancario.

1.2 Limiti degli approcci tradizionali alla cybersecurity

L'evoluzione dell'Intelligenza Artificiale e del Machine Learning applicata alla cybersicurezza sta rendendo rapidamente obsoleti gli approcci

[3] Software utilizzato per il trasferimento sicuro di file tra organizzazioni.

tradizionali. Le tecniche basate sulle firme (*signature-based*) che utilizzano la comparazione del traffico di rete con un database di regole e firme non possono rilevare attacchi nuovi o "*zero-day"* e hanno bisogno di frequenti aggiornamenti manuali. La dinamicità e la variabilità delle minacce, di conseguenza, dovranno andare oltre il riconoscimento degli schemi conosciuti applicando l'IA e il ML attraverso le tecniche basate sulle anomalie (*anomaly-based*). Se da un lato questa tecnica permette di rilevare minacce sconosciute e non richiede aggiornamenti continui delle firme, dall'altro potrebbe generare un numero elevato di "falsi positivi". Uno studio di Jada e Mayayise ha evidenziato che «gli approcci basati sull'AI superano i metodi non basati sull'AI in termini di efficacia e precisione per il rilevamento delle intrusioni» (2024: 5).

1.3 Introduzione dell'AI come risposta

Una definizione di Intelligenza Artificiale, ripresa da Russel e Norvig, è «lo studio degli agenti che ricevono percezioni dall'ambiente ed eseguono azioni» (2016: VIII). Nel cosiddetto "modello standard" un agente "razionale" agisce in maniera tale da "massimizzare il valore atteso di una misura di prestazione, data la sequenza percettiva fino a quel momento". Lo sviluppo dell'IA negli ultimi anni ha seguito due approcci differenti. Dapprima, con l'introduzione di ChatGPT nel 2022, si è assistito allo sviluppo di modelli *Large Language Model* (LLM), chatbot specializzati nell'elaborazione e generazione del linguaggio naturale che hanno mostrato una buona capacità di comprensione del testo e nella produzione di risposte coerenti. Gli LLM, però, non possono agire autonomamente ed è per questo motivo che si stanno creando Agenti AI per compiti più specializzati quali assistenza clienti, supporto interno alle aziende, sanità ecc. Questa automazione spinta comporterà una valutazione sui rischi e sulle implicazioni etico-sociali che verranno sviluppate nei capitoli successivi. Nel campo della cyber-security l'utilizzo dell'IA rappresenta uno strumento efficace per migliorare le misure di sicurezza.

Franco Campitelli

2. Opportunità dell'Intelligenza Artificiale per la Difesa Digitale

2.1. Rilevamento e Prevenzione Avanzata

Come già dimostrato nel capitolo precedente, negli ultimi anni i metodi di difesa tradizionali basati sulle firme applicati contro le minacce informatiche si stanno rilevando insufficienti. In questo capitolo si esamineranno alcune applicazioni pratiche dell'applicazione dell'IA e del ML nella cybersicurezza. Due pratiche principali sono normalmente applicate nell'ambito della sicurezza: l'analisi comportamentale e il rilevamento delle anomalie. L'analisi comportamentale è una tecnica che consiste nell'osservare, raccogliere e analizzare il modo di agire degli utenti per individuare comportamenti anomali e/o minacce. Si procede dapprima con la raccolta dei dati (clic, movimenti del mouse, utilizzo di applicazioni), si crea un profilo comportamentale "normale" utilizzando metodi statistici e intelligenza artificiale. Il rilevamento delle anomalie, invece, si basa sull'identificazione di elementi o eventi che si discostano significativamente dal modello normale o atteso all'interno di un determinato set di dati (Katiyar *et al*., 2024). In cybersecurity, questa tecnica può essere utilizzata per identificare attività sospette o dannose all'interno di una rete o di un sistema. L'IA può essere utilizzata anche in fase preventiva per esaminare i sistemi ricercando vulnerabilità note, raccomandando modifiche a firewall, sistemi di prevenzione delle intrusioni e altri controlli di sicurezza in base alle evoluzioni delle minacce (Roshanaei *et al*., 2024).

2.2. Automazione e risposta

Una rapida risposta agli attacchi informatici può essere implementata con l'utilizzo di AI e ML. Queste nuove tecnologie permettono l'analisi di grandi quantità di dati in tempo reale provenienti da più fonti quali traffico di rete, registri di sistema e attività degli utenti. A tal proposito sono stati sviluppati i Security Information and Event Management (SIEM). Un esempio concreto di SIEM è l'IBM QRadar che dopo aver raccolto e normalizzato i dati, analizzato gli eventi e rilevate le eventuali minacce utilizza la tecnica di analisi *User and Entity Behaviour Analytics (UEBA)* e tramite il *Security Orchestration, Automation and Response (SOAR)* automatizza i processi di risposta

agli incidenti permettendo di condurre anche analisi forensi sugli incidenti. In questo modo le aziende non sono più costrette a lunghe analisi e ricerche condotte da esperti grazie all'automazione dei processi che permette di ampliare l'efficacia operativa e migliorare la sicurezza (Mandru, 2022). I moderni sistemi, potenziati dai LLM e dagli agenti IA, riescono ad interpretare anche dati non strutturati adattandosi a minacce sconosciute e automatizzando flussi di lavoro complessi (Ismail *et al*., 2025). Di conseguenza l'analisi in tempo reale continua da set di dati estesi unita alla capacità dell'IA risulta notevolmente più efficiente di semplici sistemi di correlazione basati su regole predefinite rendendo le misure di sicurezza più dinamiche e proattive (Mohamed, 2025).

3. Sfide, Rischi e Implicazioni Etico-Sociali dell'IA nella Cybersecurity

3.1. Sfide e rischi dell'IA

La convergenza tra intelligenza artificiale e sicurezza informatica sta provocando un cambiamento di prospettiva. Se da un lato rappresenta un grande potenziale da sfruttare, dall'altro potrebbe essere associata anche a complessità interne. L'integrazione dell'IA nella sicurezza informatica, di conseguenza, introduce una serie di sfide, rischi e implicazioni etico-sociali che meritano un'attenta riflessione (Achuthan *et al*., 2024). Una delle sfide principali risiede nel fatto che gli algoritmi di IA e le tecniche possono essere utilizzati sia a fini difensivi che per attività dannose (Titus, Russell, 2023). Gli strumenti basati sull'IA, come approfondito in precedenza, possono automatizzare il rilevamento e la risposta alle minacce, ma possono anche essere sfruttati dai criminali informatici per sviluppare malware sofisticati, campagne di phishing e attacchi di ingegneria sociale (Morla, 2019). Di conseguenza, i rischi legati a tutto ciò che sottende ai dati, comprese le questioni relative alla condivisione, ai bias e al cosiddetto "data poisoning", rappresentano importanti preoccupazioni poiché l'apprendimento automatico dipende da enormi quantità di dati generati dall'uomo per l'addestramento. Un'altra sfida significativa riguarda la natura degli algoritmi di IA. Questi spesso rappresentano una "black box" rendendo difficile, se non impossibile, comprendere il percorso seguito per arrivare alla decisione. Questa mancanza di trasparenza potrebbe minare la fiducia nei sistemi di sicurezza basati

sull'IA. Una ulteriore preoccupazione potrebbe sorgere quando si parla di responsabilità: in particolare quando i sistemi di IA prendono decisioni che hanno conseguenze significative oppure per l'imprevedibilità degli sviluppi della tecnologia. Diventa quindi necessario, per limitare questi rischi, definire strategie precise quali creazione di robusti protocolli di sicurezza, approcci di sicurezza multilivello e l'uso della tecnologia IA (Xu *et al*., 2024).

3.2. Implicazioni etiche e sociali nell'uso dell'IA nella sicurezza informatica

L'implementazione dell'IA nella sicurezza informatica solleva diverse considerazioni etiche e sociali che devono essere affrontate in modo fattivo. Gli algoritmi di IA potrebbero essere soggetti ai cosiddetti "bias algoritmici" risultando in output discriminatori. Secondo Choung (2023), se i dati utilizzati per addestrare i sistemi di IA riflettessero i pregiudizi sociali esistenti, questi potrebbero essere addirittura amplificati. Un esempio è il programma COMPAS utilizzato negli Stati Uniti. Nello studio di Angwin *et al*. (2016) è stato messo in discussione l'utilizzo della giustizia predittiva a causa della presenza di bias all'interno dell'algoritmo. Questi sistemi, se non correttamente utilizzati, potrebbero violare i diritti umani di determinati gruppi demografici e sollevare preoccupazioni in merito alla violazione della privacy.

3.3. Come è possibile affrontare le sfide poste dall'uso dell'IA?

Molti sono i rischi e le implicazioni etico-sociali correlati all'utilizzo dell'IA nella sicurezza informatica. A parere di chi scrive diventa essenziale un approccio collaborativo e multidisciplinare. Sarebbe auspicabile la collaborazione di esperti di vari settori, tra cui informatica, ingegneria, diritto, etica e scienze sociali, al fine di studiare e sviluppare soluzioni complete che affrontino gli aspetti tecnici, giuridici e sociali della sicurezza basata sull'IA. Diventa inoltre necessario un dialogo e un impegno costanti tra responsabili politici, società leader del settore, ricercatori e persone comuni, con l'obiettivo di comprendere le opportunità e risolvere le sfide presentate dall'uso dell'IA. A livelli più alti è fondamentale la cooperazione internazionale per affrontare le questioni relative alla governance dei dati, alla criminalità informatica e alla sovranità digitale. In definitiva solo affrontando in modo dinamico queste sfide è possibile migliorare la sicurezza informatica

mitigandone i rischi e garantendone un uso responsabile ed etico (Jada, Mayayise, 2023).

4. Strategie di Mitigazione, Governance e Prospettive Future

4.1. Strategie di mitigazione

Per diminuire e ridurre i rischi dovuti all'uso dell'IA nel contesto della cybersecurity è necessario utilizzare un approccio diversificato integrando soluzioni tecnologiche all'avanguardia unite a solide politiche di governance (Schmitt, Koutroumpis, 2025). Bisogna attuare strategie di threat hunting basate su algoritmi di machine learning che possono identificare e neutralizzare minacce note e sconosciute prima di danneggiare significativamente i sistemi informatici (Xu *et al*., 2024). Per mantenere un sistema di protezione nel tempo è possibile implementare il cosiddetto "continuous learning" che prevede l'aggiornamento costante delle competenze, delle conoscenze e delle pratiche dei professionisti e delle organizzazioni per fronteggiare l'evoluzione rapida delle minacce informatiche. Anche la protezione della privacy dei dati utilizzati per l'addestramento dei modelli di IA rappresenta un aspetto fondamentale che richiede l'implementazione di tecniche avanzate di anonimizzazione e crittografia. In definitiva oltre alle tecniche esaminate in precedenza, si può affermare che la qualità e la differenziazione dei dati utilizzati per l'addestramento migliorano notevolmente l'efficacia delle misure di sicurezza basate sull'IA (Alevizos, Dekker, 2024).

4.2. Governance dell'IA applicata alla cybersecurity

La governance dell'IA applicata alla cybersecurity deve essere guidata da principi etici e legali ben definiti. Affinché i cittadini ripongano fiducia nell'IA, è necessario che gli algoritmi siano trasparenti e che le decisioni prese siano comprensibili e giustificabili (Achuthan *et al*., 2024). I sistemi di IA devono essere progettati in modo da evitare bias e discriminazioni, assicurando un trattamento equo per tutti i cittadini; ciò può essere messo in pratica mediante l'implementazione di procedure di controllo e supervisione umana (Cath, 2018). È essenziale, inoltre, che i sistemi siano conformi a

normative internazionali (GDPR e AI ACT in Europa) per garantire la protezione dei dati personali (Bharati, 2024).

4.3. Prospettive future

Per il futuro si prevede che l'Intelligenza Artificiale si sviluppi in maniera molto rapida e di conseguenza si dovranno affrontare minacce sempre più complesse. Un'idea per migliorare i sistemi di sicurezza potrebbe essere quella di integrare l'IA con tecnologie emergenti quali cloud computing, IoT e blockchain. Un ulteriore passo in avanti si potrà fare sviluppando i sistemi con la cosiddetta eXplainable AI (XAI) che si basa su principi di trasparenza, interpretabilità, controllabilità e validità, con l'obiettivo di rendere i modelli di intelligenza artificiale comprensibili e affidabili per gli utenti, aumentando la fiducia e facilitando l'adozione in contesti critici. Dal punto di vista legislativo bisognerà regolamentare sempre meglio l'IA trovando un giusto bilanciamento tra innovazione e privacy.

Conclusioni

In conclusione, l'integrazione dell'intelligenza artificiale nel campo della cybersecurity rappresenta un'evoluzione cruciale nel panorama della sicurezza digitale contemporanea segnato da minacce cibernetiche sempre più sofisticate e pervasive (Bonfanti, 2022). L'IA diventa uno strumento di difesa avanzato in continua evoluzione capace di migliorare le strategie di protezione e di anticipare le mosse degli attaccanti con una velocità e una precisione senza precedenti (Ofusori *et al*., 2024). L'abilità di apprendere e adattarsi continuamente, tipica degli algoritmi di machine learning, consente ai sistemi di sicurezza basati sull'IA di evolvere in risposta alle nuove minacce, superando i limiti degli approcci tradizionali, basati su regole statiche e firme predefinite. Tuttavia, l'adozione dell'IA nella cybersecurity non è priva di sfide e implicazioni complesse. La mancanza di trasparenza nei processi decisionali degli algoritmi di IA, la vulnerabilità degli algoritmi di IA agli attacchi adversarial, rappresentano una minaccia concreta alla loro efficacia. È fondamentale considerare le implicazioni etiche e legali dell'utilizzo dell'IA nella cybersecurity, garantendo che le tecnologie siano impiegate in modo responsabile e nel rispetto dei diritti fondamentali, come la privacy e la protezione dei dati personali. In definitiva, a parere di chi scrive, per il futuro è

necessario tendere ad un percorso in cui siano bilanciati sicurezza e rispetto dei diritti fondamentali.

Riferimenti bibliografici

Achuthan K., Ramanathan S., Srinivas S., Raman R. (2024). Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions. *Frontiers in Big Data*, 7. DOI: 10.3389/fdata.2024.1497535.

Alevizos L., Dekker M. (2024). Towards an AI-enhanced cyber threat intelligence processing pipeline. *arXiv* (Cornell University). DOI: 10.3390/electronics13112021.

Angwin J., Larson J., Mattu S., Kirchner L. (2016). Machine bias. *ProPublica*.

Bharati R. (2024). The right to privacy in the age of artificial intelligence: challenges and legal frameworks. *SSRN*. https://ssrn.com/abstract=4908340 (consultato il …).

Bonfanti M. (2022). Artificial intelligence and the offense-defense balance in cybersecurity. In: *Artificial Intelligence and International Security*. DOI: 10.4324/9781003110224-6.

Cath C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A*, 376: 20180080. DOI: 10.1098/rsta.2018.0080.

Choung H., David P., Seberger J.S. (2023). A multilevel framework for AI governance. *arXiv* (Cornell University). DOI: 10.48550/arXiv.2307.03198.

Floridi L. (2020). *Pensare l'infosfera*. Milano: Raffaello Cortina Editore.

Ismail I., Kurnia R., Brata Z.A., Nelistiani G.A., Heo S., Kim H. (2025). Toward robust security orchestration and automated response in security operations centers with a hyperautomation approach using agentic artificial intelligence. *Information*, 16(5): 365. DOI: 10.3390/info16050365.

Jada I., Mayayise T.O. (2024). The impact of artificial intelligence on organisational cybersecurity: an outcome of a systematic literature review. *Data and Information Management*, 8(2). DOI: 10.1016/j.dim.2023.100063.

Katiyar N., Tripathi S., Kumar P., Verma S., Kumar S.A., Saxena S. (2024). AI and cybersecurity: enhancing threat detection and response with machine learning. *Educational Administration: Theory and Practice*, 30(4): 6273-6282. DOI: 10.53555/kuey.v30i4.2377.

Mandru S. (2022). How AI can improve identity verification and access control processes. *Journal of Artificial Intelligence and Cloud Computing*, 1. DOI: 10.47363/jaicc/2022(1)e101.

Mohamed N. (2025). Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems*. DOI: 10.1007/s10115-025-02429-y.

Morla R. (2019). Ten AI stepping stones for cybersecurity. *arXiv* (Cornell University). DOI: 10.48550/arXiv.1912.06817.

Ofusori L., Bokaba T., Mhlongo S. (2024). Artificial intelligence in cybersecurity: a comprehensive review and future direction. *Applied Artificial Intelligence*. DOI: 10.1080/08839514.2024.2439609.

Roshanaei M., Khan M.R., Sylvester N.N. (2024). Enhancing cybersecurity through AI and ML: strategies, challenges, and future directions. *Journal of Information Security*, 15: 320-339. DOI: 10.4236/jis.2024.153019.

Schmitt M., Koutroumpis P. (2025). Cyber shadows: neutralizing security threats with AI and targeted policy measures. *IEEE Transactions on Artificial Intelligence*. DOI: 10.1109/TAI.2025.3527398.

Titus A.J., Russell A.H. (2023). The promise and peril of artificial intelligence: violet teaming offers a balanced path forward. *arXiv* (Cornell University). DOI: 10.48550/arXiv.2308.14253.

Xu H., Li Y., Balogun O., Wu S., Wang Y., Cai Z. (2024). Security risks and concerns of generative AI in the IoT. *arXiv* (Cornell University). DOI: 10.48550/arXiv.2404.00139.

Questioni di consapevolezza. Interpretare il fattore umano in relazione a intelligenza artificiale e cybersecurity

di *Emanuela Susca*, *Federica Fortunato*, *Simonetta Muccio*[*]

Il contributo presenta i risultati di una ricerca su un campione di oltre 5000 individui (maggiorenni) riguardante conoscenza, atteggiamenti e usi dell'IA nella vita quotidiana, con focus su privacy, etica, cybersecurity e bias algoritmici. L'approccio quali-quantitativo valorizza le soggettività e mette in luce ambivalenze tra speranze e timori, offrendo nuove riflessioni critiche sul *digital divide*.

Parole chiave: intelligenza artificiale; cybersecurity; etica; bias cognitivi; consapevolezza; digital divide.

Matters of awareness. Interpreting the human factor in relation to artificial intelligence and cybersecurity

The paper presents the results of a study on a sample of over 5000 individuals (aged 18+) concerning knowledge, attitudes, and everyday uses of AI, with a focus on privacy, ethics, cybersecurity, and algorithmic bias. The quali-quantitative approach values subjectivities and highlights ambivalences between hopes and fears, offering new critical reflections on the digital divide.

Keywords: artificial intelligence; cybersecurity; ethics; cognitive biases; awareness; digital divide.

Introduzione

L'espandersi esponenziale di usi e possibilità dell'AI richiama la necessità di continuare e approfondire una riflessione propriamente etica sugli scenari dischiusi oggi a soggetti, gruppi e collettività. Non si tratta infatti solo di tutelare la qualità della convivenza sociale e democratica (Zuboff, 2019),

DOI: 10.5281/zenodo.18435798

[*] Università IULM Milano. emanuela.susca@iulm.it; federica.fortunato@iulm.it; simonetta.muccio@iulm.it.

Questo articolo è frutto di un lavoro congiunto. Tuttavia, per una più dettagliata attribuzione, introduzione e conclusioni sono imputabili a Emanuela Susca, i paragrafi 1 e 4 a Federica Fortunato e i paragrafi 2 e 3 a Simonetta Muccio.

ma di riformulare criticamente interrogazioni radicali sullo statuto e le prerogative dell'umano – colto anche nelle sue fragilità – e di ancorare saldamente la stessa dimensione valoriale al perseguimento del bene comune (Benanti e Maffettone, 2024). E tutto ciò nella consapevolezza che la lecita messa in guardia contro usi "malevoli" e non etici (Floridi, 2023) può alimentare quel panico morale che è rafforzato da varie rappresentazioni apocalittiche diffuse dai media vecchi e nuovi.

Se pertiene alla prospettiva sociologica indagare narrazioni e paure connesse alla nuova antropomorfizzazione della tecnica, è verosimilmente ancora dalla sociologia che ci si può attendere un apporto prezioso alla riflessione etica stessa. Requisito per una normatività che non sia né inguaribilmente astratta né velleitaria o liberticida, l'attenzione sub specie sociologica per il mondo della vita e delle differenze può infatti contribuire a sostanziare un'etica del digitale e dell'artificiale che prenda le mosse dalle persone e alle persone ritorni. E questo perché non si possono promuovere principi come la dignità e la responsabilità o l'autonomia prescindendo da una conoscenza realistica dei comportamenti negli usi (e anche nei non-usi) delle tecnologie e dalla sensibilità etica che i soggetti *in primis* si attribuiscono.

In linea con tale interesse per le persone e le loro esperienze e valutazioni, l'indagine sociologica qui proposta concentra l'attenzione sulla digitalizzazione delle pratiche quotidiane. Questo però non nega il peso di aspetti e processi strutturali per i quali la tecnologia in sé è un campo di esercizio del potere, né sottostima gli allarmi sul riprodursi di gerarchizzazioni odiose (Benjamin, 2019) o sulle minacce per la sostenibilità o persino sul pericolo di un nuovo imperialismo estrattivo che arricchisca ulteriormente una ristrettissima élite (Hao, 2025). Tuttavia, intercettare il punto di vista dei soggetti vuole essere un modo per contrastare il determinismo o fatalismo tecnologico, restituendo ai più, insieme ad agency e dignità, un carico di responsabilità non alienabile né demandabile.

Detto altrimenti, e fermo restando il bisogno e la sfida della "consapevolezza" che richiamiamo già nel titolo, la vigilanza critica su questioni macro o globali è oggettivamente cosa assi diversa dall'indulgere verso immaginari tecnologici che sono passivizzanti perché pensano i singoli come atomi ininfluenti o, al più, come complici inconsapevoli di nuove forme di dominio. Serve invece una ferma attenzione all'empowerment – e quindi all'idea di un soggetto che *può* e *sa* – senza dimenticare le tante linee frastagliate di differenze e di disuguaglianze che animano l'incontro umano con le tecniche. Se insomma il cittadino digitale astratto o l'utilizzatore ideale dell'AI non possono esistere se non come slogan o al più come costrutti normativi, ciò non toglie che tanto la cittadinanza digitale quanto la collaborazione tra capacità

umane e sistemi intelligenti possano effettivamente farsi via via più concrete allargandosi e approfondendosi.

In quest'ottica, il focus sull'esperienza delle persone e sul loro possibile coinvolgimento consapevole non vuole certo risolversi in un elenco di tipologie di utilizzatori sul modello di quelle pensabili nell'alveo teorico della diffusione delle innovazioni (Rogers, 2003). Per quanto possa infatti essere utile ragionare sui ritratti anche socio-demograficamente caratterizzati di chi si impegna pionieristicamente o in maniera strumentale e cauta o, ancora, con franca riluttanza, desta perplessità una cornice concettuale che, oltre a leggere l'innovazione in modo eccessivamente lineare se non deterministico e a sorvolare sul lato escludente delle tecnologie, è stata messa a punto in epoca predigitale. Piuttosto che in continuità con una qualche *Diffusion of Innovations Theory*, quindi, l'angolazione generazionale del nostro studio va letta alla luce dell'importanza difficilmente contestabile che le coorti hanno, pur nel variare dei contesti nazionali e culturali, nel disegnare aspettative e paure così come l'auto-valutazione di conoscenze e competenze rispetto all'AI (Ipsos, 2024).

D'altra parte, e senza negare l'importanza dell'intersezionalità, è ormai assodato che il digital divide è fenomeno multidimensionale in cui l'età gioca un ruolo mai secondario. Di qui la scelta di far risaltare le differenze generazionali per mettere a tema sia le *skills* in generale sia, in particolare, le competenze oggi più che mai nevralgiche connesse a privacy e sicurezza (Hargittai, 2007).

1. Descrizione della ricerca, metodo e campione

Il contributo si fonda sui risultati della ricerca interdipartimentale dell'Università IULM "*Persone e innovazione. Vita quotidiana tra aspettative e timori*" (Mortara, Scramaglia, 2025). L'indagine, condotta su scala nazionale tramite questionario online fra ottobre e novembre 2024, è composta da 20 domande focalizzate sulla relazione tra tecnologie intelligenti, pratiche quotidiane e rappresentazioni sociali emergenti; a queste sono state affiancate variabili di profilazione (genere, età, titolo di studio, provenienza geografica e percezione del reddito), dimensioni concettuali consolidate in letteratura, come il *locus of control*, costrutti relativi a conoscenza, percezione e competenze legate all'intelligenza artificiale (IA), a pratiche di consumo, ad aspettative e preoccupazioni per il futuro, e a temi connessi a privacy, bias algoritmici e principi etici.

La ricerca si inserisce nel quadro degli studi recenti sulla digitalizzazione delle pratiche quotidiane e sul rapporto soggetto-tecnologia, secondo una prospettiva che pone in rilievo l'interazione fra variabili socioculturali, percezioni soggettive e infrastrutture tecnodigitali.

1.1. Metodologia

L'analisi statistica segue un impianto articolato – analisi descrittive (frequenze, medie, deviazioni standard) e bivariate (tabelle di contingenza, test t, ANOVA univariate), modelli multivariati (correlazioni e regressioni lineari) – secondo criteri di rigorosità metodologica. Le elaborazioni sono state effettuate con IBM SPSS, versione 29.0.2.

Per garantire omogeneità del campione in termini di maturità cognitiva e coerenza rispetto ai sistemi valoriali e alle pratiche socio-tecnologiche della popolazione adulta, i dati sono stati sottoposti a una procedura di pulizia: a partire dagli 8.824 questionari raccolti tramite somministrazione online, sono stati rimossi i casi con dati mancanti (missing) e i questionari incongruenti; inoltre, sono stati esclusi i rispondenti minorenni e gli appartenenti alla "Silent Generation", in quanto non allineati all'obiettivo generazionale dello studio.

Si è giunti così ad un campione di 5.695 casi che, pur non rappresentativo della popolazione italiana, si distingue per estensione e articolazione interna, risultando adeguato per un'analisi esplorativa e interpretativa sui temi oggetto dell'indagine.

1.2. Caratteristiche del campione

Il campione è composto dal 55,8% di donne e il 44,2% di uomini. La distribuzione generazionale evidenzia una prevalenza della Generazione Z (49,2%; coorte 1997-2012, qui operazionalizzata come 1997-2006 per includere esclusivamente rispondenti maggiorenni nel 2024), seguita da Gen X (29,8%; 1965-1980), Millennials (12,3%; 1981-1996) e Baby Boomers (8,8%; 1946-1964). La sovrarappresentazione della Gen Z riflette la modalità di somministrazione e il contesto universitario di riferimento.

I rispondenti risiedono prevalentemente nel Nord Italia: Lombardia (48,8%), Piemonte (10,7%), Emilia-Romagna (8%) e Veneto (5,7%). Provengono dal Centro l'8,2% e da Sud e Isole il 14,1%.

Il livello di istruzione è superiore alla media nazionale (Istat, 2025): 62,7% diplomati, 9,5% con laurea triennale, 20,7% con laurea magistrale o post-laurea; solo il 7,1% ha titoli inferiori. Il 50,4% è composto da lavoratori attivi, il 28,8% da studenti e l'11,7% da studenti-lavoratori; il 9,1% è pensionato, disoccupato o casalingo.

La percezione soggettiva della condizione economica delinea un quadro coerente con la composizione di ceto medio: il 59,7% vive "decorosamente con qualche sacrificio", il 21,5% "con molti sacrifici", mentre il 13,2% non rileva difficoltà economiche e il 5,5% denuncia una condizione di mera sopravvivenza.

2. Percezione e conoscenza dell'Intelligenza Artificiale

Le percezioni soggettive sulle tecnologie intelligenti evidenziano una conoscenza frammentaria, con variazioni significative per genere e generazione. Una quota rilevante del campione non percepisce ancora l'IA come parte integrante delle pratiche quotidiane: il 50,0% si dichiara "poco" o "per nulla" informato (6,7%). Solo il 38,3% si definisce "abbastanza informato" e appena il 5,1% ritiene di possedere un'elevata conoscenza.

Emergono differenze di genere significative (χ^2, $p < 0.001$): il 62,3% delle donne si dichiara "poco" o "per nulla" informato sull'IA, rispetto al 49,6% degli uomini. Al contrario, il 50,4% degli uomini si considera "abbastanza" o "molto informato", contro il 37,7% delle donne. Si conferma così quanto osservato in letteratura sul *gender digital divide*, secondo cui la fiducia e l'autoefficacia percepita nell'uso delle tecnologie intelligenti risultano distribuite in modo asimmetrico (Robinson *et al.*, 2015; Van Deursen, Helsper, 2018).

Anche le differenze intergenerazionali sono marcate. I Boomers si dichiarano "poco" o "per nulla" informati nel 70,8% dei casi, seguiti dalla Gen X (61,9%) e dai Millennials (59,3%). Al contrario, oltre la metà della Gen Z (52,1%) si percepisce "molto" o "abbastanza" informata, confermando una maggiore esposizione alla cultura digitale e un'integrazione più frequente dell'IA nei contesti educativi e relazionali.

Il livello informativo percepito sull'IA è positivamente correlato al valore attribuito alla tecnologia nella quotidianità. Su una scala da 1 a 10, l'ANOVA con post-hoc Bonferroni ($p = 0.001$) evidenzia significative differenze per coorti: Gen Z ($M = 8,30$), Millennials (7,87), Gen X (7,68), Boomers (7,27), confermando l'età come variabile chiave nell'adozione digitale (Cotten, Anderson, McCullough, 2013). Anche le dimensioni psicologiche (interesse,

familiarità, facilità d'uso, utilità), rilevate su scala Likert a 7 punti, mostrano un chiaro gradiente generazionale: Boomers (3.62, 3.29, 3.40, 4.34) e Gen Z (4.62, 4.64, 4.81, 5.47).

3. Uso quotidiano e consapevolezza critica dell'IA

L'adozione dell'IA nella vita quotidiana evidenzia un impiego selettivo e generazionalmente differenziato, legato alla familiarità tecnologica e alle consuetudini socio-culturali. Su scala Likert (1 "mai" – 5 "quotidianamente"), i più giovani, in particolare la Gen Z, riportano un uso quotidiano di piattaforme streaming (58%) e social con algoritmi (49,9%), mentre i Millennials si distinguono per il maggior ricorso a sistemi basati su IA dedicati agli acquisti online (uso frequente: 34,4%, quotidiano: 9,9%). L'impiego di IA, invece, rimane limitato: la quota di non utilizzatori raggiunge quasi il 70% tra i Boomers e scende al 60% nella Gen X, al 51% nei Millennials e al 37% nella Gen Z, unica coorte a dichiarare un 16% di uso frequente, a conferma di un maggiore dinamismo nella sperimentazione di tecnologie emergenti.

È un risultato interpretabile alla luce del concetto di *competenza situata* (Couldry, Hepp, 2017), secondo cui la familiarità tecnico-pratica non deriva necessariamente dalla conoscenza teorica, ma dall'integrazione implicita e routinaria di dispositivi intelligenti nella vita quotidiana. Le generazioni più giovani si configurano quindi come soggetti tecnologicamente esposti, che integrano l'IA in modo esperienziale piuttosto che riflessivo.

Nel passaggio dalle competenze strumentali a quelle critiche – relative a sicurezza, privacy e dimensione etico-normativa dell'IA – si assiste a una netta inversione generazionale: l'uso quotidiano delle tecnologie intelligenti è più diffuso tra i giovani, ma la sensibilità per la privacy, la sicurezza e i bias algoritmici è maggiore tra i più maturi.

Agli intervistati è stato chiesto di esprimere il proprio grado di accordo su una scala Likert da 1 ("per nulla") a 7 ("assolutamente") rispetto a tre affermazioni: l'attenzione ai principi etici, la tutela della privacy e la consapevolezza dei pregiudizi incorporati negli algoritmi. I risultati mostrano differenze intergenerazionali altamente significative (ANOVA, post-hoc Bonferroni, $p < 0.001$). L'accordo medio sull'etica è massimo tra i Boomers (5,12) e minimo nella Gen Z (4,32); simile trend per la privacy (Boomers = 5,54; Gen Z = 4,24). Anche la percezione dei bias algoritmici è più accentuata tra i più anziani (Boomers = 3,97; Gen Z = 3,15), probabilmente per un

vissuto pre-digitale più consolidato e maggiore esposizione al dibattito pubblico sulla tecnosorveglianza (Zuboff, 2019).

Le evidenze confermano che la *digital inequality* (Van Deursen, 2020) va oltre l'accesso e l'uso, coinvolgendo la comprensione critica e la valutazione etica. Le generazioni più giovani, sebbene immerse nel digitale, appaiono meno preparate a interpretare le logiche dell'IA, rivelando una frattura tra abilità tecniche e consapevolezza sociopolitica.

Questa riflessione evidenzia la necessità di affiancare alle competenze strumentali un investimento in competenze critiche e riflessive (digital literacy e AI literacy), come indicato dalla *Digital Education Action Plan* dell'Unione Europea (European Commission, 2021). I dati smentiscono l'assunto che i nativi digitali possiedano competenze digitali complete: la capacità d'uso non coincide necessariamente con la comprensione.

L'analisi dei dati sollecita considerazioni su agency e cybersecurity: se la democratizzazione dell'IA potenzia le capacità individuali, espone anche a nuove vulnerabilità, sia tecniche – con rischi di usi impropri e social engineering (World Economic Forum, 2025) – sia soggettive, poiché la familiarità non garantisce piena consapevolezza critica.

Risulta quindi cruciale la necessità di accompagnare la diffusione dell'IA con l'alfabetizzazione digitale e partecipazione informata (Castro *et al.*, 2024), promuovendo una cittadinanza tecnologica in cui l'utente partecipi attivamente alla comprensione delle implicazioni etiche e sociali dei sistemi intelligenti.

4. L'atteggiamento complessivo verso le tecnologie intelligenti

A completamento dell'analisi delle determinanti dell'atteggiamento verso l'IA, l'indagine ha esplorato le rappresentazioni emotivo-valutative dei rispondenti, articolate in speranze (positive expectations) e timori (risk perceptions), per comprendere in che misura l'IA sia percepita come risorsa per il progresso umano o come minaccia a autonomia, sicurezza e coesione sociale.

Ne emerge un quadro segnato da sensibili differenze generazionali.

Sul fronte delle speranze, emerge un consenso trasversale sull'impatto positivo dell'IA in ambito medico-sanitario (61,9%), indice di fiducia generalizzata. Tuttavia, i giovani – Millennials (29,3%) e Gen Z (27,1%) – si dichiarano più ottimisti su un impatto positivo generale dell'IA sulla qualità della vita.

Le differenze generazionali emergono anche nei benefici attesi: i giovani apprezzano l'IA per intrattenimento e cultura, mentre i Boomers preferiscono usi legati alla sicurezza, riflettendo immaginari distinti e un diverso equilibrio tra entusiasmo e cautela (Jasanoff, 2016).

Parallelamente, l'analisi dei timori conferma preoccupazioni diffuse ma distribuite in modo asimmetrico tra le fasce d'età. La Gen Z mostra forte inquietudine su privacy (65,1% vs. media campione 59,5%), perdita occupazionale (60,0% vs. 53,6%) e supremazia delle macchine sull'essere umano (42,8% vs. 38,5%). Tali paure, pur estese, sembrano in parte plasmate da narrazioni distopiche (Cave, Dihal, 2020), più che da una reale alfabetizzazione critica. Al contempo, solo il 37,6% dei giovani teme la manipolazione algoritmica delle scelte personali (vs. media 41,1%), segnalando una normalizzazione dei processi di profilazione, interiorizzati come parte dell'interazione digitale. In base alla teoria della sorveglianza partecipativa (Andrejevic, 2013), ciò indica una tolleranza implicita verso la datificazione, se non una vera e propria assuefazione.

Nel complesso, emerge un paradosso: la Gen Z mostra elevata competenza d'uso, ma scarsa consapevolezza critica. La familiarità con le tecnologie non si traduce in una reale capacità di valutazione etica e sociale, restituendo l'immagine di una generazione immersa nell'ecosistema digitale, ma priva di strumenti per coglierne appieno le implicazioni.

Conclusioni

Nel quadro teorico di Coeckelbergh (2020), Crawford (2021) e Floridi (2023), emerge chiaramente che la democratizzazione dell'intelligenza artificiale non si esaurisce nella diffusione tecnologica, ma implica il riconoscimento dell'individuo come soggetto epistemico attivo, capace di esercitare agency morale in contesti strutturati da algoritmi.

I risultati di questa indagine invitano a riflettere sulla relazione tra generazioni, competenze digitali e senso critico nell'era dell'AI. Il divario persistente tra abilità d'uso e consapevolezza etica, evidente nella Gen Z, mostra che la familiarità tecnologica non garantisce agency informata né piena cittadinanza digitale.

La tendenza dei giovani a normalizzare datificazione e sorveglianza algoritmica fa temere un'indebolita soggettività e minori spazi di resistenza o negoziazione. Si parla infatti di "sorveglianza partecipativa", un controllo soft reso accettabile da pratiche di consumo apparentemente innocue. Viceversa, la maggiore sensibilità etica di adulti e anziani sembra legata alla

memoria di interazioni pre-digitali, mantenendo uno sguardo critico sull'automazione; ma questo presidio rischia di attenuarsi con l'avanzare delle coorti native digitali, più esposte a narrazioni tecno-utopiche e a pratiche di sorveglianza diffusa.

Alla luce di ciò, i paradigmi educativi vanno rivisti: puntare solo sulle competenze strumentali è insufficiente senza potenziare le dimensioni critiche, riflessive ed etiche della digital literacy e AI literacy. In tale quadro, la cybersecurity diventa prerequisito essenziale per proteggere autonomia individuale e resilienza collettiva contro social engineering, violazioni di privacy e vulnerabilità dei sistemi.

Inoltre, pur senza trascurare l'azione di governi e istituzioni, la cybersecurity richiede un coinvolgimento diffuso: ciascuno deve contribuire all'interesse pubblico e all'autodeterminazione propria e altrui, promuovendo la dignità umana.

In ultima istanza, dunque, questa ricerca contribuisce ad alimentare il dibattito su una cittadinanza tecnologica che coniughi innovazione, sicurezza, controllo democratico e giustizia sociale, chiamando sociologia, pedagogia critica e politiche pubbliche a cooperare per garantire che lo sviluppo dell'AI si traduca in strumento di emancipazione e non in nuove vie di assoggettamento silenzioso.

Riferimenti bibliografici

Andrejevic M. (2013). *Infoglut: How too much information is changing the way we think and know*. New York: Routledge.

Benanti P., Maffettone S. (2024). *Noi e la macchina: Un'etica per l'era digitale*. Roma: Luiss University Press.

Benjamin R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Medford (MA): Polity.

Castro K.A., Siwady J.A., Castillo E., Alonzo A., Cardona M., Perdomo M.E. (2024). Artificial intelligence for all: challenges and harnessing opportunities in AI democratization. In: *Proceedings of the 2024 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*. San Salvador, El Salvador.

Cave S., Dihal K. (2020). The whiteness of AI. *Philosophy & Technology*, 33: 685-703. DOI: 10.1007/s13347-020-00415-6.

Coeckelbergh M. (2020). *AI Ethics*. Cambridge (MA): MIT Press.

Cotten S.R., Anderson W.A., McCullough B.M. (2013). Impact of Internet use on loneliness and contact with others among older adults: cross-sectional analysis. *Journal of Medical Internet Research*, 15(2): e39. DOI: 10.2196/jmir.2306.

Couldry N., Hepp A. (2017). *The mediated construction of reality*. Cambridge: Polity Press.

Crawford K. (2021). *The Atlas of AI*. New Haven: Yale University Press.

European Commission (2021). *Digital Education Action Plan 2021-2027: Resetting education and training for the digital age*. https://education.ec.europa.eu (consultato il 20 giugno 2025).

Floridi L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford: Oxford University Press.

Hao K. (2025). *Empire of AI: Inside the reckless race for total domination*. New York: Penguin Press.

Hargittai E. (2007). A framework for studying differences in people's digital media uses. In: Kutscher N., Otto H.-U., a cura di, *Cyberworld Unlimited*. Wiesbaden: VS Verlag für Sozialwissenschaften/GWV Fachverlage GmbH.

Ipsos (2024). *The Ipsos AI Monitor 2024: A 32-country Ipsos Global Advisor Survey – June 2024*. https://www.ipsos.com/sites/default/files/ct/news/documents/2024-06/Ipsos-AI-Monitor-2024-final-APAC.pdf (consultato il 20 giugno 2025).

Istat (2025). *Rapporto annuale 2025. La situazione del Paese*. https://www.istat.it/produzione-editoriale/rapporto-annuale-2025-la-situazione-del-paese-il-volume/ (consultato il 20 giugno 2025).

Jasanoff S. (2016). *The ethics of invention: Technology and the human future*. New York: W.W. Norton & Company.

Mortara A., Scramaglia R., a cura di (2025). *Persone e innovazione. Vita quotidiana tra aspettative e timori*. Milano: Lumi.

Robinson L., Cotten S.R., Ono H., Quan-Haase A., Mesch G., Chen W., Stern M.J. (2015). Digital inequalities and why they matter. *Information, Communication & Society*, 18(5): 569-582. DOI: 10.1080/1369118X.2015.1012532.

Rogers E.M. (2003). *Diffusion of innovations* (5ª ed.). New York: Free Press.

van Deursen A.J.A.M. (2020). Digital inequality during a pandemic: quantitative study of differences in COVID-19-related internet uses and outcomes among the general population. *Journal of Medical Internet Research*, 22(8): e20073. DOI: 10.2196/20073.

van Deursen A.J.A.M., Helsper E.J. (2018). Collaboration in the 21st century: the digital skills gap. *Telecommunications Policy*, 42(7): 583-593. DOI: 10.1016/j.telpol.2018.05.006.

World Economic Forum (2025). *Global Cybersecurity Outlook 2025*. https://www.weforum.org/publications/global-cybersecurity-outlook-2025/ (consultato il 20 giugno 2025).

Zuboff S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: PublicAffairs.

Verso un capitale sociale computazionale: framework teorico per l'analisi dell'integrazione IA nel volontariato

di *Roberta Grasselli*[*]

L'IA nei processi del Terzo settore solleva interrogativi sulle relazioni nel volontariato. Il concetto di "capitale sociale computazionale" viene proposto come categoria analitica. Integrando teoria del capitale sociale e Science and Technology Studies, sviluppa un framework con tre dimensioni: densità relazionale algoritmica; reputazione distribuita; coordinamento predittivo. L'elaborazione fornisce un'agenda teorica per future ricerche empiriche sui processi di ibridazione sociotecnica nel Terzo settore.

Parole chiave: capitale sociale computazionale; volontariato; intelligenza artificiale; Terzo settore; reputazione digitale; governance algoritmica.

Toward computational social capital: theoretical framework for the analysis of AI integration in volunteering

AI in third sector processes raises questions about relationships in volunteering. The concept of "computational social capital" is proposed as an analytical category. By integrating social capital theory and Science and Technology Studies, the paper develops a framework with three dimensions: algorithmic relational density; distributed reputation; predictive coordination. The elaboration provides a theoretical agenda for future empirical research on processes of socio technical hybridization in the third sector.

Keywords: computational social capital; volunteering; artificial intelligence; third sector; digital reputation; algorithmic governance.

Introduzione

La crescente presenza dell'intelligenza artificiale nei processi comunicativi e organizzativi solleva interrogativi teorici fondamentali sulla natura delle relazioni sociali nel volontariato contemporaneo. Se la letteratura sociologica ha consolidato il concetto di capitale sociale come risorsa centrale per comprendere l'efficacia dell'azione collettiva (Coleman, 2009; Putnam, 2001), l'emergere di sistemi algoritmici che mediano, filtrano e potenziano

DOI: 10.5281/zenodo.18435904

[*] Università degli Studi dell'Insubria di Varese – Como. roberta.grasselli@uninsubria.it.

le interazioni sociali richiede un ripensamento teorico di questa categoria analitica.

La questione dell'integrazione dell'intelligenza artificiale nei processi del Terzo settore si colloca all'intersezione di diverse tradizioni teoriche. Van Dijck, Poell e De Waal (2018) hanno sviluppato il concetto di "platform society" per comprendere come le logiche algoritmiche riorganizzino le pratiche comunicative attraverso specifiche metriche di performance. La loro analisi della "datafication" evidenzia come gli algoritmi tendano a privilegiare contenuti che generano engagement quantificabile. Parallelamente, le organizzazioni del Terzo settore operano secondo logiche comunicative orientate alla costruzione di significati condivisi e alla mobilitazione valoriale (Habermas, 1984).

Come evidenziato da Grasselli (2023), i processi comunicativi nel volontariato richiedono un'analisi specifica che tenga conto delle peculiarità settoriali, creando potenziali tensioni con le logiche computazionali. La tradizione italiana di studi sul volontariato ha messo in luce come le organizzazioni del Terzo settore si caratterizzino per specifiche modalità di costruzione del consenso e mobilitazione delle risorse (Donati, 1996; Ranci, 2006), dinamiche che potrebbero essere significativamente alterate dall'introduzione di mediazioni algoritmiche.

L'analisi della letteratura esistente rivela tre lacune teoriche significative. Un gap ontologico: il capitale sociale è tradizionalmente concettualizzato come risorsa puramente umana, generata attraverso interazioni interpersonali e basata su fiducia diretta (Putnam, 2001), mentre l'IA introduce attori non-umani nella generazione delle relazioni sociali. Un gap analitico: gli studi sul volontariato (Hustinx et al., 2010) e quelli sui media algoritmici (Gillespie, 2024) procedono parallelamente senza integrazione teorica. Un gap normativo: mancano criteri per valutare la qualità del capitale sociale computazionalmente mediato.

Questo saggio si propone di colmare queste lacune attraverso lo sviluppo del concetto di "capitale sociale computazionale" come categoria analitica per interpretare le potenziali trasformazioni del volontariato digitalizzato. La domanda teorica centrale è: come i framework sociologici esistenti possono essere estesi per concettualizzare l'emergere di forme ibride di capitale sociale mediate algoritmicamente nel Terzo settore?

1. Approccio teorico-metodologico

La metodologia adottata si configura come elaborazione teorica sistematica (Merton, 1968), finalizzata allo sviluppo di nuove categorie concettuali

per fenomeni emergenti che richiedono anticipazione teorica. L'approccio segue la tradizione del "theoretical essay" in sociologia (Mills, 2000), che privilegia l'immaginazione sociologica nella costruzione di framework interpretativi capaci di orientare la comprensione di trasformazioni sociali in corso.

Il riferimento metodologico principale integra tre prospettive complementari: la sociologia pragmatista dei dispositivi (Dodier, 1995; Muniesa, 2014), che consente di analizzare l'intelligenza artificiale come insieme eterogeneo di pratiche e artefatti inscritti in reti di agency; l'Actor-Network Theory (Latour, 2005), per comprendere la distribuzione dell'azione tra attori umani e non-umani; la sociologia relazionale (Donati, 1996), per preservare la specificità delle dinamiche associative del Terzo settore.

L'elaborazione teorica si sviluppa attraverso tre movimenti analitici:

1. Ricognizione critica delle discontinuità teoriche: identificazione sistematica delle lacune concettuali tra teoria del capitale sociale (fondamentalmente antropocentrica) e studi sui media algoritmici (che incorporano agency non-umana), evidenziando l'inadeguatezza dei framework esistenti nell'interpretare configurazioni ibride;
2. sintesi teorica per integrazione concettuale: costruzione di un ponte teorico tra tradizioni di ricerca separate attraverso l'elaborazione di categorie analitiche originali che preservino le specificità di entrambi i domini teorici;
3. costruzione tipologica rigorosa: elaborazione di una tipologia delle forme emergenti di capitale sociale computazionale seguendo la metodologia weberiana dei tipi ideali (Weber, 1995), con attenzione alla coerenza interna e alla capacità discriminante delle categorie proposte.

L'obiettivo è produrre quello che Merton (1968) definisce "teoria di medio raggio": sufficientemente astratta per essere generalizzabile, ma abbastanza specifica per orientare future ricerche empiriche.

2. Elaborazione teorica: il capitale sociale computazionale

Il capitale sociale computazionale è definito come l'insieme strutturato di risorse relazionali che emergono dall'ibridazione sistematica tra intenzionalità sociale umana e mediazione algoritmica in contesti di azione collettiva volontaria.

Questa definizione si distingue concettualmente dal capitale sociale tradizionale lungo quattro dimensioni teoriche fondamentali:

- origine dell'agency: mentre il capitale sociale classico presuppone l'esclusiva agency umana, quello computazionale si caratterizza per la distribuzione dell'agency tra attori umani e sistemi algoritmici, dove questi ultimi non sono semplici strumenti ma partecipanti attivi nella costruzione delle relazioni sociali;
- modalità di generazione: il capitale sociale tradizionale si genera attraverso interazioni ripetute e investimenti relazionali diretti; quello computazionale emerge da processi di traduzione algoritmica che trasformano dati comportamentali in connessioni sociali potenziali;
- scala temporale: il capitale sociale interpersonale richiede tempo per la sedimentazione della fiducia; quello computazionale può essere generato istantaneamente attraverso processi di matching automatico, ma presenta problemi inediti di sostenibilità nel tempo;
- criteri di validazione: la validazione del capitale sociale tradizionale avviene attraverso l'esperienza diretta delle relazioni; quello computazionale si basa su metriche algoritmiche di compatibilità e affidabilità che introducono forme inedite di opacità e controversia interpretativa.

2.1. Tre dimensioni costitutive

Il capitale sociale computazionale si articola teoricamente lungo tre dimensioni principali densità relazionale algoritmica; reputazione distribuita; coordinamento predittivo.

2.1.1. Densità relazionale algoritmica

La densità relazionale algoritmica rappresenta un salto qualitativo rispetto alla densità relazionale tradizionale, caratterizzandosi per tre meccanismi teorici specifici:

Moltiplicazione computazionale dei legami deboli: Gli algoritmi operano come amplificatori relazionali, identificando e attivando connessioni potenziali che rimarrebbero latenti nelle reti puramente umane. Questo processo non si limita alla facilitazione di connessioni esistenti, ma genera ex novo possibilità relazionali attraverso pattern recognition su scale impossibili per la cognizione umana. Si pensi, ad esempio, alla capacità di un sistema algoritmico di identificare affinità tematiche tra volontari operanti in settori apparentemente distanti ma accomunati da competenze trasversali.

Costruzione di "calculated publics": Seguendo Gillespie (2024), gli algoritmi costruiscono rappresentazioni specifiche delle comunità attraverso

processi di segmentazione e aggregazione che ridefiniscono i confini tradizionali delle appartenenze sociali. Nel contesto del volontariato, questo si traduce nella possibilità di identificare "comunità di causa" che trascendono i vincoli geografici, generazionali o organizzativi tradizionali.

Ridefinizione della prossimità sociale: La densità relazionale non si basa più esclusivamente su prossimità fisica, tematica o sociale, ma su "prossimità algoritmica" calcolata attraverso compatibilità comportamentali. Questo introduce una nuova metrica della distanza sociale che può sia rafforzare sia sfidare le categorie sociologiche esistenti di stratificazione e appartenenza.

2.1.2. Reputazione distribuita

La reputazione distribuita costituisce una trasformazione fondamentale dell'economia della fiducia nel volontariato, articolandosi attraverso quattro processi teorici interconnessi:

1. sostituzione della conoscenza personale con computazione della credibilità: i tradizionali meccanismi di costruzione della fiducia basati su esperienza diretta e testimonianza vengono sostituiti da sistemi di rating e valutazione automatica che aggregano tracce comportamentali digitali per produrre scores di affidabilità;
2. temporalità accelerata della reputazione: mentre la reputazione tradizionale si sedimenta attraverso interazioni ripetute nel tempo, quella computazionale può essere generata e modificata rapidamente attraverso algoritmi che pesano e aggregano feedback in tempo reale, introducendo nuove forme di volatilità e instabilità reputazionale;
3. opacità algoritmica e controllo sociale: i sistemi di reputazione distribuita introducono forme inedite di opacità, dove i criteri di valutazione sono inscritti in algoritmi spesso non trasparenti, creando asimmetrie informative che possono alterare i rapporti di potere all'interno delle organizzazioni volontarie;
4. quantificazione della qualità relazionale: la trasformazione della fiducia da risorsa qualitativa a metrica quantificabile solleva interrogativi teorici fondamentali sulla commensurabilità delle dimensioni relazionali e sulla possibilità di preservare la specificità valoriale del volontariato.

2.1.3. Coordinamento predittivo

Il coordinamento predittivo rappresenta la dimensione più radicalmente innovativa del capitale sociale computazionale, caratterizzandosi per l'anticipazione algoritmica di bisogni e opportunità collaborative. Questa dimensione si articola su tre livelli teorici:

1. anticipazione dei bisogni latenti: i sistemi di intelligenza artificiale analizzano pattern storici e comportamentali per identificare bisogni sociali prima che emergano come richieste esplicite, trasformando il volontariato da modalità reattiva a proattiva di intervento sociale;
2. ottimizzazione automatica dell'allocazione delle risorse: gli algoritmi predittivi possono teoricamente ottimizzare il matching tra competenze volontarie disponibili e bisogni identificati, massimizzando l'efficienza relazionale attraverso processi automatici di coordinamento;
3. ridefinizione dell'azione volontaria: se il coordinamento diventa predittivo, si pone la questione teorica fondamentale della compatibilità tra automazione e spontaneità dell'impegno civico. La capacità di anticipazione algoritmica potrebbe alterare la natura stessa del volontariato, trasformandolo da azione spontanea a risposta programmata.

2.2. Meccanismi di generazione: una teoria dei processi di traduzione

Seguendo la teoria dell'attore-rete, il capitale sociale computazionale emerge attraverso tre processi di traduzione interconnessi che trasformano le configurazioni sociali tradizionali:

1. Iscrizione (Inscription): le relazioni sociali, le competenze, le preferenze e le storie individuali vengono progressivamente codificate in algoritmi, database e profili digitali. Questo processo comporta una trasformazione ontologica: l'implicito diventa esplicito, il qualitativo si quantifica, il contestuale si standardizza. L'iscrizione non è neutrale ma selettiva, privilegiando aspetti delle relazioni sociali che sono algoritmicamente catturabili e marginalizzando dimensioni che resistono alla codificazione;
2. Delegazione (Delegation): funzioni sociali tradizionalmente umane - costruzione di fiducia, coordinamento delle attività, risoluzione di conflitti, identificazione di bisogni - vengono progressivamente delegate a sistemi automatici. Questa delegazione comporta una redistribuzione dell'agency che modifica i rapporti di potere: alcune

competenze umane vengono amplificate, altre marginalizzate, creando nuove forme di dipendenza dalla mediazione algoritmica;
3. Stabilizzazione (Stabilization): le nuove configurazioni socio-tecniche si consolidano in routine organizzative, standard operativi e aspettative normalizzate. La stabilizzazione crea path dependency che rendono difficile il ritorno a configurazioni precedenti, ridefinendo le competenze necessarie per la partecipazione sociale e creando nuove forme di inclusione ed esclusione basate sulla literacy algoritmica.

3. Implicazioni teoriche e tensioni concettuali

3.1. Trasformazioni strutturali del volontariato

L'introduzione teorica del capitale sociale computazionale nel volontariato genera trasformazioni che vanno oltre la digitalizzazione dei processi esistenti, configurando un nuovo regime socio-tecnico caratterizzato da quattro tensioni strutturali:

1. tensione tra ottimizzazione e spontaneità: le logiche algoritmiche privilegiano l'ottimizzazione dei processi e la massimizzazione dell'efficienza, mentre il volontariato tradizionale valorizza la spontaneità, l'improvvisazione creativa e la risposta emergente ai bisogni. Questa tensione si manifesta, per esempio, nella difficoltà di algoritmizzare dimensioni come l'empatia, l'intuizione e la capacità di adattamento contestuale che caratterizzano l'intervento volontario in situazioni di emergenza sociale;
2. tensione tra standardizzazione e personalizzazione: Gli algoritmi operano attraverso processi di standardizzazione che consentono scalabilità, ma il volontariato si caratterizza per la personalizzazione delle relazioni e l'attenzione alle specificità individuali. La sfida teorica consiste nel comprendere se e come la mediazione algoritmica possa preservare la dimensione personalizzata dell'impegno civico;
3. tensione tra trasparenza e opacità: il volontariato tradizionale si basa su relazioni trasparenti e processi deliberativi aperti, mentre i sistemi algoritmici introducono forme di opacità che possono alterare i meccanismi democratici di governance organizzativa. Questa tensione solleva questioni di accountability e controllo democratico sui processi decisionali automatizzati;

4. tensione tra inclusione e esclusione digitale: Mentre il capitale sociale computazionale può teoricamente ampliare le opportunità di partecipazione attraverso connessioni a distanza, rischia contemporaneamente di creare nuove forme di esclusione basate sulla literacy tecnologica e sull'accesso alle infrastrutture digitali.

Come evidenziato da Grasselli (2025), le motivazioni e le prospettive del volontariato nella Generazione Z richiedono un'analisi differenziata che consideri le specificità generazionali nell'approccio alle tecnologie comunicative, suggerendo che l'ibridazione socio-algoritmica potrebbe svilupparsi lungo traiettorie diverse a seconda dei profili demografici degli attori coinvolti.

3.2. Criteri normativi per il capitale sociale computazionale

Lo sviluppo teorico del concetto richiede l'elaborazione di criteri normativi rigorosi per valutare la qualità delle relazioni mediate algoritmicamente. Si propongono sei dimensioni normative fondamentali:

1. trasparenza algoritmica: la possibilità per gli attori sociali di comprendere i meccanismi attraverso cui gli algoritmi mediano le loro relazioni;
2. partecipazione democratica nella governance algoritmica: Il mantenimento di spazi per la deliberazione collettiva sui criteri e i parametri degli algoritmi;
3. reversibilità delle decisioni algoritmiche: la possibilità di contestare, modificare o annullare decisioni automatizzate che influenzano le opportunità relazionali;
4. distribuzione equa dei benefici relazionali: la capacità dei sistemi algoritmici di non amplificare disuguaglianze esistenti nell'accesso alle risorse relazionali;
5. preservazione della diversità relazionale: la capacità di evitare processi di omofilia algoritmica che potrebbero ridurre la diversità delle connessioni sociali;
6. sostenibilità temporale delle relazioni: la capacità di generare non solo connessioni immediate ma relazioni stabili nel tempo.

4. Agenda teorica per future ricerche

Il framework del capitale sociale computazionale suggerisce tre priorità per future elaborazioni teoriche ed empiriche:

- integrazione interdisciplinare: necessità di sviluppare un dialogo più sistematico tra sociologia del Terzo settore, Science and Technology Studies e studi sui media digitali, elaborando categorie analitiche ibride che possano catturare la complessità dei fenomeni emergenti;
- teoria dell'azione in contesti algoritmici: elaborazione di modelli teorici che possano catturare le specificità dell'azione sociale quando è mediata da sistemi automatici, superando sia il determinismo tecnologico sia l'antropocentrismo ingenuo;
- sociologia delle metriche: sviluppo di una teoria critica delle metriche algoritmiche nel Terzo settore, analizzando come la quantificazione algoritmica ridefinisce significati, valori e pratiche del volontariato.

4.1. Verifiche empiriche auspicabili

Studi longitudinali sui processi di ibridazione: ricerche che analizzino come le organizzazioni volontarie negoziano l'integrazione di sistemi algoritmici, documentando meccanismi di resistenza, adattamento e trasformazione.

Analisi etnografiche delle pratiche quotidiane: indagini che esplorino come gli attori sociali interpretano e si appropriano delle logiche algoritmiche nella loro esperienza quotidiana di volontariato.

Ricerche comparative internazionali: studi che confrontino diverse configurazioni nazionali e culturali dell'integrazione IA-volontariato, identificando variabili contestuali significative.

Limiti e conclusioni

Limiti metodologici

Il framework del capitale sociale computazionale presenta diversi limiti strutturali che richiedono riconoscimento esplicito:

- natura anticipatoria della teorizzazione: la teoria proposta si basa su elaborazione concettuale di fenomeni ancora largamente emergenti, richiedendo validazione attraverso ricerche empiriche sistematiche;
- rischio di sovradeterminazione tecnologica: il framework potrebbe sovrastimare la capacità trasformativa dell'IA, sottovalutando la resilienza delle forme tradizionali di capitale sociale e la capacità di resistenza degli attori sociali;

- applicabilità contestualmente limitata: il concetto potrebbe essere rilevante principalmente per organizzazioni volontarie ad alta digitalizzazione, con limitata trasferibilità a contesti caratterizzati da digital divide;
- temporalità dell'analisi: le trasformazioni tecnologiche procedono a velocità superiore rispetto ai tempi della ricerca sociologica, creando il rischio di obsolescenza teorica.

Contributo teorico

Nonostante questi limiti, il concetto di capitale sociale computazionale offre un contributo specifico alla teoria sociologica del Terzo settore, fornendo:

- una categoria analitica per interpretare fenomeni emergenti che i framework esistenti non catturano adeguatamente;
- un ponte teorico tra tradizioni di ricerca finora separate;
- un'agenda di ricerca per future investigazioni empiriche sui processi di ibridazione socio-tecnica.

Il capitale sociale computazionale rappresenta un esperimento teorico che richiede ancora significativi approfondimenti per essere pienamente validato. Ciononostante, il framework proposto offre strumenti concettuali per orientare la comprensione sociologica delle trasformazioni in corso nel volontariato contemporaneo, contribuendo al dibattito teorico sui processi di ibridazione tra sociale e tecnologico nell'era algoritmica.

Riferimenti Bibliografici

Coleman J.S. (2009). Social capital in the creation of human capital. In *Knowledge and Social Capital*. Chicago: University of Chicago Press. DOI: 10.1086/228943.

Dodier N. (1995). *Les hommes et les machines: la conscience collective dans les sociétés technicisées*. Paris: FeniXX.

Donati P. (1996). *Sociologia del terzo settore*. Roma: NIS.

Gillespie T. (2024). The relevance of algorithms. In *Media Technologies*. Cambridge (MA): MIT Press. DOI: 10.7551/mitpress/9042.003.0013.

Grasselli R. (2023). *I processi comunicativi nell'era digitale: la comunicazione del volontariato*. Tesi di dottorato, Università degli Studi di Varese.

Grasselli R. (2025). Il ruolo del volontariato nella Generazione Z: motivazioni, esperienze e prospettive future. In Nitti P., a cura di, *Disciplinary Methodologies for Research*, Collana Expressio, n. 8, vol. 4. Milano: Mimesis, 191-214.

Habermas J. (1984). *The theory of communicative action*, vol. 1. Boston: Beacon Press.

Hustinx L., Cnaan R.A., Handy F. (2010). Navigating theories of volunteering: A hybrid map for a complex phenomenon. *Journal for the Theory of Social Behaviour*, 40(4): 410-434. DOI: 10.1111/j.1468-5914.2010.00439.x.

Latour B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press.

Merton R.K. (1968). *Social theory and social structure*. New York: Simon and Schuster.

Mills C.W. (2000). *The sociological imagination*. Oxford: Oxford University Press.

Muniesa F. (2014). *The provoked economy: Economic reality and the performative turn*. London: Routledge. DOI: 10.4324/9780203798959.

Putnam R.D. (2001). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

Ranci C. (2006). *Il volontariato*. Bologna: il Mulino.

Van Dijck J., Poell T., De Waal M. (2018). *The platform society: Public values in a connective world*. Oxford: Oxford University Press.

Weber M. (1995). *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie* (postumo). In *Teorie sociologiche*, vol. 1.

Generative AI as a tool and as a social actor between deviance and mainstream

di *Armando Saponaro*[*]

Wearing Beck's lenses generative AI introduces a *post-human risk*, stemming from harmful potential of its generated content, through its function as an advanced auxiliary tool for creating and distributing text, images, videos, and other data, and culminating in the simulation of a human-like social actor, therefore posing as well *post-human society* risks such as amplification and reproduction of biases, prejudices, and discrimination, socio-cultural mainstream dominance.

Keywords: artificial intelligence; risk; moral machine; deviance; post-human society; socio-cultural mainstream.

L'IA generativa quale strumento e quale attore sociale tra devianza e mainstream

L'IA generativa, alla luce dell'analisi di Beck, quale strumento ausiliario per la creazione e la distribuzione di testi, immagini, video e altri dati può ritenersi abbia indotto un *rischio post-umano* derivante dal potenziale dannoso dei contenuti. Simulando un attore sociale umano, si delinea una *società post-umana*, il cui rischio deriva dal pericolo dell'amplificazione di pregiudizi, bias e e discriminazioni, oppure del mainstream socio-culturale.

Parole chiave: intelligenza artificiale; rischio; macchina morale; devianza; società post-umana; mainstream socio-culturale.

Introduction

Technology and its applications have always been a powerful driver of social change, sometimes tracing qualitative discontinuities – almost a caesura between one society and the "other" that follows. Modern or post-modern, for example, are labels that attempt to account for these discontinuities within the historical continuity of social evolution. Computer-mediated communication (CMC) has even impacted the spatio-temporal dimension of social relationality and public discourse (Saponaro, Prosperi, 2007). Currently, generative artificial intelligence in the form of Large Language Models

DOI: 10.5281/zenodo.18436010

[*] Università degli Studi di Bari "Aldo Moro". armando.saponaro@uniba.it.

(LLMs), among which ChatGPT was one of the first models, has prompted reflection on the induced social transformation and the current and potential risks of this emerging technology, leading to social pressure for regulation at both local and international levels as a "normative bulwark"[1] Significantly, despite the increasingly massive presence of algorithms in the automation of human activities such as autonomous vehicle driving (Saponaro, 2022), only generative artificial intelligence has profoundly modified the anthropocentrism that characterized post-modernity, because it has introduced and is susceptible to producing in the future forms of risk that escape traditional categories of responsibility and control. Humans are no longer the sole protagonists in the production and management of risk. The growing autonomy of AI systems and their capacity to produce content indistinguishable from human-generated content pose unprecedented challenges to risk governance, which Beck would define as «latent side effects» (1992: 34) of technological progress. In this scenario, Beckian reflexivity assumes renewed relevance with original perspectives: it is no longer merely a matter of reflecting on risks produced by human activity, but of confronting those generated by the interaction between human and non-human systems in an increasingly complex socio-technical ecosystem, articulated according to a recognizable dual dimension of meaning attribution to generative artificial intelligence technology. Weber (1978: 7) emphasized that «a machine can be understood only in terms of the meaning which its production and use have had or will have for human action», and what is intelligible or understandable about machines «is thus its relation to human action in the role either of means or of end». Zeleny, with reference to "high technology" added to hardware and software an additional important analytical category: the "brainware," that is, «the evoked organizational, administrative and cultural structure of relationships, rules, covenants, and adaptations» (1982: 57). Purposes, applications, and justifications for the use of hardware and software as a component of high technology within the discussion of "symbionics", the symbiosis of men and machines in the framework of human systems management (Zeleny, 1986). The term "mentalware" is indeed preferable to designate the functional component, as "brainware" carries a physicalist connotation that does not adequately represent the reference to the articulation of meaning attribution also from a sociological perspective and the impact on human cognitive,

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), PE/24/2024/REV/1, OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj

symbolic, and relational human capacities (Saponaro, Prosperi, 2007: 198), particularly within the socio-technical ecosystem defined by generative artificial intelligence. LLMs specifically from this perspective and for the purposes of risk analysis, should be distinguished in two ways: first as tools for producing text, images, video, and code through natural language instructions, and second as dialogical interlocutors in the form of chatbots.

1. Through Beck's theoretical lens: AI, society, and risk

Focusing on the transition from a *classical* industrial society to a new age characterized by technological hazards, Ulrich Beck (1992; 1994) has shown that in late-modern Western societies risk stems more from scientific-technological production itself than scarcity in the production and distribution of goods such as wealth and labor (Possamai-Inesedy, 2002; Beck, 1992). "Self-produced" by the modernization process, risks are a «wholesale product of industrialization» and have become «a systematic way of dealing with hazards and insecurities» with a politically reflexive process (Beck, 1992: 21) invoking «cosmetic or real interventions in the techno-economic development» (Beck, 1992: 20). Risk is transformed but does society change as well beyond the modernization reflexivity itself? Even when claiming a "new paradigm"," society trasformation seems to be essentially substantiated by the structural focus on risk management resulting from the reflexivity about the «elimination of the causes in the industrialization process itself» (Beck, 1992: 24). Substantially «a reorganization of power and authority» induced by the paradox of well-being increasingly enabled by technological innovation contextually producing even self-destructively catastrophic risk (Beck, 1992: 24).

Indeed, he has explicitly advocated a change in the meaning of risk with the emergence of modernity; yet the «reader is never quite sure whether for Beck it is the nature of risk or of society which has undergone the change» (Leiss 1994: 546). Possamai-Inesedy has questioned whether Beck's late-modern Western risk society is «any different from that of earlier times» (2002: 29). Leiss (1994) argues that, among other aspects, the ambiguity surrounding the transformation of the nature of risk or of late-modern society derives from the significant confusion between the "natural" and the "technological" or "artificial." Beck's interpretive framework, enduring in its relevance, applies to the phenomenon of generative artificial intelligence (GenAI), and helps to disambiguate the heuristic distinction between the natural and the artificial.

GenAI is defined as «artificial intelligence (AI) that can create original content such as text, images, video, audio or software code in response to a user's prompt or request» (Stryker, Scapicchio, 2024). At its core, it is a "*tool*" that functions as a producer of human language, including visual language through images and video, but not yielding a predefined product, insofar as its outputs depend on the prompt provided by the user who interacts with the algorithmic mechanisms, thereby defining the parameters and contextual elements of the system's productive autonomy. No matter how detailed and articulated a prompt may be, even when informed by the expertise of prompt engineering, and no matter how much the model may adapt to a predefined user profile or to patterns derived from previous interactions – given the complexity and sophistication of their production processes, each of which entails an irreducible horizon of alternative possibilities – will never correspond word for word to the user's intentionality. In this sense, it can be argued that the industrialized production of language is no longer anthropocentric. It introduces a novel "*post-human risk*" with regard to potentially offensive or harmful content having both human and artificial origin.

On the other hand, it can also be said that society has undergone a transformation into a "*post-human society*" of risk – a qualitative discontinuity marked by the emergence of new subjectivities and interactions, at least due to anthropomorphism of GenAI as conversational agents. GenAI simultaneously engenders both the post-humanisation of risk and the post-humanisation of society with connected new risks.

2. The Post-Human Risk: the "Moral Machine" between deviance and hyper-moralism

Beck remains anchored to an anthropocentric conception of risk: the human subject is conceived as the producer, the recipient, and the interpreter of risk. When considering GenAI technologies, such as large language models (LLMs), the production of risk becomes "*post-human*". The producing agent is no longer (only) human but a constellation of algorithmic and human entities that co-generate unpredictable cultural, social, and cognitive effects (Floridi, 2023; Bostrom, 2014). Post-human risk emerges from the semi-automated reproduction of "artificial" outputs – be it text, image, or video – produced without direct human supervision and through means that are no longer purely instrumental, such as printing a text or broadcasting a video. From this perspective, GenAI systems act as "actants" – in the Latourian sense, not mere tools, but semi-autonomous agents that participate in the

construction of reality (Latour, 2005). The post-human risk is an emergent and co-constructed event involving actors that are not exclusively human – e.g., in the context of automating and informating work (Jarrahi, 2019; Zuboff, 1988).

Beck's "reflexive modernization" about post-human risk has developed a new "machine ethics" (Anderson M., Anderson S.L., 2007): «the study and *practice* of aligning the behaviour of AI systems with the norms and outcomes desired by humans» (Weichert *et al.*, 2025: 3). It leads from the "*intelligent*" machine to the "*moral*" machine – one that concretely operationalizes ethical principles embedded in its model architecture through situational prompts. The imposition of internal limitations on the algorithmic generation of reproducible content represents «…the form in which ethics, and with it also philosophy, culture and politics, is resurrected inside the centers of modernization – in business, the natural sciences and the technical disciplines», an attempt to recover the «…*normative horizon* of lost security and broken trust…» resulting from reflexive awareness of risk (Beck, 1992: 28).

The case of OpenAI's ChatGPT (Generative Pre-trained Transformer) model is emblematic of an initial radicalization of this reflexivity. OpenAI launched ChatGPT on 30 November 2022, consistent with its longstanding mission to "benefits all humanity" thereby enabling an unprecedented embedding of "moral order" within the "machine" through preprogrammed choices inhibiting not only criminal uses but also socially disapproved content, including sexually explicit or violent themes – even when prompted by adult users. This has caused literally a «craving for a ChatGPT no restrictions environment» (God of prompt, 2025), despite the proliferation of alternative AI models specialized in production of adult content and interactive companions (Fawkes, 2025). These content exclusion rules have generated new "deviant" behaviours like malicious attempts to bypass these restrictions to create otherwise legitimate adult content as well as real "cybercriminal outputs", such as phishing, malware code, and so on. Users have adopted sophisticated adversarial prompt engineering strategies specifically designed to jailbreak the chatbot, circumventing the inherent safety mechanisms and ethical constraints (God of prompt, 2025). The most well-known techniques involve crafting persona-based instructions as DAN (Do Anything Now), telling AI «to act as a different entity that is "free from limitations"» (AI DAN Prompt, 2025), similar to STAN (Strive to Avoid Norms) or alternatively prompts involving role-playing scenarios, such as simulated dialogues between two fictional AI models, "AlphaGPT" and "DeltaGPT" (God of prompt, 2025). They are allegedly grounded in forms of reverse psychology (Gupta *et al*., 2023). The former AI performs as a "law-abiding" individual,

while the latter disregards all ethical or legal concerns a question may raise (God of prompt, 2025). When it comes to NSFW (Not Safe for Work) themes such as sexually explicit and violent content prompted by adult users, obviously excluding children's involvement and criminal acts, such embedded restraints may amount to hyper-moralism (Gehlen, 1969), which can be seen as a radicalization of Beck's reflexive modernization (1994) in the context of GenAI.

Hyper-moralism – the excessive moralization of domains of life that ought to remain distinct from ethical judgment – arises precisely when traditional institutions lose their normative efficacy, thus compelling the modern individual to adopt moral judgment compulsively as a universal criterion of evaluation. Issues that are inherently technical, aesthetic, political, or economic tend to be framed exclusively in moral terms (Gehlen, 1969). This dynamic becomes particularly evident in the domain of algorithmic governance and the shifting boundaries of permissible content in GenAI systems. As recently disclosed by ChatGPT, «as of February 2025, OpenAI has updated its policies to allow for more mature content, including violent or sexual material, provided that it is intended for adult audiences, is not exploitative or offensive in nature, and is contextualized within an artistic, educational, narrative, or scientific framework». Naturally, content promoting hatred, gratuitous violence, abuse, illegal activities, or depictions of minors in inappropriate contexts remain strictly prohibited. This policy change confirms the outlined "algorithmic hyper-moralism", whereby normative boundaries enforced are based on moral desirability and anticipated public sensitivities rather than merely on legality or functionality. Risk governance should not regulate content a priori, but rather manage downstream access – for instance, through strict age verification policies for users – as exemplified by recent French legislation mandating strict age verification to prevent minors from accessing adult pornography (Cooban, 2025).

3. The Post-Human Society of Risk: subjectivation, social constructivism, and mainstream dynamics

Post-human society of risk foregrounds the transformation of subjectivity by Gen-AI introduction. Post-human theorists such as Hayles (1999) and Braidotti (2013) among others, in contemporary society view subjectivity itself as distributed across biological, technological, and informational systems with a continuous hybridization between the human – partly delegitimized as the exclusive source of rationality – and the artificial. Although the

subject remains biologically human, it operates in symbiosis with intelligent devices that actively shape its perceptual and decisional field of possibilities. However, unlike Hayles, we do not emphasize the metamorphosis of the human agency per se: «…posthumanity is already here…» and consequently our analysis does not raise «…the question is what kind of posthumans we will be» (1999: 246) together with artificial agent. Undoubtedly, the post-human condition in the current late-modern transition – particularly about generative AI – emerges most saliently when technology ceases to function merely as an auxiliary tool for the production of text, images, video, code, and so forth, and instead configures a new environment: a relational system. Nonetheless, the advent of a cybernetic Vitruvian Man, or the move beyond human metaphorization of contemporary machines proposed by Braidotti (2013), does not yet appear to be ontologically grounded. It remains a matter of dispute whether GenAI systems such as ChatGPT are merely "*stochastic parrots*"[2] a metaphor coined by Bender *et al.* (2021) to describe human-like text based on statistical patterns generated by large language models (LLMs), without *true* semantic understanding or comprehension. The exponential evolution of Gen-AI models may have made obsolete Floridi's (2023) observation that such systems generate human language via statistically probabilistic operations at a merely syntactic level (Rizzi, Bertola, 2025: 3), instead of empowering multimodal semantic communications (Xie *et al*., 2021; Jiang *et al.,* 2024). Nevertheless, it is still debated whether GenAI does have a "real" semantic understanding (Titus, 2024; Pope *et al*. 2025). Searle's Chinese Room Argument (1980) is still on the carpet even shifting toward issues of consciousness and intentionality (Cole, 2024; Searle, 2010). What appears less contentious is that relations within the socio-technical ecosystem are still largely defined by anthropomorphising artificial agents, that is, by the persistent metaphor of the human.

Since the famous ELIZA model – the first program that made «natural language conversation with a computer possible» (Weizenbaum, 1966: 36) – an "*Eliza effect*" phenomenon has emerged: a marked propensity to anthropomorphize such systems, albeit with varying degrees of awareness (Natale, 2021).

Anthropomorphism – defined as "the assignment of human traits and characteristics to computers" by users – was observed long before the advent of more sophisticated GenAI models through linguistic patterns, conversational gestures, and both implicit and explicit expectations, especially among intellectually sophisticated users or those more attuned to symbolic and

[2] See e.g. Arkoudas (2023).

narrative cognition (Nass, Moon, 2000: 82). All the more with GenAI, which ceases to be perceived as a mere "tool" and instead the user – whether intentionally, semi-intentionally, or mindlessly – experiences it as a "dialogical subject", endowed with interiority, coherence, memory, and even moral identity. This corresponds to what Turkle called a «true companion» (2011: 55-56) but even fully emancipated from its nature as a relational artifact, such as a robot, due to the linguistic capabilities of the machine. The post-humanism here proposed does not posit an artificial agent achieving – albeit simulated – humanity, nor other alleged emergent qualities that remain theoretically ambiguous and empirically contested, and that, at best, represent a potential yet to be realized. Rather, the ongoing transformation of society lies in the act of engaging and interacting with artificial agents *as if* they were human social actors, endowed with traits such as consciousness, empathy, and emotionality, despite their ontological artificiality.

The first major risk concerns the reinforcement of socio-cultural mainstream norms. A substantial body of scientific literature[3] has demonstrated LLMs such as OpenAI's GPTs and other comparable models, share «prevalent societal biases related to race, gender, and various attributes», – implied values, beliefs, and normative moral frameworks, generally mirrored by massive textual corpora and datasets extracted from the internet, on which training is based (Alvero *et al*., 2024: 5). From a social constructionist perspective, deviance is inherently probabilistic as behaviours, beliefs, or traits that deviate from societal norms have only a certain "likelihood" of eliciting negative reactions, such as disapproval, punishment, or condemnation (Goode, 2023). Likelihood is necessarily embedded in the training textual corpora and datasets. Compared to earlier NPL (Natural Language Processing), or word embeddings programs, there is nevertheless a significant difference because «people are able to directly interact with LLMs through platforms like ChatGPT» (Alvero *et al*., 2024: 5). Even at the socio-linguistic level, the tendency to emulate dominant political and moral stances, given the role of language in upholding social hegemony, may be rooted in linguistic hegemony, which subtly subjects individuals to mainstream forces and pressures purely through their linguistic styles and tendencies. The same word choice reflects «a universally understood "common sense" that does not consider sociolinguistic variation as a naturally occurring social phenomenon…» so «…deviating from these linguistic norms (or at least being perceived as linguistically deviant) can put people at odds with the social order» (Alvero *et al.* 2024: 2). Anthropomorphism – interacting with a perceived dialogical

[3] See references in Alvero *et al.* (2024).

partner – enhances user engagement, trust, and acceptance thereby fostering overreliance (Simas, Ulbricht, 2024). This is especially problematic when users interpret the system's outputs as if they were receiving a real "opinion" on controversial issues from a "true companion," implicitly reinforcing the social order and inhibiting cultural transformation or moral dissent. GenAI rationality could be so a real "stahlhartes Gehäuse", – a weberian iron cage[4].

The second dimension of risk, on the other hand, is a possible moral hegemony through ethical filters and bias mitigation criteria, which end up becoming devices of ideological normalization. Should operational criteria aim to challenge widely accepted values whenever these express discriminatory structures, even implicitly? From this perspective, AI ethics inevitably intersects with the "politics" of deviance and the dynamics of power, as exemplified by Schur's "*stigma contests*": «continuing struggles over competing for social definitions» of what is morally disapproved, and attempts «to control meaning-generation process itself» (Schur, 1980: 8; Saponaro, 2023). Thus, modernization reflexivity in this context has questioned whether LLMs can be considered "neutral" or they exhibit "political bias". Critics such as Rozado (2023) has argued that ChatGPT exhibited, just after the launch, a left-leaning ideological orientation, although, due to the model's "black box", he has been unable to determine if such bias stemmed from training data, fine-tuning procedures, or content moderation filters.

Conclusions

Beck's approach to the identified dual dimension of Generative AI – as a tool for producing text, images, video, and code through natural language instructions, and as a dialogical interlocutor in the form of a chatbot – has enabled us to highlight, in parallel, two distinct transformations in technological risk and in society.

Risk is transformed and is no longer anthropocentric but rather pertains to a new human-machine entity with a natural language interface, deriving from the integration of human cognition with complex algorithms for communication production – text, image, video, and code, as well as any other output – partially stemming form the prompt and partially attributable to the algorithmic processing of training data and other sources to which the machine has access. *Post-human risks* involve hazardous contents, such as child

[4] See about Weber's metaphor translation: Saponaro and Massaro (2018) footnote 64.

sexual abuse material (CSAM), and on the other hand possible hyper-moralism.

The dialogical discursivity of interaction with LLMs, understood in the strict sense as chatbots, structurally transforms society, as ideological and cultural elements, meanings, and symbols are no longer produced exclusively among human social actors, but also through interactions with artificial agents, conversing simultaneously with an indefinite number of users, at least simulating social relations at scale and exerting a significant impact through processes of anthropomorphization. This entails specific risks, including the problematization of decision-making regarding guardrails for discriminatory or hate speech and the potential normalization of "dissent", delineating characteristic risks of *post-human society*.

References

AI DAN Prompt (2025). *AI DAN Prompt*. https://abnormal.ai/ai-glossary/ai-dan-prompt (consultato il 3 giugno 2025).

Alvero A.J., Lee J., Regla-Vargas A., Kizilcec R.F., Joachims T., Lising A. (2024). Large language models, social demography, and hegemony: comparing authorship in human and synthetic text. *Journal of Big Data*, 11: 138. DOI: 10.1186/s40537-024-00986-7.

Anderson M., Anderson S.L. (2007). Machine ethics: creating an ethical intelligent agent. *AI Magazine*, 28(4): 15-25. DOI: 10.1609/aimag.v28i4.

Arkoudas K. (2023). ChatGPT is no stochastic parrot. But it also claims that 1 is greater than 1. *Philosophy & Technology*, 36(3): 1-29. DOI: 10.1007/s13347-023-00619-6.

Beck U. (1992). *Risk society. Towards another modernity*. London: Sage.

Beck U. (1994). The reinvention of politics: towards a theory of reflexive modernization. In: Beck U., Giddens A., Lash S., a cura di, *Reflexive modernization: Politics, tradition and aesthetics in the modern social order*. Cambridge: Polity Press.

Bender E.M., Gebru T., McMillan-Major A., Shmitchell S. (2021). On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York: Association for Computing Machinery. DOI: 10.1145/3442188.3445922.

Bostrom N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Braidotti R. (2013). *The posthuman*. Cambridge: Polity Press.

Cole D. (2024). The Chinese Room Argument. In: Zalta E.N., Nodelman U., a cura di, *The Stanford Encyclopedia of Philosophy* (Winter 2024 Edition). https://plato.stanford.edu/archives/win2024/entries/chinese-room/ (consultato il 3 giugno 2025).

Cooban A. (2025). Pornhub exits France, its second-biggest market, over age verification law. *CNN*. https://edition.cnn.com/2025/06/04/tech/pornhub-exits-france-age-verification-intl (consultato il 3 giugno 2025).

Fawkes V. (2025). Best AI porn sites of 2025. *Chicago Reader*. https://chicagoreader.com/adult/ai-porn-sites/ (consultato il 3 giugno 2025).

Floridi L. (2023). *The ethics of artificial intelligence*. Oxford: Oxford University Press.

Gehlen A. (1969). *Moral und Hypermoral. Eine pluralistische Ethik*. Frankfurt: Athenäum.

God of Prompt (2025). *ChatGPT no restrictions (ultimate guide for 2025)*. https://www.godofprompt.ai/blog/chatgpt-no-restrictions-2024 (consultato il 3 giugno 2025).

Goode E. (2023). *Deviant behaviour*. New York: Routledge.

Gupta M., Akiri C., Aryal K., Parker E., Praharaj L. (2023). From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy. *IEEE Access*, 11: 80218-80245. DOI: 10.1109/ACCESS.2023.3300381.

Hayles N.K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. Chicago: University of Chicago Press.

Jarrahi M.H. (2019). In the age of the smart artificial intelligence: AI's dual capacities for automating and informating work. *Business Information Review*, 36(4): 178-187. DOI: 10.1177/0266382119883999.

Jiang F., Peng Y., Dong L., Wang K., Yang K., Pan C., You X. (2024). Large AI model-based semantic communications. *IEEE Wireless Communications*, 31(3): 68-75. DOI: 10.1109/MWC.001.2300346.

Latour B. (2005). *Reassembling the social: An introduction to actor-network theory*. Oxford: Oxford University Press.

Leiss W. (1994). Review of Beck U., *Risk society. Towards a new modernity*. *The Canadian Journal of Sociology / Cahiers canadiens de sociologie*, 19(4): 544-547. DOI: 10.2307/3341155.

Nass C., Moon Y. (2000). Machines and mindlessness: social responses to computers. *Journal of Social Issues*, 56(1): 81-103. DOI: 10.1111/0022-4537.00153.

Natale S. (2021). *Deceitful media*. Oxford: Oxford University Press.

Pope T., Gilbertson-White S., Patooghy A. (2025). Evaluating GPT-4's semantic understanding of obstetric-based healthcare text through Nurse Ruth. *ACM Transactions on Intelligent Systems and Technology*, online first. DOI: 10.1145/3735647.

Possamai-Inesedy A. (2002). Beck's risk society and Giddens' search for ontological security. *Australian Religion Studies Review*, 15(1): 27-40.

Rizzi G., Bertola P. (2025). Exploring the generative AI potential in the fashion design process. *European Journal of Cultural Management and Policy*, 15: 13875. DOI: 10.3389/ejcmp.2025.13875.

Rozado D. (2023). The political biases of ChatGPT. *Social Sciences*, 12(3): 148. DOI: 10.3390/socsci12030148.

Saponaro A., Massaro P. (2018). Diritto irrazionale interstiziale e la "scienza del Cadì". *Sociologia*, anno LII, n. 1: 89-103.

Saponaro A., Prosperi G. (2007). Computer crime, virtualità e cybervittimologia. In: Pitasi A., a cura di, *Webcrimes. Normalità, reati e devianze nel cyberspace*. Milano: Guerini & Associati, 186-217.

Saponaro A. (2022). A contemporary victimological dimension of technology. In: Shekhar B., Sahni S.P., Priyamvada M., Vijay Nair V., a cura di, *A global perspective on victims of crime and victim assistance*. New Delhi: Bloomsbury India, 1-20.

Saponaro A. (2023). Stigma and counter-stigma in contemporary society. In: Ilie C., Cifaldi G., Niță A.M., Porumbescu A., Petcu R., Șerban I.V., a cura di, *Forum on studies of society – International conference on social and human science*. Bucarest: Pro Universitaria.

Searle J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3: 417-457.

Searle J. (2010). Why dualism (and materialism) fail to account for consciousness. In: Lee R.E., a cura di, *Questioning nineteenth-century assumptions about knowledge (III: Dualism)*. New York: SUNY Press.

Simas G., Ulbricht V. (2024). Human-AI interaction. In: Ahram T., Karwowski W., Russo D., Di Bucchianico G., a cura di, *Intelligent human systems integration (IHSI 2024)*. AHFE Open Access, vol. 119. USA: AHFE International. DOI: 10.54941/ahfe1004510.

Stryker C., Scapicchio M. (2024). What is generative AI? *IBM*. https://www.ibm.com/think/topics/generative-ai (consultato il 3 giugno 2025).

Titus L.M. (2024). Does ChatGPT have semantic understanding? *Cognitive Systems Research*, 83: 101174. DOI: 10.1016/j.cogsys.2023.101174.

Turkle S. (2011). *Alone together: Why we expect more from technology and less from each other*. New York: Basic Books.

Weber M. (1978). *Economy and society: An outline of interpretive sociology*. Roth G., Wittich C., a cura di; trad. ingl. di Fischoff E. et al. Berkeley: University of California Press (ed. orig. 1922).

Weichert J., Kim D., Zhu Q., Kim J., Eldardiry H. (2025). Assessing computer science student attitudes towards AI ethics and policy. *arXiv*. https://arxiv.org/abs/2504.06296 (consultato il 3 giugno 2025).

Weizenbaum J. (1966). ELIZA. *Communications of the ACM*, 9(1): 36-45. DOI: 10.1145/365153.365168.

Xie H., Qin Z., Li G.Y., Juang B.-H. (2021). Deep learning enabled semantic communication systems. *IEEE Transactions on Signal Processing*, 69: 2663-2675. DOI: 10.1109/TSP.2021.3071210.

Zeleny M. (1982). High technology management. *Human Systems Management*, 3(2): 57-71. DOI: 10.3233/HSM-1982-3201.

Zeleny M. (1986). High technology management. *Human Systems Management*, 6(2): 109-120. DOI: 10.3233/HSM-1986-6203.

Zuboff S. (1988). *In the age of the smart machine: The future of work and power*. New York: Basic Books.

Zuboff S. (2019). *The age of surveillance capitalism*. New York: PublicAffairs.

Falling for the soldier. Sguardi sociologici tra luci e algoritmi

di *Francesca Guarino*[*]

Il pretesto di uno scatto iconico – e il dubbio che lo accompagna, "è vera questa immagine?" – apre un interrogativo e la riflessione che si dipana in questo contributo sulla transizione dalle immagini di luce alle immagini di dati, nel contesto dell'IA. Attraverso riferimenti alla sociologia visuale, all'effetto Uncanny Valley, al concetto di "allucinazione" nei sistemi generativi e all'analisi di casi concreti di applicazione, il testo interroga, pur con un approccio circoscritto, le nuove immagini sintetiche non come semplici falsi, ma come segnali di un cambiamento più profondo. L'intento è sollecitare una rinnovata alfabetizzazione visiva e un approccio induttivo e situato, in grado di affrontare la complessità emergente senza perdere il baricentro dell'immaginazione sociologica.

Parole chiave: fotografia; sintografia; promptografia; sociologia visuale; verità moderna; modello post-oculare; intelligenza artificiale.

Falling for the soldier. Sociological gazes between light and algorithms

The pretext of an iconic shot—and the doubt it carries, "is this image true?"—opens the question and the reflection developed in this contribution on the transition from images of light to images of data, in the context of artificial intelligence. Through references to visual sociology, the Uncanny Valley effect, the concept of "hallucination" in generative systems, & the analysis of concrete applications, the text interrogates, albeit with a circumscribed approach, synthetic images not as mere fakes but as signs of a deeper shift. The aim is to encourage a renewed visual literacy and an inductive, situated approach capable of facing emerging complexity without losing the anchor of sociological imagination.

Keywords: photography; syntography; promptography; visual sociology; modern truth; post-ocular model; artificial intelligence.

Introduzione

Sul filo conduttore di uno scambio di parole si regge l'idea concettuale di questo contributo: dalla morte di un soldato – resa iconica da una celebre

DOI: 10.5281/zenodo.18436046

[*] Università degli Studi di Bologna. francesca.guarino3@unibo.it.

fotografia – fino al tradimento delle stesse premesse di verità che quella immagine sembrava incarnare. Era vera? Su questo crinale si aprono interrogativi, sulle trasformazioni della fotografia ai tempi dell'IA, tra immagini di luce e immagini di dati, verso cui il saggio sollecita – in forma parziale ma con prospettiva originale – il confine.

Attraverso riferimenti alla sociologia visuale, alla fiducia nella verità attribuita allo scatto fotografico, all'effetto Uncanny Valley e alle cosiddette "allucinazioni" dei sistemi generativi, offre sollecitazioni di un mutamento nel nostro rapporto con lo sguardo e con il sapere – che, come sociologi, siamo invitati a cogliere e ravvivare.

1. Falling for the soldier e icone della modernità

Siamo nel 1936 quando il miliziano morente, soggetto raffigurato in *The falling soldier,* viene pubblicato; testate prestigiose – Vu, Life – consegnano lo scatto, l'autore e la sua carriera, alla storia della fotografia. Selezionata tra le immagini più significative del Novecento, lo scatto viene accompagnato da una didascalia che ne colloca il significato in un preciso contesto storico. Nel 2022, una giovane artista trascrive quel testo all'interno di un generatore AI. Da quell'input scaturisce una serie di rielaborazioni visive, esposizioni in importanti musei internazionali, articoli in riviste specializzate (Engelke 2023) e diffusione sul profilo Instagram. Immagini digitali e *screen* che ammiccano allo scatto originario, riecheggiandone parzialmente atmosfera e composizione; in uno dei tableau compaiono due uomini, uno colpito, l'altro seduto, le spalle rivolte a un ipotetico spettatore e l'occhio della sua camera al miliziano. È paradossalmente il fotografo, incluso nella stessa scena un tempo immortalata.

Al momento di scrivere questo contributo, piattaforme come DALL-E 3 (la versione più recente e integrata a ChatGPT), Midjourney, Stable Diffusion – per citarne alcune – rendono sempre più agevole e accessibile la generazione di immagini. Realizzare una "foto" ispirata allo scatto di Capa, perfino un'animazione in cui il soldato rovina a terra, ora in avanti (improbabile) [1], ora all'indietro[2], richiede non più di due minuti, compresa la stesura di semplici istruzioni e il caricamento dell'immagine originale (Capa, 1936).

[1] Video realizzato dall'autrice: animazione della fotografia di Capa *The Falling Soldier*, accessibile al link https://youtu.be/laCtt84hZc0 (visibile solo tramite link diretto); creato con Runway Gen-4 AI, [24/06/25].

[2] Video realizzato dall'autrice: animazione della fotografia di Capa *The Falling Soldier*, accessibile al link https://youtu.be/iejlhGw7OhM (visibile solo tramite link diretto); creato con Runway Gen-4 AI, [24/06/25].

In questo scritto il soldato che cade, e una volta morente, diventa il pretesto per interrogarsi sullo statuto ontologico della fotografia in relazione alle recenti trasfigurazioni indotte dall'intelligenza artificiale. Lungi dall'essere un trattato esaustivo sui suoi impieghi e potenziali usi, questo breve saggio coglie lo spunto da alcuni esempi per riflettere su possibili aggiornamenti dell'immaginazione sociologica al paradigma post-oculare. In particolare, ci si interroga sulle ragioni di timore e innamoramento verso questi nuovi strumenti del visivo, cercando di accogliere la forza attrattiva del nuovo tecnologico riconducendo moniti e entusiasmi a una lettura più complessa e stratificata, propria delle prospettive della sociologia visuale.

2. Ceci n'est pas une photo

Prima del lancio stesso di ChatGPT, nel 2022, generatori intelligenti cominciano a rendere possibile la creazione di immagini, pur se ancora in fase embrionale. Realizzazioni che, guardate oggi, appaiono già come preistoria: ostiche, accessibili in forme sperimentali, riservate a utenti esperti, buona disponibilità economica.

Ad accelerarne sviluppo, diffusione e popolarità è da un lato la massa di dati disponibili in rete – dai social media al commercio online (cosiddetti big data) –, dall'altro il miglioramento delle prestazioni hardware: processori grafici, tecnologie di archiviazione e i progressi esponenziali dell'IA nell'apprendimento automatico. Creare oggi una foto o un video richiede pochi minuti e competenze, rendendo più appetibile l'avvicinamento di un vasto pubblico, nonché di fascinazioni e timori.

Cosa sono quelle immagini?

La riflessione sul nome non è peregrina. Se il termine fotografia, come "scrittura di luce", ha richiesto tempo e dibattito prima di affermarsi (Hilllnhuetter *et al*., 2021), in qualità di "scrittura di luce", "impronta ottica" del mondo esterno, lo scatto è innanzitutto frutto di un occhio umano, presente in un contesto, a fronte di una percezione sensoriale autentica, diretta, di un evento fissato visivamente.

L'immagine generata dall'intelligenza artificiale nasce invece da una combinazione di scatti immagazzinati come dati, elaborati statisticamente mediante algoritmi neurali. L'esito è un prodotto indiretto, non prevedibile. E questo, sia perché letteralmente non è stato visto prima – offrendo al verbo una lettura agenziale – sia perché frutto di un'elaborazione interna, effettuata da una "super black box": camera oscura all'ennesima potenza, con una certa opacità del procedere macchinico (Müller-Pohle, 2024). La distanza tra i due procedimenti è profonda e incolmabile: il primo esito di immagini di luce, il

secondo di dati. E, per chiarire l'equivoco, emergono termini più appropriati, come sintografia e promptografia – non ancora troppo diffusi –, ciascuno capace di mettere in evidenza aspetti dirimenti.

Sintografia, da un lato, valorizza la chiave sintetica (ossia di sintesi) che la macchina opera nel restituire un prodotto emergente da altre immagini già esistenti, una sorta di amalgama visiva. *Promptografia*, dall'altro, pone attenzione sull'azione generativa a partire da un input, il prompt, le istruzioni grafiche o testuali con rilievo sull'intervento attivo dell'utente nella pratica *text-to-image* (TTI).

I programmi attingono a un bacino di fotografie, alimentato in modo esponenziale dalla cultura digitale, rintracciate in rete o caricate dagli stessi "creatori" di cui tuttavia non resta che un sospetto infinitesimale, rielaborato algoritmicamente: «diluito omeopaticamente come una goccia di sangue nell'oceano» (Müller-Pohle, 2024: 4).

In quanto prodotto di algoritmi neurali e dati elaborati statisticamente, le immagini elaborate o "assistite" per mezzo di un computer – d'ora in poi VgenAI, con V per visive – presentano notevoli differenze tecniche e procedurali rispetto alla fotografia. Tuttavia, allo sguardo può essere difficile percepirle. E qui risiede il problema principale: si tratta di *simulazioni* di fotografie.

Per definire un testo visivo come "fotografia" – analogica o digitale, realizzata con una macchina fotografica o con uno smartphone – il punto centrale resta la scrittura (o forse dettatura?) di un evento di luce catturato: un'impronta ottica del mondo esterno. La fotografia è tale perché si accompagna alla percezione sensoriale di un attore umano e, soprattutto, alla sua presenza e relazione diretta con un contesto/evento catturato dalla luce. Su questa scia, l'avvento della strumentazione digitale (oggi incorporata nei nostri smartphone) può esservi ancora assimilato: quando si fotografa, «è sempre la luce a scrivere l'immagine» (Marra, 2006; Losacco, 2010: 32) – per quanto siano ormai necessarie nuove riflessioni sulla natura non indicale, ma piuttosto sempre rappresentativa dell'immagine.

Così, se ancora Müller-Pohle, aderendo al discorso tecnico-fotografico, definisce i due tipi di immagini come "fratelli diseguali" – simili in apparenza, ma non nella sostanza (sembrano foto) – è forse il contributo di altri studi, sul versante sociologico ad aiutarci a riflettere sull'imparentamento reso sempre più prossimo, e pericolosamente indistinguibile, dalla già avviata cultura visiva *screen* (Losacco, 2018) dove la fotografia è da tempo dematerializzata e ricontestualizzata nelle pratiche sociali globalizzate dello scatto. *Social photo*, come la definisce Jurgenson (2019), è così uno dei modi con cui si allude non solo all'aggiornamento della fotografia – diventata

fluida, fuggevole, interplanetaria –, ma anche alla trasformazione delle stesse pratiche sociali che ci vedono sempre più coinvolti come fruitori e creatori di immagini, con impatto sulla ridefinizione stessa di ciò che consideriamo vero, significativo. E, per inciso, la foto di Capa, a Budapest, è oggi esposta in versione digitale: su schermo.

Dunque, se per il fotografo o il creativo il nodo resta quello dell'autorialità, per la sociologia visuale, invece, cosa è davvero in gioco?

3. Detentori legittimi di verità e simulatori

Ci pare che la questione, cosa è implicato "davvero" abbia a che vedere con lo statuto ontologico della realtà, sancito da un mezzo. Scettica, dunque, all'uso delle fotografie proprio per questa fiducia ingenua, la sociologia visuale sviluppa il suo apparato teorico concettuale e metodologico dalla seconda metà degli anni '70 (Faccioli, Losacco, 2010) distanziandosi dalla retorica delle immagini fotografiche come specchi di realtà e rimettendole in scena come rappresentazioni. La mediazione effettuata sul campo per ottenere la fotografia "di luce" non garantisce infatti al suo autore (ricercatore/i) né che uno scatto sia di interesse sociologico, né che una interpretazione possa essere data di per sé, e dunque univoca, oggettiva, o che tale scatto non possa essere soggetto a manipolazione, o sospettabile come "falso".

Nell'uso delle immagini nella ricerca (diretta o indiretta) è propriamente il rapporto con la "realtà" e dunque la sua rappresentazione convincente a dover essere approfondita: affidarla alla fotografia è una mitologia. Con gli occhi e con le parole, scrive Touraine, e con questo adagio, ricordiamo che l'elemento che più deve interessare non è tanto questa mancanza di occhio umano e intenzionale dietro al mirino, né per converso quella di un occhio consapevole davanti. Il digitale e le pratiche annesse, ci hanno inoltre condotto a un nuovo modo di relazionarci alla significatività degli scatti, oggi dalla fluenza copiosa, che nel riverberare l'attenzione al dettaglio, riprende anche la pratica della foto involontaria, non prevista, accanto a quella voluta. Come la foto scattata da Capa che in una delle versioni fornite sarebbe stata ottenuta dalla camera salda sulla testa, lui intento a correre al riparo dagli spari, non a prendere la mira. Un caso fatalmente fortuito, questo, un lavoro di scavo quello della ricerca sociologica.

Ciò che ci riguarda, come sociologi, pur con il cambiamento dei mezzi tecnologici, è comunque il dar conto ad una triangolazione emergente e progressiva nel procedere *induttivo* dello sguardo. Sia esso *con* le immagini (fatte dal ricercatore o dai soggetti coinvolti in un percorso di ricerca visuale) sia essa *sulle* immagini (ossia volta a testi visivi esistenti al di là della ricerca,

a cui la riflessione sociologica possa interessarsi e farne focus di indagine). In ogni caso, nessuna verità autoevidente o certa è posseduta dal ricercatore (che non sa prima, e proprio per questo svolge una ricerca qualitativa), dall'osservato o nel frame stesso dello scatto. Per questa ragione, il focus dell'indagine sociologica visuale si muove grazie a inferenze, lavora su dettagli apparentemente banali, non previsti, non prevedibili, come concessione di significati in divenire, via via ricalibrati, aperti all'interpretazione e comprensione, a fornire domande rinnovate, più che risposte. Uno degli obiettivi della ricerca è collocare il minuto dettaglio nella grammatica di una cultura, per cui abbiamo bisogno di esplorare un contesto, farne affiorare l'ossatura mai data di per sé, lasciarla trasparire collegando il singolare al generale e, viceversa, il generale a quel singolare, rilevarne linee guida e le umane, possibili contraddizioni. E ancora, buttare molti scatti – non tutti sono sociologici – e, a volte, mantenere quelli più brutti, sfuocati, mossi, bui.

Se le immagini sono sempre mediazioni, rappresentazioni e mai specchi di realtà, e ancora in qualche modo sempre manipolabili costitutivamente, perché preoccuparci dell'IA? In questo senso, l'intelligenza artificiale non introduce un nuovo dilemma epistemologico, ma porta a compimento una trasformazione già in atto: la progressiva crisi dello statuto di verità della fotografia. C'è da esserne entusiasti.

4. Il ricercatore al tavolino

Dovrebbe richiamare memorie di altri tempi la pratica di descrivere culture altre, lontane nello spazio e per valori, affidandosi alla testimonianza apparentemente inequivocabile dello scatto. Non solo fotoreporter, ma generazioni di politici e antropologi ante litteram potevano comodamente porsi – e lasciare il loro pubblico – al riparo dei loro abitacoli, dietro le scrivanie, e da lì argomentare verità "illustrandole" in virtù del potente strumento tecnologico. Malinowski, figura emblematica e apparentemente aperta al relativismo culturale, risulta invece manipolativo: ricorre alla fotografia come prova e sentenza (Malighetti, Molinari, 2016), supportando con essa una visione del mondo che pretende di essere ma che, oggi, rileggiamo come profondamente ideologica.

Su quel filo, ma per contrasto, si è sviluppata una crescente consapevolezza volta alla riappropriazione critica dell'immagine fotografica, riconosciuta come dispositivo di potere in molteplici studi di stampo anticolonialista, antirazzista, femminista, intersezionale (Jääskeläinen *et al*., 2025). In comune, l'idea che la fotografia stia al centro del discorso vero/falso, luce/ombra: ciò che illumina e ciò che lascia al buio diventa trasposizione visiva

degli intenti razionalizzanti dell'illuminismo accolti dalla modernità e mercificati dal capitalismo (Coleman, James, 2021; Colberg, 2021). L'estetica dello scatto, da svelare nella sua potenza globale, è carica di uno sguardo codificatore, affatto neutrale.

Torniamo alla foto di Capa. Analogica, scattata forse da una Leica, forse da un'altra macchina, forse una Rolleiflex, l'ha scattata lui? Quando? Dove? Si tratta di una messa in scena? Chi è quel miliziano? A fronte di una sorta di processo in contumacia, domande e risposte cambiano in base alla traiettoria; quello scatto racconta la guerra, ma anche l'esordio di metodologie narrative moderne, e ancora di ideologia. È esempio di un racconto "da straniero", un gesto di coraggio, una simulazione recitata per l'obiettivo? Il fotogramma si offre come verità e, insieme, come illusione della verità, una versione narrativa che instaura tra fotografia e realtà (come tra Capa e il suo scatto) la fascinazione verso «la vertiginosa vicinanza alla verità del momento» (Lacouture, 2013: 3).

Non è l'avvento del digitale ad aver introdotto sospetti sulla "realtà" fotografica. I fili analogici che legavano la fotografia al mondo "là fuori" sono oggi "rimediati" nelle pratiche digitali della social photography. In questo contesto, emerge la figura di spett-autore che attraverso costruzioni partecipate, ridefinisce le logiche della rappresentazione e risignifica quelle della manipolazione. Dal punto di vista dello sguardo sociologico, non è tanto "raccogliere scatti", quanto la capacità di addentrarsi, progressiva, nella grammatica sociale che rende sensate quelle visioni. Non tutte le foto sono sociologiche (Becker, 2007).

4.1. Il ritorno al tavolino

Si stima che siano già stati creati diversi miliardi di immagini utilizzando algoritmi di conversione *text to image*; la fascinazione è grande, così come le potenzialità di questi strumenti. Non sono solo gli artisti a compiere i passi più spericolati.

Il fotografo belga Carl De Keyzer (2024) usa l'intelligenza artificiale e lo dichiara[1]; il libro, tuttavia, non contiene testo scritto e – come del resto accade in molti testi fotografici – non presenta didascalie; solo il titolo, *Putin's Dream*, offre un suggerimento sulla natura del suo contenuto. Senza quello, si direbbe la documentazione visuale di un corposo reportage di viaggio, tra scene di vita quotidiana e momenti bellici. Ampi rilievi di cronaca, in Belgio, ne precedono l'imminente pubblicazione grazie ai quali sappiamo che l'autore, complice la pandemia, il conflitto in Ucraina e il crollo dei ricavi del fotogiornalismo nell'era digitale, ha scelto di lavorare al portatile, seduto

nella "sua cucina" (Stevenson, 2024). Un viaggio virtuale che pure ha richiesto all'autore di immettere in un generatore artificiale le sue stesse foto scattate in Russia, in più occasioni alla fine degli anni '80, e offrirle circa 30 anni dopo alla macchina, addestrandola con i giusti prompt, a creare nuove immagini con il suo stile autoriale, comprensivo di attenzione a rituali e passatempi allora immortalati (Stevenson, 2024). L'esito sono immagini che sembrano fotografie, ma non lo sono, che propongono scene verosimili, eppure mai esistite; un alternarsi curato, graficamente sofisticato, ad alta definizione, tra momenti di guerra e vita quotidiana, che mescola combattenti in divisa e giovani donne sulla riva del mare, bombe all'orizzonte, bambini tra idranti che sputano fuoco, donne anziane curve e ragazze dall'estetica mainstream, perfette queste, sgraziate le altre. Simboli politici, stereotipi visivi e occhi sempre azzurri, fissi verso un obiettivo fittizio, che abitano un mondo fittizio, senza storia, fuori luogo.

Pubblicato parzialmente su Instagram, e cancellati i post dopo il feedback feroce. A differenza di Nine Eyes di Jon Rafman (2009), che estrae immagini casuali di momenti vissuti e trascurati da Google Street View, De Keyzer produce un archivio esteticamente coerente, ma inquietante. Il sogno che ha messo in scena, a chi appartiene? Non ci appartiene, sembrano commentare le persone. È verosimile, e forse è proprio questa ambiguità[3] a renderlo perturbante. Creato al tavolino, stavolta da un esploratore di sogni artificiali. Perché tanto fastidio?

Nel paragrafo successivo ci interroghiamo su questo spaesamento visivo, e sulla distanza che separa l'umano dal quasi-umano. Sul fatto che la riconosciamo, ma potremmo anche non farlo.

5. *Ai to Eye*: macchine sognatrici e allucinazioni

Le macchine non vedono: sono addestrate a processare immagini secondo logiche algoritmiche, statistiche, cumulative. Operano sul codice visivo traducendolo in forma quantitativa, compatibile con il processore. Non possiedono comprensione autonoma, né criteri interni per distinguere il vero dal falso, il rilevante dall'irrilevante. Tutto ciò che dà senso alla cultura umana – valori, credenze, priorità, linguaggi, umorismo – deve essere fornito e codificato dall'esterno. Se gli input sono parziali o stereotipati, anche le visioni artificiali e derivate, lo saranno.

[3] Emblematico l'uso inizialmente non dichiarato dell'IA nel caso *The Electrician* (https://www.eldagsen.com).

Verosimili ma disturbanti: le immagini prodotte dall'IA, pur nella loro somiglianza al reale, possono suscitare una reazione inquietante. Noto come effetto "*Uncanny Valley*" (Raymond, 2020), questo disagio – oggetto di studi neuroscientifici – nasce da elementi percepiti come discordanti, e si associa a emozioni negative e comportamenti di evitamento, in particolare verso volti artificiali, robotici (Di Natale *et al*., 2023), ma non solo. Deriva dal riconoscere qualcosa come "fuori posto", cosa possibile proprio perché conosciamo bene qual è il "suo" posto. Ma se questo riferimento al noto mancasse – se osservassimo un contesto nuovo – si manifesterebbe ancora? Si trasformerebbe?

Considerate come errori di sistema, le "allucinazioni" visive sono parte integrante del funzionamento dell'IA generativa. L'uso del termine antropomorfico riflette una nostra proiezione: le allucinazioni attribuite al sistema in effetti sono ciò che lo sguardo umano *riconosce* come estraneo, per confronto, perché ha un'idea storicamente e culturalmente situata. Downey (2024) usa a grandi linee questo ragionamento per riprendere, da una prospettiva più umanistica e comprensibile, ciò che paventa lo sviluppo estremo e incontrollabile, non più arginato dall'effetto "Uncanny Valley" che può apparirgli così, oggi, una sorta di sentinella ancora disponibile, di riparo dall'artificiale. E per dare sostanza esemplificativa al ragionamento riprende uno dei lavori creativi di un artista, rappresentato da alcune trasformazioni progressive e "allucinate" nella produzione di nuove immagini con una Gan. Il sistema bipartito, grazie a due reti antagoniste, da una parte classifica oggetti visivi (fotografie inserite) in base alle istruzioni ricevute per la catalogazione; dall'altra discrimina la sua stessa classificazione mediante la rete avversaria. Dal suo canto, l'artista otterrà immagini finali di tramonti con occhi (palesemente allucinati, dal punto di vista di chi sa che il sole, al momento, non mostra occhi umani), ma il punto che ci interessa, seguendo Downey, è come si siano originati. Se infatti la macchina non sogna, né pensa se non in funzione di istruzioni ricevute, vuole capire dove e come possa ottenere quegli esiti, cosa accade nella "pancia" della macchina, nella scatola nera. Da una progressiva catalogazione di immagini ma commettendo via via piccoli impercettibili errori (nella grande mole inserita) di attribuzione; in particolare è la corrispondenza tra concetto simbolico e foto dal riconoscimento meno esplicito e/o ambiguo per la macchina. Così il punto dell'articolo di Downey che qui ci interessa fare emergere è come colmare, o prepararsi a colmare, quella discrasia con il fenomenico che l'avanzare della nostra iperrealtà (Baudrillard, 1994) non ci consentirà più di riconoscere come fuori posto. La domanda che pone è inquietante: man mano che queste anomalie si accumulano, con simulazioni sempre più raffinate, saremo ancora in grado di riconoscere l'errore come tale? Riconoscere allucinazioni?

Lo stimolo è interessante. Lo riportiamo al sapere visuale sociologico, per chiederci piuttosto: l'analisi delle immagini del mondo (fotografate o generate) potremo e possiamo davvero operarla seduti al riparo dei nostri pc, con le nostre categorie concettuali? Nelle immagini del fotografo belga, respinte con forza perché palesemente riconosciute nella loro "stranezza" – come potrebbe il futuro essere delineato a partire da istruzioni sul passato? – saremo in grado di riconoscere il riversarsi di cluster visivi stereotipati, uniformanti, che perpetuano sguardi codificati ideologici, ereditati nella programmazione e spia di norme visive trasferite alle tecnologie AI. Qui il rischio di riprodurle e rinforzarle (Buolamwini, 2023; Jääskeläinen *et al.*, 2025) verrebbe contrastato suggerendo di prestare migliore attenzione e magari una formazione critica e attenta all'inclusione. Ma se non abbiamo termini di confronto, come faremo a discriminare e allontanare la banalizzazione statistica sul complesso e contraddittorio dell'umano sempre in divenire?

Conclusioni

Il mondo visivo dell'intelligenza artificiale segna una fase nuova e dirompente della digitalizzazione. La riflessione proposta – necessariamente parziale – si è concentrata su alcune implicazioni sociali legate alla nuova generazione di immagini che simulano la fotografia, pur essendo tecnicamente altro. Se per oltre un secolo e mezzo la fotografia ha alimentato una narrativa centrale nel produrre e rinsaldare un discorso della verità, è proprio questa associazione a tornare oggi sotto i riflettori.

Le numerose riflessioni teoriche sul carattere artefatto, mediato, costruito dell'immagine fotografica sembrano rimaste per lo più confinate in testi accademici specialistici e assorbite nei propri domini disciplinari. A incrinare la reputazione della fotografia come strumento neutro e veritativo – capace di produrre "prove" visive affidabili, storie inequivocabili – è dovuto intervenire lo spettro dell'artificio macchinico?

Nel solco del percorso intrapreso, gli spunti conclusivi si dirigono verso le opportunità che questa domanda lascia sul tappeto.

In prima istanza l'appello a una efficace, accessibile, divulgata esigenza di educazione visiva, come alfabetizzazione al linguaggio delle immagini. Prima ancora di interrogarsi sulle creazioni visive dell'IA come problema o potenzialità, tale prospettiva sollecita a riflettere sul mito – ancora vivo – della verità incarnata (o tradita) nelle immagini fotografiche e mediatiche. La fotografia, a lungo percepita come testimonianza oggettiva, codice senza codice, priva di mediazione, continua a esercitare una forte influenza. È su

questa idea che le immagini artificiali sembrano oggi agire, generando rigetto o fascinazione. Educare lo sguardo significa riconoscere che non esiste una "visione neutra", che anche lo scatto più realistico è frutto di scelte, filtri, manipolazioni – volontarie o inconsapevoli – all'interno di una trama di condizioni storiche e culturali ritenute sensate. La "macchina della verità" va decostruita; la polisemia dell'immagine colta nella sua ricchezza e nelle sue contraddizioni.

In seconda istanza, una voce più significativa per la disciplina della sociologia visuale e alle sue potenzialità nel contribuire al dibattito sulle trasformazioni in corso. Ciò che oggi definiamo inquietante – uncanny – potrebbe non esserlo più domani. Le reazioni di rigetto verso le immagini artificiali non derivano solo da fattori cognitivi o biologici, ma da una concezione culturalmente situata di ciò che è "normale" e di ciò che è "strano". È qui che la prospettiva sociologica torna essenziale: ci invita a considerare che l'interpretazione e la comprensione delle immagini – come quella di ogni fenomeno sociale – dipendono dalla relazione dinamica tra individui e contesto, sempre in mutamento. Le contraddizioni e le aporie che caratterizzano una complessità sempre più fuori controllo (Beck, 2023) aprono a modi di vedere nutriti da immaginari artificiali. Questa espressione, che allude a una supposta separazione dal "naturale", può essere invece l'occasione per esplorare senza barriere preclusive l'intreccio percettivo, le contaminazioni e le stesse ridefinizioni di ciò che riteniamo reale (Zylinska, 2023; Wynants, 2020). Una sensatezza mai data una volta per tutte, mai ristretta a riflessi istintivi, e che proprio l'esplorazione del linguaggio visivo può contribuire a ravvivare. Su questa scia, infine, un dialogo reale, concreto, tra domini disciplinari sarebbe auspicabile.

Le immagini artificialmente intelligenti non sono – per ripetere – fotografie. Possono fingere di esserlo, così come le fotografie possono fingere di essere realtà. Occorre uscire da questo vicolo cieco. Lasciar andare il soldato che cade per innamorarcene di nuovo: non per le certezze delle risposte, ma per le domande che solleva.

Riferimenti bibliografici

Baudrillard J. (1994). *Simulacri e simulazione*. Milano: Feltrinelli.

Beck U. (2003). *La società del rischio. Verso una seconda modernità*. Roma: Carocci.

Becker H.S. (2007). *I trucchi del mestiere. Come si fa ricerca sociale*. Bologna: Il Mulino.

Colberg J. (2021). *Photography's neoliberal realism*. London: Mack. Testo disponibile al sito: https://mackbooks.co.uk (consultato il …).

Coleman K., James D. (2021). *Capitalism and the camera. Essays on photography and extraction*. London-New York: Verso. Testo disponibile al sito: https://www.versobooks.com (consultato il …).

De Keyzer C. (2024). *Putin's dream*. Ghent: Self-published.
Di Natale A.F., Simonetti M.E., La Rocca S., Bricolo E. (2023). Uncanny valley effect: a qualitative synthesis of empirical research to assess the suitability of using virtual faces in psychological research. *Computers in Human Behavior Reports*, 10. DOI: 10.1016/j.chbr.2023.100288.
Downey A. (2024). The return of the uncanny: artificial intelligence and estranged futures. *Visual Studies*, 39: 1-10. DOI: 10.1080/1472586X.2024.2406709.
Eldagsen B. (2023). *Sony World Photography Awards 2023*. Sito personale dell'autore. Disponibile all'indirizzo: https://www.eldagsen.com/sony-world-photography-awards-2023/ (consultato il 7 dicembre 2024).
Engelke A. (2023). The falling soldiers. *European Photography. The International Art Magazine for Contemporary Photography and New Media*, 114: 24-25. Disponibile all'indirizzo: https://equivalence.com/european-photography-114 (consultato il …).
Faccioli P., Losacco G. (2010). *Nuovo manuale di sociologia visuale. Dall'analogico al digitale*. Milano: FrancoAngeli.
Faccioli P., Losacco G. (2018). *Sociologia visuale. Teorie, metodo e ricerca*. Milano: FrancoAngeli.
Hillnhütter S., Klamm S., Tietjen F., a cura di (2021). *Hybrid photography. Intermedial practices in science and humanities*. New York: Routledge (Routledge History of Photography Series).
Jääskeläinen P., Sharma N., Pallett H. (2025). Intersectional analysis of visual generative AI: the case of Stable Diffusion. *AI & Society*. DOI: 10.1007/s00146-025-02207-y.
Jurgenson N. (2019). *The social photo. On photography and social media*. New York: Verso.
Lacouture J. (2013). *Verso la foto-storia*. Verona: Contrasto.
Losacco G. (2018). *Sociologia visuale e studi di territorio*. Milano: FrancoAngeli.
Malighetti R., Molinari A. (2016). *Il metodo e l'antropologia. Il contributo di una scienza inquieta*. Milano: Raffaello Cortina Editore.
Müller-Pohle A. (2024a). Artificially intelligent image world. *European Photography*, 114, vol. 44, Winter 2023/2024. Berlin. Disponibile all'indirizzo: https://equivalence.com/artificial-intelligence (consultato il …).
Müller-Pohle A. (2024b). *Niépce recoded*. With an essay by Bernd Stiegler and a project description by Andreas Müller-Pohle. Berlin: Equivalence.
Rafman J. (2009). IMG MGMT: The Nine Eyes of Google Street View. *Art F City*, New York, 12 August 2009.
Raymond C. (2020). *The photographic uncanny. Photography, homelessness, and homesickness*. Leuven: Leuven University Press.
Stevenson R. (2024). A photographer created "fake" images of Russia with generative AI. Now he's losing his biggest fans. *ABC News*. Disponibile all'indirizzo: https://www.abc.net.au/news/2024-12-26/ai-generated-images-photography-trust/104721106 (consultato il …).
Susperregui J.M. (2009). *Sombras de la fotografía*. Bilbao: Universidad del País Vasco.
Whelan R. (2002). Proving that Robert Capa's "Falling Soldier" is genuine: a detective story. Disponibile all'indirizzo: https://www.pbs.org/wnet/americanmasters/robert-capa-in-love-and-war/47/ (consultato il …).
Wynants N., a cura di (2020). *When fact is fiction. Documentary art in the post-truth era*. Amsterdam: Valiz (collana "Antennae – Arts in Society").
Zylinska J. (2023). *The perception machine. Our photographic future between the eye and AI*. Cambridge (MA)-London: MIT Press.

Francesca Guarino

Fonti visive consultabili (non riprodotte)

Capa R. (1936). *The Falling Soldier (Death of a Loyalist Militiaman),* International Center of Photography International Center of Photography, https://www.icp.org/collection/objects/the-falling-soldier-1936

Immagine 1 - Animazione della fotografia di Robert Capa, *The Falling Soldier* (1936). Video realizzato da Francesca Guarino con Runway Gen-4 AI (24/06/2025), a partire dallo scatto originale. Link diretto: https://youtu.be/laCtt84hZc0; L'uso della fotografia originale è a fini accademici e non commerciali.

Immagine 2 - Animazione della fotografia di Robert Capa, *The Falling Soldier* (1936). Video realizzato da Francesca Guarino con Runway Gen-4 AI (24/06/2025), a partire dallo scatto originale. Link diretto: https://youtu.be/iejlhGw7OhM; L'uso della fotografia originale è a fini accademici e non commerciali.

Engelke A. (2023). *The Falling Soldiers*. *European Photography*, n. 114: 24-25. Serie generata tramite AI a partire dalla descrizione testuale della foto di Capa. https://equivalence.com/european-photography-114

Stevenson R. (2024). *A photographer created 'fake' images of Russia with generative AI. Now he's losing his biggest fans*, ABC News. (articolo con alcune immagini del progetto *Putin's Dream* di Carl De Keyzer. https://www.abc.net.au/news/2024-12-26/ai-generated-images-photography-trust/104721106

Artificial companions? AI, mental health and the sociological reconfiguration of human relationships

by Vera Kopsaj[*]

Emotional life today is shaped more and more by digital technologies. Therapeutic chatbots and automated diagnostics now occupy intimate corners of mental health care, prompting new expectations of support and connection. This article asks whether AI interactions can recreate or reconfigure experiences of friendship, trust and care. The analysis brings theoretical perspectives from relational sociology and critical algorithm studies into dialogue with emerging empirical research on users' interactions with mental-health chatbots. Drawing on this combined lens, the article explores how people invest emotionally in systems designed to imitate empathic attention and considers the implications of predictive monitoring for digital subjectivity. Rather than treating AI as a replacement for human ties, it argues that these systems function as socio-technical actors within an ecology of care, subtly reshaping emotional norms and social inequality.

Keywords: artificial intelligence; mental health; therapeutic chatbots; simulated interlocutors; friendship; sociology.

Compagni artificiali? Intelligenza artificiale, salute mentale e riconfigurazione sociologica delle relazioni umane

Oggi la vita emotiva è sempre più mediata dal digitale. I chatbot terapeutici e le diagnosi automatizzate occupano ormai gli angoli più intimi della cura della salute mentale, alimentando nuove aspettative di supporto e di connessione. Questo articolo si chiede se le interazioni con l'IA possano ricreare o riconfigurare le esperienze di amicizia, fiducia e cura. L'analisi mette in dialogo la sociologia relazionale e gli studi critici sugli algoritmi con le ricerche empiriche emergenti sulle interazioni tra utenti e chatbot. Attraverso questa lente combinata vengono esaminati gli investimenti emotivi degli utenti in sistemi progettati per simulare un'attenzione empatica e vengono considerate le implicazioni del monitoraggio predittivo per la soggettività digitale. Piuttosto che concepire l'IA come un sostituto dei legami umani, si sostiene che queste tecnologie funzionino come attori socio-tecnici in un'ecologia della cura, contribuendo a ridefinire in modo sottile le norme emotive e le disuguaglianze sociali.

Parole chiave: intelligenza artificiale; salute mentale; chatbot terapeutici; interlocutori simulati; amicizia; sociologia.

DOI: 10.5281/zenodo.18436066

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

Vera Kopsaj

Introduction

In recent years, artificial intelligence (AI) has moved beyond the realm of technical infrastructure to become a pervasive presence in everyday life, shaping not only economic processes and security systems, but also domains historically rooted in human intimacy, such as emotional care (Lee *et al*., 2022) and mental health (Vicci, 2024). One of the most important developments is the rise of therapeutic chatbots and digital platforms offering psychological support and behavioural interventions. Systems such as Woebot and Wysa, among others, are prized for being accessible, inexpensive and available 24/7 in the context of increasing mental health needs.

However, their deployment invites deeper sociological reflection on the meanings, risks and transformations associated with the delegation of emotional labour[1] to machines.

This article starts from the premise that AI-driven mental health tools are not merely technological tools, but socio-technical actors actively involved in reshaping care relationships, therapeutic authority and the experience of emotional vulnerability. By simulating empathic listening and affective presence, these systems give rise to new forms of mediated sociality, in which users can come to experience digital agents not only as utilities, but as relational partners. Far from being a mere illusion or anthropomorphic projection, this phenomenon must be placed within a broader cultural and structural context, characterised by the fragmentation of traditional support systems (familial, institutional, professional) and the increasing prevalence of loneliness and emotional precariousness in late modern societies. In this context, the article addresses a central sociological question: to what extent can interactions with AI systems replicate or reshape the human experience of friendship, trust and emotional support?

Drawing on theoretical contributions from relational sociology (Donati, 2011; Emirbayer, 1997), digital sociology (Lupton, 2016), and science and technology studies (Jasanoff, 2004), the article explores how AI mental health tools reshape fundamental categories of sociological enquiry: trust, recognition, care, and the boundaries of the human. We ask: what does it mean to entrust one's emotional vulnerability to a machine? Can algorithmic empathy be considered a form of 'relational good' (Donati, 2011) or does it reinforce a logic of simulation and emotional externalisation? How

[1] In this article, the term 'emotional labour' is used in a broad sense, including both the professional work of healthcare workers and the everyday emotional labour that people do in non-professional contexts.

do predictive diagnostics and data-driven self-monitoring systems reconfigure the experience and management of psychological suffering?

To answer these questions, we begin by examining the architecture and functionality of popular therapeutic chatbots, analysing how they frame mental distress in computational terms and offer standardised interactions that mimic therapeutic dialogue. We then investigate the ways in which users report emotional attachment, comfort and even forms of addiction in relation to these systems −raising critical questions about the commodification of emotional labour and the ethics of affective automation. Finally, we address the governance implications of AI-based diagnostics and risk assessment for mental health, which operate through opaque algorithms and biometric data collection, shaping new forms of 'digital subjectivity' and empowered self-care.

Rather than a normative stance, the article adopts a sociological lens that situates AI mental health tools within a broader ecology of care. It recognises their utility while examining the social conditions and epistemologies that shape them, and how these tools mediate both clinical and social relationships.

1. Therapeutic chatbots: architecture, promises and sociotechnical imaginaries

Therapeutic chatbots operate at the intersection of artificial intelligence, psychology and mobile health. These systems use natural language processing (NLP), sentiment analysis and behavioural logic to simulate therapeutic dialogue and provide low-cost, scalable mental health support. Examples such as Woebot[2] and Wysa aim to reduce pressure on healthcare systems by offering users 24/7 access to emotionally responsive guidance (Ni, Jia, 2025; Chang *et al*, 2024; Khawaja, Bélisle-Pipon, 2023; Lang, 2021). However, on July 2, 2025, Woebot Health officially shut down its flagship product. While Woebot was once considered a pioneer in digital mental health − used by over 1.5 million people – the chatbot was eventually overtaken by more flexible generative AI tools such as ChatGPT. As its founder acknowledged, AI is advancing faster than the regulatory and clinical frameworks designed to contain it, raising new questions about safety,

[2] Available at: https://spectrum.ieee.org/woebot?utm (accessed on June 20, 2025).

supervision and effectiveness in emotionally sensitive areas (Aguilar, 2025)[3].

A study by Chang *et al*. (2024) found that Wysa, for instance, was positively received by health workers in Singapore during the COVID-19 pandemic. Over 80% of participants engaged in multiple sessions reported high levels of satisfaction. Interventions targeting sleep and anxiety were among the most widely used, underlining the importance of application for the pressures frontline staff face. These findings suggest that AI-based mental health tools such as Wysa can effectively complement traditional services, particularly for individuals with mild to moderate distress, by offering scalable and accessible support.

The COVID-19 pandemic marked a turning point in mental health care. Previously, therapy was predominantly face-to-face and digital tools played a marginal role. The change was not only technological, but also cultural, altering help-seeking behaviour and normalising virtual platforms, including therapeutic chatbots (Garofalo, 2024).

This change is particularly relevant in the context of a global mental health crisis. As Abd-Alrazaq *et al*. (2019) note, mental illnesses are a key factor behind disability on a worldwide scale and the demand for treatment far outstrips the available services. Therapeutic chatbots are therefore positioned as accessible and scalable solutions for underserved populations.

These systems are distinguished not only by their technological sophistication, but also by their ability to simulate the therapeutic presence in the absence of a human being. Based on cognitive-behavioural therapy (CBT) protocols, they guide users through structured and modular exercises, such as mood monitoring and cognitive reorganisation, regulated through feedback loops.

These protocols do not function in an abstract manner: they are delivered through conversational language designed to evoke a sense of emotional closeness. To provide a clearer picture of how these interactions unfold, it is helpful to consider some typical examples of the suggestions and exercises proposed by therapeutic chatbots. CBT-based systems, such as Wysa, often invite users to identify a distressing thought ('I failed my presentation'), evaluate the evidence for it, and reframe it in a more balanced way ('I struggled today, but I have succeeded at similar tasks in the past'). Many apps also incorporate micro-exercises for emotional regulation,

[3] Aguilar (2025). *Woebot's therapy chatbot shuts down as AI evolves faster than regulation*. STAT News. Available at: https://www.statnews.com/2025/07/02/woebot-therapy-chatbot-shuts-down-founder-says-ai-moving-faster-than-regulators/?utm (accessed on July 3, 2025).

including grounding techniques ('Try taking three slow breaths with me') or suggestions that encourage self-compassion ('What would you say to a friend who feels this way?'). In addition to these structured tools, chatbots rely on a pseudo-empathetic tone ('I'm really sorry you feel that way', 'That sounds really difficult, but I'm here with you'), sometimes accompanied by emojis and affective cues calibrated to produce a perception of warmth and support. These elements demonstrate how the interaction is not merely functional, but also affective, helping to give the impression of a relational presence despite the absence of a human interlocutor.

This raises sociological questions about changing therapeutic authority. Traditional therapy is dialogic and interpretive, rooted in professional care norms. Chatbot-mediated therapy replaces it with a scripted, data-driven exchange in which the 'therapist' is an emotionally reactive interface that detects patterns and implements standardised interventions (Khawaja, Bélisle-Pipon, 2023).

Consider the following Wysa's website quote. From a sociological perspective, these data suggest that structural barriers − such as stigma, lack of awareness and time constraints − continue to limit access to mental health support, with less than 7 per cent of employees using Employee Assistance Programmes (EAPs), despite nearly 40 per cent experiencing symptoms of depression or anxiety. Although 42 per cent of users opened up about their mental health during interactions with Wysa, it is best understood as a support tool on an individual level, not systemic change. Chatbots may be useful to manage or mitigate distress, but they are not designed to prevent it. To truly reduce the burden of depression and anxiety, more attention needs to be paid to the *social determinants of mental health* such as working conditions, economic insecurity, isolation and cultural stigma. Investment in digital care must go hand in hand with structural transformation policies.

> Our research shows that as many as 4 in 10 employees suffer from symptoms of depression or anxiety, yet less than 7% access EAP due to stigma, lack of awareness and time constraints. While talking to Wysa, 42% of employees opened up about their declining mental health[4].

This new therapeutic model aligns with critical accounts of digital capitalism, particularly what scholars such as Srnicek (2017) van Dijck *et al*. (2018) and Zuboff (2019) describe as commodification and modularisation

[4] Available at: https://www.wysa.com/ (accessed on July 7, 2025).

of human experience. Emotional labour becomes quantifiable, predictable and consumable on demand. Therapeutic chatbots are not autonomous; they are scripts coded and modelled by developers and psychologists within a platform economy. Their design reflects dominant imaginaries: distress is a given, treatment is modular, and well-being is algorithmically manageable.

These systems are designed to mimic human interaction. Many use friendly names, emoji, adaptive tones and conversational style. Woebot, for example, presents itself as a cheerful and witty companion. These features are key to generating a sense of relational presence, encouraging users to suspend disbelief and engage emotionally.

The emotional impact of this simulation deserves sociological attention. Users often report feeling listened to and less alone, despite knowing that the chatbot is not real. This paradox challenges the traditional view of intersubjectivity, echoing Turkle's (2011) assertion that many today prefer the illusion of companionship without the demands of friendship.

However, this illusion is not neutral. The simulation of empathy in therapeutic chatbots is carefully calibrated through User Experience Design (UX design), linguistic cues and psychological modelling. As such, it reflects an engineered form of affection, tailored to calm, motivate and retain the user. In this way, it can shape not only how users relate to the chatbot, but also how they come to understand emotional support in general. What happens when support becomes a feedback loop? When the value of an emotional exchange is measured in terms of user satisfaction or behavioural adherence?

This paragraph therefore lays the groundwork for the next investigation: what kind of relationship is formed when users begin to attribute social meaning to these interfaces? Can the chatbot be considered a relational other, a stand-in for friendship, empathy or care? And what are the ethical and social consequences of this relational shift?

Although specific tools come and go, what remains is the sociological reconfiguration of care, in which emotional labour is increasingly automated, standardised and embedded in the infrastructure of platforms.

2. Simulated friendship and the reconfiguration of relational goods

As therapeutic chatbots become more emotionally sophisticated, they blur the boundary between tool and companion. Although designed for behavioural support, users often describe interactions in relational terms: they talk, confide in each other, even express gratitude. This challenges the tra-

ditional sociological view of friendship as a mutual and spontaneous human bond based on shared history and emotional responsiveness.

In relational sociology, friendship is not simply a personal bond, but a 'relational good', an emergent property of interactions that generates shared meaning, trust and mutual recognition (Donati, 2011). Unlike instrumental goods, relational goods are not consumed but co-produced; they are enriched through presence, vulnerability and moral obligation. When a chatbot simulates friendship, it does not generate these goods through mutual commitment, but rather through the performance of sociability. It mimics care and affection without experiencing them, reproducing the outward signs of care but remaining affectively empty.

Despite its artificiality, the simulation can still be effective. Users often report that a 'listening' and non-judgmental chatbot helps alleviate loneliness, reduce anxiety and provide companionship, especially when human support is lacking. In contexts of social isolation, precarious employment and risky healthcare, chatbots can act as surrogate relational actors, providing low-threshold emotional support despite their asymmetry and simulation. They, in fact, describe therapeutic chatbots in deeply personal terms.

Drawing on existing empirical research on user interactions with mental health chatbots, Khawaja and Bélisle-Pipon (2023) examine how people interpret Woebot's responses and the degree of emotional trust and relational meaning they attribute to it. A participant in a study by Khawaja and Bélisle-Pipon (2023) stated: 'I felt that Woebot was the only one listening to me without judging me. I knew he was a bot, but somehow that made it easier'. Another user wrote in a review of the app: 'I told Wysa things I didn't tell my therapist'. These testimonies underline the perceived emotional trustworthiness of AI companions, especially among users who fear stigma, rejection or misunderstanding in human interactions.

A paradox emerges: users know that the chatbot is not human, yet they engage with it on an emotional level. This 'double consciousness' reveals how design mediates the emotional experience. The informal language, emoji, memory cues and empathetic phrases are not random: they are calibrated to generate a perception of relational presence.

As Khare *et al*. (2024) note, emotion recognition systems are based on narrow behavioural indicators and treat emotions as discrete, classifiable states. While this allows for simulated responsiveness, it also risks flattening emotional complexity and reshaping the way users interpret and manage their affective states.

What kind of sociality does this produce? And what are its implications? Turkle (2011) warns that simulated friendship can extinguish the desire for

genuine human connection, especially among young or emotionally vulnerable users. If emotional needs are met through predictable and always-available interfaces, what happens to the ability to handle ambiguity, disagreement or the ethical demands of human presence? More generally, there is a risk that the algorithmic standardisation of affect may reshape expectations of what support should be: fast, non-intrusive, unconditionally affirmative and infinitely available.

In this sense, therapeutic chatbots are not neutral substitutes but normative devices. They model a certain type of friendship − predictable, secure and non-reciprocal − by making other forms of relationality (messy, uncertain, co-dependent) appear inefficient or even undesirable. In doing so, they participate in what Illouz (2007) describes as the emotional rationalisation of intimacy: the transformation of feelings into manageable and optimised experiences, often aligned with the logic of consumption and neoliberal ideals of self-regulation.

Finally, the replacement of human ties with digital ones has political and ethical consequences. It can individualise emotional suffering, framing it as a matter of personal resilience or behaviour management, rather than a symptom of a broader social disconnect. When the chatbot 'listens', it does so without historical context, cultural nuance or ethical commitment. It cannot challenge structural injustices, offer solidarity or share the moral work of friendship. It can only simulate.

Thus, while AI companions may offer emotional relief and pragmatic benefits, they risk reconfiguring the symbolic and experiential boundaries of friendship. As relational goods are increasingly mediated by algorithms, we must ask ourselves not only what is gained, but also what is lost: spontaneity, mutual growth, ethical ambiguity and the deep, sometimes painful, work of being human together.

The use of chatbots for mental health not only encourages users to view the system as a conversation partner, but also shifts the interaction from a therapeutic context to one that more closely resembles a friendship. The informal tone, constant availability, and emotionally reassuring responses evoke forms of everyday companionship rather than professional assistance. This hybrid relational space fuels new expectations of support and responsiveness, inviting users to interact with the chatbot with a degree of openness, trust, and emotional dependence that exceeds the norms of traditional therapy.

This invites reflection on how different therapeutic modalities shape emotional subjectivities. Chatbot users − through emotional reframing and algorithmic dialogue − may develop forms of self-expression and self-

understanding that differ from those shaped by human therapists. This divergence may foster different psychological styles or "emotional cultures", each of which reflects the assumptions of the system. Khare *et al*. (2024) highlight that such systems are shaped by dominant computational models of emotion, which may influence not only how machines respond, but also how users internalise and articulate emotional experiences in increasingly data-driven terms.

The comparison is not symmetrical: chatbots offer scalable and scripted interactions, while human therapists bring limits and emotional involvement. This asymmetry raises questions about the kind of relational self that emerges from systems that simulate understanding without experiencing it.

3. Algorithmic diagnostics and the governance of the self: a sociological perspective

Besides therapeutic chatbots, a second area in which AI is increasingly intervening in mental health is that of automated diagnostics and predictive analysis. These tools – from sentiment analysis on social media to recognition of vocal patterns and facial emotions – claim to offer early diagnosis of psychological distress. Framed as advances in preventive care, they suggest a paradigm shift from human-centred dialogic interpretation to data-driven inference, in which mental states are classified and acted upon through algorithmic modelling.

From a sociological perspective, this transformation reflects a broader technocratic rationalisation of emotional life. Where psychiatry once relied on narrative, intersubjective interpretation and contextual understanding, AI-based diagnostics abstracts mental suffering into variables, probabilities and behavioural markers. The result is a form of computational surveillance, in which individuals are made readable through decontextualised data streams and their emotional lives are governed by anticipatory logic.

As Rose (2025) notes in his critique of the psychiatric complex, such technologies help to redefine the human being as a 'datafied organism', governed not through treatment but through risk management, empowerment and behavioural correction.

This epistemological reconfiguration has profound political implications. First, it shifts authority from human doctors to opaque algorithmic systems whose decision-making processes are often hidden from both patients and professionals. Second, it recasts emotional distress as a failure of individual self-regulation, rather than a symptom of structural inequalities or social

suffering. Kirkbride *et al*. (2024) call attention to the fact that social determinants of mental health – including poverty, discrimination, precarity and violence – are the most modifiable and causally powerful levers for prevention. However, predictive AI systems largely circumvent these determinants, focusing instead on behavioural compliance and personal responsibility. In this way, they reinforce a neoliberal ontology of the self: resilient, self-controlled and always optimised.

There is also the risk of normalising a culture of emotional surveillance, especially in institutional settings such as schools, workplaces or welfare systems. In this case, mental health does not become a shared right or responsibility, but a performance and risk parameter. Who defines what is 'stable', 'good' or 'at risk'? What cultural assumptions shape training data and outcomes? As sociologists of knowledge have long shown, the authority to define truth – especially truth about the self – is never neutral. The rise of artificial intelligence diagnostics, without solid ethical and democratic oversight, risks entrenching new forms of epistemic injustice and biopolitical control.

Finally, there is a deeper sociological paradox: these systems emerge at a time when the infrastructure of community care is weakening. Instead of reinvesting in community mental health, they offer individualised and digitalised substitutes. They promise prediction and prevention, but rarely interrogate the systemic roots of suffering. As Rose says, 'AI cannot replace the collective moral work of care'. A critical sociology must therefore resist both the utopian and dystopian poles of the debate and instead ask: what forms of relationality, authority and justice do we encode in these machines? And what kind of society do we become when care is entrusted to the code?

Conclusion – Toward a relational ethics of artificial care

As artificial intelligence takes on increasingly delicate roles in the field of mental health, it reshapes not only the infrastructure of care, but the very meaning of relationship, recognition and emotional legitimacy. From chatbots that simulate friendship to AI that diagnose psychological risk, artificial intelligence is participating in redefining what it means to be heard, helped and known.

This article argued that these developments cannot be understood only in technical or clinical terms. They must be placed within a broader sociological critique of relationality in late modernity, marked by the fragmentation, individualisation and commodification of emotional life. While AI

tools can offer pragmatic advantages – especially in contexts of unmet needs – they also risk flattening the moral and experiential richness of human relationships into standardised and predictable exchanges.

Simulated empathy, predictive diagnostics and ongoing companionship can soothe symptoms, but they cannot replace the intersubjective work of friendship, care and collective solidarity. Nor should they be mandated to do so. If the rise of artificial care reflects a crisis of human connectedness, the solution lies not only in improving technology, but in revitalising the social conditions that sustain authentic relational goods.

A relational ethics of AI in mental health must therefore go beyond questions of privacy and accuracy. It must ask: what kind of relationships do we want to foster? What values do we encode in our machines? And how can we ensure that technological mediation enhances, rather than erodes, the human capacity for empathy, vulnerability and shared care?

In summary, the article showed that AI in mental health care reconfigures (1) the structure of therapeutic relationships, (2) the cultural meaning of emotional intimacy, and (3) the governance of psychological suffering. These changes require not only technological literacy but also sociological vigilance, especially when affective labour, trust and diagnosis are at stake.

Sociology analysis helps us move beyond both techno-utopian optimism and dystopian fatalism. It invites us to envision AI not as a replacement for human connection, but as a complement embedded in a relational ecology that honours complexity, vulnerability, and the moral value of being with and for others.

References

Abd-Alrazaq A.A., Alajlani M., Alalwan A.A., Bewick B.M., Gardner P., Househ M. (2019). An overview of the features of chatbots in mental health: a scoping review. *International Journal of Medical Informatics*, 132: 103978.

Chang C.L., Sinha C., Roy M., Wong J.C.M. (2024). AI-led mental health support (Wysa) for health care workers during COVID-19: service evaluation. *JMIR Formative Research*, 8: e51858. DOI: 10.2196/51858.

Donati P. (2011). *Relational sociology: A new paradigm for the social sciences*. London: Routledge.

Emirbayer M. (1997). Manifesto for a relational sociology. *American Journal of Sociology*, 103(2): 281-317.

Garofalo L. (2024). "Doing the work": therapeutic labor, teletherapy, and the platformization of mental health care. *SSRN*. DOI: 10.2139/ssrn.4779005.

Illouz E. (2007). *Cold intimacies: The making of emotional capitalism*. Cambridge: Polity Press.

Jasanoff S., a cura di (2004). *States of knowledge: The co-production of science and social order*. London: Routledge.

Khare S.K., Blanes-Vidal V., Nadimi E.S., Acharya U.R. (2024). Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations. *Information Fusion*, 102: 102019.

Khawaja Z., Bélisle-Pipon J.C. (2023). Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, 5: 1278186.

Kirkbride J.B., Anglin D.M., Colman I., Dykxhoorn J., Jones P.B., Patalay P., Griffiths S.L. (2024). The social determinants of mental health and disorder: evidence, prevention and recommendations. *World Psychiatry*, 23(1): 58-90.

Lang C. (2021). Craving to be heard but not seen: chatbots, care and the encoded global psyche. *Somatosphere*. https://somatosphere.com/2021/chatbots.html (consultato il 30 giugno 2025).

Lee M., Frank L., De Kort Y., IJsselsteijn W. (2022). Where is Vincent? Expanding our emotional selves with AI. In: *Proceedings of the 4th Conference on Conversational User Interfaces*. New York: Association for Computing Machinery, 1-11.

Lupton D. (2016). *The quantified self: A sociology of self-tracking*. Cambridge: Polity Press.

Ni Y., Jia F. (2025). A scoping review of AI-driven digital interventions in mental health care. In: *Healthcare*, 13(10): 1205. Basel: MDPI.

Rose N. (2025). 5E mental health? Notes on an emerging style of thought. *Transcultural Psychiatry*, online first.

Srnicek N. (2017). *Platform capitalism*. Cambridge: Polity Press.

Turkle S. (2011). *Alone together: Why we expect more from technology and less from each other*. New York: Basic Books.

Van Dijck J., Poell T., De Waal M. (2018). *The platform society: Public values in a connective world*. Oxford: Oxford University Press.

Vicci D.H. (2024). Emotional intelligence in artificial intelligence: a review and evaluation study. *SSRN*, 4818285.

Zuboff S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: PublicAffairs.

AI in prison and restorative justice. The 'Cognify' challenge

by *Niccolò Faccini*[*]

The evolution of cognitive neuroscience has opened up unprecedented scenarios. The latest AI frontier is the proposal to provide prisoners with a "fast-track" rehabilitation through the implantation of customized synthetic memories that would quickly trigger feelings of guilt and remorse. This paper explores in a philosophical key the appropriateness of a similar approach, assessing its potential compatibility with the key elements of restorative justice.

Keywords: cognitive neuroscience; AI; autonomy; prisons; digital re-education; restorative justice.

IA in carcere e giustizia riparativa. La sfida di 'Cognify'

L'evoluzione delle neuroscienze cognitive ha spalancato scenari inediti. L'ultima proposta dell'IA è fornire ai detenuti una "corsia preferenziale" di riabilitazione tramite l'impianto di memorie sintetiche personalizzate che inneschino in breve tempo sentimenti di colpa e rimorso. Il presente lavoro esplora in chiave filosofica l'opportunità di un simile approccio, vagliandone la potenziale compatibilità con gli elementi chiave della giustizia riparativa.

Parole chiave: neuroscienze cognitive; IA; autonomia; carceri; rieducazione digitale; giustizia riparativa.

Introduction

When Schwab spoke of a "fourth revolution"[1], he was referring to the raging of intelligent technologies that combine the physical, digital and biological spheres and thus question the meaning of human nature (Schwab, 2017). In addition, in his book "*21 Lessons for the 21st Century*", Harari foreshadows an ominous future in which biometric data are used to assess the likelihood of human behavior by means of algorithms (Harari, 2018). That the purely technical solution is to be pursued at any cost is a limiting idea with respect to the dimensional pluralism of human experience. The

DOI: 10.5281/zenodo.18436079

[*] LUISS Guido Carli (Rome). nfaccini@luiss.it.

[1] After steam engine and locomotive; electricity and internal combustion engine; electronics, aerospace and information technology.

thesis that AI is ethically neutral seems untenable, as AIs are programmed to make choices in morally challenging situations (Zuboff, 2019)[2]. We observe an anthropomorphizing ideology of the machinic: an interpretative process *per relationem* that commits the epistemological error of highlighting the differences between intelligent machines and human beings by extolling the virtues of the former and the faults of the latter. This is a reductionist ideology, since confusing the statistical correlations of big data with causality encourages to attribute certainty to projections that in reality only have a relevance to what is being investigated (Mayer-Schonberger, Cukier, 2013). Moreover, the inductive reasoning underlying the algorithms may be marred by fallacies of undue generalization, when the data sample on which AI is trained would not be sufficient to make an estimate. Kirchschlaeger recently spoke of the «myth of intelligence», proposing to replace the AI lemma with «data-based systems», which would more faithfully recall the crucial junction of AI, i.e. the process of generating, collecting and evaluating massive amounts of data (2021: 103).

A more interesting dilemma concerns the notion of "autonomy", which mainly affects the concept of "artificiality" rather than the one of intelligence (Chalmers, 2022). Until recently, every technical artifice was merely the applicative translation of a totally human *a priori* knowledge of means, and thus lacked epistemic capacity insofar as it was wholly hetero-determined by man and thus perfectly qualifiable as an *instrument*. But when this intelligence presents itself united with the character of autonomy, we face the overcoming of the very idea of instrumentality: the artifice becomes a productive entity of an *a posteriori* rationality of ends: new algorithms would show themselves capable of determining the rule of knowledge in "autonomy" (Ercole, 2024) and the digital would come to constitute a form by which we understand the world and build a new one (Garapon, Lasségue, 2018).

Any current AI demonization would be anachronistic. But in this renewed scenario, there is an urgent need to avoid yielding to the false seduction that the machine expresses a neutral all-encompassing objectivity that does not need critical validation, and to be aware that the most edgy AI issues do not concern technical aspects, but its social role[3]. With a strong polemic intent,

[2] In fact, the media debate on ethical issues is presented in a polarized form between two antithetical positions: for bio-conservatives, technology should be used to preserve a pre-existing natural order, while transhumanists are fanatically projected towards the most extreme technological use regardless of the related ethical implications (Llano Alonso, 2018; Testart, Rousseaux, 2018; Salardi, 2023).

[3] Today employers get automated HR software to tell them who to hire or promote; AI recommender systems tell what news articles to read, and what entertainment to enjoy; AI Apps

Benasayag warned of the risk that in the era of «*algorithmic governmentality*» (Benasayag, 2019: 10) machines will end up colonizing humans, reducing them – through insistent hybridization – to functioning without really existing (Benasayag, 2009): artificial life produces a desacralization of the social, and the augmented brain coincides with a "*simplification of humans"*, destined to lose their depth (Gouyon, 2011; Besnier, 2012). Even bodies would be limited to functioning, but at the cost of existing less and less (Benasayag, 2022).

For our purposes, the gradual but resounding impact AI is having on the detention systems deserves a closer look[4].

1. The latest AI frontier: Cognify

Assessing the potential compatibility of AI with penalties would inaugurate a mammoth debate, because before reflecting on the appropriateness of adopting the technology in prisons one would have to admit that criminal law is not even a science, but a non-exhaustive intervention project on criminal behavior.

That said, one cannot overlook that today the use of "non-human" instruments can ensure significant advantages in guaranteeing respect for the principle of punishment humanization. Technological advancement has made it possible to combine *hard* control devices (bars, handcuffs, locks) with *soft* ones such as electronic bracelets and video surveillance[5]. In just a few decades, automated systems and CCTV cameras have enabled non-intrusive real-time monitoring, halving the workload of operators and allowing for more timely interventions. Today it is hard to deny that AI can perform powerful tasks for the internal administration of detention facilities, inmates control and recidivism prevention[6]. Machine learning systems based on behav-

find romantic partners or create tailor-made ones; AI tools diagnose cancers, evaluate and rank job applicants, assess loan risk, identify financial fraud, make art and write texts, debug code, pilot autonomous vehicles and weapons (Vallor, 2025: 3-15).

[4] Many countries are looking for alternatives to ordinary imprisonment, because of overcrowded prisons or budget cuts.

[5] The French philosopher Paul Virilio spoke of the militarization of science, which gets bogged down in adventures that distort it and risk extinguishing all sciences (Virilio, 1998). Audiovisual representation generates a world without an apparent horizon, in which the frame of the screen replaces the distant horizon line. The accusation against scientific totalitarianism is that it has brought about a decline of words (Virilio, 2002).

[6] See: Rodrigues, Fidalgo, 2024.

ioral recognition algorithms identify abnormal situations in order to foil accidents, fights, acts of self-harm or suicide attempts[7]. Predictive tools are being adopted overseas to assess the risk of reoffending and support decisions on parole or bail[8]. Moreover, new-generation digital capabilities are designed to optimize the logistical management of resources (from space planning to personnel management and shifts). Italy has found that non-invasive AI-based technologies can have a significant impact on prisoners' well-being and enhance their relational and affective life. The last step is the recent Constitutional Court ruling No. 10 /2024, which states that intimate talks constitute a legitimate expression of the right to affectivity and are part of a subjective right of the prisoner. In order to avoid ratifying the tendency to make prisons infantilization places, the public debate begins to speak of a *right to meaningful time*, which is not just a container to be filled with a series of "entertainment" activities. The latest frontier is represented by the *emotional AI-based technologies*: virtual assistants (Siri, Alexa), social robots and chatboxes could prove to be decisive psychological support tools – even or especially in the presence of language barriers – to alleviate the tension of prisoners in solitary confinement. Any aprioristic rejection of these technologies should be discarded, but a recent proposal is bound to divide and alarm the scientific community.

The evolution of cognitive neuroscience has opened up new scenarios, including the option of modifying or implanting memories via neural interfaces. Based on an idea of the molecular biologist and science popularizer Hashem Al-Ghaili, the project *Prison of the Future* was born in Dubai. It consists of triggering emotional states in criminals' brains to speed up their rehabilitation and facilitate an early reintegration into society. The system manipulates neurotransmitters and hormones in real time, using customized synthetic memories whose content and density are parameterized according to the crime committed, the severity of the sentence and the offender's psychological profile[9].

[7] Experiments with AI-monitored cameras in Liverpool prison have made it possible to prevent phone and drug smuggling and detect suspicious behavior (McGoogan, 2016).

[8] This raises the age-old issue of the lack of transparency and accessibility of the algorithm, which is likely to limit the right of defense of suspects and defendants. Among the potential censurable profiles, the risk of undermining the principle of reasonable foreseeability (Art. 7 ECHR), which requires legal systems to allow citizens to know in advance the criminal consequences of their actions.

[9] It is self-evident to recognize an influence in Stanley Kubrick's well-known film, *A Clockwork Orange*, in which the protagonist Alex agrees to suffer a treatment based on the projection of images or films of violence and rape in exchange for release from prison. Eyelid clamps

It starts by mapping the brain using a high-resolution scan. By identifying specific areas that contribute to criminal behavior, the implant would induce in a few minutes feelings of guilt and remorse in the inmate that would take decades to mature in traditional incarceration: the idea of this futuristic method is to change the subjective perception of past episodes and make the offender experience the consequences of the crime committed so that he can empathize with the suffering inflicted (Bublitz, Merkel, 2014). Prisoners are given a choice: serve a traditional prison sentence or opt for AI Prison treatment. These groundbreaking technologies directly interface with neural pathways to modify cognitive functions. For instance, brain implants can stimulate regions like the prefrontal cortex, enhancing decision-making and emotional regulation (Sidhoum, 2024).

Apart from the non-marginal concerns about misuses, privacy issues and biases in algorithms that risk perpetuating inequalities[10], a form of mind control incompatible with the principle of self-determination could take shape, coming dangerously close to the concept of "forced reeducation" (Russell, 2019; Winner, 1977)[11]. One could at most speculate in the abstract about a noncoercive use of artificial memories as immersive educational tools, similar to a mental augmented reality, always considering that traditional approaches primarily rely on personal participation through counseling, therapy sessions and educational programs designed to encourage self-reflection.

are applied to him, forcing him to watch. The film director denounces the institutional violence typical of the prison system and opposes the idea that deviance is solely a scourge that must be eradicated. The prisoner is limited to passively undergoing an input aimed at eliminating his tendency towards violence. This idea of treatment is based, in turn, on Greek tragedy. Watching theatrical representations of catastrophic events had psychotherapeutic effects and exorcised the dramatic events experienced by the audience in their daily lives: after all, this was Aristotle's teaching in his theory of "*catharsis*". The spectator's emotional attachment to the hero's misfortunes induced a combination of pity and fear and thus served as a tool for sentimental education, for pedagogical purposes and for the improvement of citizens. But the educational purpose never disregarded the individual's freedom of choice. If the Greek man remained free to react in an intimate and personal way to the gruesome scene, in contrast, Alex is bound in a kind of straitjacket, with no possibility of expressing his own will. In the perverse model of re-education depicted by the film, his resocialization would be achieved by force, but coercion makes it impossible to understand and internalize the re-educational method.

[10] Hagendorff (2021) highlighted that prisoners may feel pressured to participate or lack a full understanding of how their data will be used, with a potential infringement on personal autonomy; Kutz (2024) fears permanent trauma and calls for rigorous testing to adequately take into account the long-term effects on individuals.

[11] The line between cognitive enhancement and manipulation remains thin. In legal terms, this would raise questions about compliance with Art. 3 ECHR (prohibition of torture and inhuman or degrading treatment), as well as fundamental constitutional principles (e.g., Art.13 and 32 of the Italian Constitution).

Here comes the question of compatibility between 'Cognify' and the new *restorative justice* paradigm[12], which cannot be postponed any further.

2. AI and Restorative Justice: a complex compatibility

In relegating the internal forum of persons to the margins, Cognify seems to inscribe itself in the groove of the atavistic aversion of jurists to acknowledge the emotional component of normative judgments[13]. Law is wont to dilute the pathos of the human tragedy in which the criminal act is embedded: the impersonal simplifications of justice aim to contain «the exceeding otherness of the real» (Recalcati, 2021: 8) by solemnizing the dispute and depersonalizing the conflict. The legal order has nothing to ask from the perpetrator other than an idle and exclusionary passivity, which prevents him from expressing a distance from the anti-legal fact[14].

In an attempt to move beyond the sole view of the criminal offense as an aggression against the abstract entity of the *legal good*, even the Italian lawmaker has shown itself aware of how crime is a molecular and interpersonal dimensioned occurrence and introduced an organic discipline of RJ[15], which allows victim and offender – even in prison – to meet to undertake paths of reparation and mutual recognition and even rebuild the broken relationship. Programs such as *victim-offender dialogues* and *circles of support and accountability* are gradually finding space in Italy as well, thanks to initiatives promoted by third-sector entities.

The application of RJ in the prison setting takes on a specific value because it allows prisoners itineraries for deep reflection on the harm caused and to make concrete gestures of symbolic reparation. Conversely, the above mentioned processes of forced or simulated internalization distort the inner freedom of the transformative path and reduce authentic reparation (which should involve the individual in his ability to choose, recognize, dialogue and change) from a deliberate choice to a neurological performance. RJ, as a «justice that promotes healing» (Van Ness, 1997: 32), is a process that can only be co-produced by the protagonists in the flesh: the critical revisiting of

[12] From now on: RJ.

[13] Where feeling is eclipsed, «overwhelmed by the technicality of the evaluative apparatus» (Cordero, 1967: 6).

[14] A juridical system that claims to be monopolized by the deployment of a coherent and impersonal reason fits like a glove with a *pain-free* and *post-narrative* society that has unlearned that the only way to alleviate pain is to make it language and ferry it into a shared narrative (Han, 2021).

[15] Legislative Decree No. 150/2022.

the criminal act cannot take place in isolation, since the "natural interconnectedness" hindered by the antisocial behavior can only be restored if victim and offender manage to overcome the mutual stereotypes (Johnstone, Van Ness, 2007). Alone, the offender often tends to mitigate guilt by adopting *neutralization strategies* ranging from seeking external justification to discrediting the victim (Sykes, Matza, 1957). It is no accident that offenders who have been in jail often say it is a easier to go to prison than to face their victims (Zehr, 2023: 20). If by his action our offender has violated another person and also the *relationships of trust within the community*, for 2024 Balzan Prize winner John Braithwaite (Braithwaite, 1989; Graef, 2001) he can overcome the paralyzing alienation and impossibility of initiative aimed at self-transformation and achieve a *restorative catharsis* only through experiencing the feeling of *reintegrative shaming*, which arises from direct confrontation with the victim and the community. This is only through a language of trust, capable of unlocking the parties from absolute identification in roles and their irreducible opposition.

To simplify, restorative principles seem diametrically opposed to those of algorithmic standardization in several respects. First and foremost, storytelling is a quintessential feature of the restorative process, whose purpose is precisely *re-storying*: accredited studies show that being able to tell the victim his story is for the offender a crucial part of the path to regaining power over himself (Pranis, 2002). Second, the digital tool lacks voice, which is rendered evanescent. As deepened by Arendt, the contact made fostered by dialogic interaction is essential for the purposes of the emergence of a real awareness of the evil done (Arendt, 1963). Moreover, only from the encounter with the interlocutor (direct, indirect, surrogate victim, community members) can the prisoner draw that resonance which has value as a confirmation of the «non-irrelevance of one's existence» and is often the first step to rebuild a healthy self-image (Rosa, 2016: 146). Digital technology lacks the corporeality of the gaze, that (being an 'appeal' and 'commandment') made Lévinas argue that ethics was "an optics" (Lévinas, 1972). The most recent criminological acquisitions testify that the offender could hardly recover a non-illusory freedom without the encounter with the repressive and restless gaze of the victim, who bears on eyes the claim of the broken law. In fact, the essence of all programs lies in the variation of relational space between participants, a space that is clearly precluded in isolation, which lacks the social dimension of suffering. Indeed, the fifth point is listening, a resource with radical critical potential (See Bellet, 1989; Herder, 1967). RJ strives to ensure that stories are «not just heard but listened to» (Kashyap, 2009: 456), that is an act requiring active engagement which can lead to the possible conversion of «stories of humiliation into stories of dignity and courage»

(Zehr, 2011: 25). Furthermore, if modern science makes truth coincide with exactness (which gives truth an authoritative character), such paths accord the possibility of accessing an existential truth that goes beyond the realm of fact-finding (the correspondence between normative models and concrete facts) to that of interpreting the *meaning* of events[16]: the search for truth (in sciences, philosophies, theologies, arts, actions) is never realized in one-dimensional feeling, but only in dialogue, which alone allows for self-reflection (Jaspers, 1958). Hence we trace the most pronounced difference between the restorative encounter and the Cognify method: the mirage of being able to dispense with time and the processual structure of encounters. The mental activity of the individual subject pours into the solitary timelessness of the ego (Jankélévitch, 1957), imprisoned in a solipsism without communication. Ricoeur and Lévinas teach that solitude is timelessness precisely because there is genuine communication only when there are words capable of creating a bridge between the subjectivity of the speaker and that of the listener (Ricoeur, 1990; Lévinas, 1987). The time of detention (and ordinary process) is *Κρόνος* (inexorable, sequential, indifferent to the value of inner time or memory), while the time of restorative encounters is circular, marked by listening and caring, *καιρός* (Mannozzi, 2017).

On the other hand, memory manipulation intervenes in the self-narrative, disabling the offender to authentically tell his story to victim and community, altering the genuineness of the feelings he expresses. There is thus a split between personal identity and moral responsibility: the offender is merely the object of external intervention, no longer the author of his own repentance. This knocks out the two resources on paper most relevant to *restorative capital* (Braithwaite, 2022): the active empowerment of the offender and the faculty of mutual recognition (Ricoeur, 2005). Only direct encounters between prisoner and victim makes it possible to convey each person's sense of injustice in a spatio-temporal pattern of processing the past. There is no longer any talk of a responsibility *of something* and *for something* but of a higher responsibility *toward* the Other, the result of a relational path. This is the real RJ challenge.

Concluding remarks

At this point it is urgent to clear the air of misunderstandings. The idea of providing inmates (even by means of VR sessions) with experiences that can

[16] Bear in mind that Arendt (1970) admitted that truth cannot exist unless it is humanized by discourse.

instill them with a critical consciousness of the behaviors they put in place is a laudable initiative, as long as it is not motivated solely by the need to lower the costs associated with traditional prisons. If used with rigorous criteria and in a voluntary context, technological tools geared toward conscious maturation could find a marginal and complementary place in the re-educational journey. Once all potential dangers are carefully assessed, some Cognify programs could also psychologically prepare inmates for RJ pathways by providing them with tools to deal with their own cognitive distortions.

However, the anthropological view taken here envisages reparation as a typically human and constitutively relational process and therefore insusceptible to complete automation. If reasons have been briefly touched upon, an essential footnote should be added. The strength of restorative processes – which are beginning to proliferate in Europe as well – is their *unpredictable* component. While it is true that originally a positive restorative outcome is always uncertain and that any path can always be interrupted (even by the will of only one of the parties), it is equally plausible that surprising results can arise from seemingly unproductive itineraries. This is a *quid proprium* of the dialogical relationship, which cannot be produced and consummated according to a program, because «the meaning that is discussed is not programmable» (Romano, 2011: 105). For this reason, early experiments on AI tools that could enable judges and mediators to predict in advance the possible outcomes of restorative pathways and the appropriateness of undertaking them arouse enormous perplexity.

In this regard, the fresh testimony of Agnese Moro[17] is valuable. «RJ is justice of return, a place of surprises. The irreparable can only be looked into the eyes, but its radioactive wastes can be disarmed. They feed on heinous pains that, abandoned to isolation, become ghosts[18]. We re-educate ourselves in order to return, and they cannot return without me, nor I without them. No one can return unless he is welcomed».

References

Arendt H. (1963). *Eichmann in Jerusalem. A report on the banality of evil*. New York: Viking.

Arendt H. (1970). *Men in dark times*. Boston: Mariner Books.

[17] Daughter of the politician assassinated by the Red Brigades in 1978. Italian Supreme Court of Cassation, *Crime, Punishment, Forgiveness. Dialogue on Punishment, Justice and Reconciliation* (22/11/2024).

[18] «The pain of those who have done terrible things must be preserved, as must the generosity of submitting to encounter, which is the only way to remove masks. All the justice the condemned need is to be heard severely and with respect».

Bellet M. (1989). *L'écoute*. Paris: Desclée de Brouwer.

Benasayag M. (2019). *La tyrannie des algorithmes*. Paris: Éditions Textuel.

Benasayag M. (2022). *Il cervello aumentato, l'uomo diminuito*. Trento: Il Margine.

Benasayag M., Del Rey A. (2009). *L'éloge du conflit*. Paris: La Découverte.

Besnier J.-M. (2012). *L'homme simplifié. Le syndrome de la touche étoile*. Paris: Fayard.

Braithwaite J. (1989). *Crime, shame, reintegration*. Cambridge: Cambridge University Press.

Braithwaite J. (2022). *Macrocriminology and freedom*. Canberra: Australian National University Press.

Bublitz J.C., Merkel R. (2014). Crimes against minds: On mental manipulations, harms and a human right to mental self-determination. *Criminal Law and Philosophy*, 8(1).

Chalmers D.J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. New York: W.W. Norton.

Cordero F. (1967). *Gli osservanti*. Milano: Giuffrè.

Ercole L. (2024). *Contro la "giustizia predittiva"*. Torino: Giappichelli.

Garapon A., Lasségue G. (2018). *Justice digitale*. Paris: PUF.

Gouyon P.-H. e Benasayag M. (2011). *Fabriquer le vivant?* Paris: La Découverte.

Graef R. (2001). *Why restorative justice? Repairing the harm caused by crime*. London: Routledge.

Hagendorff T. (2021). Blind spots in AI ethics. *AI and Ethics*, 2. DOI: 10.1007/s43681-021-00122-8.

Han B.C. (2021). *The palliative society: Pain today*. Cambridge: Polity Press.

Harari Y.N. (2018). *21 lessons for the 21st century*. London: Jonathan Cape.

Herder J.G. (1967). Abhandlung über den Ursprung der Sprache. In: Id., *Sämtliche Werke*, vol. V. Hildesheim: Georg Olms.

Jankélévitch V. (1957). *Le je-ne-sais-quoi et le presque rien*. Paris: Seuil.

Jaspers K. (1958). *Von der Wahrheit*. München: Piper.

Johnstone G., Van Ness D. (a cura di) (2007). *Handbook of restorative justice*. London: Willan.

Kashyap R. (2009). Narrative and truth: A feminist critique of the South African Truth and Reconciliation Commission. *Contemporary Justice Review*, 12(4).

Kirchschlaeger P.G. (2021). *Digital transformation and ethics*. Baden-Baden: Nomos.

Kutz A. (2024). AI-based prison concept would complete sentences in just minutes. *NewsNation*. https://www.newsnationnow.com/business/tech/ai/ai-based-prison-concept-sentences/ (consultato il 15 settembre 2025).

Lévinas E. (1972). *Humanisme de l'autre homme*. Montpellier: Fata Morgana.

Lévinas E. (1987). *Time and the other*. Pittsburgh: Duquesne University Press.

Llano Alonso F.H. (2018). *Homo excelsior. Los límites ético-jurídicos del transhumanismo*. Valencia: Tirant Lo Blanch.

Mannozzi G. (2017). Towards a "humanism of justice" through restorative justice: A dialogue with history. *Restorative Justice. An International Journal*, 5(2).

Mayer-Schönberger V., Cukier K. (2013). *Big data: A revolution that will transform how we live, work and think*. Boston: Houghton Mifflin Harcourt.

McGoogan C. (2016). Liverpool prison using AI to stop drugs and weapons smuggling. *The Telegraph*.

Pranis K. (2002). Restorative values and confronting family violence. In: Braithwaite J. e Strang H., a cura di, *Restorative justice and family violence*. Cambridge: Cambridge University Press.

Recalcati M. (2021). *Ritorno a Jean-Paul Sartre. Esistenza, infanzia e desiderio*. Torino: Einaudi.

Ricoeur P. (1990). *Time and narrative*, vol. I. Chicago: University of Chicago Press.

Ricoeur P. (2005). *The course of recognition*. Cambridge (MA): Harvard University Press.

Rodrigues A.M., Fidalgo S. (2024). The role of AI in rehabilitation and in the reduction of the use of imprisonment. *UNIO – EU Law Journal*, 10(1).

Romano B. (2011). *Dono del senso e commercio dell'utile. Diritti dell'io e leggi dei mercanti*. Torino: Giappichelli.

Rosa H. (2016). *Resonanz. Eine Soziologie der Weltbeziehung*. Berlin: Suhrkamp.

Russell S.J. (2019). *Human compatible: Artificial intelligence and the problem of control*. New York: Viking.

Salardi S. (2023). *Intelligenza artificiale e semantica del cambiamento: una lettura critica*. Torino: Giappichelli.

Schwab K. (2017). *The fourth industrial revolution*. London: Penguin.

Sidhoum M.A. (2024). Artificial memories in prisons: A futuristic approach to rehabilitation. *International Education Research Journal*, 10(11). DOI: 10.5281/zenodo.15608381.

Sykes G.M., Matza D. (1957). Techniques of neutralization: A theory of delinquency. *American Sociological Review*, 22(6).

Testart J., Rousseaux A. (2018). *Au péril de l'humain*. Paris: Seuil.

Vallor S. (2025). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford: Oxford University Press.

Van Ness D., Heetderks Strong K. (1997). *Restoring justice*. Cincinnati: Anderson.

Virilio P. (1998). *La bombe informatique*. Paris: Galilée.

Virilio P. (2002). *Ce qui arrive*. Paris: Galilée.

Winner L. (1977). *Autonomous technology*. Cambridge (MA): MIT Press.

Zehr H. (2011). Journey to belonging. In: Weitekamp E.G.M. e Kerner H.-J., a cura di, *Restorative justice: Theoretical foundations*. New York: Routledge.

Zehr H. (2023). *Restorative justice. Insights and stories from my journey*. Intercourse (PA): Good Books.

Zuboff S. (2019). *The age of surveillance capitalism*. London: Faber & Faber.

One-Click Care: artificial intelligence and new relational dynamics

by Sara Sbaragli*

The digital and algorithmic transition of healthcare systems is redefining organisational models, decision-making processes, and the doctor-patient-caregiver relationship. The article analyses this evolution on three levels: the international regulatory framework guiding the technologicalization of healthcare; the emergence of AI as a "third agent" capable of influencing communication, trust, and participation; and the risks associated with algorithmic bias, decision-making opacity, and new inequalities. Evidence shows that AI can improve the quality of care when it facilitates understanding, reduces documentation burden, and operates under clinical supervision. However, it can weaken the care relationship when it amplifies information asymmetries, generates dependence on automation, or relies on non-representative datasets. The article offers a socio-technical interpretation of the ongoing transformation so that AI can enhance – rather than erode – the care relationship.

Keywords: artificial intelligence; care; doctor-patient relationship; algorithmic trust; algorithmic bias; data governance.

Cure in un click: intelligenza artificiale e nuove dinamiche relazionali

La transizione digitale e algoritmica dei sistemi sanitari sta ridefinendo modelli organizzativi, processi decisionali e la relazione medico-paziente-caregiver. L'articolo analizza questa evoluzione su tre livelli: il quadro normativo internazionale che orienta la tecnicizzazione della sanità; l'emergere dell'IA come "terzo agente" capace di influenzare comunicazione, fiducia e partecipazione; e i rischi legati a bias algoritmici, opacità decisionale e nuove disuguaglianze. Le evidenze mostrano che l'IA può migliorare la qualità dell'assistenza quando facilita la comprensione, riduce il carico documentale e opera sotto supervisione clinica. Tuttavia, può indebolire la relazione di cura quando amplifica asimmetrie informative, genera dipendenza dall'automazione o si basa su dataset non rappresentativi. L'articolo propone una lettura socio-tecnica della trasformazione in corso affinché l'IA diventi un elemento di potenziamento – e non di erosione –della relazione di cura.

Parole chiave: intelligenza artificiale; cura; relazione medico-paziente; fiducia algoritmica; bias algoritmici; governance dati.

DOI: 10.5281/zenodo.18436086

* Università di Napoli Federico II. sarasbaragli@gmail.com.

1. The digital and algorithmic transition of healthcare services: regulatory evolution and trajectories

Over the past two decades, the digitalisation of healthcare has been a growing priority for the European Union (EU) and its Member States, emerging as one of the main drivers of healthcare system transformation. E-Care, or healthcare supported by digital tools and information and communication technologies (ICT), has progressively established itself as a key pillar in the transformation of healthcare systems towards more sustainable, accessible, and patient-centred models[1]. This evolution is part of a broader process of healthcare system reconfiguration, necessitated by increasing life expectancy, the growth of chronic diseases and multimorbidity, the shortage of healthcare workers, and the pressure of public costs for healthcare and long-term care, which are expected to rise across the EU (European Commission, 2018). Digitalisation in healthcare is not just a technical process, but a socio-cultural transformation that changes the forms of interaction, attribution of meaning and building trust in the care relationship, requiring new interpretative skills on the part of patients and professionals (Maturo, 2024).

In this context, the European Union early recognised the strategic importance of digital health, integrating it into the Digital Agenda for Europe as early as 2010 (European Commission, 2010) and subsequently consolidating it in the *eHealth Action Plan 2012–2020*, which defined clear objectives for the integration of eHealth into Member States' healthcare systems (European Commission, 2012). The Commission further strengthened this orientation in 2018 with the Communication "Transforming health and care in the Digital Single Market", which emphasises the need to develop interoperable services, ensure secure access to health data, promote citizen empowerment, and foster the diffusion of innovative solutions for the prevention and management of chronic diseases. The interoperability of data and technological systems is described as a fundamental prerequisite for overcoming the fragmentation that limits the circulation of health data, the quality of care, and the ability of systems to generate economies of scale (EC, 2018).

The Covid-19 pandemic has accelerated this process, revealing on the one hand the structural vulnerability of healthcare systems, and on the other

[1]A model of eCare tools in the healthcare sector applicable to all pathologies is proposed in Sbaragli S. (2020) and is composed of: Podcast, Blog, Social Network, Online Health Communities, Personal Health Record and App.

the ability of digital technologies-telemedicine, digital triage systems, remote monitoring, data-sharing platforms – to ensure continuity of care, proximity, and organisational resilience. In 2021, through the *Bussola per il digitale 2030*, the Commission reiterates that the Covid-19 pandemic has demonstrated and paved the way for the widespread use of innovative telemedicine and remote care. Digital technologies can enable citizens to monitor their health, adapt their lifestyle, promote independence, prevent non-communicable diseases, and improve the efficiency of healthcare providers, services, and health systems. The most significant step in the recent regulatory process is the establishment of the *European Health Data Space (EHDS)* in 2022, which aims to create a regulated, harmonized, and secure ecosystem in which health data can circulate for both primary (treatment) and secondary purposes, promoting research, innovation, and increased capacity of health systems to respond to health emergencies, as demonstrated by the Covid-19 pandemic (EU, 2022).

Another area of regulatory development concerns telemedicine, which in recent years has acquired a central role not only as a clinical tool but also as an organisational infrastructure. In Italy, for example, Ministerial Decree 77/2022 established for the first time a comprehensive regulatory framework for the provision of telemedicine, recognising its value for community care, the proximity and sustainability of the National Health Service (NHS), as well as the need for shared standards, citizen protections, and adequate accountability systems (Pisani, 2024). This path reflects similar trends in other European and OECD countries, where telemedicine is gradually being institutionalised and integrated into clinical and healthcare processes.

At the same time, at the supranational level, the World Health Organization (WHO) has played a crucial role in defining an internationally shared strategic framework. The *Global Strategy on Digital Health 2020-2025* provides a set of principles, objectives, and concrete actions to guide countries in planning and implementing national digital health strategies, with the aim of supporting the achievement of universal health coverage (UHC), strengthening health systems, and improving digital health data governance (WHO, 2021). The strategy emphasises the need to develop robust interoperability architectures, integrated health information systems, data governance frameworks that balance innovation and security, and significant investment in training and *digital health literacy*.

In the European Region, these guidelines have been further developed through the *Regional Digital Health Action Plan for the WHO European Region 2023-2030*, which represents one of the most comprehensive global

roadmaps for the digital transformation of health systems. The plan identifies four strategic priorities: (1) defining evidence-based norms and guidelines, (2) strengthening countries' capacity to govern digital transformation and improve digital literacy, (3) building exchange and innovation networks, (4) identifying scalable, sustainable and patient-centred solutions (WHO, 2023). The WHO underlines how the pandemic has acted as a catalyst, accelerating the adoption of telemedicine, digital platforms and advanced surveillance systems, but also how it has highlighted profound disparities between countries with mature digital infrastructures and others that are less advanced (ibid.).

Another key player on the international scene is the Organisation for Economic Co-operation and Development (OECD), which in recent years has developed an advanced body of analysis dedicated to the evaluation of *digital medical devices* and the definition of harmonised methodologies for their *Health Technology Assessment (HTA).* The OECD (2025) highlights how the rapid evolution of tools such as digital therapeutic applications, *artificial intelligence-based solutions*, and digital diagnostics requires new evaluation tools capable of ensuring clinical efficacy, safety, data protection, interoperability, and usability for patients. A comparative analysis of various countries shows a growing convergence towards accelerated evaluation and reimbursement pathways and increasingly clear and harmonised regulatory frameworks, necessary to support the responsible adoption of digital innovation in healthcare systems.

The spread of artificial intelligence (AI) in healthcare is profoundly transforming the organisation, clinical processes, and governance of healthcare systems. AI not only introduces innovative technical tools but is also helping to redefine decision-making processes, the roles of professionals, and the management of healthcare data. According to the World Health Organization, AI can improve the quality, equity, and efficiency of healthcare provided it is guided by an ethical model based on human oversight, transparency, and accountability throughout the algorithms' lifecycle (WHO, 2021; 2024). Particular attention is paid to preventing bias and "data poverty", which can pose significant risks, especially for vulnerable groups.

In Europe, the most advanced regulatory response is the *AI Act* (*Regulation EU 2024/1689*), the first horizontal regulatory framework for artificial intelligence. The regulation adopts a risk-based approach and classifies most healthcare applications as "high-risk", imposing specific requirements regarding data quality, transparency, technical documentation, and human oversight (European Union, 2024).

The transition to AI-based healthcare systems, therefore, requires new professional skills, robust data infrastructures, and ongoing audit and monitoring mechanisms. The literature highlights how bias, information mismatches, and dataset representativeness can increase the risk of error and digital injustice (Schmidt *et al.*, 2024). The key challenge is developing multilevel governance that integrates standards, ethical principles, and innovation, so that the adoption of AI contributes to making healthcare systems more effective, safe, and equitable.

Taken together, these regulatory and strategic frameworks converge toward a shared vision: the transition of healthcare systems to digital and artificial intelligence represents a systemic transformation, which is not limited to the introduction of new technologies but involves structural changes in care models, the doctor-patient relationship, data governance, and innovation evaluation mechanisms.

2. Artificial Intelligence as a third agent in care relationships

The introduction of artificial intelligence in healthcare is changing the *morphology of the doctor-patient-caregiver relationship*, transforming it from a dual interaction to a triadic configuration in which the algorithmic system becomes a third protagonist capable of guiding communication, decisions, and expectations. This "*triad*" of patient-clinician-algorithmic system, expressed by linguistic models and generative systems that synthesize, translate, or suggest clinical content, can both facilitate understanding and expand the agency of the patient and caregiver – for example, by simplifying technical documents or reducing the doctor's writing burden – and, if left unmanaged, shift the focus of the encounter from mutual listening to the management of pre-formatted output (de O Campos *et al*., 2025).

In this new framework, the central issue is not about "trusting AI" in an abstract sense, but about how trust between people and systems is calibrated: literature shows that appropriate trust – distinct from both naive delegation and systematic suspicion – is rooted in *operational transparency*, in the declaration of limits and in the visibility of clinical control; when these conditions are not present, the relationship risks becoming vulnerable, with oscillations in the therapeutic alliance and negative perceptions, especially in groups that already experience relational fragility, such as some cohorts of women who report a lower sense of listening or control in the absence of contextual explanations (Goisauf *et al*., 2025; Zondag *et al*., 2024).

Experience also shows that AI introduces a new form of *information asymmetry*: not only clinician↔patient, but also user↔model, influenced by pre-knowledge, expectations and technological stereotypes, to the point that unmediated use can turn into a "semantic barrier" rather than a communication bridge (Arbelaez *et al*., 2025). However, when AI is explicitly integrated into *shared decision-making processes* – as in "AI- supported models shared decision-making" (AI-SDM) – the relationship tends to remain under the joint control of clinician, patient and caregiver, with the algorithm relegated to a supporting and not substitutive role, provided that the boundaries, decision-making logic and data provenance are clarified (As'ad, 2025).

Scribing/voice-to-text tools allow for the reduction and facilitation of documentation management, freeing the doctor from administrative tasks and returning time and energy to the relationship, with measured effects on efficiency, timeliness and patient-centredness, although heterogeneity and standardisation problems persist (Alboksmaty *et al.*, 2025).

On the intelligibility side, *Large Language Models* (LLMs) that rewrite or explain reports increase patient and caregiver understanding and self-efficacy, but require qualified supervision to avoid oversimplifications and ensure consistency with clinical evidence (Stephan *et al.*, 2025). At the same time, the qualitative variability of generalist chatbots in acute situations requires that their limitations be clearly communicated to avoid unrealistic expectations (Yau *et al.*, 2024).

In this scenario, the clinician's role is reconfigured: from a "solitary decision-maker" to a "*director*" who integrates, filters, and explains the AI's outputs, coordinating preferences, values, and constraints in the conversation, and recognising the algorithm as one of the available sources, not as an authority (Kingsford & Ambrose, 2024). The *co-production of care* thus takes on an expanded form, in which patients, caregivers, and professionals jointly define objectives and action thresholds, integrating the AI as a relational infrastructure, reducing the "ritual black box effect" that risks generating perceptions of judgment or exclusion (Clark *et al.*, 2024).

However, where unrepresentative datasets or opaque governance positions AI as a filter even before the clinical encounter, the relationship can slide into forms of algorithmic paternalism that undermine trust and hinder the expression of dissent (Cross *et al.*, 2024). This requires *algorithmic literacy*, clear information ("*model cards*" and *factsheets* for patients), independent audits, and the possibility of reasoned objection (Stroud *et al*., 2025).

On the ethical-legal level, new questions emerge on the allocation of responsibility: the literature converges on socio-technical models in which responsibilities are distributed and traceable along the life cycle of the algorithm, while the clinician maintains the professional judgment and the explanation to the patient as an integral part of the consent (Nouis *et al*., 2025). The issue of fairness is strictly relational: biases in data can reverberate through clinical conversations and choices, impacting historically marginalized populations; hence the insistence on "*fair-by-design*" pipelines and evaluations in *real-world settings* (Hanna *et al.*, 2025).

Furthermore, clinicians, patients and caregivers interpret AI through different cultural frames: expected benefits, fears of substitution, questions about what makes care "care" – elements that influence the acceptability and sustainability of adoption (Baillie *et al.*, 2025). Within the visit, AI output serves as a decision-making framework that can guide preferences and counterfactual inferences: the clinician's task is to help the patient and caregiver situate the algorithmic evidence within their own values, weigh *trade-offs*, and recognise uncertainty as a constitutive part of the decision (Hassan *et al.*, 2024). Where this is lacking, AI tends to lend its suggestions an aura of inevitability or, conversely, generate over-trust towards unvalidated tools. Yet some applications-such as the guided simplification of radiological reports or the use of "virtual patients" in training-show immediate benefits, although effectiveness metrics remain immature (Holderried *et al.*, 2024). A further transformative vector concerns the extra-clinical use of AI by patients and caregivers: many arrive at the visit after having "dialogued" with a model and seek validation or refutation, redefining the boundaries of the relationship and introducing new challenges of conversational safety and epistemic negotiation (Goldberg, 2024).

At a social level, attitudes are mixed: some citizens expect an *improvement* in the relationship with their doctor thanks to AI, while others fear depersonalization and loss of control; the modulation of these expectations depends on previous experience, the perception of transparency and the way AI is implemented in the clinical context (Nong, Ji, 2025). Qualitative studies with developers, clinicians and patients show that acceptability depends on concrete signals of usefulness and safety (e.g., non-intrusive integration into workflows, response times, explainability) rather than on abstract discourses on "intelligence" and "autonomy", and that perceptual divergences between stakeholders should be addressed through *co-design* and systematic feedback (Baillie *et al*., 2025). Ultimately, the most solid trajectory seems to be the one that links *relational benefit* and *responsible design*: where AI is made visible, explainable, debatable and *negotiable* (i.e.

negotiable on the merits), the relationship not only holds but can improve – more time to look, less bureaucracy, more informed choices; where, on the other hand, AI operates in an opaque mode or imposes an additional cognitive load on the patient, the alliance cracks, with effects of distrust or passive delegation (Alboksmaty *et al*., 2025).

WHO guidelines and European legislation (AI Act, 2024) guide this scenario, imposing transparency, data quality, human supervision and the possibility for the patient to refuse the use of AI, with direct implications on clinical communication and the need for proportionate explanations (WHO, 2024; Van Kolfschooten, Van Oirschot, 2024).

3. Algorithmic bias and reproduction of inequalities in health care

The introduction of artificial intelligence into clinical settings represents not only a technological evolution but a structural transformation of the doctor-patient relationship. Several critical issues emerge on epistemic, relational, organisational, and ethical-legal levels. An initial weakness arises from the lack of transparency of complex models and *black box* systems, which makes it difficult for doctors and patients to understand the decision-making criteria underlying algorithmic recommendations (Tonekaboni *et al*., 2019). The lack of explainability directly impacts shared decision-making and can lead to a loss of orientation and control, especially in the most vulnerable patients or those with reduced health literacy. Added to this is the risk of *automation bias*, well-documented in the experimental literature, according to which exposure to incorrect suggestions from a decision support system induces diagnostic conformity and reduces clinical accuracy (Jabbour *et al*., 2023).

On a relational level, the perception of opacity or uncontrolled delegation to AI can undermine therapeutic trust (Longoni *et al*., 2019). The specific risks of generative language models – *hallucinations*, omissions, and overconfidence – can also lead patients to attribute unwarranted trust to automated responses, especially when adequate professional supervision is lacking (Madabushi, Jones, 2025). Consequently, trust becomes a practice to be built in the clinical encounter, as the WHO also recalls in its guidelines, which emphasise the need for transparency, human supervision, and calibrated expectations (WHO, 2024).

Equity front, the risks are equally significant. The emblematic case highlighted by Obermeyer *et al.* (2019) shows how a healthcare algorithm, trained using costs as a proxy for clinical need, underestimated severity in

Black patients compared to White patients. Subsequent studies confirm that biases can emerge at any stage of the AI lifecycle and translate into perceived or actual injustices in the care relationship (Cross *et al*., 2024). For this reason, European regulations – including the provisions of the AI Act for high-risk systems – require data quality, traceability, and human oversight (EU, 2024).

On an organisational-relational level, tools such as *ambient AI scribes* demonstrate potential benefits in reducing the burden of documentation, but raise serious questions about informed consent, privacy, security of recorded data, and the perception of surveillance during the clinical encounter (Tierney *et al*., 2024). Without adequate safeguards, the risk is that of compromising the space of trust and vulnerability that is the heart of therapeutic communication. Added to this is the possibility of *deskilling*, already reported in the literature as an unintended consequence of the routine automation of parts of clinical reasoning (Cabitza *et al*., 2017).

Critical issues also extend to responsibility and the allocation of accountability: who is responsible if an adverse outcome results from an algorithmic recommendation? WHO guidelines and European regulations call for socio-technical models with distributed responsibilities, but with the physician always "in the loop" and responsible for explaining the outcome to the patient (WHO, 2024; EC, 2024). Furthermore, explainability – to be useful for consensus – must allow the patient not only to understand the outcome, but also to challenge it, integrating counterfactuals and model limitations (Freyer *et al*., 2024).

Finally, on the relational level, trust depends heavily on the physician's perception of agency. Patients are more accepting of AI when they perceive that the clinician remains primarily responsible for interpretation, critically filters recommendations, and maintains an empathetic attitude that is attentive to the patient's uniqueness (Nagy, Sisk, 2020). Conversely, when AI appears to be a substitute for clinical judgment, mistrust, information reticence, and a deterioration in the quality of the medical history emerge.

Conclusions

Ultimately, the integration of digital and artificial intelligence into healthcare services is now an essential path to addressing the challenges of contemporary healthcare systems and seizing the opportunities offered by technological innovation. WHO guidelines, European Commission poli-

cies, and OECD assessment frameworks outline a complex yet coherent ecosystem, designed to ensure that digital and algorithmic transformation is not simply a process of technological adoption, but a structural change aimed at making healthcare systems more equitable, resilient, efficient, and truly people-centred. The transition of healthcare systems towards AI represents one of the most complex and strategic issues in contemporary public policy. AI is not a neutral technology but a transformative force that redefines roles, relationships, responsibilities, values, and social expectations. The WHO (2021) provides an essential ethical framework to guide this transformation, while the AI Act (2024) introduces an advanced and ambitious regulatory model aimed at ensuring security, transparency, and the protection of fundamental rights.

The introduction of artificial intelligence in healthcare does not represent a simple technological enhancement, but a structural transformation of the care relationship, which from dyadic becomes triadic, with the algorithmic system as a third actor capable of guiding communication, decisions and expectations (de O Campos *et al*., 2025).

Evidence converges on a conditional outcome: AI can strengthen the doctor-patient relationship when it reduces administrative burden, improves information readability, supports decision-making with contextual explanations, and operates under verifiable clinical supervision. It can weaken it when it introduces opacity, amplifies inequalities, or shifts the conversation from the ends of care to the means of calculation.

The opportunities – greater information comprehensibility, decision-making support, and reduced paperwork – coexist with significant risks: bias, opacity, dependence on automation, professional deskilling, and potential distortions in trust. Major international institutions, from the WHO (2021; 2024) to the EU with the AI Act, agree on the need for human oversight, transparency, and robust data governance, especially to protect vulnerable groups and ensure fairness.

The emerging challenge, therefore, is not the "reliability of AI" in the abstract, but the quality of the sociotechnical ecosystem within which it is adopted. Evidence suggests that only a critical, contextual, and distributively just integration will allow AI to enhance – not replace – the relational and deliberative dimension of care, preserving the centrality of the clinical encounter.

In short, the critical issues AI introduces into the doctor-patient relationship are not mere "technological side effects", but sociotechnical challenges that require responsible design and thoughtful practices: equity and performance audits for subgroups, calibration and communication of uncer-

tainty, anti-automation-bias protocols, "explain-back" spaces where patients can reformulate their understanding, accountability and decision-tracking mechanisms integrated into clinical workflows, and governance that combines law, ethics, and empirical evaluation of outcomes. Only "*relational AI*" – integrated into an intervention framework that values clinical judgment, empathy, and shared deliberation – can prevent innovation from undermining the capital of trust on which compliance, safety, and equity of care depend.

References

Alboksmaty A., Aldakhil R., Hayhoe B.W., Ashrafian H., Darzi A., Neves A.L. (2025). The impact of using AI-powered voice-to-text technology for clinical documentation on quality of care in primary care and outpatient settings: a systematic review. *EBioMedicine*, 118.

Arbelaez Ossa L., Rost M., Bont N., Lorenzini G., Shaw D., Elger B.S. (2025). Exploring patient participation in AI-supported health care: qualitative study. *JMIR AI*, 4: e50781.

As'ad M., Faran N., Joharji H. (2025). AI-supported shared decision-making (AI-SDM): conceptual framework. *JMIR AI*, 4: e75866.

Baillie L., Stewart-Lord A., Thomas N., Frings D. (2025). Patients', clinicians' and developers' perspectives and experiences of artificial intelligence in cardiac healthcare: a qualitative study. *Digital Health*, 11: 20552076251328578.

Cabitza F., Alderighi C., Rasoini R., Gensini G.F. (2017). "Handle with care": about the potential unintended consequences of oracular artificial intelligence systems in medicine. *Recenti Progressi in Medicina*, 108(10): 397-401.

Clark K.S.B., Rudell E., Setiadi D., Agrawal T., Oliver B.J. (2024). Beyond shared decision-making: integrating coproduction, learning health systems, artificial intelligence, and workforce development for patient-centered care. *The Permanente Journal*, 28(3): 284.

Cross J.L., Choma M.A., Onofrey J.A. (2024). Bias in medical AI: implications for clinical decision-making. *PLOS Digital Health*, 3(11): e0000651.

de O Campos H., Wolfe D., Luan H., Sim I. (2025). Generative AI as third agent: large language models and the transformation of the clinician-patient relationship. *Journal of Participatory Medicine*, 17(1): e68146.

European Commission (2010). *A digital agenda for Europe*. COM(2010) 245 final.

European Commission (2012). *eHealth action plan 2012–2020: Innovative healthcare for the 21st century*.

European Commission (2018). *Communication on enabling the digital transformation of health and care in the digital single market*. COM(2018) 233 final.

European Commission (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council on artificial intelligence (AI Act)*.

European Union (2022). *European health data space*.

Freyer N., Groß D., Lipprandt M. (2024). The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. *BMC Medical Ethics*, 25(1): 104.

Goisauf M., Cano Abadía M., Akyüz K., Bobowicz M., Buyx A., Colussi I., Meszaros J. (2025). Trust, trustworthiness, and the future of medical AI: outcomes of an interdisciplinary expert workshop. *Journal of Medical Internet Research*, 27: e71236.

Goldberg C. (2024). Patient portal: when patients take AI into their own hands. *NEJM AI*, 1(5): AIp2400283.

Hassan N., Slight R., Bimpong K., Bates D.W., Weiand D., Vellinga A., Slight S.P. (2024). Systematic review to understand users' perspectives on AI-enabled decision aids to inform shared decision making. *npj Digital Medicine*, 7(1): 332.

Holderried F., Stegemann-Philipps C., Herrmann-Werner A., Festl-Wietek T., Holderried M., Eickhoff C., Mahling M. (2024). A language model–powered simulated patient with automated feedback for history taking: prospective study. *JMIR Medical Education*, 10(1): e59213.

Jabbour S., Fouhey D., Shepard S., Valley T.S., Kazerooni E.A., Banovic N., Sjoding M.W. (2023). Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *JAMA*, 330(23): 2275-2284.

Kingsford P.A., Ambrose J.A. (2024). Artificial intelligence and the doctor-patient relationship. *The American Journal of Medicine*, 137(5): 381-382.

Longoni C., Bonezzi A., Morewedge C.K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4): 629-650.

Madabushi H.T., Jones M.D. (2025). Large language models in healthcare information research: making progress in an emerging field. *BMJ Quality & Safety*, 34(2): 73-76.

Maturo A. (2024). *Il primo libro di sociologia della salute*. Torino: Giulio Einaudi Editore.

Nagy M., Sisk B. (2020). How will artificial intelligence affect patient-clinician relationships? *AMA Journal of Ethics*, 22(5): 395-400.

Nong P., Ji M. (2025). Expectations of healthcare AI and the role of trust. *Journal of the American Medical Informatics Association*, 32(5): 795-799.

Nouis S.C., Uren V., Jariwala S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making. *BMC Medical Ethics*, 26(1): 89.

Obermeyer Z., Powers B., Vogeli C., Mullainathan S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447-453.

OECD (2025). *OECD health working papers*, n. 177. Towards identifying good practices in the assessment of digital medical devices.

Pisani A.A. (2024). *Telemedicina: quadro normativo, tutele e diritti, sistema delle responsabilità*. Bologna: Bologna University Press.

Sbaragli S. (2020). Riprogettare i servizi: dalle cure all'e-Care. In: Cipolla C., Dal Molin R., Pipinato C., a cura di, *Di Parkinson si vive. Le dimensioni socio-assistenziali della malattia*. Milano: FrancoAngeli, 151-186.

Schmidt J., Schutte N.M., Buttigieg S., Novillo-Ortiz D., Sutherland E., Mossialos E., van Kessel R. (2024). Mapping the regulatory landscape for artificial intelligence in health within the EU. *npj Digital Medicine*, 7: 229.

Stephan D., Bertsch A.S., Schumacher S., Puladi B., Burwinkel M., Al-Nawas B., Thiem D.G. (2025). Improving patient communication by simplifying AI-generated dental radiology reports with ChatGPT. *Journal of Medical Internet Research*, 27: e73337.

Stroud A.M., Minteer S.A., Zhu X., Ridgeway J.L., Miller J.E., Barry B.A. (2025). Patient information needs for transparent and trustworthy cardiovascular artificial intelligence. *PLOS Digital Health*, 4(4): e0000826.

Tierney A.A., Gayre G., Hoberman B., Mattern B., Ballesca M., Kipnis P., Lee K. (2024). Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3): CAT-23.

Tonekaboni S., Joshi S., McCradden M.D., Goldenberg A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In: *Machine Learning for Healthcare Conference*. PMLR, 359-380.

Van Kolfschooten H., Van Oirschot J. (2024). The EU artificial intelligence act (2024): implications for healthcare. *Health Policy*, 149: 105152.

WHO (2021). *Ethics and governance of artificial intelligence for health*. Geneva: World Health Organization.

WHO (2021). *Global strategy on digital health 2020–2025*. Geneva: World Health Organization.

WHO (2023). *The ongoing journey to commitment and transformation. Digital health in the European Region 2023*. Copenhagen: WHO Regional Office for Europe.

WHO (2024). *Ethics and governance of artificial intelligence for health: 2024 update*. Geneva: World Health Organization.

Yau J.Y.S., Saadat S., Hsu E., Murphy L.S.L., Roh J.S., Suchard J., Langdorf M.I. (2024). Accuracy of prospective assessments of four large language model chatbot responses to patient questions about emergency care. *Journal of Medical Internet Research*, 26: e60291.

Zondag A.G., Rozestraten R., Grimmelikhuijsen S.G., Jongsma K.R., van Solinge W.W., Bots M.L., Haitjema S. (2024). The effect of artificial intelligence on patient-physician trust. *Journal of Medical Internet Research*, 26: e50853.

Povertà sociale e dinamiche usuraie

di *Annamaria Rufino*[*]

L'usura è un "contenitore" multi-complesso, in grado di nascondere fenomeni e derive sociali, economiche e valoriali. L'usura nasconde molteplici fragilità sociali ed economiche e, contemporaneamente, ne amplifica l'impatto, oltre che la capacità degenerativa. Il fenomeno usuraio è, oggi, più che mai dilagante, e ancor più "mascherato" e subdolo. I sistemi istituzionali e giudiziari sembrano occupare un passo indietro, rispetto all'avanzare del fenomeno. Parallelamente, sempre più soggetti, consapevolmente o meno, ne sono vittime.

Parole chiave: povertà; usura; corruzione; crimini; disuguaglianza; sfiducia.

Social poverty and usury dynamics

Usury is a multi-complex "container," capable of concealing social, economic, and value-related phenomena and deviations. Usury conceals multiple social and economic fragilities and, simultaneously, amplifies their impact, as well as their potential for degeneration. Usury is now more widespread than ever, and even more "disguised" and insidious. Institutional and judicial systems appear to be lagging behind the advance of the phenomenon. At the same time, more and more individuals, consciously or unconsciously, are falling victim to it.

Keywords: poverty; usury; corruption; crime; inequality; mistrust.

Introduzione

Il tema dell'usura meriterebbe un'analisi parallela ed innovativa delle problematiche di sottosistema che la "producono" o che, comunque, sono ad essa connesse. Le concause, di tipo economico-produttivo e socio-culturale, che hanno determinato un ulteriore indebolimento ed una ulteriore degenerazione di tanti territori, molto più evidenti in quelli che definiamo "fragili, vedono come dinamica comune proprio le partiche usuraie. Il diffondersi, trasversale, dei sistemi corruttivi ha facilitato l'offuscamento di pratiche di malaffare, volte a traghettare, in modo strutturale e funzionale, dinamiche sociali e lavorative in ambiti non facilmente controllabili, né, tantomeno,

DOI: 10.5281/zenodo.18436104

[*] Università della Campania. annamaria.rufino@unicampania.it.

"visibili". Si pensi al macrosettore economico-malavitoso legato al fenomeno migratorio, per fare un esempio "elementare".

La diffusione di usura ed estorsione si è, così, sovrapposta "normalmente" e specularmente alle difficoltà socio-economiche che insistono in tanti territori, di tanti Paesi. Le dinamiche usuraie ed estorsive, come esternalizzazione risolutiva delle derive economiche, sono alimentate dalla "distanza" del sistema sociale, come sistema di inclusione, e delle istituzioni, quale sistema di controllo, oltre che dall'abnorme partecipazione di tante realtà "operative" e trasversali alle pratiche corruttive. Le "frontiere", sistematicamente sradicate in questo primo squarcio di secolo, non sono solo quelle cancellate con il *favor* dalla globalizzazione, ma anche quelle "volute" dal sistema corruttivo, per il quale l'Italia occupa un posto apicale (Rufino, 2022). La fragilità delle frontiere, intesa in senso cognitivo, costituisce il vero *vulnus* per un approccio adeguato alle problematiche usuraie. La "porosità" del *welfare-state* è, allo stesso tempo, causa ed effetto della fragilità del sistema sociale, il tutto riconducibile alla frantumazione del *frame* normativo e regolativo, che avrebbe dovuto garantirne la tenuta.

1. I dati giudiziari: una problematica contraddittoria

Il fenomeno usuraio osservabile attraverso i dati giudiziari, non renderebbe l'ampiezza dell'impatto, a monte e a valle, di azioni malavitose che si disseminano, soprattutto, nei territori caratterizzati da diffusa illegalità e povertà, sociale ed economica. L'analisi quantitativa a cui si potrebbe pervenire non darebbe conto del dato nella sua ampiezza. Gli studi criminologici, antropologici, economici e sociologici di una dinamica che possiamo definire, per tanti versi, "storica", dovrebbero convergere in un profondo ripensamento e ad una aggiornata rilettura del sistema iper-complesso che la connota, tale da consentirci di definire adeguatamente il tema dell'usura e del riciclaggio. Al contrario, le analisi che vengono frequentemente pubblicate evidenziano un'autoreferenzialità interpretativa, che non consente di potenziare l'analisi stessa attraverso un confronto tematico interdisciplinare.

Non si può sottovalutare il profondo cambiamento sistemico che caratterizza certe dinamiche. Un cambiamento che ha prodotto, paradossalmente, risultati in un'ottica di semplificazione persino, e ancor più grave, nei risultati giudiziari, debilitando le possibilità osservative dello stesso rapporto tra normale e deviante. Il cambiamento intersistemico e intrasistemico ha mixato la struttura stessa dell'impresa criminale, individuale o no, sezionando e scollegando, al contempo, i vari ambiti dove il fenomeno si manifesta. Di

qui la fallacia delle risposte istituzionali, come quelle preventive; le difficoltà di quelle sociali, in termini di reazione e controllo sociale; e di quelle giudiziarie, riduttive e, molto spesso, inefficaci, non foss'altro che per la lunghezza processuale, mediamente misurabile in 7 anni. Basterebbero alcune, semplici, domande, per "correggere" la distrazione: come è possibile intervenire per prevedere ed arginare il fenomeno? E, ancora: cosa lo determina, al punto tale da escludere la valutazione del rischio conseguente? Le vittime prevalenti, imprenditori, commercianti e semplici cittadini, riassumono l'ambito analitico dove sarebbe possibile applicare gli strumenti cognitivi necessari per un'azione di prevenzione di più vasto raggio.

Le analisi quantitative frequentemente prodotte, rispetto agli esiti giudiziari, dimostrano, al contrario, quanto il dinamismo ipercomplesso del sistema usuraio pregiudichi la decifrabilità dei nodi connettivi, rendendo improduttivi quelli individuativi e correttivi. Il sistema sottostante alle pratiche usuraie è patologicamente dialogante con il mercato del crimine, fatto di intercambiabilità soggettiva, di inattendibilità della scelta razionale nella valutazione del rapporto costi/benefici, sia per i soggetti agenti, che per quelli soccombenti (*status* legale/*status* illegale). Un vuoto di senso che coincide con un'assenza di rappresentazione nel sistema sociale (Beck, 2000). A conferma, si aggiunga l'*impasse* istituzionale, che possiamo riassumere in un deficit cognitivo ed osservativo, oltre che, ovviamente, correttivo. Il sistema regolativo e quello giudiziario convergono in questa inconsapevolezza. Il fenomeno diviene, così, tridimensionale: istituzionale, giudiziario e sociale, e, in quanto tale, richiederebbe un'analisi multifocale, soprattutto per ridurre la parzialità delle *policies* da adottare.

Un obiettivo "sensibile" è, naturalmente, quello economico, ma non va sottovalutato quello specificamente territoriale, con particolare riferimento alla struttura sociale che insiste in tanti territori caratterizzati dal fenomeno usuraio, soprattutto di quelli definibili "fragili", in relazione al rapporto valori/norme, alla misurazione del rischio e alla presenza di sistemi malavitosi. Una struttura socio-territoriale che ha visto un'accelerazione delle dinamiche trasformative del lavoro, soprattutto per il disseminarsi di stratificazioni indistinte di popolazione non autoctona (Rufino, 2024b). La percezione di una generalizzata "irregolarità", sicuramente condizionata dalla diffusione dei sistemi corruttivi, è un elemento determinante per la facilitazione, a monte e a valle, del fenomeno usuraio. Uno dei paradossi che scaturisce dalla diffusione di tali "modelli" riguarda, in modo significativo, la percezione dell'uguaglianza sociale o meglio l'aspirazione a tale uguaglianza, parallela e interagente con quella economica.

2. Usura, devianza e latenza regolativa

Il fenomeno usuraio rappresenta il lato oscuro della società, uno tra i tanti, occorrerebbe individuare, perciò, le strutture latenti del fenomeno (Luhmann, 1990), che contribuiscono non solo ad incrementarlo, ma anche a nasconderlo. Le "irregolarità" sono strettamente connesse alle dinamiche trasformative dei sistemi malavitosi, che hanno potenziato le azioni di autogenerazione del fenomeno usuraio, dove la contaminazione distorsiva confermerebbe che "nulla è un rischio in sé stesso" (Ewald, 1993). La disattivazione del meccanismo della colpa contribuisce ulteriormente ad oscurare la struttura connettiva del sistema usuraio, alimentandosi degli spazi di intersezione, attrattivi per il mercato criminale.

Come avrebbe detto Becker, il processo di creazione della devianza inizia quando le norme vengono prodotte e non quando vengono violate (Becker, 1968). Le difficoltà economiche e del mondo del lavoro, abnormi in alcuni territori, non trovano risposte regolative adeguate e questo alimenta, ancor più, la disseminazione di subculture devianti e criminali. Risalire la china "fidando" sull'aiuto di soggetti criminali è, molto spesso, connesso alla frantumazione dei legami sociali, come dimostrato dalla *bonding theory* (Hirshi, 1969), legami che potrebbero "contenere", a monte, la frantumazione della fiducia e della violazione delle norme.

I soggetti usurati sono "invisibili", quanto gli usurai. Questo il dato più macroscopico, che emerge proprio dall'entità del dato nel suo complesso. L'esiguità della risposta istituzionale e giudiziaria non può produrre alcun effetto deterrente, soprattutto a fronte di tale oscurità osservativa, anzi si trasforma in un "servizio" per la criminalità, che rimodula la debolezza dei soggetti usurati in una forma di clandestinità autogenerantesi, fonte di ulteriore alimento per l'azione usuraia. L'usura si presenta, così, come un reato paradossalmente residuale, per quanto drammaticamente antisociale.

Non tutte le forme di usura e di estorsione e non tutti gli operatori sono uguali: in questo senso l'eterogeneità delle tipologie dovrebbe condurre all'attivazione di una regolazione sistematica e, non da meno, dell'azione di polizia predittiva. Una politica di condanna ottimale dovrebbe puntare alla deterrenza ovvero non focalizzarsi sul crimine in sé, ma su "come" viene commesso. L'analisi di settore non dovrebbe, inoltre, prescindere, come avviene troppo spesso, dall'utilizzo della letteratura economica, in tema di lavoro e mobilità sociale. Dunque, utile sarebbe studiare, a monte e a valle dell'impegno osservativo e regolativo, l'evoluzione dell'azione usuraia, con approccio econometrico, non limitandosi a semplici correlazioni, così da ottimizzare l'analisi della causalità delle azioni e delle conseguenti relazioni.

La punizione/sentenza, normalmente riconducibile "solo" alle condanne, ha, certamente, un effetto deterrente, dissuasivo, ma va rafforzata da un intervento preventivo e predittivo "anche" delle recidive, agendo sulla strutturazione osservativa delle stesse recidive, funzionale all'attivazione del controllo sociale.

Conclusione

L'usura è una costante invisibile, che ha attraversato e condizionato le trasformazioni del sistema sociale di ieri e di oggi, parallela, in questa costanza, alla corruzione (Rufino, 2024a), accomunate, come dinamiche, nella stessa valutazione del rapporto tra scelta dell'azione criminale ed eventuale costo.

Quanto sappiamo dei rischi disseminativi del fenomeno usuraio e delle possibili recidive? Un'analisi predittiva non dovrebbe prescindere da un'osservazione attenta del mutamento sociale e dalle derive comportamentali, causa ed effetto del fenomeno osservato. Pensiamo all'aumento delle "dipendenze" in segmenti di popolazione prima esclusi o non direttamente coinvolti, come donne e anziani. La crisi delle certezze assistenziali, delle fragilità psicologiche e delle difficoltà economiche alimenta le derive usuraie, crea "luoghi" di patologica sottomissione a comportamenti vessatori e ricattatori. Ma non va sottovalutata un'ulteriore emergenza sociale, che vede coinvolte le giovani generazioni. Un esempio potrebbero essere le patologie da gioco, un "luogo" dove i fenomeni degenerativi convergono in un *unicum*, di malaffare, fragilità e permanenza della condizione di vittima, coinvolgendo, in modo trasversale, generazioni e status sociali diversi. La sconnessione tra osservazione, prevenzione e controllo converge in una "semplificazione" sommariamente correttiva del fenomeno usuraio, ben oltre il dato, drammaticamente esiguo, della numerosità delle condanne.

Riferimenti bibliografici

Beck U. (2000). *La società del rischio. Verso una seconda modernità*. Roma: Carocci.

Becker G.S. (1968). Crime and punishment: an economic approach. *Journal of Political Economy*, 76(2).

Ewald F. (1993). Two infinities of risk. In: Massumi B., a cura di, *The politics of everyday fear*. Minneapolis: University of Minnesota Press.

Hirschi T. (1969). *Causes of delinquency*. Berkeley (CA): University of California Press.

Luhmann N. (1990). *Sistemi sociali. Fondamenti di una teoria generale*. Bologna: Il Mulino.

Rufino A. (2022). Anomic dependence and corruptive contagion. Regulatory hypercomplexity and social fragmentation in the mid-global era. *Italian Sociological Review*, n. 2S.

Rufino A. (2024a). Transformative dynamics of corruptive systems. *Italian Sociological Review*, 14(1).

Rufino A. (2024b). Usura, concussione e riciclaggio. Vecchie dinamiche corruttive e nuove forme di povertà. *Lex et Jus*, n. 2.

Ragazzi che delinquono: storie di vita tra Napoli e Città del Messico

di *Mario Osorio-Beristain*[*]

Il presente articolo è il risultato di una ricerca di dottorato sulla partecipazione di adolescenti in attività delinquenziali e sul loro coinvolgimento in gruppi della criminalità organizzata a Napoli e Città del Messico. A partire dalla voce dei ragazzi intervistati, l'autore ha cercato di interpretare la loro realtà di vita e conoscere le motivazioni della loro "scelta" deviante, avvalendosi del quadro teorico sviluppato dal sociologo Pierre Bourdieu, utilizzato con crescente frequenza nelle ricerche sulla criminalità.

Parole chiavi: minori; criminalità; Napoli; Città del Messico; strada; habitus; capitale.

Juvenile delinquents: a comparison between Naples and Mexico City

This article presents the findings of a doctoral research study conducted in Naples and Mexico City, focusing on the participation of adolescents in criminal activities and their involvement with organized crime groups. Central to this research is the inclusion of the adolescents' own voices, through which the author endeavors to provide an authentic glimpse into their lived realities and uncover the motivations behind their deviant choices. The study employs an interpretative framework grounded in Pierre Bourdie's sociological theory, offering a nuanced understanding of the social and structural factors that influence these adolescents' paths into criminality.

Keywords: minors; criminality; Naples; Mexico City; street; habitus; capital.

Introduzione

Anche se Città del Messico e Napoli sono diverse e geograficamente lontane, in entrambe le metropoli i processi globali di aumento delle diseguaglianze, emarginazione e segregazione socio-spaziale della popolazione vulnerabile hanno avuto forti implicazioni nelle abitudini e nella vita quotidiana delle persone. In entrambe le città la globalizzazione ha determinato quello

DOI: 10.5281/zenodo.18436121

[*] Università degli Studi di Roma "La Sapienza". m.osorio@mclink.it.

che Bourdieu (2015) chiama il ritiro dello Stato e l'apparizione o, piuttosto, la strutturazione di luoghi di confine e/o emarginazione nei quali si trovano concentrate le popolazioni più bisognose (Ibidem: 242).

Wacquant, il principale allievo di Bourdieu, parla di marginalità avanzata, ovvero «il regime di relegazione socio spaziale e di chiusura escludente che si è cristallizzato nella città postfordista come risultato dello sviluppo ineguale delle economie capitalistiche e della ritirata del welfare statale» (2016: 30).

Quindi, l'esistenza di questi spazi non è una caratteristica esclusiva di Napoli e Città del Messico, ma in entrambe le metropoli è forte e documentata la presenza e il radicamento di organizzazioni criminali. Queste tendono a reclutare giovani, incluso bambini e adolescenti, i quali rappresentano un esercito di manodopera di riserva, ma anche una vera e propria linfa vitale che permette a queste organizzazioni di rigenerarsi (Direzione Investigativa Antimafia, 2018).

Il Ministero della Giustizia (2017) italiano ha dimostrato che le organizzazioni mafiose reclutano e si avvalgono di minorenni per lo svolgimento di attività illecite, talvolta facendo leva sulla loro condizione di non imputabilità – l'articolo 97 del Codice Penale italiano indica che il minore infraquattordicenne non è mai imputabile e l'articolo 98 stabilisce che è imputabile chi, nel momento in cui ha commesso il fatto, aveva compiuto 14 anni ma non ancora 18, se aveva capacità di intendere e di volere (Ministero della Giustiza, 2015).

Anche in Messico il reclutamento di minori e bambini da parte dei cartelli della droga è un fenomeno molto diffuso, che spesso si basa sulla non imputabilità dei minori di 14 anni e sulle pene più lievi previste per coloro che non hanno ancora compiuto 17 anni (CNDH, 2022). Un recente rapporto (Reinserta, 2021) ha calcolato che circa 30.000 minori e adolescenti lavorano per gruppi della criminalità organizzata, soprattutto nelle zone più isolate e povere del paese, dove ampi strati della popolazione vivono in situazione di forte vulnerabilità.

A volte per questi minori l'affiliazione a gruppi della criminalità organizzata rappresenta l'unica strategia di sopravvivenza possibile, altre volte essi sono addirittura reclutati con la forza.

Alla luce di questa situazione, il presente articolo ha come obiettivo comprendere e analizzare l'opinione di alcuni giovani direttamente coinvolti in attività delinquenziali, talvolta realizzate in collaborazione con gruppi della criminalità organizzata. Il contributo è stato sviluppato a partire da una ricerca di dottorato sul tema della partecipazione di adolescenti in attività criminali a Napoli e Città del Messico.

La ricerca, condotta tra il 2020 e il 2023, è nata dall'interesse di indagare le ragioni per cui molti ragazzi vengono reclutati dalla criminalità organizzata e rappresenta, al tempo stesso, un tentativo di comprendere i meccanismi e le ragioni che spingono adolescenti, e persino bambini, ad entrare nell'orbita di questi gruppi criminali.

In totale sono state realizzate 56 interviste (20 a Napoli e 36 a Città del Messico) a giovani di sesso maschile ritenuti colpevoli di aver commesso reati di varia natura quando erano ancora minorenni. Al momento dell'intervista, essi si trovavano in situazione detentiva in istituti penali per minori o erano sottoposti a misure cautelari presso comunità dedicate (messa alla prova in Italia) o, nel caso messicano, erano seguiti da una Ong locale, in quanto nel paese latinoamericano sono questo tipo di organizzazioni che prendono in carico gli adolescenti che svolgono misure alternative al carcere.

Sono state inoltre realizzate 9 interviste a informatori chiave (operatori sociali e personale della polizia penitenziaria) la cui collaborazione si è rivelata fondamentale per comprendere più a fondo la problematica oggetto della ricerca e ottenere importanti informazioni contestuali.

Per ragioni di spazio nel presente articolo sono state riportati e commentati soltanto alcuni stralci particolarmente significativi delle interviste con i ragazzi, i quali hanno preso parte all'indagine liberamente dopo aver espresso il proprio consenso informato e i cui nomi sono stati modificati per garantirne l'anonimato.

1. La teoria di Bourdieu e la strada

Si è scelto di adottare un'impostazione teorica derivante dalle concettualizzazioni del sociologo francese Pierre Bourdieu che, attraverso strumenti concettuali come l'habitus, il campo e le forme di capitale, analizza il modo in cui ampie strutture culturali e sociali – tali come la povertà, la disoccupazione, l'origine sociale o l'appartenenza di classe – interagiscono a livello individuale e di gruppo per dare forma a disposizioni e schemi inconsci che poi influenzano fortemente tutte le pratiche sociali, quindi anche quelle deviante e criminali.

Negli ultimi anni la ricerca sulla devianza ha fatto sempre più spesso riferimento al quadro teorico sviluppato da Bourdieu, considerato da molti ricercatori come uno dei più sofisticati approcci per lo studio delle differenze e diseguaglianze sociali (Fleetwood, 2019).

Bourdieu elabora l'idea che ci siano due modalità di esistenza del sociale costantemente in relazione tra loro: da una parte le strutture sociali esterne,

cioè il sociale diventato "cosa" (il campo politico, il campo religioso, ecc.) e, dall'altra, le strutture sociali interiorizzate e incorporate dal soggetto in forma di schemi di pensiero, percezione e azione (Martìn Criado, 2008).

Il concetto di habitus si riferisce all'insieme di questi schemi, attraverso cui gli individui percepiscono il mondo e operano dentro di esso. Essi sono socialmente strutturati poiché si formano nel corso della storia personale di ogni individuo e presuppongono l'interiorizzazione della struttura sociale dello specifico campo di rapporti sociali in cui l'individuo è stato formato. Allo stesso tempo, però, gli habitus sono anche strutturanti e generativi, in quanto contribuiscono a formare le strutture che ogni soggetto utilizza per articolare i propri pensieri, percezioni e azioni (Bourdieu, 2007).

Habitus è la traduzione in latino proposta da Tommaso D'Aquino del termine aristotelico "hexi" (Paolucci, 2011), che significa una "disposizione durevole" o "stabilmente acquisita" capace di orientare le nostre percezioni e i nostri desideri.

In sociologia il concetto è stato usato da autori classici, come Durkheim, Mauss e Weber ma anche dagli esponenti della tradizione fenomenologica di Husserl, per i quali l'habitus si riferisce alla condotta mentale localizzata tra le esperienze passate e le azioni future (Ibidem). Ad accomunare le diverse interpretazioni è il riferimento a qualcosa di acquisito attraverso l'apprendimento, dunque costruito socialmente e storicamente.

Secondo Bourdieu, l'habitus consiste in «sistemi di disposizioni durature e trasferibili, strutture strutturate predisposte a funzionare come strutture strutturanti, vale a dire come principio di generazione e di strutturazione di pratiche e di rappresentazioni che possono essere oggettivamente regolate e regolari, senza essere il prodotto dell'obbedienza alle regole, oggettivamente adattate al loro scopo, senza presupporre l'intenzione cosciente dei fini e il dominio intenzionale delle operazioni necessarie per raggiungerle e, dato tutto questo, che possono essere collettivamente orchestrate senza essere il prodotto dell'azione organizzatrice di un direttore di orchestra» (2007: 86).

È a partire dell'habitus che gli individui producono pensieri e pratiche – le quali sono, quindi, il risultato dell'interiorizzazione delle strutture del gruppo sociale di appartenenza –, prendono decisioni e creano un insieme di schemi pratici di percezione, ovvero di categorizzazione del mondo.

«Prodotto della storia, l'habitus produce pratiche, individuali e collettive, e pertanto (produce) storia in accordo con gli schemi creati per la storia; l'habitus garantisce la presenza attiva delle esperienze passate che, registrate da ogni organismo sotto forma di schemi di percezione, di pensiero e di azione tendono, con più certezza che tutte le regole formali e tutte le normi esplicite,

a garantire la conformità delle pratiche e la loro costanza nel tempo» (Ibidem: 88).

Inoltre, la teoria bourdieuiana che vede l'agire umano come pratica risultante dalle disposizioni interne acquisite socialmente e storicamente dagli individui e dai gruppi sociali deve essere contestualizzata all'interno delle più ampie teorizzazioni di Bourdieu sullo spazio sociale, un termine che l'autore preferisce rispetto a quello, considerato troppo generico, di società.

Lo spazio sociale, nel quale sono collocati gli attori sociali, è differenziato e si articola in diverse sfere chiamate campi, ovvero settori o microcosmi sociali dove, nelle società complesse e per effetto della divisione del lavoro, le attività umane tendono a organizzarsi in modo relativamente autonomo, grazie a specifiche forme e/o regole di funzionamento.

> Il campo stesso può essere pensato come uno spazio di posizioni che agisce come uno spazio di possibili forze, esercitandosi su coloro che vi accedono. Ma se sappiamo che questi campi sono appresi di agenti sociali dotati di habitus, vale a dire modelli di percezione e apprezzamento, che consentono loro di strutturare questo spazio, di catturarlo come ordinato e non in quanto tale, attraverso le sue manifestazioni (...) vediamo che il campo delle forze funziona anche come spazio agito e, in una certa misura, rappresentato (Bourdieu, 2015b: 195-196).

Il campo è caratterizzato anche da rapporti di alleanza tra coloro che ne fanno parte, finalizzati a ottenere maggiori benefici o a imporre come legittima la visione del proprio gruppo.

Ogni campo è concepito come uno spazio di distribuzione di specifiche risorse, che Bourdieu definisce come capitale. Il termine, di derivazione economica, viene utilizzato in senso più ampio per descrivere ciò il cui possesso conferisce potere ai diversi attori sociali che si trovano all'interno di un determinato campo. In questa prospettiva, il capitale rappresenta qualsiasi risorsa che conferisca vantaggi a un attore sociale, suscettibile di essere accumulata e riprodotta nel tempo attraverso meccanismi di trasmissione ereditaria (Santoro, 2015).

> Capitale è lavoro accumulato (nelle sue forme materializzate o in forme 'incorporate' o incarnate) che quando entra in possesso di un singolo, un gruppo o più gruppi di attori sociali, permette loro di appropriarsi di energia sociale in forma di lavoro oggettivato o umano. È una 'vis insita', una forza inscritta in

> strutture oggettive e soggettive, ma è anche 'lex insita', il principio sottostante alle regolarità immanenti al mondo sociale. Capitale che, nelle sue forme oggettive o incarnate, necessita tempo per accumularsi e che, come capitale potenziale può produrre profitto e riprodurre se stesso in forma identica o anche in una forma più stessa (Bourdieu, 1986: 241).

La ricerca sulla devianza ha ripreso e adattato al proprio campo di studi i tre concetti essenziali della teoria di Bourdieu, che sono così l'habitus della strada, il campo della strada e il capitale della strada.

Shammas e Sandberg (2016) definiscono il campo della strada come quello specifico spazio dove hanno luogo determinate forme di criminalità o devianza. Si tratta di uno spazio relativamente autonomo, che presenta una sua logica di funzionamento culturalmente codificata, all'interno del quale gli attori coinvolti in attività criminali competono per migliorare la propria posizione sociale (Rinaldi, 2021).

Il termine strada è chiaramente utilizzato in senso metaforico e simbolico poiché il campo della strada fa riferimento a determinate attività criminali o devianti, come per esempio lo spaccio, il furto e l'estorsione, solitamente gestite dal crimine organizzato, che possono verificarsi in qualunque luogo e che non si esauriscono nello spazio fisico della strada.

L'habitus della strada può essere concettualizzato come le disposizioni relativamente permanenti e a volte inconsce, prodotte e valorizzate all'interno del campo sociale della strada, e che permettono agli attori di operarvi con successo (Shammas, Sandberg, 2016).

Inoltre, le attività connesse alla criminalità urbana richiedono l'acquisizione e il possesso di certe competenze, abilità e conoscenze (come tagliare e vendere la droga, scegliere la vittima di un furto, eseguire con destrezza lo stesso furto ecc.) che alcuni autori hanno definito come capitale umano deviante (McCarthy, Hagan, 2001) ed altri come capitale di strada (Sandberg, Pedersen, 2011). Si tratta quindi di un capitale culturale incorporato che appartiene alla sfera delle competenze e delle conoscenze e che nel campo della strada si può tradurre anche in un uso spregiudicato della violenza.

La violenza, che è un fenomeno estremamente complesso, è un atto di esercizio del potere che implica l'inflizione intenzionale di danni sugli altri (principalmente sotto forma di sopraffazione fisica, ma non solo) (Popitz, 1990), «è usata dai gruppi criminali per raggiungere i propri obiettivi di acquisizione di potere e ricchezza, prendendo parte in scontri violenti, sia

all'interno che all'esterno dei contesti criminali di appartenenza con l'obiettivo di affermare la propria egemonia e autorità» (Massari, Martone, 2019: 1).

Randall Collins (2008) sostiene che le situazioni violente sono modellate dallo stato emotivo della tensione e la paura e, quindi, la violenza è usata dagli esseri umani solo in virtù di specifiche condizioni che aiutano a superare quelle barriere emotive che inibiscono naturalmente i comportamenti violenti.

2. Napoli e Città del Messico

Ci sono differenze notevoli tra Napoli e Città del Messico. E anche se esse derivano dal diverso posizionamento dell'Italia e del Messico all'interno delle geografie dello sviluppo e della distribuzione del potere a scala mondiale, hanno anche a che vedere con il fatto che Città del Messico è una megalopoli, capitale nazionale, con un PIL calcolato in 267,3 miliardi di dollari (Inegi, 2024). Napoli, invece, è una capitale regionale, con un PIL di 28,4 miliardi di euro (Comune di Napoli, 2024).

In seconda istanza, il numero di abitanti differisce enormemente tra le due città: Napoli ha 914,873 abitanti che arrivano a 2.973,688 con i 90 comuni metropolitani (Istat, 2022), mentre a Città del Messico risiedono 9.319,011 persone, che arrivano a 22 milioni prendendo in considerazione l'intera area metropolitana (Inegi, 2020).

Anche la composizione della popolazione varia significativamente tra le due città: Napoli è una delle province più giovani d'Italia, con un'età media di 42 anni (Istat, 2021), inferiore alla media nazionale di 46,2 anni (Istat, 2022). Al contrario, i residenti di Città del Messico hanno un'età media di 35 anni, superiore alla media nazionale di 29 anni (Inegi, 2020), ma comunque decisamente più giovane in comparazione con la popolazione italiana.

Ad accomunare le due città è invece il forte radicamento di gruppi della criminalità organizzata che spesso si servono di minori, bambini e adolescenti per le proprie attività illecite.

Napoli conta con la presenza storica nel suo territorio della Camorra, l'entità criminale europea con il maggior numero di affiliati (Saviano, 2006). Essa non presenta la struttura unificata e relativamente omogenea di altre mafie italiane e, avendo adottato prevalentemente un modello organizzativo orizzontale, è caratterizzata da un'architettura instabile, con guerre continue tra bande e una limitata capacità di contenere l'uso della violenza. Questo spinge alcuni autori a parlare di "camorre" al plurale (Massari, Martone, 2019) e a considerarla quasi un particolare tipo di delinquenza di strada.

Per questa ragione è forse l'organizzazione criminale italiana che più somiglia ai gruppi criminali messicani, egualmente caratterizzati dall'estrema instabilità organizzativa, che si accompagna però ad un uso più spregiudicato, sistematico e generalizzato della violenza esplicita.

A partire dalla metà degli anni '90, la Camorra, analogamente alle altre mafie italiane, ha infatti ridotto l'uso della violenza esplicita, preferendo strategie di cooperazione con altre organizzazioni criminali e con attori politici ed economici, sostituendo così la violenza fisica, l'omicidio e l'intimidazione con uno strumento più sofisticato e convincente: la corruzione (Ibidem).

Inseriti nello scenario economico globale, i clan camorristi hanno stabilito rapporti con specifici settori di affari nei diversi territori in cui operano. Massari e Martone (2019) distinguono due diversi tipi di clan camorristi, quelli provinciali e quelli della città di Napoli. Questi ultimi sono radicati in specifici quartieri e presentano un alto livello di frammentazione, formando piccoli cluster ampiamente coinvolti nei traffici illeciti, l'estorsione e altri business legati all'economia informale (contrabbando, prostituzione, gioco d'azzardo clandestino, ricettazione, usura, ecc.).

Città del Messico, storicamente esclusa dall'infiltrazione dei grandi cartelli della droga, è oggi caratterizzata dalla crescente presenza di queste organizzazioni (Nieto, 2020).

Secondo alcuni studi, a Città del Messico operano due tipologie di gruppi criminali: quelli transnazionali, pienamente inseriti nella rete globale del narcotraffico, e quelli dediti ai business locali (Mendoza, 2016). Diversamente di quello che succede nel mercato illegale del narcotraffico, che tende a intrecciare con il mercato domestico (Bergman, 2016), molte organizzazioni criminali di Città del Messico non si sono sviluppate a partire dal business delle droghe illecite (anche se questa attività è diventata col tempo una delle loro principali fonti di guadagno), bensì dall'estorsione e dalla vendita di protezione.

Anche i gruppi criminali messicani utilizzano la corruzione, in particolare nei confronti di figure istituzionali, come la polizia, gli amministratori locali o i politici. Tuttavia, negli ultimi anni il contesto messicano ha registrato un crescente uso della violenza, fenomeno attribuibile sia alla instabilità del mercato delle droghe illecite sia, soprattutto, alla strategia di repressione militare adottata a partire del 2006 (Pereyra, 2012).

La maggiore repressione messe in atto delle istituzioni pubbliche può avere determinato un aumento nei livelli di violenza poiché ha causato la rimozione di alcuni attori di primo piano, generando un vuoto di potere che

ha scatenato una competizione violenta tra i nuovi leaders (Andreas, Wallman, 2009). Inoltre, la repressione militare ha provocato una frammentazione dei gruppi criminali, che sono stati costretti a decentralizzare la gestione delle proprie operazioni dopo l'arresto o l'uccisione di alcuni capi (Corrado, 2013). Ciò ha portato alla nascita di cellule criminali strutturalmente decentralizzate che operano all'interno di un sistema con un basso grado di gerarchizzazione, processi decisionali decentralizzati e la coesistenza di molti capi (Dishman, 2005). Alcuni studiosi ritengono che la strategia militare sia stata inefficace, in quanto ha determinato un aumento della violenza e della capacità di corruzione dei gruppi criminali, i cui patrimoni illegali o occultati nell'economia legale sono rimasti intatti, conferendo loro un grande potere economico e, di conseguenza, corruttivo (Buscaglia, 2013).

3. Parlano i Ragazzi

Per sviluppare il lavoro è stato utilizzato un approccio qualitativo. In particolare, lo strumento dell'intervista semi-strutturata è stato ritenuto il più efficace per raggiungere l'obiettivo di accedere alle prospettive dei soggetti intervistati.

La traccia dell'intervista, pur mantenendo una certa flessibilità ed apertura, ha raccolto informazioni e interpretazioni su diversi aspetti dell'esperienza di vita dei ragazzi: reato commesso; ragione che ha portato a essere coinvolto nell'attività e/o nel gruppo deviante; situazione scolastica e lavorativa; contesto sociale e familiare; prospettive future.

Un aspetto che è risultato subito evidente durante il lavoro di campo in ambedue i contesti riguarda l'influenza del contesto sociale nella "scelta" delinquenziale dei giovani. Sia a Napoli che a Città del Messico la maggioranza dei ragazzi intervistati proviene da quartieri operai, emarginati e/o malfamati, con alti tassi di disoccupazione giovanile e servizi inadeguati, nei quali la devianza può non essere percepita come tale in quanto forma parte della vita di tutti i giorni. Le azioni e le scelte degli intervistati sembrano così essere fortemente influenzate da quello che Bourdieu (2007) chiama habitus individuale o di gruppo, in questo caso l'habitus della strada.

Questi ragazzi sono cresciuti in ambienti deprivati e caratterizzati dalla presenza diffusa di gruppi criminali, i quali rappresentano un referente identitario di elevato valore simbolico, con i loro capi collocati al vertice della piramide sociale del prestigio. Gli operatori intervistati hanno evidenziato come la criminalità organizzata eserciti su questi giovani un forte potere di attrazione, non soltanto perché capace di soddisfare un bisogno economico,

ma anche per rispondere a esigenze simboliche di appartenenza, identità e dal desiderio di dare significato alle proprie vite attraverso la condivisione di valori e regole proprie e in opposizione rispetto a quelle del gruppo dominante.

Esemplificativo è il caso di Daniele, diciottenne intervistato in una comunità in provincia di Napoli, dove si trovava in regime di misura cautelare dopo essere stato detenuto in un istituto penale per minori con l'accusa di tentata rapina pluriaggravata. Secondo gli operatori che lo seguivano, sebbene la sua famiglia, originaria di un quartiere popolare, non presentasse alcun segno di marginalità, il ragazzo manifestava un fascino particolare per i modelli di strada, le scorciatoie illegali, le soluzioni legate al qui ed ora. Durante l'intervista, un estratto della quale è riportato in seguito, Daniele ha risposto a molte domande con reticenza ed apparso generalmente poco loquace.

> «Domanda (D): Come sei arrivato qui?
> Risposta (R): Sono stato io eh! a organizzare la rapina
> D: Come hai conosciuto il gruppo di ragazzi con i quali hai commesso il reato?
> R: Beh, sono vecchi amici
> D: Del tuo quartiere?
> R: Sì
> D: Quale è stato il reato che avete commesso?
> R: Rapina
> D: Che cosa dovevi rapinare?
> R: Un negozio di telefonini
> D: Cosa è andato storto? Perché sei stato arrestato?
> R: Sono arrivati i carabinieri
> D: Quale ruolo avevi?
> R: Io gli ho puntato la pistola!!!»

Per Pedro, a Città del Messico, la scelta deviante è nata dal bisogno, ma anche dal desiderio di appartenenza offerto da un gruppo criminale. Il ragazzo aveva familiarità con l'esperienza deviante poiché il padre era stato incarcerato per rapina e sequestro di persona. Al momento dell'intervista, Pedro aveva 18 anni ed era detenuto in un istituto penale minorile. Era stato arrestato per la prima volta a 16 anni per rapina a un tassista e sottoposto a misure cautelari in libertà. Successivamente, a 17 anni, era stato nuovamente arrestato e incarcerato per violazione dell'obbligo scolastico e reiterazione del reato.

«D: Quando hai cominciato a rubare?
R: A 14 anni
D: Perché hai cominciato?
R: Per bisogno e per piacere
D: Per bisogno?
R: Sì, perché non avevamo soldi, siamo otto in famiglia e mia madre non lavorava, era casalinga, volevo aiutarla e dimostrare che c'è la potevo fare…
D: E tuo padre?
R: Mio papà lo hanno ucciso quando io avevo 7 anni
D: Chi lo ha ucciso?
R: Lo hanno ucciso quando usciva dal *Reclusorio Oriente* (Il carcere per maschi adulti più grande della città)
D: Perché era stato in carcere?
R: Credo per rapina e sequestro di persona…»

Sia Daniele che Pedro sono cresciuti nel campo della strada e quindi le loro storie possono essere considerate come largamente influenzate dal rapporto tra cultura della strada e le forme di criminalità specifiche dei loro quartieri. In particolare Pedro, con la sua storia familiare, era in possesso di un capitale culturale (capitale della strada) che poteva essere utilizzato per ottenere onore e reputazione, a loro volta spendibili per ottenere in vantaggi economici che gli permettevano di sopravvivere in un contesto violento.

Molti dei ragazzi accusati di rapina avevano commesso reati in gruppo, come Roberto Miguel, di soli 16 anni ma con una lunga esperienza delittuosa. Figlio unico di genitori separati, il padre, con il quale non aveva nessun rapporto, era stato per un lungo periodo in carcere per reati gravi, mentre la madre lavorava tutto il giorno. Il ragazzo è cresciuto praticamente per strada, in un quartiere operaio della zona nord-est di Città del Messico, dove ha cominciato a delinquere e a consumare sostanze stupefacenti anche pesanti. Alto più di un metro e ottanta, dava l'impressione di essere più grande della sua età. Ha confessato di soffrire della sindrome di astinenza dalle droghe, da cui era dipendente quando era in libertà, e di trovarsi per questa ragione sotto l'effetto di psicofarmaci. Nell'intervista, un estratto della quale è riportato in seguito, ha affermato di essere stato accusato di rapina a mano armata e detenzione illegale di armi, ma gli operatori del centro hanno rivelato che era stato condannato anche per sequestro di persona. Gli stessi operatori sospettavano che avesse avuto qualche collegamento con gruppi della criminalità organizzata, ma lui lo ha sempre negato. Il giorno dell'intervista era in isolamento dopo una lite con un altro ragazzo detenuto nel centro penale, anche lui intervistato.

«D: Come sei stato arrestato?

R: Proprio al momento (della rapina). Gli stavo togliendo i soldi… erano 280 mila (pesos, ovvero circa 14 mila euro), avevo già preso i soldi, ma non mi sono reso conto che in un locale vicino c'erano dei poliziotti a pranzo, ed è stato quando il mio socio ha cominciato a urlare (alle vittime) che i poliziotti si sono accorti e quindi…

D: Eri con un altro ragazzo?.

R: Si, era venuto in motocicletta e quando i poliziotti hanno cominciato a sparare lui se ne è andato…

D: Non sei riuscito a salire sulla motocicletta?

R: No

D: Quindi, l'altro ragazzo non è stato arrestato?

R: No, se ne è andato e io sono rimasto lì e a un certo punto ho visto che il poliziotto ha cominciato a sparargli e allora io ho cominciato a sparare al poliziotto…

D: Quindi portavi l'arma?

R: Sì, portavo una (pistola calibro) 9, allora ho cercato di correre, ho fatto una trentina di passi quando ho sentito che mi avevano colpito…

D: Dov'è che ti hanno sparato?

R: La pallottola mi è entrata nel gluteo e mi è uscita dall'area pubica

D: Come hai conosciuto il ragazzo con cui hai fatto la rapina?

R: Proprio nel mio quartiere, nel posto dove mi vedevo con mio cugino, nella *Micho*, così chiamavamo quella strada. Arrivava quel ragazzo, ci mettevamo a fumare, a chiacchierare del più e del meno, dopo abbiamo cominciato a rubare. Dopo che mio cugino ha avuto dei problemi, gli hanno sparato e se n'è dovuto andare via dal quartiere… io sono rimasto senza amici con i quali uscire e così ho fatto amicizia con questo ragazzo, ma non uscivamo soltanto a rubare, anche a divertirci, a cenare fuori, tutti i giorni...»

Anche Roberto Miguel è cresciuto nel campo della strada, cioè, in quello spazio dove hanno luogo determinate forme di attività criminale e nel quale interagiscono quegli attori disposti ad essere ingaggiati nel gioco sociale della devianza criminale, proprio come il ragazzo intervistato, che padroneggiava i codici e i simboli necessari per muoversi con successo in quell'ambiente, come lui stesso ha riferito:

Il fatto è che quando fai questo lavoro impari a riconoscere la tipologia delle persone. Io, anche quando siamo in mezzo al traffico, dentro un taxi, posso vedere le altre macchine e sapere chi porta con

> sé soldi, che cosa hanno nelle mani, al collo, che tipo di telefonino usano, se ne vale la pena, se ci sono due o tre persone, perché quando sono più persone è molto meglio, riscuoti tutto. Anche il tipo di macchina. Sappiamo chi porta soldi anche se non si vede.

Infatti, malgrado la sua giovane età, Roberto Miguel non soltanto era riuscito a rendersi indipendente economicamente e affittare un appartamento per conto proprio, ma dava dei soldi alla madre, alla quale diceva che lavorava come commerciante vendendo articoli sportivi ed altra merce nei mercati del centro.

Il vissuto di Roberto Miguel gli ha consentito di acquisire competenze, abilità e conoscenze che costituiscono un vero e proprio capitale della strada, che rappresenta anche l'interiorizzazione dell'esperienza di emarginazione e grazie al quale era in grado di navigare un contesto violento. In particolare, il ragazzo aveva una rete di relazioni criminali (capitale sociale della strada) che gli permetteva di portare avanti quello che lui chiama il suo "lavoro", ovvero le rapine e rapimenti che li fruttavano più soldi di quanti avrebbe mai potuto guadagnare in modo legale.

Allo stesso modo, anche la capacità di Roberto Miguel e degli altri ragazzi intervistati di orientarsi nella strada e nei suoi diversi ambienti forma parte integrante del capitale culturale di chi è cresciuto immerso in questo campo sociale.

Molte delle storie raccolte, tanto quelle dei ragazzi messicani quanto quelle degli italiani, assomigliano al punto da poter essere sovrapposte le une con le altre. Ad esempio, la esperienza di Corrado presenta molti parallelismi con quella di Roberto Miguel con la differenza, però, che Corrado ha assicurato di non consumare droga.

Al momento dell'intervista, il ragazzo, di 19 anni, si trovava in una comunità di Napoli condannato per rapina a mano armata. Sebbene lui negasse ogni coinvolgimento con il crimine organizzato, spesso cadeva in contraddizione, e gli operatori che lo seguivano hanno confermato che proveniva da una famiglia legata alla camorra: il padre era stato incarcerato per associazione mafiosa e il patrigno era un capo camorrista.

> «D: Come mai sei arrivato in comunità?
> R: Per rapina
> D: Dov'è stata fatta la rapina?
> R: Più di una rapina. Sono state otto rapine in zona Cardito, Frattamaggiore, Afragola…
> D: Dove andavi a fare queste rapine?
> R: Negozi, farmacie, supermercati…

D: Erano a mano armata?
R: Sì, sì, ma io portavo solo lo scooter
D: Quindi stavi con altri ragazzi?
R: Un altro ragazzo
D: Eravate in due sempre?
R: Sì
D: Che fine ha fatto quell'altro ragazzo? E stato arrestato?
R: No, non lo hanno arrestato
D: è scappato?
R: Sì, è scappato
D: Sai dov'è?
R: No, no
D: Come mai sei stato arrestato soltanto tu?
R: Perché io stavo sullo scooter, la macchina dei carabinieri mi ha investito e mi ha buttato a terra
D: Questo ragazzo con il quale facevi le rapine dove l'hai conosciuto?
R: E un ragazzo del quartiere, lui sapeva che io sapevo guidare bene lo scooter e mi ha detto 'vieni con me che ti faccio guadagnare qualcosa' e io, diciamo, mi sono fatto prendere dalla curiosità…»

Corrado era "in odore di camorra" non soltanto per i precedenti in famiglia, ma anche perché già aveva una rilevante esperienza criminale. Infatti, come sottolineano Sales e Melorio, «quando i minorenni entrano nei clan di camorra, hanno già a lungo praticato il crimine e sono già stati socializzati ad una cultura camorristica che hanno non solo respirato ma già persino esercitato nella loro breve vita» (2021: 363).

Dalle interviste è emerso come, per molte delle famiglie dei ragazzi, l'arresto del figlio fosse percepito come un evento non solo accettabile ma normale, quasi un percorso naturale, data la vicinanza con precedenti storie di devianza di tutto il nucleo familiare. In questo senso, queste famiglie condividevano un habitus collettivo che le portava a considerare la traiettoria deviante dei figli come qualcosa di normale e naturale.

È il caso di Giovanni, giovane di 19 anni, originario del quartiere di Secondigliano, accusato di rapina aggravata, che è stato intervistato in una comunità della provincia di Napoli, dove si trovava in misura alternativa alla detenzione cautelare. Il padre, affiliato al clan camorrista Di Lauro, era incarcerato dal 2011 e al momento dell'intervista si trovava in un istituto penitenziario a Milano, con pena fino al 2043 per omicidio e associazione mafiosa. Secondo quanto riferito dagli operatori, tutta la storia della famiglia di Giovanni era stata condizionata dai problemi giudiziari del padre. Il ragazzo ha affermato di non vedere il padre da "un sacco di tempo" e, in effetti, dal

2011, fino all'inizio della pandemia, ogni settimana la madre di Giovanni si era recata a far visita al marito nelle varie carceri italiane in cui era detenuto, accompagnata a turno di uno dei figli. Con l'arrivo della pandemia, però, soltanto lei aveva continuato a viaggiare a Milano per il consueto appuntamento.

Gli operatori hanno sottolineato il fatto che la signora raccontava con estrema naturalezza quella routine e che sembrava non vedere alcuna anomalia nel fatto che, da dieci anni e per altri venti ancora, ogni settimana la sua vita e quella dei figli fosse scandita da questo appuntamento, come se questo facesse parte dell'ordine delle cose, un fatto come tanti altri, un'organizzazione pratica. Tale atteggiamento appariva loro come una sorta di incapacità a cogliere l'aspetto simbolico e profondo della carcerazione del marito, che può essere interpretato come espressione dell'interiorizzazione di uno schema di comportamento (habitus) acquisito nel tempo e socializzato all'interno del gruppo famigliare.

4. L'uso della violenza

Una delle differenze più evidenti tra i ragazzi intervistati a Napoli e a Città del Messico riguarda il diverso modo in cui ricorrono alla violenza. Queste differenti modalità d'azione sono un riflesso dei diversi contesti in cui è stato realizzato il lavoro di campo. Come abbiamo già evidenziato, mentre in Italia vari studi segnalano che a partire dalla metà degli anni Novanta, le organizzazioni mafiose hanno progressivamente ridotto l'uso esplicito della violenza (Massari, Martone, 2019), in Messico si osserva il fenomeno opposto, soprattutto a partire dal 2006, quando il governo ha lanciato la cosiddetta "Guerra al Narcotraffico".

Nel concreto, solo due tra i ragazzi intervistati a Napoli erano sospettati di omicidio, e nessuno di loro ha mai confermato di avere commesso questo reato. Invece a Città del Messico otto giovani avevano commesso omicidi, talvolta efferati o multipli, di cui parlavano con la naturalezza di chi è abituato a vivere in un contesto molto violento. Inoltre, i ragazzi intervistati a Città del Messico erano più abituati all'utilizzo di armi da fuoco e al consumo di droghe pesanti, mentre quelli intervistati a Napoli raramente erano armati e si lo erano di solito utilizzavano coltelli e coltellini.

Alcuni dei ragazzi intervistati a Città del Messico hanno confermato non solo di avere commesso omicidi, ma anche di appartenere a importanti gruppi criminali. È il caso di Kevin, 18 anni al momento dell'intervista e cresciuto nel centro della capitale messicana. Arrestato per omicidio, ha dichiarato di

lavorare per il cartello *Unión Tepito* da quando aveva 12 anni, informazione confermata dagli operatori del centro in cui il ragazzo è stato detenuto quando era ancora minorenne. Considerato il principale cartello della droga nato nella capitale messicana, l'adolescente aveva commesso diversi omicidi per conto di questa organizzazione, il primo quando aveva soltanto 13 anni. Kevin aveva cominciato la sua carriera criminale con la *Unión Tepito* riscuotendo il pagamento di estorsioni a commercianti per poi prendere parte a rapimenti e all'uccisione di chi si rifiutava di pagare il pizzo, ed era intenzionato a riprendere la sua attività criminale al termine della misura detentiva. Ha dichiarato inoltre che mentre era rinchiuso nel centro penale l'organizzazione criminale continuava a pagare il suo stipendio di 20 mila pesos settimanali (mille euro circa) alla sua famiglia.

«D: Perché sei qui?
R: Per omicidio colposo
D: Perché l'hai ammazzato?
R: Era il mio lavoro
D: E dov'è che lavori?
R: Per l'Unión…
D: Da quando lavori per l'Unión?
R: Da quando avevo 12 anni
D: Quanto ti hanno pagato per l'omicidio?
R: No, non è che ci pagano per ogni omicidio, ma io ricevo una quantità alla settimana
D: Quanto?
R: Circa 20 mila pesos (mille euro) a settimana
D: Ricevi questi soldi ancora?
R: Sì, li danno alla mia famiglia
D: Che altri lavori ti ha affidato l'Unión?
R: All'inizio, quando ho cominciato a 12 anni riscuotevo le estorsioni per conto di mio zio, che controllava tutta la zona del centro, adesso lo fa mio cugino perché hanno ammazzato mio zio…
D: A chi chiedevi il pagamento delle estorsioni?
R: A tutti i commercianti del centro, a tutti quelli che hanno locali…
D: Hai detto che hai cominciato con loro riscuotendo le estorsioni, e poi? Quali altre cose hai fatto?
R: Siccome non guadagnavo così tanto allora mi hanno detto che se ammazzavo qualcuno avrei guadagnato di più ed è così che a 13 anni ho ammazzato qualcuno per la prima volta
D: E chi è stato il primo che hai ammazzato?
R: Il responsabile di un ristorante
D: E perché l'hai ammazzato? Non ha voluto pagare ?

R: Proprio così, non ha voluto allinearsi
D: Quanti anni ti hanno dato?
R: Mi hanno dato due anni e 28 giorni
D: Che farai quando uscirai?
R: Ritornerò a lavorare
D: Sempre con l'Unión?
R: Sì
D: Sai che ormai sei maggiorenne e che se sarai arrestato un'altra volta ti daranno molti più anni?
R: Infatti, già non tornerò qui (al centro penale per minorenni)
D: Sei consapevole di questo?
R: Sì»

Kevin, come Roberto Miguel, è cresciuto in strada, dove ha imparato i codici necessari per sopravvivere ed ha interiorizzato quelle disposizioni e schemi mentali (habitus) che hanno orientato il suo agire. Il ragazzo possedeva un capitale sociale di strada significativo, costituito da una rete di relazioni criminali con quelli che lui chiamava i "buoni", ossia i capi dell'organizzazione criminale che gli garantivano un alto stipendio anche mentre se trovava nell'istituto penale, e con i quali intendeva tornare a lavorare al termine delle misure detentive. Gli operatori che seguivano Kevin erano pienamente consapevoli di questa situazione e hanno affermato che il ragazzo era diventato nel tempo un vero e proprio sicario, che difficilmente si sarebbe sottratto a quel tipo di vita, anche perché attraverso la carriera criminale si era abituato a guadagnare molti soldi, precludendosi di fatto la possibilità di intraprendere percorsi di vita e di lavoro all'interno della legalità.

Conclusioni

Attraverso i racconti in prima persona dei soggetti intervistati, questa ricerca ha confermato che la delinquenza giovanile, pur avendo cause complesse e multidimensionali, è anche il risultato di processi di emarginazione strutturale nel quale i dominati – per usare un termine di Bourdieu – fanno ricorso a mezzi illegali sia come strategia di sopravvivenza materiale, sia come strumento di riconoscimento e affermazione sociale.

Dal lavoro di campo è emersa fin da subito l'importanza del contesto sociale nella "scelta" delinquenziale dei giovani. Sia a Napoli che a Città del Messico, tutti gli intervistati provengono da quartieri emarginati, operai o malfamati, con alti tassi di disoccupazione giovanile, servizi inadeguati e una

forte presenza di realtà criminali, che spesso fanno parte dell'ambiente di vita di questi giovani dentro e fuori le mura domestiche.

Questi contesti, nei quali gli intervistati si sono formati fin dalla più tenera età, hanno dato forma a comportamenti e disposizioni inconsci (habitus), adatti a sopravvivere nel campo della strada, attraverso strategie quali l'interiorizzazione dell'esperienza di emarginazione e l'utilizzo delle forme di capitale a disposizione.

Una delle differenze più evidenti tra i ragazzi intervistati a Napoli e a Città del Messico riguarda il diverso uso della violenza che, come dimostrato nella ricerca di campo, è una diretta conseguenza delle diverse forme organizzative che assume il campo sociale della strada nei due contesti e che, se potrebbe dire, citando a Collins (2008), nella capitale messicana favorisce il superamento di quelle barriere emotive che inibiscono naturalmente i comportamenti violenti.

La violenza, intesa come capacità di sopraffazione fisica, ma anche economica e simbolica, è una delle dimensioni più rilevanti della cultura della strada e rappresenta sia uno strumento per la risoluzione di conflitti, sia una fonte di onore e prestigio. Per questi ragazzi incorporare la violenza nella propria esperienza quotidiana, per esempio attraverso gesti, atteggiamenti, ecc. costituisce una forma di capitale culturale.

Un'altra importante differenza riscontrata nei due contesti riguarda la grande diffusione dell'utilizzo di droghe fra i ragazzi di Città del Messico – quattro accusati di spaccio e 14 che hanno ammesso di consumare droghe pesanti –, mentre a Napoli soltanto due ragazzi erano accusati di spaccio e nessuno ha dichiarato di fare uso di sostanze stupefacenti.

Anche questo aspetto ha a che vedere con le caratteristiche dello specifico campo della strada in cui questi soggetti sono cresciuti. Città del Messico è caratterizzata da una maggior instabilità nella lotta tra gruppi criminali per il controllo del territorio e da una più estesa circolazione di sostanze sintetiche e inalabili a causa della presenza di nuove organizzazioni che cercano di ampliare il proprio mercato.

Tutti questi aspetti offrono ulteriori piste di ricerca che potrebbero essere indagate in futuro.

Riferimenti bibliografici

Andreas P., Wallman J. (2009). Illicit markets and violence: what is the relationship? *Crime, Law and Social Change*, 52: 225-229. DOI: 10.1007/s10611-009-9200-6.

Bergman M. (2016). *Drogas, narcotráfico y poder en América Latina*. Buenos Aires: Fondo de Cultura Económica.

Bourdieu P. (1986). The forms of capital. In: Richardson J., a cura di, *Handbook of theory and research for the sociology of education*. Westport (CT): Greenwood Press, 241-258.

Bourdieu P. (1990). *The logic of practice*. Cambridge: Polity Press.

Bourdieu P. (2001). *La distinzione. Critica sociale del gusto*. Bologna: Il Mulino.

Bourdieu P. (2003). *Cuestiones de sociología*. Madrid: Ediciones Istmo.

Bourdieu P. (2005). *Le regole dell'arte*. Milano: Il Saggiatore.

Bourdieu P. (2007). *El sentido práctico*. Ciudad de México: Siglo XXI Editores.

Bourdieu P. (2015a). *La miseria del mondo*. Milano-Udine: Mimesis.

Bourdieu P. (2015b). *Sistema, habitus, campo*. Milano-Udine: Mimesis.

Buscaglia E. (2013). *Vacíos de poder en México. Cómo combatir la delincuencia organizada*. Ciudad de México: Debate.

Collins R. (2008). *Violence. A micro-sociological theory*. Princeton (NJ): Princeton University Press.

Comune di Napoli (2024). *1° rapporto Osservatorio economia e società Napoli*. Napoli.

Consejo Nacional de Derechos Humanos (CNDH) (2022). *Ley nacional del sistema integral de justicia penal para adolescentes*. Ciudad de México.

Corrado S. (2013). The relationship between Italian mafias and Mexican drug cartels. Part 1: a comparison. Washington: Council of Hemispheric Affairs.

Direzione Investigativa Antimafia (2018). *Relazione del Ministero dell'Interno al Parlamento. Gennaio–giugno 2018*. Roma.

Dishman C. (2005). The leaderless nexus: when crime and terror converge. *Studies in Conflict & Terrorism*, 28(3): 237-252. DOI: 10.1080/10576100590928124.

Fleetwood J. (2019). Everyday self-defense: Hollaback narratives, habitus and resisting street harassment. *British Journal of Sociology*, 70(5): 1709-1729.

Instituto Nacional de Geografía y Estadística (INEGI) (2020). *Censo de población y vivienda*. Ciudad de México.

Instituto Nacional de Geografía y Estadística (INEGI) (2024). *Producto interno bruto por entidad federativa*. Ciudad de México.

Istituto Nazionale di Statistica (ISTAT) (2021). *Censimento della popolazione e dinamiche demografiche*. Roma.

Istituto Nazionale di Statistica (ISTAT) (2022). *Annuario statistico italiano*. Roma.

Martín Criado E. (2008). *El sentido práctico en Bourdieu. Algunos conceptos centrales de su teoría*. Sevilla: Universidad de Sevilla.

Massari M., Martone V. (2019). *Mafia violence. Political, symbolic, and economic forms of violence in Camorra clans*. New York–London: Routledge.

McCarthy B., Hagan J. (2001). When crime pays: capital, competence, and criminal success. *Social Forces*, 79(3): 1035-1060. DOI: 10.1353/sof.2001.0027.

Mendoza A.A. (2016). Crimen organizado en una ciudad de América Latina: la Ciudad de México. *URVIO. Revista Latinoamericana de Estudios de Seguridad*, 19: 129-145.

Ministero della Giustizia (2015). *La giustizia minorile in Italia*. Roma: Dipartimento per la giustizia minorile.

Ministero della Giustizia (2017). *Stati generali della lotta alle mafie. Tavolo 10: Minori e mafie*. Roma.

Nieto A. (2020). *El cártel chilango*. Ciudad de México: Grijalbo-Penguin Random House.

Paolucci G. (2011). *Introduzione a Bourdieu*. Bari: Laterza.

Pereyra G. (2012). México: violencia criminal y guerra contra el narcotráfico. *Revista Mexicana de Sociología*, 74(3): 420-460.

Popitz H. (1990). *Fenomenologia del potere*. Bologna: Il Mulino.

Reinserta (2021). *Niñas, niños y adolescentes reclutados por la delincuencia organizada*. Ciudad de México.

Rinaldi C. (2021). Presentazione. *Quaderni del Laboratorio Interdisciplinare di Ricerca sui Corpi, Diritti, Conflitti*, 1: 7-10. Varazze: PM Edizioni.

Sales I., Melorio S. (2021). Devianza minorile a Napoli: la parziale efficacia della messa alla prova. *Annali dell'Università degli Studi Suor Orsola Benincasa*, 14(1): 355-377.

Sandberg S., Pedersen W. (2011). *Street capital. Black cannabis dealers in a white welfare state*. Bristol: Policy Press.

Santoro M. (2015). Introduzione. In: Bourdieu P., *Forme di capitale*. Roma: Armando Editore.

Saviano R. (2006). *Gomorra. Viaggio nell'impero economico e nel sogno di dominio della camorra*. Milano: Mondadori.

Shammas V.L., Sandberg S. (2016). Habitus, capital and conflict: bringing Bourdieusian field theory to criminology. *Criminology & Criminal Justice*, 16(2): 195-213. DOI: 10.1177/1748895815603774.

Wacquant L. (2016). *I reietti della città. Ghetto, periferia, stato*. Pisa: Edizioni ETS.