TEORIA

T

Rivista di filosofia
fondata da Vittorio Sainati
XLVI/2026/1 (Terza serie XXI/1)

Legal and Ethical Design for Trustworthy AI Systems

Criteri giuridici ed etici per progettare sistemi affidabili di intelligenza artificiale

Edizioni ETS

TEORIA

Rivista di filosofia
fondata da Vittorio Sainati
XLVI/2026/1 (Terza serie XXI/1)

Iscritto al Reg. della stampa presso la Canc. del Trib. di Pisa n° 10/81 del 23.5.1981

Direzione e Redazione
Dipartimento di civiltà e forme del sapere dell'Università di Pisa, via P. Paoli 15, 56126 Pisa,
tel. (050) 2215400 - www.cfs.unipi.it

Direttore Responsabile
Adriano Fabris

Periodico semestrale

Abbonamento (cartaceo, privato): Italia € 55,00; Europa € 65,00; mondo € 75,00
Abbonamento (cartaceo, istituzionale): Italia € 65,00; Europa € 75,00; mondo € 85,00
PDF (online, stampabile): privato (accesso individuale) € 50,00
PDF (online, stampabile): istituzionale (fino a 20 range IP) € 60,00
Abbonamento cartaceo + PDF: privato € 85,00; istituzionale (fino a 20 range IP) € 95,00

Bonifico bancario intestato a
Edizioni ETS - Banca C.R. Firenze, Sede centrale, Corso Italia 2, Pisa
IBAN IT 21 U 03069 14010 100000001781
BIC/SWIFT BCITITMM
causale: abbonamento «Teoria»

«Teoria» è indicizzata ISI Arts&Humanities Citation Index e SCOPUS, e ha ottenuto la classificazione "A" ANVUR per i settori 11/C1-C2-C3-C4-C5. La versione elettronica di questo numero è disponibile sul sito: www.rivistateoria.eu

L'indice dei fascicoli di «Teoria» può essere consultato all'indirizzo: www.rivistateoria.eu
Qui è possibile acquistare un singolo articolo o l'intero numero in formato PDF, e anche l'intero numero in versione cartacea.

I numeri della rivista sono monografici. Gli scritti proposti per la pubblicazione sono double blind peer reviewed. I testi devono essere conformi alle norme editoriali indicate nel sito.

Expenditure fully financed under the PNRR project - PE00000013 named FAIR - Future Artificial Intelligence Research - M.4-C.2-I 1.3 - Notice 341/2022 CUO: I53C22001380006, funded by European Commission under the NextGeneration EU programme. The views and opinions expressed in this document are solely those of the authors and do not necessarily reflect the official position of the European Union nor the European Commission, NEither the European Union nor the European Commission can be held responsible for them.

Palazzo Roncioni - Lungarno Mediceo, 16, I-56127 Pisa - info@edizioniets.com - www.edizioniets.com
Distribuzione
Messaggerie Libri SPA - Sede legale: via G. Verdi 8 - 20090 Assago (MI)
Promozione
PDE PROMOZIONE SRL - via Zago 2/2 - 40128 Bologna
ISBN 978-884677440-8 ISSN 1122-1259

Contents / Indice

T

Premise / Premessa

This issue of «Teoria» brings together the results of research carried out on the topics of deontology, ethics, and the law of artificial intelligence by six Italian research groups – based at the Universities of Pisa, Trento, Naples "Federico II", Bari "Aldo Moro", Turin Politecnico, and Bologna – as part of the FAIR (Future Artificial Intelligence Research) project. The project was funded by the European Commission under the NextGeneration EU programme – National Recovery and Resilience Plan (PNRR) – Project: M4 C2 – Investment 1.3, Extended Partnership PE00000013, CUP: I53C22001380006.

The description of the project and the Foundation that supports it, as well as an outline of the specific topics specifically addressed by the six research groups, are provided in the first essay of this issue. The remaining essays, which were also presented for public debate during a conference held on October 30, 2025, explore some of the topics that were the subject of specific investigations.

We would like to thank the coordinators of the various Spokes and, above all, the president of the FAIR Foundation, Marco Conti, for all the work done to achieve the project's objectives. Special thanks go to the young researchers from the various research groups who, by working together and sharing their expertise, have made possible the attainment of the results presented here.

Questo fascicolo di «Teoria» raccoglie i risultati della ricerca compiuta sui temi della deontologia, dell'etica e del diritto dell'intelligenza artificiale da sei gruppi di ricerca italiani – appartenenti alle Università di Pisa, Trento, Napoli "Federico II", Bari "Aldo Moro", Torino Politecnico, Bolo-

gna – nell'ambito del progetto FAIR (Future Artificial Intelligence Research). Il progetto è stato finanziato dalla Commissione Europea nell'ambito del programma NextGeneration EU – Piano Nazionale di Ripresa e Resilienza (PNRR) – Progetto: M4 C2 – Investimento 1.3, Parternariato Esteso PE00000013, CUP: I53C22001380006.

La descrizione di questo progetto e della Fondazione che lo sostiene, nonché l'articolazione dei temi specificamente affrontati dai sei gruppi di ricerca, sono forniti nel primo saggio del fascicolo. I restanti saggi, presentati anche al dibattito pubblico nel corso di un convegno svoltosi il 30 ottobre 2025, approfondiscono alcuni temi oggetto di specifiche indagini.

Desideriamo ringraziare i coordinatori dei vari Spoke e, soprattutto, il presidente della Fondazione FAIR, Marco Conti, per tutto il lavoro svolto ai fini del raggiungimento degli obbiettivi del progetto. Un grazie speciale va ai giovani ricercatori dei vari gruppi di ricerca che, lavorando insieme e condividendo le proprie competenze, hanno contribuito all'acquisizione dei risultati che qui presentiamo.

Adriano Fabris
Università di Pisa
adriano.fabris@unipi.it

Carlo Casonato
Università di Trento
carlo.casonato@unitn.it

T

Adriano Fabris e Carlo Casonato

Quali criteri etici e giuridici per l'Intelligenza Artificiale? Il lavoro del TP 1 del progetto FAIR

1. *Presentazione generale del progetto FAIR*

La Fondazione FAIR – Future Artificial Intelligence Research è un ente senza fini di lucro istituito con l'obiettivo di attuare gli interventi previsti dal Piano Nazionale di Ripresa e Resilienza (PNRR) e da eventuali successivi programmi di finanziamento dedicati al settore dell'Intelligenza Artificiale. Essa rappresenta un'infrastruttura strategica nazionale per la promozione della ricerca scientifica e dell'innovazione tecnologica, il cui primo obiettivo è quello di coordinare il Partenariato Esteso FAIR, concepito per rafforzare la cooperazione tra mondo accademico, enti di ricerca e sistema produttivo.

L'organizzazione del partenariato si fonda sul modello Hub & Spoke, nel quale la Fondazione agisce come hub, ovvero soggetto attuatore e referente unico del progetto, responsabile della gestione, del coordinamento e del monitoraggio complessivo delle attività. Gli spoke, invece, rappresentano i soggetti esecutori, incaricati di sviluppare e implementare le linee di ricerca in collaborazione con lo hub, in un'ottica di integrazione e sinergia tra competenze differenti.

La Fondazione riunisce alcune tra le più autorevoli istituzioni di ricerca e università italiane, affiancate da importanti realtà industriali. Ne fanno parte quattro enti di ricerca pubblici – il Consiglio Nazionale delle Ricerche, la Fondazione Bruno Kessler, l'Istituto Nazionale di Fisica Nucleare e l'Istituto Italiano di Tecnologia – e tredici università italiane di eccellenza. A questi si aggiungono cinque aziende private – Bracco, Expert.ai, Intesa Sanpaolo, Leonardo e Lutech – e il Consorzio Interuniversitario Nazionale per l'Informatica.

 DOI: 10.4454/aphm7q94

In coerenza con la Strategia Nazionale Italiana per l'Intelligenza Artificiale, la Fondazione FAIR mira a promuovere una ricerca di frontiera capace di coniugare l'eccellenza scientifica con l'innovazione tecnologica e con un approccio etico e sostenibile. La sua azione intende superare la frammentazione della ricerca italiana nel campo dell'IA, favorendo la creazione di una rete unitaria e interdisciplinare che valorizzi le competenze diffuse sul territorio nazionale e ne incrementi la visibilità internazionale. Allo stesso tempo, la Fondazione si propone di sviluppare un'intelligenza artificiale antropocentrica, affidabile e sostenibile, promuovendo la formazione e l'attrazione di talenti e contribuendo alla costruzione di un ecosistema di ricerca e innovazione capace di garantire continuità e sostenibilità nel lungo periodo.

Il Progetto FAIR si fonda su una visione olistica e multidisciplinare dell'intelligenza artificiale, intesa non soltanto come tecnologia, ma come campo di conoscenza destinato a incidere profondamente sulla società. Il progetto affronta in modo integrato le dimensioni teoriche, modellistiche e ingegneristiche dell'IA, includendo anche la riflessione sulle implicazioni etiche, giuridiche e ambientali connesse al suo sviluppo. L'obiettivo è la creazione di sistemi intelligenti capaci di interagire e collaborare con gli esseri umani, di adattarsi a contesti in continuo mutamento, di operare in modo consapevole dei propri limiti e di rispettare i principi di sicurezza, trasparenza e fiducia.

La governance del progetto si articola in dieci spoke tematici, coordinati da otto università statali e due istituzioni di ricerca, ai quali partecipano ulteriori istituzioni affiliate, tra cui università, enti vigilati dal Ministero dell'Università e della Ricerca e soggetti privati. Attraverso questa struttura, la Fondazione FAIR si configura come un modello innovativo di collaborazione tra ricerca pubblica e industria, volto a consolidare il ruolo dell'Italia nel panorama scientifico internazionale e a promuovere uno sviluppo tecnologico etico, sostenibile e orientato al benessere collettivo.

Le aree tematiche affrontate dagli spoke di FAIR hanno implicato diversi livelli di possibile interazione, in quanto il progetto aveva come obbiettivo quello di risolvere problemi complessi che attraversano ambiti differenti. Per conseguire questo obiettivo, è stato necessario condividere approcci, materiali e informazioni tra i vari team di ricerca. La cooperazione tra le aree tematiche all'interno di FAIR è stata pertanto garantita dai Progetti Trasversali (TP), gruppi di lavoro dinamici che hanno riunito ricercatori impegnati su sfide comuni e che hanno operato trasversalmente ai singoli *spoke*, o gruppi di ricerca. I TP hanno favorito sinergie tra i diversi ambiti di ricerca e gettato le basi per ulteriori collaborazioni, coordinando attività e conoscenze condivise.

2. *IL TP1*

In questo contesto, il TP1 – coordinato a livello nazionale da Carlo Casonato, Adriano Fabris e Giovanni Sartor – si era proposto di sviluppare metodologie di progettazione e contenuti volti alla realizzazione di sistemi di intelligenza artificiale (IA) affidabili, lungo le diverse dimensioni della «trustworthiness» così come delineate dai documenti europei sull'intelligenza artificiale (IA) e dall'*AI Act*. L'obiettivo principale è stato quello di verificare la conformità dei sistemi AI ai valori etici e giuridici fondamentali – quali privacy, equità, spiegabilità e controllo umano – al fine di accrescerne l'affidabilità e promuovere una maggiore fiducia nell'adozione di tali tecnologie.

Le entità computazionali pervasive, autonome e intelligenti costituiscono ormai parte integrante dell'ecologia sociale contemporanea. Attualmente, la maggior parte dei dispositivi connessi è abilitato all'IA, fenomeno che sottolinea l'urgenza di sviluppare sistemi in grado di rispettare principi etici e giuridici coerenti con una visione *human-centered*. La conformità a tali principi, lungi dall'essere un vincolo accessorio, diviene quindi un elemento fondativo dell'IA stessa, contribuendo al rafforzamento dei valori etici e giuridici nella società.

Da questa prospettiva, il rapporto tra IA e principi etico-legali rappresenta una questione trasversale che coinvolge tutti gli *spoke* del progetto. Ciascun *work package*, seppur in misura differente, è stato chiamato a confrontarsi con problematiche giuridiche ed etiche, rendendo necessaria un'attenta coordinazione tra i diversi *work packages*. A tal fine, sono stati previsti negli anni di svolgimento del progetto numerosi workshop tematici di verifica, oltre a quello conclusivo che ora presentiamo. Essi sono stati finalizzati ad approfondire tematiche quali: l'IA affidabile e *human-centered* nei diversi contesti progettuali (by-design, in-design, per i progettisti); lo sviluppo di modelli etici per IA affidabile in scenari incerti e multi-agente; l'analisi delle implicazioni etiche degli algoritmi in settori e scenari paradigmatici, tra cui IA resiliente, adattiva, simbiotica ed *edge-exascale*; i fondamenti etici per l'autonomia regolabile; i metodi per diritto ed etica computabili, mediante tecnologie innovative per la governance giuridica ed etica dell'IA; e infine la realizzabilità dell'*AI Act*, con la verifica delle attività progettuali rispetto ai principi della proposta di regolamento europeo. Tale approccio ha voluto garantire che la progettazione e l'implementazione dei sistemi di IA fossero coerenti con un quadro etico e normativo solido, promuovendo al contempo un ecosistema tecnologico affidabile, responsabile e

centrato sull'essere umano, valorizzando la cooperazione trasversale tra gli *spoke* e l'efficacia del Progetto Trasversale.

Più precisamente, i diversi *spoke* che hanno cooperato nel TP1 hanno indagato le dimensioni giuridiche, etiche e tecniche dell'intelligenza artificiale (IA) in una vasta gamma di ambiti applicativi. Pur mantenendo ciascuno una prospettiva disciplinare e settoriale ben determinata – dalla sanità alla giustizia, dal mercato del lavoro al monitoraggio ambientale – sono emerse alcune linee di ricerca comuni, segno di una crescente convergenza attorno alla necessità di sviluppare sistemi di IA conformi alla normativa vigente, eticamente solidi e tecnicamente affidabili. In particolare, i partecipanti al TP1, nel periodo della ricerca, si sono concentrati su quattro principali direttrici condivise:

1. *L'AI Act: interpretazione, applicazione e criticità*

L'analisi e l'applicazione del *Regolamento Europeo sull'Intelligenza Artificiale* (*AI Act*), all'interno del complesso ecosistema normativo dedicato alla IA, ha rappresentato uno dei temi centrali delle attività di ricerca del TP1. I ricercatori hanno approfondito il testo da prospettive giuridiche, etiche e tecniche, esaminandone vantaggi, limiti e implicazioni pratiche. Lo Spoke 1 si è occupato di analizzare il *risk-based approach* adottato dal regolamento e, connesso a ciò, il ruolo della nozione di «vulnerabilità» in correlazione con esso. Particolare attenzione è stata dedicata a questioni cruciali quali i *bias* e la trasparenza (Spoke 2), le proibizioni previste dall'articolo 5 (Spoke 3) e i meccanismi di governance, come la *coregolamentazione*, gli standard armonizzati e i *sandbox* per l'IA (Spoke 8).

La ricerca ha inoltre considerato diversi contesti di applicazione, come la Pubblica Amministrazione (Spoke 3), il settore sanitario (Spoke 2), i sistemi giudiziari e di *law enforcement*, nonché l'interazione dell'AI Act con altri strumenti regolatori europei, in particolare il GDPR e il DMA (Spoke 8). L'obiettivo comune è stato quello di valutare come l'*AI Act* possa essere correttamente interpretato e efficacemente implementato, garantendo un equilibrio tra innovazione e tutela dei diritti fondamentali.

2. *Sistemi di IA etici e valutazione dei rischi*

Un altro filone trasversale della ricerca ha riguardato la costruzione di sistemi di IA eticamente accettabili e giuridicamente responsabili. Diversi *spoke* hanno affrontato il tema con approcci complementari: lo Spoke

6 ha introdotto modelli di *Assumption-Based Argumentation* e di rischio etico, finalizzati a formalizzare un ragionamento «value-sensitive» nei sistemi di IA. Lo Spoke 3 ha invece analizzato i rischi occupazionali legati all'uso dell'IA nella gestione del lavoro e nella valutazione delle prestazioni, mentre lo Spoke 1 ha proposto un quadro teorico ispirato al «pensiero lento» e all'umanesimo digitale, volto a riflettere sulle trasformazioni culturali e morali indotte dall'automazione e dall'uso crescente di algoritmi decisionali.

Queste ricerche convergono nella definizione di un approccio integrato che consideri congiuntamente il rischio legale e l'accettabilità etica delle soluzioni di IA.

3. *Trasparenza, spiegabilità e* accountability

La necessità di garantire trasparenza, spiegabilità e responsabilità nei sistemi di IA rappresenta un punto di convergenza fondamentale tra gli *spokes*. Sul piano tecnico, lo Spoke 7 ha sviluppato controllori basati su *Deep Reinforcement Learning* interpretabili, capaci di trasformare la logica «black box» in sistemi decisionali fondati su regole esplicite. Sul versante giuridico e sociale, lo Spoke 8 ha invece elaborato modelli di ragionamento giuridico basati su regole e schemi di argomentazione causale per affrontare situazioni di incertezza e conflitto normativo. Parallelamente, lo Spoke 2 ha evidenziato come la trasparenza costituisca un requisito non solo tecnico, ma anche etico e legale, in linea con i principi dell'*AI Act*, mentre lo Spoke 1 ha approfondito gli effetti dell'IA generativa sulla percezione e sulla responsabilità, segnalando i rischi legati alla manipolazione delle immagini e alla distorsione della realtà.

4. *L'impiego dell'IA in domini sensibili*

Un ulteriore tema comune ha riguardato l'uso dell'intelligenza artificiale in ambiti ad alta sensibilità giuridica ed etica. Lo Spoke 6 ha analizzato le applicazioni giudiziarie e di *law enforcement*, concentrandosi in particolare sui rischi derivanti dalle classificazioni preventive e dall'uso di logiche predittive nei settori fiscale e penale. Lo Spoke 3 ha studiato la gestione algoritmica del lavoro, i diritti dei lavoratori e i rischi professionali, nonché l'uso dell'IA in agricoltura, nelle campagne elettorali e nella governance regionale. Gli Spoke 1 e 2 si sono focalizzati sui sistemi di IA in ambito sanitario, valutandone gli aspetti regolatori e di policy, mentre gli Spoke

7 e 8 hanno indagato le applicazioni legate alla gestione ambientale, alla conformità legale e al settore assicurativo.

Nel complesso, queste linee di ricerca hanno delineato una visione comune: quella volta a promuovere un ecosistema europeo dell'IA che sia non solo innovativo, ma anche giusto, trasparente e rispettoso dei valori fondamentali dell'Unione.

3. *Le ricerche dei vari* spoke

Andando ancora più nello specifico, e considerando le attività di ogni singolo *spoke*, i gruppi di ricerca del TP1 hanno approfondito i seguenti temi:

Spoke 1

Il gruppo di ricerca dell'Università di Pisa, coordinato per quanto riguarda il profilo etico da Adriano Fabris, ha concentrato i propri studi sull'etica dell'intelligenza artificiale lungo tre linee tra loro interconnesse: i fondamenti metaetici dell'IA, la trasformazione della comunicazione nell'era digitale e le implicazioni etiche delle immagini artificiali. L'obiettivo complessivo è stato quello di promuovere una *human-centered AI*, capace di potenziare la capacità umana di decidere e agire responsabilmente, valorizzando sia l'individuo, sia la collettività. Ciò implica la necessità di ripensare le basi stesse dell'IA per orientarla verso una progettazione affidabile ed etica *by design*, fondata su co-progettazione, validazione e uso consapevole dei programmi e dei dispositivi.

Sul piano metaetico, la ricerca ha approfondito i concetti di «ambiente digitale», «rischio» e «vulnerabilità». L'ambiente digitale è stato interpretato come un sistema di relazioni tra esseri umani e tecnologie, che può assumere forme strumentali, mediali o immersive. Da questa prospettiva nasce il concetto di *ecologia dell'IA*, che mira a gestire le vulnerabilità generate dall'interazione con sistemi intelligenti, applicando il principio di vulnerabilità come fondamento normativo per costruire relazioni basate su cura, responsabilità e consapevolezza. In tale contesto, il gruppo ha proposto una filosofia del digitale che promuove un «pensiero lento» e riflessivo, capace di bilanciare la velocità e l'astrazione dei processi algoritmici, in vista di un nuovo umanesimo che unisca immaginazione tecnica, etica e responsabilità.

Un'attenzione particolare è stata poi rivolta alla comunicazione nell'era *post-verità*, segnata dall'erosione della verità condivisa e dalla mediazione

algoritmica dell'informazione. Il gruppo ha analizzato modelli etici inclusivi e umanocentrici per la comunicazione mediata dall'IA, al fine di preservare il dialogo democratico e la qualità del dibattito pubblico.

Infine, la ricerca ha affrontato l'etica delle immagini artificiali, analizzando come l'IA generativa trasformi immaginazione e percezione visiva, specialmente in ambiti come pubblicità, giornalismo e medicina. Sono stati esaminati i rischi legati a *bias*, *deepfake* e manipolazioni digitali, evidenziando la necessità di integrare principi etici e regolatori europei per prevenire la «defattualizzazione» della realtà e favorire un uso consapevole delle tecnologie visive emergenti.

Spoke 2

Il gruppo di ricerca dell'Università di Trento, coordinato per quanto riguarda il profilo giuridico da Carlo Casonato, ha approfondito in modo sistematico l'analisi del quadro normativo dell'AI Act, concentrandosi sia sulle sue potenzialità sia sulle criticità legate all'attuazione. Particolare attenzione è stata dedicata alle vulnerabilità, alla trasparenza e al principio del controllo umano, oltre che alla regolamentazione dei modelli generali di IA, con un focus sulle implicazioni etiche e giuridiche del loro impiego in ambito medico. In tale contesto, sono state esaminate anche linee guida e codici deontologici.

L'attività si è caratterizzata per un forte approccio interdisciplinare, grazie alla collaborazione tra giuristi, eticisti, medici e informatici, che ha consentito di affrontare in modo concreto temi quali i *bias* algoritmici, la trasparenza e l'evoluzione del quadro regolatorio europeo. Oltre alla partecipazione a diversi convegni nazionali, alcuni membri del gruppo hanno da ultimo partecipato a un seminario scientifico a Bilbao, dedicato proprio a questi aspetti, contribuendo al dibattito internazionale sull'uso responsabile dell'IA in ambito sanitario.

Parallelamente, l'Università di Trento ha svolto un ruolo di riferimento nell'organizzazione della serie biennale di webinar *AI Act 1.01* (marzo-aprile 2024 e giugno 2025), che ha favorito la diffusione e la discussione dei risultati della ricerca: infatti, le registrazioni dei webinar, rese disponibili nella piattaforma YouTube, hanno complessivamente raccolto circa 3000 visualizzazioni. Gli incontri hanno inoltre goduto di un'ampia partecipazione e del contributo di relatori provenienti da discipline e *spoke* diversi. Inoltre, la Fondazione Bruno Kessler (FBK) ha sviluppato una piattaforma di *AI-enactment* basata su infrastruttura MLOps, concepita per supportare

le Pubbliche Amministrazioni nello sviluppo sostenibile di applicazioni IA, integrando strumenti di *accountability*, metriche di qualità e algoritmi per la valutazione etica dei dati e dei modelli.

Spoke 3

Lo Spoke 3, dell'Università di Napoli «Federico II», coordinato da Giovanna De Minico, ha condotto un organico programma di ricerca sulla regolazione europea e nazionale, con particolare attenzione al diritto costituzionale e a quello del lavoro. Esso si è articolato in due aree principali distinte, ma strettamente coordinate tra loro.

A) Nel campo del diritto costituzionale, la De Minico ha approfondito l'evoluzione delle fonti giuridiche nell'*e-society* dall'eteronomia classica, passando per la *co-regulation*, quindi, gli standard armonizzati fino alla tecnica *as source of law*: su questo tema la coordinatrice ha già pubblicato alcuni suoi saggi in riviste italiane e internazionali. Inoltre, il gruppo ha analizzato l'applicazione iniziale del Regolamento (UE) 2024/1689 con particolare riferimento alle proibizioni dell'art. 5 e alla *Comunicazione della Commissione sulle pratiche di IA vietate* (C(2025) 884 final), tema questo al quale si è dedicata particolarmente la Tuozzo. Infine, è stato esaminato il ruolo della governance dell'IA all'interno delle istituzioni italiane ed europee, anche attraverso il contributo al Board sull'Intelligenza Artificiale, istituito dall'AGCOM, di cui la De Minico è componente.

A completamento dei temi generali prima indicati, si menzionano gli studi intorno alla regolamentazione della comunicazione politica e delle campagne elettorali, con attenzione alle decisioni dei tribunali europei e rumeni (Tuozzo); la sostenibilità dell'IA in agricoltura e nella governance dei dati, collegando i diritti individuali e collettivi con quelli sociali e con il mercato unico (De Tullio); l'impatto dell'IA sulla governance regionale e sui rapporti Stato-Regioni (Grossi); l'uso pubblico dell'IA e la necessità di modelli di *Responsible AI* (Palomba); la libertà di espressione e le sfide etiche poste dall'IA generativa nella comunicazione e nella creatività artistica (Parente).

Quanto agli studi in proiezione futura si indica quello sui *neuroright*, che sta sviluppando la De Minico, avvalendosi peraltro di un periodo di studio, svolto all'*University of Pennsylvania*, dove ha contribuito con *speech* e Relazioni ai lavori del *Center of Neuroscience and Society*, in https://neuroethics.upenn.edu/about-us/people/visiting-scholars/. Il secondo studio in atto riguarda invece la recente legge italiana sull'IA n.132/2025, condotto principalmente da Palomba e Parente.

B) Il secondo fronte è quello del diritto del lavoro, la cui ricerca, coordinata da Laura Tebano, ha analizzato l'impatto dei sistemi di gestione algoritmica sul lavoro e sui diritti collettivi, nel contesto dell'*AI Act* e della normativa europea. Sono stati esaminati i sistemi di IA ad alto rischio impiegati nei luoghi di lavoro, inclusi quelli di riconoscimento delle emozioni, con particolare attenzione alla trasparenza, alla tutela della privacy e alla sicurezza dei lavoratori. Ulteriori approfondimenti hanno riguardato le politiche attive per l'occupazione, la prevenzione dei rischi professionali nel lavoro da remoto e i diritti fondamentali degli atleti di *e-sports* in relazione alle normative antidoping.

Spoke 6

Il gruppo di ricerca dell'Università di Bari «Aldo Moro» coordinato da Francesca Alessandra Lisi ha sviluppato un articolato programma di ricerca sul tema dell'accettabilità della Symbiotic Artificial Intelligence (SAI), contribuendo in modo significativo agli obiettivi di TP1. Le attività si sono concentrate su tre dimensioni principali: legale, etico-filosofica e applicativa.

Sul piano giuridico, a partire dall'*AI Act*, i ricercatori del gruppo hanno analizzato i sistemi di IA ad alto rischio destinati all'uso giudiziario e di *law enforcement*, approfondendo in particolare i rischi connessi all'impiego di logiche algoritmiche e di decisioni predittive nei settori della fiscalità e del diritto penale. La riflessione ha messo in luce la debolezza concettuale di un approccio basato su una valutazione *ex ante* del rischio in un sistema, come quello giudiziario, che per sua natura opera *ex post*. Inoltre, sono stati esplorati scenari sensibili, come l'uso dell'IA nei casi di violenza di genere, con l'obiettivo di garantire maggiore tutela e trasparenza.

Dal punto di vista etico-filosofico, la ricerca ha affrontato questioni centrali quali vulnerabilità, inclusione, coscienza artificiale, libero arbitrio e *IA sociale*, proponendo un approccio multilivello alla comprensione del rapporto umano-macchina e alla decostruzione della centralità antropocentrica. Sono stati inoltre condotti studi sui *bias* algoritmici e sulla *fairness*, con particolare attenzione alla prospettiva di genere e ai processi di esclusione o discriminazione insiti nei sistemi automatizzati.

Sul piano applicativo, in collaborazione con l'Università del Salento (progetto eDef-AI), è stato sviluppato un modello di *Assumption-Based Argumentation* arricchito da pesi e calcolato tramite *Answer Set Programming*, utile per il ragionamento etico in ambito sanitario. È stato inoltre proposto un metodo formale per la specifica e la validazione di modelli decisionali etici basati

su *fuzzy Petri nets*, illustrato attraverso un caso di studio in ambito medico (care robot per il monitoraggio del paziente). Parallelamente, lo Spoke ha contribuito alla diffusione dei risultati in ambiti interdisciplinari e nel dibattito pubblico, promuovendo una cultura tecnologica più riflessiva e inclusiva.

Spoke 7

Lo Spoke 7, attivo al Politecnico di Torino, è stato coordinato da Barbara Caputo. Al suo interno, il WP7.6, (Ethical, Legal, Economical and Societal issues in Edge and Exascale AI), coordinato da Andrea Maria Lingua ha sviluppato un articolato programma di ricerca sull'applicazione dell'intelligenza artificiale a dati geospaziali, ambientali e 3D, con particolare attenzione a privacy, equità e interpretabilità. Le attività si sono articolate in più linee di ricerca interconnesse.

Un primo gruppo di ricercatori ha condotto numerosi audit di sistemi di IA (in campo assicurativo, in sistemi di classificazione volti, e in LLM) e studi empirici sulla documentazione dei dataset come mezzo per favorire trasparenza e accountability. Inoltre, sono state condotte tredici interviste con stakeholder per individuare criticità e indicazioni utili alla definizione di un quadro metodologico e di un protocollo condiviso per la generazione di dati sintetici, basato su modelli di ultima generazione adattabili a diversi contesti. È stato inoltre elaborato un sistema di codifica e avviata l'analisi delle trascrizioni delle interviste.

Un altro gruppo ha integrato modelli AI per la segmentazione semantica e la classificazione automatica di immagini a supporto del processo peritale assicurativo. Tali modelli si basano su ontologie di dominio finalizzate a guidare l'etichettatura automatica delle immagini, consentendo una rappresentazione coerente e interpretabile delle variabili decisionali chiave visibili nelle immagini. I risultati sperimentali sono stati discussi con esperti del settore, al fine di validare le raccomandazioni del sistema e di verificarne l'efficacia nel migliorare la trasparenza e l'uniformità del processo decisionale. Parallelamente, sono stati sviluppati modelli generativi e sistemi basati su *PointLLM* per la segmentazione semantica di ambienti urbani e *indoor*, il riconoscimento di specie marine (dataset SardinIA) e la classificazione automatica di immagini satellitari e iperspettrali per l'individuazione del degrado nei beni storici. Inoltre, sono stati sperimentati approcci *NeRF* e *Gaussian Splatting* per la ricostruzione 3D del patrimonio culturale subacqueo e approcci neuro-simbolici per migliorare la segmentazione semantica delle nuvole di punti architettoniche.

Infine, una terza linea di ricerca ha riguardato lo sviluppo di un sistema di controllo avanzato per impianti HVAC, basato su *Deep Reinforcement Learning*. I modelli addestrati sono stati trasformati in controllori interpretabili tramite alberi decisionali, capaci di garantire prestazioni elevate, efficienza energetica e trasparenza operativa. Le attività sono state accompagnate da workshop e corsi universitari sull'uso responsabile dell'IA nei dati geospaziali e ambientali.

Spoke 8

Lo Spoke 8 dell'Università di Bologna coordinato da Giovanni Sartor ha condotto infine un'ampia attività di ricerca dedicata agli aspetti regolatori, giuridici e applicativi dell'intelligenza artificiale, con un approccio interdisciplinare che unisce diritto, etica e tecnologia.

Sul piano regolatorio, il gruppo ha analizzato in profondità il ruolo dell'Unione Europea nel panorama normativo internazionale dell'IA, esaminando gli sviluppi multilaterali e bilaterali e il loro impatto sull'agenda digitale europea. Particolare attenzione è stata dedicata all'interazione tra *AI Act*, *Digital Markets Act* (DMA) e GDPR, con riferimento al consenso degli interessati e alle sue implicazioni giuridiche. È stato inoltre elaborato un framework concettuale per la valutazione qualitativa dei rischi legali dei sistemi di IA, pensato per garantire conformità normativa e tutela dei diritti fondamentali, soprattutto in relazione ai sistemi ad alto rischio e ai modelli di IA generali. La ricerca ha affrontato anche la «*twin transition*» – digitale ed ecologica – studiando l'impatto dell'IA sui modelli occupazionali e sulle trasformazioni del lavoro, tema che sarà approfondito in vista del World Congress 2027.

Nell'ambito della governance dell'IA, sono state esaminate strategie per l'implementazione di *AI regulatory sandboxes* a livello europeo, anche attraverso il progetto EUSAiR, coordinato da A. Rotolo, al fine di promuovere la cooperazione tra Stati membri e standardizzare i processi di sperimentazione regolatoria.

Sul versante più tecnico-giuridico, lo Spoke ha sviluppato modelli di ragionamento giuridico basati su regole, trattando il *legal reasoning* come un compito di classificazione binaria. Sono stati integrati la logica deontica e l'*argumentation theory* per gestire obblighi conflittuali e causalità legale, e proposti strumenti di *causal model checking* per analizzare la responsabilità giuridica. Infine, la ricerca ha esplorato l'uso dei Large Language Models (LLM) nel diritto, valutandone il potenziale nel rilevamento

di clausole vessatorie e nell'allineamento tra logica legale, linguaggio naturale e standard normativi, come il Codice della Strada.

English title: What ethical and legal criteria should apply to Artificial Intelligence? The work of WP 1 of the FAIR project

Abstract

The contribution outlines the work carried out and the results achieved within the Transversal Project (TP) of the PNRR FAIR Project. It begins by providing an overall presentation of the project, followed by an in-depth analysis of the specific features of the TP in all its components, namely the research groups from various Italian universities that belong to some of the Spokes. The TP's ethical and legal focus is addressed by these groups through different approaches, all aimed at creating a digital ecosystem that is ethical, legally compliant, and human-centered.

Keywords: AI; FAIR prject; trasversal project; ethical and legal values.

Adriano Fabris
Università di Pisa
adriano.fabris@unipi.it

Carlo Casonato
Università di Trento
carlo.casonato@unitn.it

T

Marco Billi e Antonino Rotolo

Valutazione della conformità ai sensi dell'articolo 57 dell'AI Act: sinergia con gli spazi di sperimentazione normativa*

Introduzione

Il Regolamento (UE) 2024/1689 sull'intelligenza artificiale (*AI Act*), pubblicato nella Gazzetta Ufficiale dell'Unione europea il 12 luglio 2024 ed entrato in vigore il 1° agosto 2024, segna l'avvio di una nuova fase della regolazione tecnologica europea. La maggior parte delle sue disposizioni diventerà applicabile dal 2 agosto 2026, introducendo un quadro uniforme di requisiti e procedure per lo sviluppo, l'immissione sul mercato e l'utilizzo dei sistemi di intelligenza artificiale nell'Unione.

La portata del Regolamento è ampia: esso si applica a fornitori, distributori, importatori, rappresentanti autorizzati, produttori (*providers*) di prodotti che integrano sistemi di IA e utilizzatori (*deployers*) che operano nel mercato dell'UE o i cui sistemi incidono su soggetti europei, anche se sviluppati al di fuori del territorio unionale (art. 2).

La valutazione di conformità (*conformity assessment*) è disciplinata in particolare dal Capitolo III, Sezione 2, che stabilisce i requisiti per i sistemi di IA ad alto rischio. Secondo l'articolo 3, paragrafo (20), essa è definita come «il processo volto a dimostrare se i requisiti stabiliti nella Sezione 2 del Capitolo III, relativi a un sistema di IA ad alto rischio, siano stati soddisfatti».

La conformità all'AI Act consiste in una valutazione multilivello che dipende da diversi fattori[1]:

* Il presente contributo è stato scritto prima della pubblicazione della bozza dell'atto di esecuzione per l'istituzione di sandbox per l'IA ai sensi dell'AI Act.

[1] Una disamina completa della metodologia da seguire per valutare la conformità di un si-

 DOI: 10.4454/02g1gb68

– La natura del sistema – determinare se il sistema rientri nella definizione di «sistema di intelligenza artificiale» di cui all'art. 3(1), ossia se operi con un certo grado di autonomia e produca output (predizioni, raccomandazioni, decisioni) che influenzano ambienti o persone.
– Il ruolo dell'attore economico – il Regolamento distingue obblighi specifici per fornitori, *deployers*, importatori e distributori. L'individuazione del ruolo è essenziale per definire le responsabilità giuridiche e operative (Capo III).
– L'ambito di applicazione – è necessario verificare se il sistema sia escluso dal campo di applicazione (ad esempio, sistemi destinati esclusivamente a fini militari, di ricerca o uso personale).
– La classificazione del rischio – gli obblighi derivanti dalla classificazione basata sul rischio (*risk-based approach*). I sistemi sono classificati in quattro categorie:
• pratiche vietate (art. 5).
• sistemi ad alto rischio (artt. 6-49 e Allegati I-III).
• sistemi a rischio limitato (art. 50).
• sistemi a rischio minimo o nullo, per i quali non è previsto alcun obbligo specifico.

Nel caso dei sistemi ad alto rischio, la conformità al Regolamento diventa oggetto di valutazione formale, e deve essere completata prima della messa sul mercato o della messa in servizio del sistema.

Questa valutazione mira a dimostrare che il sistema soddisfa i requisiti essenziali previsti dal Capo III, Sezione 2 (affidabilità dei dati, trasparenza, gestione del rischio, sorveglianza umana, robustezza, sicurezza informatica, accuratezza e tracciabilità).

Il processo di valutazione di conformità secondo l'AI Act

La valutazione di conformità rappresenta il processo attraverso il quale un fornitore dimostra che i requisiti per i sistemi di intelligenza artificiale ad alto rischio sono stati pienamente soddisfatti. L'AI Act prevede due principali procedure, la cui scelta o obbligatorietà dipende dalle caratteristiche specifiche del sistema e dall'applicazione di standard armonizzati[2].

stema di IA è fuori dall'obiettivo del presente contributo. La seguente divisione è suggerita da M. Quaranta *et al.*, *AI Compliance Frameworks: A Handbook For Compliance Tools Based on AI Acts' Requirements* (October 02, 2025). Al momento in formato preprint.

[2] Sul ruolo, requisiti, e procedure per verificare la conformità di un sistema di IA, si rimanda a E. Thelisson *et al.*, *Conformity assessment under the EU AI act general approach*, in «AI and

a) Autovalutazione interna (*internal control*)

Questa procedura non richiede l'intervento di un organismo notificato. Il fornitore è responsabile dello svolgimento dell'autovalutazione, che comporta una verifica approfondita del proprio Sistema di Gestione della Qualità (QMS), in conformità con l'articolo 17 dell'AI Act. Il QMS deve garantire la copertura di tutti gli aspetti previsti dal Regolamento, tra cui: conformità normativa, specifiche tecniche, controlli di qualità, testing, gestione dei dati, gestione dei rischi e monitoraggio continuo.

b) Valutazione di terza parte (*third-party assessment*)

Questa procedura, disciplinata dall'articolo 43(1)(b) dell'AI Act, prevede il coinvolgimento di un organismo notificato, al quale il fornitore deve presentare una domanda completa di valutazione del proprio QMS (ex art. 17) e della documentazione tecnica del sistema di IA. L'organismo notificato esegue audit periodici per verificare che il QMS sia mantenuto e applicato correttamente, può effettuare test aggiuntivi sul sistema e, in caso di esito positivo, rilascia un certificato di valutazione valido a livello dell'Unione.

Il processo richiede inoltre che gli operatori economici mantengano una documentazione tecnica completa e aggiornata (art. 11) e applichino un QMS robusto (art. 17), garantendo la tracciabilità di ogni decisione e aggiornamento del sistema. A ciò si aggiunge l'obbligo di registrazione dei sistemi ad alto rischio nella banca dati europea (art. 49) e di sorveglianza post-commercializzazione (art. 72)[3].

Per agevolare l'adempimento di tali obblighi e promuovere l'innovazione regolata, l'AI Act introduce lo strumento delle AI Regulatory Sandboxes (AIRS) (art. 57)[4]. Le sandbox rappresentano ambienti controllati di sperimentazione regolatoria, nei quali fornitori e autorità competenti collaborano alla verifica, sperimentazione e messa a punto di sistemi di IA innovativi in condizioni reali o quasi-reali.

Le sandbox rappresentano dunque un collegamento operativo tra la regolazione sperimentale e la conformità formale, e costituiscono un labo-

Ethics», 4.1 (2024), pp. 113-121; e J. Mökander *et al. Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation*, in «Minds and Machines» 32.2 (2022), pp. 241-268.

[3] La dottrina sta cominciando a proporre metodologie e strumenti per disegnare un QMS compliant, si veda ad esempio H. Mustroph *et al.*, *Design of a quality management system based on the eu ai act*, in «Legal Knowledge and Information Systems», IOS Press, 2024, pp. 314-320.

[4] Per una contestualizzazione delle AIRS si veda S. Ranchordas, *Experimental regulations for AI: sandboxes for morals and mores*, in «Morals & Machines», 1.1 (2021), pp. 86-100.

ratorio di regulatory learning in cui la conoscenza maturata nel contesto sperimentale può essere tradotta in procedure, standard e prassi conformi al diritto dell'Unione.

Infine, bisogna ricordare che il Regolamento UE sull'intelligenza artificiale (AI Act) introduce un requisito obbligatorio: ciascuno Stato membro deve istituire almeno un sandbox regolatorio per l'IA (operativo entro il 2 agosto 2026)[5]. Questi ambienti controllati consentono ai fornitori di sistemi IA di sviluppare e testare i prototipi sotto la supervisione delle autorità, con l'obiettivo di conciliare innovazione e conformità.

Obiettivi delle sandboxes secondo l'AI Act

Gli articoli 57 e seguenti dell'AI Act delineano gli scopi principali secondo il legislatore europeo, ed in particolare sono[6]:

- Certezza del diritto: aumentare la chiarezza normativa per agevolare la conformità delle imprese.
- Condivisione di *best practice*: favorire lo scambio di esperienze e prassi ottimali tra autorità e operatori.
- Innovazione e competitività: stimolare nuovi servizi IA e rafforzare l'ecosistema europeo dell'IA.
- Apprendimento normativo basato sui dati: raccogliere evidenze concrete dai test per affinare regolamenti e linee guida.
- Accesso accelerato al mercato UE: facilitare l'immissione sul mercato di sistemi IA innovativi (in particolare di PMI e start-up) riducendo gli ostacoli iniziali.

Questi obiettivi si allineano agli scopi di analoghe iniziative nelle fintech e in altri settori, dove i sandbox sono stati introdotti proprio per bilanciare flessibilità applicativa e tutela dei consumatori[7]. Per esempio, nei sandbox in materia di fintech le autorità vigilanti collaborano con banche e start-up

[5] Articolo 57, comma 1.

[6] Tra i tanti, per approfondire si vedano H.J. Allen, *Regulatory sandboxes*, in «Geo. Wash. L. Rev.», 87 (2019), p. 579; e C. Novelli *et al.*, *Getting Regulatory Sandboxes Right: Design and Governance Under the AI Act*, in «Available at SSRN», 5332161 (2025).

[7] Per uno stato delle arte sulle sandbox in materia di fintech, si veda J.J. Goo *et al.*, *The impact of the regulatory sandbox on the fintech industry, with a discussion on the relation between regulatory sandboxes and open innovation*, in «Journal of Open Innovation: Technology, Market, and Complexity», 6.2 (2020), p. 43.

per testare nuovi servizi digitali: i regolatori ottengono informazioni sulle criticità normative delle tecnologie emergenti, mentre gli operatori godono di un ambiente normativo più flessibile durante la sperimentazione.

Ruolo delle sandbox regolatorie nel processo di conformità

Per quanto riguarda nello specifico la valutazione di conformità, le AIRS possono avere un ruolo complementare e di supporto[8]. Come previsto dall'articolo 57, paragrafo 7 dell'AI Act, la partecipazione a una sandbox regolatoria non sostituisce la valutazione di conformità formale, ma la affianca in modo strutturato e complementare. Le autorità competenti forniscono ai fornitori e ai potenziali fornitori orientamento sulle aspettative regolatorie e sulle modalità di adempimento ai requisiti e agli obblighi stabiliti dal Regolamento. Al termine della sperimentazione, esse rilasciano – su richiesta – una prova scritta delle attività completate con successo e un rapporto di uscita che documenta nel dettaglio i test svolti, i risultati ottenuti e gli esiti di apprendimento (*learning outcomes*)[9].

Tali documenti, riconosciuti formalmente dall'articolo 57(7), possono essere utilizzati dai fornitori come evidenze di conformità all'interno del successivo processo di valutazione o nell'ambito delle attività di sorveglianza del mercato. Il medesimo articolo stabilisce che i *Written Proofs* e gli *Exit Reports* devono essere positivamente presi in considerazione dalle autorità di vigilanza del mercato e dagli organismi notificati (*Notified Bodies*), con l'obiettivo di semplificare e accelerare le procedure di valutazione della conformità dei sistemi di IA ad alto rischio[10].

La natura giuridica, secondo il dettame della normativa europea, di questi output è quindi di tipo preparatorio: essi hanno valore probatorio ma non vincolante. La funzione delle sandbox è quindi duplice, operano come

[8] Il legislatore non ha avuto ancora l'occasione di specificare quale sarà il ruolo preciso che le sandbox dovranno avere. In attesa dell'uscita degli standard armonizzati, la dottrina si sta interrogando sul ruolo che questi potranno giocare nel sistema delle sandbox, ad esempio A. Tartaro, E. Panai, *Leveraging technical standards within AI regulatory sandboxes: Challenges and opportunities*, in F. Bagni, F. Seferi (eds), *Regulatory sandboxes for AI and cybersecurity: Questions and answers for stakeholders*, in «National Cybersecurity Lab», 2025, pp. 116-129.

[9] Sempre con riferimento all'articolo 57, comma 7, «The competent authority shall also provide an exit report detailing the activities carried out in the sandbox and the related results and learning outcomes».

[10] Letteralmente «shall be taken positively into account by market surveillance authorities and notified bodies, with a view to accelerating conformity assessment procedures to a reasonable extent».

piattaforme di pre-valutazione, aiutando i fornitori ad allinearsi ai requisiti di conformità, e allo stesso tempo promuovono buone pratiche di sperimentazione regolatoria, favorendo la diffusione di un approccio *compliance-by-design*[11].

L'attività nella sandbox non sostituisce quindi la procedura formale di conformità, ma ne costituisce un passaggio preparatorio dotato di peso documentale significativo, pur privo di valore conclusivo.

Se la sandbox regolatoria fosse interpretata come un'alternativa pienamente sostitutiva alla valutazione di conformità, si correrebbe il rischio di minare l'autorità e le garanzie procedurali proprie del sistema di conformità, oltre a esporre le autorità competenti e gli stakeholder a potenziali responsabilità qualora, in fase di sorveglianza post-mercato, emergesse la non conformità di un sistema precedentemente sperimentato all'interno della sandbox. Pertanto, bisogna domandarsi fino a che punto le sandbox regolatorie devono essere considerate strumenti di supporto e accelerazione, ma non di sostituzione, della valutazione formale.

Funzionamento operativo e buone prassi

Come evidenzia la letteratura, ogni sandbox necessita di obiettivi e parametri di misurazione chiari sin dall'inizio[12]. È importante definire *scopi precisi*, criteri di ingresso/uscita e responsabilità dei partecipanti. Ad esempio, nel processo preparatorio le autorità devono rispondere a domande chiave: quale problema normativo intendono esplorare? Quali stakeholder includere? Che dati utilizzare e come garantire sicurezza e riservatezza? Poiché i sandbox possono coinvolgere più ambiti regolatori (ad es. privacy, sicurezza, leggi settoriali), è spesso necessario associare più autorità competenti[13]. In sintesi, una progettazione trasparente – e comunicata chiaramente agli innovatori – è essenziale affinché l'esperimento produca risultati validi e generalizzabili.

[11] Come già anticipato, questa sembra l'interpretazione più plausibile allo stato attuale della disciplina.

[12] Abbiamo già citato C. Novelli e S. Ranchordas come contributi che chiariscono questo punto.

[13] Il problema è affrontato da S. Ranchordas *et al.*, *Regulatory sandboxes and innovation-friendly regulation: between collaboration and capture*, in «Italian Journal of Public Law», 16 (2024), p. 107.

Il ruolo dei Notified Bodies

Occorre quindi prendere in considerazione il ruolo e le responsabilità degli organismi adibiti a svolgere la funzione di certificatori di conformità secondo l'AI Act, la *Notifying Authority* ed i *Notified Bodies*. Ai sensi dell'articolo 3, punto (21) dell'AI Act, il *conformity assessment body* è definito come «un organismo che svolge attività di valutazione della conformità di terza parte, incluse prove, certificazioni e ispezioni».

Questi organismi notificati (*Notified Bodies* - NB), accreditati secondo il Regolamento (UE) 2019/1020, rappresentano un attore fondamentale nel garantire l'indipendenza e la solidità delle verifiche sui sistemi di IA ad alto rischio. La loro funzione è quindi quella di assicurare che il processo di valutazione sia indipendente, trasparente e ripetibile, fungendo da intermediari tra innovazione tecnologica e tutela dell'interesse pubblico[14].

Il processo di certificazione dei sistemi di intelligenza artificiale ad alto rischio rappresenta l'atto conclusivo del percorso di valutazione di conformità previsto dall'AI Act. Ai sensi della normativa, la certificazione può essere rilasciata unicamente da organismi notificati (i NB) che siano stati accreditati da un'autorità notificante (*notifying authority*) nazionale, la quale ne verifica la competenza, l'indipendenza e la conformità ai requisiti previsti dall'AI Act[15].

È fondamentale chiarire, come precedentemente accennato, che le AI Regulatory Sandboxes non costituiscono in alcun modo un meccanismo di certificazione. Esse possono svolgere una funzione di preparazione alla conformità (*conformity assessment preparation*), fornendo evidenze, analisi e risultati di testing che saranno poi presi in considerazione da un organismo notificato o da un'autorità di vigilanza del mercato, ma non rilasciano né attestati di conformità né certificati di validazione tecnica.

Nel contesto delle sandbox regolatorie, la collaborazione tra autorità competenti e organismi notificati può quindi assumere un valore strategico. Le evidenze raccolte durante la sperimentazione – Written Proof ed Exit Report – possono costituire materiale preliminare di audit e contribuire a ridurre tempi e costi nelle successive fasi di certificazione, a condizione che tali risultati siano riconosciuti o avallati da un NB.

Sebbene l'articolo 31, paragrafo 5 dell'AI Act vieti ai NB di essere direttamente coinvolti nella progettazione, sviluppo, commercializzazione o utilizzo di sistemi di IA ad alto rischio – al fine di preservarne l'indipendenza di giu-

[14] L'AI Act, articoli 28 e seguenti, disciplina ruolo e competenze delle Notifying Authorities e Notifying Bodies

[15] Questa procedura si trova tra gli articoli 28 e 32.

dizio e l'integrità – la loro partecipazione indiretta, in qualità di osservatori tecnici o validatori metodologici, non contrasta con tale principio[16]. In questa prospettiva, i NB potrebbero assumere un ruolo di supervisione o validazione dei protocolli di test adottati nei sandbox, garantendo che le prove sperimentali rispettino criteri di qualità, tracciabilità e rappresentatività conformi agli standard europei armonizzati.

Qualora la partecipazione dei NB fosse formalmente ammessa e regolamentata, le attività svolte nei sandbox acquisirebbero un valore aggiunto all'interno del procedimento di conformità, potendo essere considerate pre-assessment evidence a tutti gli effetti. Anche in assenza di un coinvolgimento diretto, è comunque auspicabile che i test condotti nelle sandbox siano preventivamente avallati o riconosciuti dai NB, in modo da assicurare continuità metodologica tra la fase sperimentale e la successiva valutazione di conformità[17].

Tuttavia, qualora il processo sperimentale di un sandbox sia formalmente accettato da un organismo notificato o da un'autorità notificante, e i test siano condotti dal *main partner* del progetto (ad esempio un *Testing and Experimentation Facility* – TEF), i risultati di tali prove possono essere riconosciuti come input qualificato nel successivo processo di certificazione. In tal caso, il contributo della sandbox diventa una fase di pre-validazione che aumenta l'efficienza della certificazione, riducendo la duplicazione dei test e la ridondanza dei controlli.

Poiché gli organismi notificati operano sul mercato come soggetti economici, l'inclusione di questi nel ciclo di sperimentazione pre-market pone problemi di natura economica, soprattutto in ambito di concorrenza e sostenibilità. Gli NB hanno il diritto esclusivo di condurre le attività di valutazione di terza parte, ma se vengono coinvolti in fase anticipata nei sandbox – offrendo servizi non gratuiti o selettivi – vi è il rischio di distorsione della concorrenza[18]. Per evitare asimmetrie competitive tra NB, la partecipazione deve essere gestita attraverso criteri di trasparenza e rotazione, e possibilmente in coordinamento con le autorità notificanti nazionali e la Commissione europea.

[16] Almeno questo è quanto si può dedurre se si presuppone un'assenza di un conflitto di interessi, ad esempio in assenza in una remunerazione per l'attività svolta che proviene dai provider e dai partecipanti stessi nelle sandbox.

[17] Anche quando questi test non fossero quindi portati a termine dai NB stessi.

[18] Pensiamo, ad esempio, al caso in cui non vengano applicati criteri trasparenti di selezione o accesso: in tale situazione, gli organismi notificati coinvolti all'interno di una sandbox potrebbero ottenere un vantaggio competitivo ingiustificato rispetto a quelli esterni. Ciò contrasterebbe con i principi di indipendenza, imparzialità e parità di trattamento stabiliti dagli articoli 31 e 32 dell'AI Act, che richiedono ai NB di operare senza legami economici o interessi che possano compromettere la neutralità delle valutazioni di conformità.

Una possibile soluzione, ancora da chiarire sul piano giuridico, potrebbe arrivare dall'atto di esecuzione (*implementing act*) ex articolo 58 dell'AI Act, se quest'ultimo disciplinerà:

– le modalità con cui gli NB possono essere coinvolti nei sandbox senza violare il principio di neutralità del mercato.
– i limiti alle tariffe o ai costi che possono essere applicati in fase sperimentale.
– le regole di accesso non discriminatorio per tutti gli NB interessati, anche con riferimento ai settori verticali (sanità, mobilità, finanza, sicurezza).

Dal punto di vista economico, emerge inoltre un problema di sostenibilità: se un fornitore paga sia per i servizi sandbox (pre-market testing) sia per la certificazione formale da parte dell'NB, si genera un rischio di doppio pagamento, che potrebbe disincentivare la partecipazione alle sandbox e aumentare i costi complessivi di accesso al mercato.

Ciò potrebbe perturbare il mercato e innalzare i costi di transazione per i provider di IA, soprattutto per le PMI, riducendo la competitività del sistema europeo rispetto ad altri contesti regolatori meno onerosi. Questo contributo intende anche domandarsi se questa concezione di divisione delle responsabilità tra sandbox e NB possa venir superata dagli interessi economici e di sviluppo tecnologico che regolano il mondo dell'Intelligenza Artificiale.

Il progetto EUSAiR

Ora esaminiamo come questa sfida viene affrontata a livello operativo, e in particolare come il progetto EUSAiR[19] sta contribuendo a sostenere e coordinare gli sforzi dell'Unione Europea nello sviluppo delle sandbox regolatorie per l'intelligenza artificiale.

EUSAiR (European Sandboxes for AI Regulation) è un'iniziativa biennale finanziata dal programma Digital Europe che mira a creare un quadro armonizzato per la sperimentazione regolatoria in ambito IA. Il progetto supporta gli Stati membri nell'attuazione delle sandbox, fornendo modelli e strumenti comuni per la documentazione, la valutazione e il monitoraggio dei progetti, al fine di favorire l'innovazione, la certezza e la conformità all'AI Act.

[19] Si rimanda al sito del progetto https://eusair-project.eu/.

Aspetti operativi e tecnici

Dal punto di vista operativo, tre elementi risultano determinanti per massimizzare l'utilità dei risultati prodotti dalle sandbox regolatorie[20]. In primo luogo, il livello di maturità tecnologica del sistema, o Technology Readiness Level (TRL), incide direttamente sulla solidità delle evidenze raccolte e sulla preparazione alla conformità: sistemi più maturi offrono esiti più affidabili e immediatamente utilizzabili nel percorso regolatorio, mentre quelli in fase iniziale traggono maggiore beneficio dal confronto consulenziale e formativo. In secondo luogo, un approccio modulare alla conformità consente di impiegare la sandbox sia come strumento pre-certificativo dedicato a specifici requisiti, sia come spazio di valutazione preliminare più ampia e olistica, utile alla pianificazione delle verifiche successive. Infine, l'adozione di standard armonizzati e di formati interoperabili, come previsto dagli articoli 17 e 40 dell'AI Act, garantisce coerenza documentale e facilita la collaborazione tra autorità nazionali, organismi notificati e operatori economici.

Un ulteriore aspetto chiave riguarda i test in condizioni reali, riconosciuti dall'articolo 60 dell'AI Act come strumento essenziale per verificare l'efficacia dei controlli tecnici e procedurali in contesti rappresentativi dell'uso effettivo dei sistemi di IA. Nel quadro delle Testing and Experimentation Facilities e delle AI Regulatory Sandboxes, queste attività consentono di valutare il grado di conformità ai requisiti normativi e di misurare la maturità regolatoria dei partecipanti attraverso modelli di avanzamento progressivo. Il progetto EUSAiR ha predisposto un framework di sandbox multilivello, un primo basato sulle attività base di supporto documentale ai provider, ed un secondo, denominato AIRS Add-ons, che offre inoltre servizi specializzati per il supporto alla conformità normativa, alla sicurezza informatica e alla protezione dei dati, integrando la collaborazione con le Autorità per la Protezione dei Dati e con gli organismi di valutazione della conformità. Per garantire la qualità metodologica e la credibilità dei risultati, è necessario il coinvolgimento di esperti altamente qualificati, tra cui specialisti di protezione dei dati, coordinatori di test in ambienti reali e professionisti della conformità dotati di competenze specifiche sugli standard armonizzati e sulle metodologie di valutazione.

All'interno di questo quadro operativo, due strumenti documentali assumono un ruolo centrale: la Written Proof e l'Exit Report[21]. La Written

[20] Per i risultati del progetto si rimanda ai deliverable, che presto saranno pubblici.

[21] I seguenti modelli di Exit Report e Written Proof preparati da EUSAiR vanno a com-

Proof, rilasciata dall'autorità competente al termine della partecipazione, costituisce la prova ufficiale delle attività svolte e dei risultati conseguiti. Essa descrive in modo sintetico le attività completate, i requisiti normativi affrontati e lo stato finale del progetto rispetto agli obiettivi fissati nel piano della sandbox. Il progetto EUSAiR ha sviluppato un modello standardizzato di Written Proof per garantire coerenza tra Stati membri e facilitare l'utilizzo di questo documento all'interno dei processi di conformità e sorveglianza di mercato.

L'Exit Report, invece, rappresenta la documentazione più estesa e analitica dell'intera esperienza in sandbox. Esso riassume i servizi di consulenza e supporto forniti, le sfide regolatorie incontrate, la valutazione dei risultati ottenuti rispetto agli indicatori di performance, nonché eventuali incidenti, reclami o segnalazioni registrate durante la sperimentazione. Include infine una riflessione sugli insegnamenti tratti (*lesson learned)* e sui contributi forniti all'evoluzione delle pratiche regolatorie europee. Secondo quanto previsto dall'articolo 57, paragrafo 7, dell'AI Act, le autorità nazionali dovrebbero fornire la versione destinata ai partecipanti entro un mese dalla conclusione della sandbox e quella destinata alla diffusione pubblica entro due mesi. Il modello operativo elaborato da EUSAiR per la redazione di questi report contribuisce a rendere omogenee le procedure documentali e a consolidare la base probatoria utile per la successiva valutazione di conformità.

Written Proof

La Written Proof ha una duplice natura: da un lato, costituisce un documento tecnico-normativo che registra in modo sintetico le attività e i risultati conseguiti durante la sandbox; dall'altro, è uno strumento di trasferimento di conoscenze e di raccordo tra la sperimentazione e la valutazione formale di conformità.

Il modello elaborato da EUSAiR si articola in cinque sezioni principali. La Parte A, dedicata alle informazioni generali, identifica il partecipante, la durata e il riferimento amministrativo della sperimentazione, garantendo la tracciabilità amministrativa del progetto. La Parte B descrive in modo analitico le attività tecniche condotte – sviluppo, addestramento, validazione e test –

plementare quanto descritto nella bozza dell'atto di implementazione (https://digital-strategy.ec.europa.eu/en/consultations/commission-seeks-feedback-draft-implementing-act-establish-ai-regulatory-sandboxes-under-ai-act).

fornendo la base empirica per la valutazione della robustezza del sistema di IA e per la successiva compilazione del fascicolo tecnico. La Parte C è dedicata alle questioni regolatorie affrontate e ai requisiti dell'AI Act oggetto di analisi o verifica, documentando le aree di incertezza interpretativa chiarite durante la sperimentazione e contribuendo all'apprendimento istituzionale.

La Parte D riassume lo stato di avanzamento del progetto, con una valutazione sintetica dell'idoneità del sistema a proseguire verso la fase di conformità formale. Questa sezione rappresenta un punto di raccordo tra la dimensione sperimentale e quella regolatoria, offrendo alle autorità elementi oggettivi per stimare la maturità regolatoria del partecipante. Infine, la Parte E raccoglie gli elementi formali di approvazione e sottoscrizione, assicurando la validità e l'autenticità del documento ai fini del riconoscimento ufficiale.

La Written Proof, così strutturata, assolve una funzione di ponte tra sperimentazione e conformità. Per i fornitori, costituisce una prova della solidità del proprio sistema di gestione della qualità e del rischio; per le autorità, rappresenta un documento interoperabile che può essere utilizzato come evidenza tecnica nel processo di conformity assessment, in linea con l'articolo 57, paragrafo 7, dell'AI Act. La sua efficacia dipende dal mantenimento di un equilibrio costante tra valore probatorio e valore conoscitivo: se intesa unicamente come certificazione preliminare, ridurrebbe la sandbox a un audit; se considerata solo come esercizio di apprendimento, perderebbe la propria rilevanza regolatoria. Trovare la giusta complementarità tra queste due dimensioni sarà oggetto di test nel corso del progetto.

Exit Report

L'Exit Report rappresenta il secondo asse documentale della fase post-sperimentazione nelle sandbox e ha una funzione più estesa rispetto alla Written Proof. È concepito per raccogliere, interpretare e sistematizzare tutte le evidenze emerse durante la partecipazione alla sandbox, offrendo una visione completa del percorso sperimentale e dei risultati ottenuti.

Anche in questo caso, il modello sviluppato da EUSAiR adotta una struttura articolata. La Parte A include le informazioni generali, garantendo la tracciabilità amministrativa e temporale. La Parte B descrive i servizi e il supporto forniti all'interno della sandbox – inclusi consulenza regolatoria, accesso a infrastrutture europee e assistenza tecnica – documentando il grado di interazione tra autorità e partecipante. La Parte C riassume le attività

svolte e i risultati raggiunti, con particolare attenzione ai rischi gestiti nelle fasi di sviluppo, addestramento e test, e quindi fornisce elementi di prova sulla sicurezza e affidabilità del sistema.

La Parte D analizza le sfide regolatorie incontrate e le soluzioni adottate, delineando un quadro delle barriere normative affrontate e delle strategie di adattamento sperimentate. Questa sezione contribuisce direttamente alla chiarificazione dei requisiti regolatori, offrendo un patrimonio di conoscenze utile sia alle autorità sia ai legislatori. La Parte E valuta le performance del progetto sulla base di indicatori chiave, misurando il rispetto dei tempi, la qualità dei risultati e la coerenza con gli obiettivi progettuali: tali elementi costituiscono un riferimento quantitativo per la verifica di conformità.

La Parte F raccoglie eventuali incidenti, reclami o malfunzionamenti, insieme alle azioni correttive intraprese. La loro documentazione introduce una dimensione di *learning by error* che anticipa, in ambiente controllato, il meccanismo di monitoraggio post-market. La Parte G sintetizza le *lesson learned* e gli insegnamenti regolatori, costituendo il punto di contatto tra la sperimentazione e la normazione: qui l'esperienza del singolo progetto si trasforma in conoscenza sistemica utile al miglioramento del quadro normativo europeo. La Parte H, infine, contiene gli elementi formali di approvazione e sottoscrizione, che assicurano la validità istituzionale del documento e ne consentono l'utilizzo ufficiale ai fini della conformità.

Grazie a questa articolazione, l'Exit Report svolge una funzione di raccordo strutturale tra apprendimento e conformità. Da un lato, fornisce alle autorità un archivio tecnico e organizzativo che semplifica le verifiche successive, riducendo i tempi di valutazione e migliorando la qualità del processo di *conformity assessment*. Dall'altro, alimenta il *regulatory learning*, consentendo di individuare aree di ambiguità normativa e di orientare la futura evoluzione degli standard armonizzati e delle prassi di supervisione.

Complementarità con audit e monitoraggio post-commercializzazione

Nel quadro complessivo di vigilanza, il percorso di sandbox si collega sempre alla logica *ex-ante* di valutazione di conformità. Come già affermato, l'AI Act conferma che i sistemi IA ad alto rischio potranno essere immessi sul mercato solo se sottoposti ed aver passato con successo una valutazione di conformità (interno o con organismo notificato). Tuttavia, la fase successiva all'entrata in commercio prevede un monitoraggio post-mercato che integra e completa la conformità iniziale.

Dal punto di vista dei controlli, la letteratura distingue tre approcci di auditing IA[22]: *auditing di funzionalità* (verifica degli scopi e delle motivazioni d'uso del sistema), *auditing del codice* (analisi del software sottostante) e *auditing d'impatto* (indagine sugli effetti reali del sistema). Le procedure di conformità previste dall'AI Act combinano elementi di auditing di funzionalità e di codice (es. analisi della documentazione, test di robustness, verifiche di governance). Il monitoraggio post-commercializzazione, per sua natura, aggiunge invece la dimensione dell'audit d'impatto (valutazione di outcome effettivi, bias, rischi emergenti). Questo è particolarmente importante per sistemi IA che continuano ad apprendere o a evolversi dopo il rilascio: il post-market monitoring rileva eventuali derive non previste e alimenta interventi correttivi basati sui dati reali di utilizzo. In altre parole, la fase dopo il sandbox e l'entrata in servizio non diminuisce le responsabilità del fornitore, ma garantisce che l'aderenza alle norme rimanga costante lungo tutto il ciclo di vita del sistema.

I sistemi di intelligenza artificiale ad alto rischio devono inoltre rispettare rigorosi requisiti di cybersecurity, come previsto dall'AI Act e in linea con le indicazioni di ENISA e degli standard europei emergenti. La sicurezza informatica è un elemento critico del conformity assessment, influenzando direttamente la robustezza, l'affidabilità e la tracciabilità dei sistemi IA.

All'interno dei sandbox regolatori, la verifica dei requisiti di cybersecurity può essere condotta in modalità simulata o controllata, consentendo ai fornitori di identificare vulnerabilità, testare misure di mitigazione e integrare protocolli di protezione già nella fase di sviluppo. Tuttavia, tali test preliminari non sostituiscono le attività formali di audit dei NB, che rimangono necessarie per il rilascio della certificazione.

Un aspetto innovativo è l'opportunità di creare un regulatory learning per gli NB in materia di cybersecurity. Partecipando a sandbox che simulano attacchi informatici, gestione di incidenti e resilienza dei sistemi, gli organismi notificati possono acquisire familiarità con le minacce emergenti, aggiornare le checklist di valutazione e affinare le linee guida per la certificazione. Questo processo può ridurre i tempi di audit e aumentare la qualità complessiva delle valutazioni, contribuendo a un ecosistema più sicuro e affidabile.

[22] In questo caso si fa riferimento soprattutto agli approcci previsti da J. Mökander, *Auditing of AI: legal, ethical and technical approaches*, in «Digital Society» 2.3 (2023), p. 49 e D. Hartmann *et al.*, *Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society*, in «AI and Ethics», 5.4 (2025), pp. 3617-3638.

Dal punto di vista operativo, è essenziale definire una roadmap integrata tra sandbox, TEF e organismi notificati per, in primis, garantire che i test di cybersecurity eseguiti in ambiente controllato siano statisticamente rappresentativi dei test richiesti per la conformità; ed in secundis, allineare i criteri di vulnerabilità e mitigazione agli standard europei armonizzati.

In pratica, un sistema IA che ha superato con successo i test di cybersecurity all'interno del sandbox, e che mantiene condizioni di *compliance* con la documentazione tecnica richiesta, sarà altamente predisposto a superare la successiva valutazione di conformità formale. Questo approccio riduce i rischi di errori procedurali, minimizza i costi di ripetizione dei test e contribuisce alla sostenibilità del processo di certificazione, senza generare conflitti con i requisiti di mercato o duplicazioni economiche.

In sintesi, l'integrazione dei test di cybersecurity nei sandbox regolatori costituisce un collegamento operativo e cognitivo tra sperimentazione e certificazione formale, rafforzando sia la robustezza tecnica dei sistemi IA sia l'efficacia complessiva del compliance ecosystem europeo.

Una delle criticità più rilevanti è la frammentazione istituzionale. L'idea di un «*one-stop shop*» per la valutazione della conformità dell'IA appare difficilmente realizzabile nel breve termine: i partner coinvolti – NB, ENISA, TEF, autorità nazionali, AI Office – operano con mandati distinti e competenze differenziate.

Per mitigare questi rischi e promuovere una cooperazione, è auspicabile che le sandbox stabiliscano meccanismi di coordinamento stabile con:

- gli organismi notificati (NB), per l'allineamento metodologico e la mutua riconoscibilità dei risultati di test.
- ENISA, per la definizione dei requisiti di cybersicurezza dei sistemi di IA (ancora in fase di consolidamento).
- i TEF settoriali, per la conduzione di prove tecniche in condizioni reali e l'utilizzo di dataset di riferimento.
- le autorità nazionali competenti (NCA) e l'AI Office, per assicurare legittimità istituzionale.

Tale cooperazione non si esaurisce nella sola raccolta di evidenze tecniche, ma può dar vita a un vero e proprio regulatory learning condiviso. Anche gli NB possono trarre beneficio da questa interazione, acquisendo familiarità con le nuove categorie di sistemi di IA, le metriche di rischio, le architetture di rete neurale o i modelli general-purpose (GPAI). In tal senso, la sandbox non è solo un *testing ground* per le imprese, ma anche un

laboratorio di apprendimento regolatorio per le autorità e gli organismi di valutazione della conformità.

La cooperazione deve quindi basarsi su protocolli di interoperabilità e formati standardizzati di reporting (ad esempio per *Written Proof* ed *Exit Report*), che consentano di ridurre la distanza tra test in ambiente controllato e valutazione formale.

Conclusioni

I sandbox regolatori per l'IA incarnano un approccio agile di compliance by design: integrando supporto normativo e verifica tecnica in un contesto controllato, aiutano i fornitori ad allineare precocemente i propri sistemi ai requisiti UE. La disciplina dell'AI Act valorizza questa esperienza: le prove documentali dei sandbox (report e attestati) sono ufficialmente riconosciute come evidenza di conformità, con vantaggi concreti nel successivo processo di certificazione. Nel complesso, il modello sandbox consolida la sinergia tra innovazione e regolazione: da un lato facilita l'accesso precoce al mercato europeo, e dall'altro offre alle autorità insight pratici per ottimizzare le norme. In tal modo, si costruisce gradualmente un ecosistema di audit e vigilanza IA più informato e partecipato, capace di gestire in modo sostenibile i rischi delle tecnologie emergenti.

Con l'entrata in piena applicazione dell'AI Act nel 2026, la valutazione di conformità diventerà un asse portante della governance tecnologica europea. L'integrazione tra sandbox, organismi notificati e autorità competenti segnerà il passaggio da un modello puramente reattivo a uno proattivo e cooperativo.

In tale contesto, le sandbox regolatorie non solo accelerano la certificazione dei sistemi ad alto rischio, ma contribuiscono a un ecosistema di fiducia multilivello, in cui la conformità è il risultato di un processo di apprendimento reciproco tra industria, ricerca e regolazione.

English title: Conformity assessment under article 57 AI Act: Synergy with Regulatory Sandboxes

Abstract

This article examines the evolving architecture of conformity assessment under the EU Artificial Intelligence Act (AI Act), focusing on the interaction between regulatory sandboxes, notified bodies, and the formal certification process. It explores how Written Proofs and Exit Reports, developed within sandbox environments, serve as structured documentation bridging experimental testing and market conformity procedures. The analysis highlights the dual nature of these instruments: as technical evidence supporting compliance and tools for regulatory learning that inform institutional guidance and standards development. Particular attention is given to the role of notified bodies, their potential engagement within sandbox testing under Article 31(5), and the risks of overlap or market distortion. The study argues for a balanced integration of sandbox outcomes into conformity workflows, proposing a cooperative model where regulatory experimentation closely interacts with conformity procedures.

Keywords: regulatory sandboxes; conformity assessment; notified bodies; exit report; cybersecurity.

Marco Billi
University of Bologna
marco.billi3@unibo.it

Antonino Rotolo
University of Bologna
antonio.rotolo@unibo.it

La presente ricerca è stata svolta nell'ambito del progetto EUSAIR. Le opinioni espresse nel presente contributo sono esclusivamente degli autori e non riflettono necessariamente la posizione dell'Unione Europea, che non può essere ritenuta responsabile dei contenuti qui riportati. Grant Agreement n. 101195535.

T

Silvia Dadà

From Risk Society to Digital Risk Society: Systemic Risk and the Challenges of the EU AI Act

1. *History of the concept of risk*

The notion of risk has undergone numerous transformations over the centuries, while preserving its fundamental function: an attempt to control and manage the disorder inherent in reality. This dimension was particularly emphasized by Mary Douglas, an anthropologist who explored the political and cultural significance of the concept of risk. Closely tied to the themes of contamination and purity, risk belongs to that set of symbols through which societies assign responsibility, identify blame, and create social cohesion:

> Whose fault? is the first question. Then, what action? Which means, what damages? what compensation? what restitution? and the preventive action is to improve the coding of risk in the domain which has turned out to be inadequately covered. Under the banner of risk reduction, a new blaming system has replaced the former combination of moralistic condemning the victim and opportunistic condemning the victim's incompetence (Douglas, 1992: 15-16).

Humanity has always confronted the future and its uncertainty, though not all cultures have adopted the same strategies for doing so. In antiquity, for instance, concerns over unpredictability were entrusted to the relationship with the divine and to the interpretation of divine will, without recourse to the notion of risk or to the techniques of calculation, management, and assessment associated with it. As Peter Bernstein (1996) observes, the turning point lies in the shift from reliance on the gods to reliance on rational and strategic calculation, which becomes the defining feature of the modern notion of risk. There is, therefore, a close connection between the concept of risk and the process of secularization. As Ulrich Beck puts it: «When

 DOI: 10.4454/3cz9p720

Nietzsche announces: God is dead, then that has the – ironic – consequence that from now on human beings must find (or invent) their own explanations and justifications for the disasters which threaten them» (Beck, 2008).

Although the origins of the term remain uncertain (possibly Arabic), the Neo-Latin form *risicum* is already attested in the medieval period (Luhman, 1996). Its earliest uses are primarily found in connection with maritime trade and ship insurance, where it referred to adverse events – such as floods, epidemics, or earthquakes – that could jeopardize the success of a voyage. These phenomena were not dependent on human action, nor were they considered calculable. This excluded any connection between this early idea of risk and human responsibility. The scope for action and prediction was thus extremely limited. In this initial phase, risk essentially denoted the danger of suffering misfortune, over which human intervention could exert little influence.

With the advent of modernity in the eighteenth century, however, the meaning of risk expanded to encompass elements attributable to human agency. Within this context, risk acquired a calculable and objective character. The distinction between risk and danger became more sharply defined: «[…] in the case of risk/danger in the fact that only in the case of risk does decision making (that is to say contingency) play a role. One is exposed to dangers. Of course, the behaviour of those concerned» (Luhmann, 1993: 23).

A series of factors contributed to the development of this new understanding. The Enlightenment fostered a general faith in human progress and in the rational, objective comprehension of the world. Moreover, advances in the fields of probability and statistics made it possible to refine techniques of prediction on a mathematical basis[1].

The modern sense of "risk" is therefore characterized by its distinction, on the one hand, from "danger" (since it depends on human decision), and on the other, from "uncertainty" (because of its rational and measurable nature). It is also defined by its ambivalence, encompassing both positive and negative connotations. Especially in financial and insurance contexts, the assumption of risks may indeed entail loss, but it may equally represent opportunity and gain.

[1] On the relationship between risk and statistics see Bernstein (1996).

2. *Global risk in the risk society*

Things change significantly in the transition from modernity to postmodernity, or late modernity, the stage that marks the entry into what Ulrich Beck aptly called the "risk society" (1992). With the growing mistrust of scientific certainty and objectivity, the fragmentation of grand narratives and traditions (accompanied by suspicion toward authority and institutions), the concept of risk gradually shifts from the dimension of control to that of uncertainty. As Anthony Giddens argues:

> If risk has always been conceived as a way of dealing with the future, of managing it and bringing it under our control, this is no longer the case today: our attempts to control the future tend to rebound against us, forcing us to consider alternative ways of engaging with uncertainty (Giddens, 2000: 40, our translation).

A pervasive sense of insecurity troubles society, primarily because the very status of the dangers we face is changing. Deborah Lupton, (1999) in her work dedicated to this topic, identifies six types of risks that characterize our time: (1) environmental risks; (2) lifestyle risks; (3) health risks; (4) risks in interpersonal relationships; (5) economic risks; and (6) risks of crime. Although these risks manifest in distinct domains, their common point of origin lies in their connection with the development of new technologies. As Luhmann (2002) emphasizes, technology "transforms dangers into risks simply because it creates possibilities for decision-making that previously did not exist". Thus, one can distinguish between *technical risks* – adverse events caused by the structure of the technology itself (for example, an accident due to a system malfunction) – and *sociotechnical risks*, which result from the deliberate use of technology and the power relations guiding such use (for example, the detonation of a bomb).

It is therefore no coincidence that this evolution of the concept of risk has become even more evident in our era of rapid technological progress. On the one hand, calculating and predictive capacities – as well as life-enhancing tools – have expanded considerably; on the other, margins of instability and unpredictability regarding the consequences of human action have increased proportionately. Even the distinction between natural and artificial is blurred: the looming catastrophes threatening our planet are increasingly caused by human activity and the use of technology, to the point that François Ewald has spoken of genuine "artificial catastrophes".

In Beck's postmodern society, risks acquire a *global* character (Beck, 2008) which comprises three main features: (1) the *delocalization* (spatial,

temporal, and social) of their causes and consequences; (2) the *incalculability* of their outcomes and impacts, rendering every decision grounded in a fundamental not-knowing; and (3) their *non-compensability*, which makes the logic of indemnification impossible.

These characteristics make calculability and control especially difficult, since human action now exerts effects so extensive across space and time that the causal chain of responsibility is often uncertain and difficult – if not impossible – to reconstruct. Responsibility for present action increasingly conceals the threat of unimaginable and unforeseen consequences. In this context, risk becomes uncertainty, undermining the possibilities of compensation, damage limitation, security, and calculation: «Risk society is a catastrophic society. In it the exceptional condition threatens to become the norm» (Beck, 1999: 24). In our age, therefore, the weight and importance of each decision is magnified, as is the desire to identify those responsible. Yet this identification grows ever more difficult, due to both uncertainty and systemic complexity. The specialization and division of labor in industrialized society have fragmented responsibility to the point of near dissolution. As Anders perceptively observed:

> The 'technification' of our being: the fact that today it is possible that unknowingly and indirectly, like screws in a machine, we can be used in actions, the effects of which are beyond the horizon of our eyes and imagination, and of which, could we imagine them, we could not approve – this fact has changed the very foundations of our moral existence. Thus, we can become 'guiltlessly guilty', a condition which had not existed in the technically less advanced times of our fathers (Anders, 1962: 1).

Thus, the modern idea of risk gives way to a far more unstable and elusive concept, one that no longer allows us fully to contain and control our future and the consequences of our actions, though it still strives to «calculate the incalculable» (Dean, 1999). As Mark Coeckelbergh (2015a) has argued, risk – or more precisely, *being-at-risk* – becomes an existential category, describing humanity's condition of exposure and uncertainty in the age of new technologies.

The purpose of risk forecasting and calculation has always been to devise strategies for eliminating or mitigating threats. Covello and Mumpower (1985) suggest that, historically, the main techniques have been: (1) avoiding risk through prohibitions; (2) regulating and modifying human activities to reduce the magnitude of risk; (3) reducing the vulnerability of exposed populations; (4) developing interventions after events to mitigate impact; and (5) compensating for damages through insurance mechanisms.

Today, however, in the face of such unprecedented power and unpredictability, our risk management techniques – as Beck vividly puts it – resemble bicycle brakes mounted on an intercontinental rocket. Precautionary measures for prevention and damage reduction often prove ineffective, since the catastrophic consequences of our actions are not fully foreseeable (consider pollution and climate change), while *ex ante* solutions such as intervention and insurance appear neither sufficient nor proportionate. After all, how and whom can we compensate in the face of an irreversible global catastrophe, such as the explosion of a nuclear power plant?

Beyond its historical evolution (with the significant shifts in meaning already traced), the concept of risk also embodies a plurality of interpretations. Although the landscape is broad and varied, two principal approaches may be distinguished: one that conceives risk in an essentially objective and rational way, and another that views it as a social and cultural construct, not dependent on factuality itself but on how it is managed and represented.

The first perspective is adopted primarily by disciplines such as engineering, actuarial mathematics, statistics, and epidemiology, and is characterized by a technical-scientific approach. Here, risk is understood as something inherent in reality itself; what is subjective is not the risk but rather its perception, which differs between experts and laypeople and is influenced by cognitive mechanisms that undermine rational evaluation. Risk is thus treated as an objectively measurable fact, the product of the probability of an event and the severity of its consequences: magnitude ($R = P \times D$). Attempts at containment and mitigation focus on altering one of these two variables. Both the identification of risk and the search for solutions are generally reduced to technical factors.

The second perspective, by contrast, is sociocultural in nature and developed largely by philosophical, anthropological, and sociological reflection. Its basic assumption is that risk, before being an objective datum, is the interpretive category through which modern society is governed, and by which it defines its relationship with otherness, with knowledge, and with the future. This confers upon the category of risk (and upon the decisions deriving from it) a political dimension, whereby specific strategies of governance and responsibility distribution are elaborated in response to risks. While it is important to distinguish between the various interpretations, as well as between risk and its perception, it must nevertheless be recognized that the two remain deeply interdependent: «scientific rationality without social rationality remains empty, but social rationality without scientific rationality remais blind» (Beck, 1992: 30).

3. *Systemic Risk*

The transition to the late-modern era has thus given rise to a different conception of risk, which Beck has defined as *global*, insofar as it is de-localised, incalculable, and non-compensable. Within this theoretical framework, the early 2000s saw growing scholarly attention to a more specific dimension of risk: *systemic risk*. This concept, widely adopted across various scientific and disciplinary domains, offers a more nuanced understanding of how certain contemporary risks have emerged and evolved. To grasp this concept, it is first necessary to briefly clarify the meaning of "system". Drawing on philosophical reflection, cybernetics, and complex system science, Terenzio (2025) offers a detailed analysis of the term, identifying its core features. Chief among these is the relational nature of its elements, articulated through a dynamic interplay between parts and the whole. A system thus emerges as an organized and autonomous configuration of components, sustained by nonlinear dynamics. Yet, the order generated within systems is neither absolute nor permanent; rather, it constitutes a fragile equilibrium, continually exposed to disruptions and emergent phenomena capable of transforming individual elements and reconfiguring the system as a whole. It is precisely this interconnection and inherent instability that give rise to what are known as systemic risks.

It is challenging to establish a single, universally accepted definition of this concept, given the breadth of its theoretical and practical applications. One of the most widely cited definitions is offered by Kaufman and Scott (2003, p. 372) who state: «Systemic risk refers to the risk or probability of breakdowns in an entire system, as opposed to breakdowns in individual parts or components, and is evidenced by co-movements (correlation) among most or all parts». A defining feature of systemic risk, therefore, is its tendency to affect multiple components of a system simultaneously. In a similar vein, the definition provided by the Organization for Economic Co-operation and Development introduces the concept of systemic risk as referring to those "risks that threaten society's essential systems, such as infrastructure, health care and telecommunications" (OECD, 2003). A key emphasis in these studies lies in the transmission and scope of risk across the interconnected components of the system (Poledna *et al.*, 2020)[2].

[2] This does not imply that the scope of risk is necessarily global. As Aven and Renn (2019) emphasize, the extent of risk may be regional, national, or global. What is crucial, however, is the internal relationship among the components of the system, regardless of its geographical scale.

Systemic risks can be understood as threats whereby localized failures, accidents, or disruptions have the potential to affect an entire system through contagion mechanisms. In highly interconnected environments – such as health, ecosystems, infrastructure, and food – characterized by complex causal structures and non-linear relationships between causes and effects risk does not remain confined to isolated events but spreads across networks of mutual dependence. The limited understanding of these interconnections, combined with the inherent complexity of such systems, poses significant challenges to both prevention and effective risk management. Systemic risk is characterized by cascading effects that spread across interconnected systems through the movement of people, goods, capital, and information across borders. These dynamics can lead to widespread disruption or even systemic collapse.

The paradigmatic example of systemic risk – indeed, the event after which the very term gained widespread currency – is the global financial crisis of 2007[3]. Prior to this crisis, banking regulation was largely microprudential in orientation, concentrating on individual institutions and the risks apparent in their balance sheets. This approach rested on the assumption that constraining excessive risk-taking at the level of each bank would suffice to prevent the build-up of systemic vulnerabilities across the financial system. The crisis, however, exposed the shortcomings of this framework, demonstrating its inability to capture the intricate interconnections among financial institutions and markets (Allen and Carletti, 2013). More recently, the COVID-19 pandemic (Trump *et al.*, 2021) has triggered a global crisis, clearly revealing the deeply interconnected nature of key systems such as healthcare, finance, food supply chains, and labor markets. This event underscored how disruptions in one domain can rapidly cascade across others, amplifying vulnerabilities and challenging the resilience of complex socioeconomic structures.

Renn *et al.* (2022, 1904-1905) identify several defining properties of systemic risks. First, *complexity*, which refers to the difficulty of tracing and reconstructing the causal relationships within the system. Second, *uncertainty*, stemming from the indeterminacy of causes and characteristics of the phenomena, which in turn undermines confidence in both analysis and decision-making. Third, *ambiguity*, understood as the coexistence of mul-

[3] Other illustrative examples of these dynamics include the desertification and collapse of the Aral Sea, pandemics, cybersecurity threats, global climate change, and the depletion of fish stocks (IRGC, 2018).

tiple, and sometimes conflicting, interpretations of the same phenomenon. Finally, the *ripple effect,* whereby a disruption generates cascading consequences that extend far beyond the initial source of risk.

A valuable overview of systemic risk research is provided by a recent Briefing Note of the UN Office for Disaster Risk Reduction (Sillmann *et al.*, 2022), which offers an integrated perspective on climate, environmental, and disaster risk science. It examines the evolution of systemic risk across disciplines, identifying common conceptual threads without enforcing a singular definition. The Note outlines key attributes of systemic risk and discusses the types of data and information needed to enhance its practical understanding. In the report, systemic risks are examined along several dimensions that help capture their complexity and scope. The first concerns their *scale*: such risks may unfold at the global, national, regional, or local level, yet in all cases they tend to generate repercussions that transcend their initial boundaries. A second dimension relates to the *relations* within and across systems, characterized by feedback loops, interactions, and interdependencies. These intertwined connections amplify the potential for risk propagation and make it difficult to address threats in isolation. Third, systemic risks are marked by the *comprehensibility of the system*. They often involve a lack of knowledge, unpredictability, uncertainty, and ambiguity, which may lead to an underestimation of both their causes and their potential consequences. This opacity undermines the ability to anticipate developments or to design timely responses. A fourth dimension concerns the *transboundary effects* that such risks produce. These include contagion, cascading dynamics, and non-linear processes, through which a localized disruption can trigger a ripple effect, spreading across sectors and jurisdictions that initially appear unrelated.

Finally, systemic risks are defined by their *possible outcomes*, ranging from breakdowns and disruptions to the collapse of entire economic, social, or environmental systems. It is this capacity to destabilize or disintegrate complex structures that makes systemic risks particularly difficult to prevent and to govern.

When considered in light of these dimensions, the concept of systemic risk emerges as a powerful analytical category that extends and refines the sociological theorization of the "risk society" proposed by Beck. Systemic risk does not simply supplant the conventional risks of industrial modernity with those of a globalized and technologically mediated world; rather, it reconfigures the landscape of risk by juxtaposing two coexisting dimensions. The first comprises risks that remain bounded, relatively predictable, and

therefore more susceptible to regulatory control and mitigation strategies, such as regulation and control mechanisms[4]. The second consists of risks that are profoundly complex, marked by interdependencies, cascading effects, and epistemic opacity, which resist both prediction and governance.

From this perspective, systemic risk can be seen as a conceptual bridge between Beck's diagnosis of the reflexive modernity of late industrial societies and contemporary debates on global vulnerabilities, interconnectivity, and the fragility of socio-technical systems. It highlights how the dynamics of globalization, digitalization, and environmental change have intensified the conditions that Beck identified, producing a risk environment that is not only more pervasive but also qualitatively different. Thus, systemic risk serves not merely as a descriptive category but as a normative and operational framework, compelling institutions to rethink strategies of governance, resilience, and adaptation in a world where uncertainty is not an exception but a constitutive feature of social life.

4. *Digital risk society*

After reconstructing the notion of risk from antiquity to the contemporary age, we can now examine how the concept applies to the risks associated with AI systems, focusing on their specific characteristics and the ways in which they are addressed.

It is widely acknowledged that artificial intelligence (AI) poses significant risks. These threats can be categorized according to various criteria. The study by Hendrycks *et al.* (2023) provides an overview of the main sources of catastrophic risk associated with artificial intelligence, organizing them into four categories: 1) *malicious use*, the deliberate deployment of AI systems to cause harm; 2) *AI race*, the adoption of unsafe systems or the relinquishment of human control to machines driven by competitive pressures; 3) *organizational risks*, the increased likelihood of catastrophic failures resulting from system complexity; 4) *loss of control*, the inherent difficulty in governing agents that may become significantly more intelligent than humans.

These risks stem both from the *inherent nature* of AI models – particularly those developed through machine learning algorithms – and from their

[4] Common examples include bicycle theft, foodborne illnesses, and traffic accidents – events that, while harmful, remain localized and can be effectively addressed using existing tools and procedures.

potential misuse. For example, AI can be employed for real-time individual recognition or user profiling, raising serious ethical and privacy concerns.

The first category concerns *technical risks*, namely errors in outputs linked to the quality of data and the reliability of algorithms. These issues are primarily related to accuracy and robustness. Since such systems are fueled by vast quantities of data – often gathered from our online activities – it is essential that datasets faithfully reflect the reality they are meant to represent. If datasets lack accuracy, consistency, completeness, or correctness, the resulting outputs will be flawed, and when used for decision-making, they may expose individuals to significant harm. Likewise, the algorithms processing these datasets must operate according to logical procedures aligned with their intended purposes. Yet how can the accuracy of data and algorithms be verified? The greatest difficulty lies in detecting and identifying errors, especially in systems based on deep learning and neural networks, which are often opaque and difficult to interpret. This opacity problem exacerbates the challenge, as it prevents human understanding of outputs and their validity.

A second category comprises *sociotechnical risks*, which arise from the interaction between humans and AI systems. One notable example is cyberattacks, where generative AI is exploited for phishing, malware distribution, deepfakes, and other forms of social engineering. Sociotechnical risks also emerge from the legitimate use of AI, where certain applications may nonetheless expose individuals or groups to vulnerabilities – for instance, biometric tracking and facial recognition technologies, which may compromise individual freedoms as a side effect of their deployment[5]. Both technical and sociotechnical risks ultimately affect human beings, but their causes differ: the former stem from technical deficiencies, while the latter arise from the human-machine relationship.

Beyond this distinction based on causation, risks may also be classified by their object. One major category concerns privacy risks, related to the management of personal data, which may be used by companies for purposes beyond those to which individuals have consented. Such practices blur the traditional boundary between public and private life by enabling the disclosure of information – such as health status, financial standing, or political preferences – that was once strictly personal. This proliferation of data can create a pervasive sense of surveillance, as individuals are con-

[5] In their study, Guan and colleagues (2022) distinguish between «technical risk» and «management risk», defining the former as encompassing «algorithm risk, data risk, and technology risk», and the latter as including «management risk and decision risk».

tinuously monitored through their devices (computers, smartphones, smartwatches) (Lupton, 2016).

Human autonomy is also at stake, since increasingly sophisticated profiling techniques influence our choices, not only in consumer behavior but also in political opinion-formation. The Cambridge Analytica scandal demonstrated the severe consequences of such practices, but AI is now routinely employed for electoral and propagandistic purposes (O'Neill, 2018; Hinds *et al.*, 2020). This trend is especially visible in online communities – our new digital *agorà* – where opinion exchange has become polarized and self-referential, generating phenomena such as filter bubbles (Parisier, 2011) and echo chambers (Cinelli *et al.*, 2020). These dynamics threaten democratic deliberation processes and compromise the quality of information, which is increasingly polluted by fake news and other forms of truth manipulation (Giusti and Piras, 2020).

Another pressing issue is algorithmic bias, which undermines fairness and justice in automated decision-making. Biases may originate from humans who embed them into systems, or they may be produced autonomously by the systems themselves. Datasets often conceal gender, ethnic, economic, or cultural prejudices that can profoundly harm vulnerable groups. A notable example is the use of the COMPAS software in the United States criminal justice system, which, in assessing recidivism risk, systematically assigned higher scores to African American defendants (Angwin *et al.*, 2023). Similar issues have arisen in medicine, where diagnostic accuracy for minority populations has proven lower due to underrepresentation in training data, leading to discriminatory outcomes (Guerrero *et al.*, 2018).

It is also necessary to consider the so-called *existential risks* (Cappelen *et al.*, 2025) – future scenarios in which AI becomes extremely powerful, evolving into a form of "superintelligence" (Bostrom, 2014) capable of subjugating or even destroying humanity. While some view this possibility as an unlikely dystopian projection, the rapid advancement of generative AI compels us not to dismiss such concerns lightly.

Finally, AI generates *externalities* (Hagendorff, 2022), i.e., consequences seemingly unrelated to its use but directly stemming from it, particularly affecting the environment and labor. Data development, collection, and storage entail a significant ecological footprint, exacerbating the planet's precarious condition. In the labor market, AI displaces human work in specific sectors, with some professions expected to disappear, while others – such as those of so-called *crowdkworkers* – emerge in precarious conditions with little legal protection.

This brief overview of the new threats associated with AI invites reflection in relation to our broader theme. We observe both elements of continuity with past risks and significant elements of novelty. As in earlier times, risks often assume a global dimension and may lead to catastrophic outcomes. In the case of AI, such globalized risks derive primarily from the massive scale of data involved and the pervasiveness of the technology.

Sundberg (2023) conceptualizes this emerging landscape as the *digital risk society*, a context defined by the pervasive presence of intangible technologies which, although frequently presented as solutions, simultaneously generate new and multifaceted risks. This society is further characterized by processes of dehumanization, as an ever-growing range of tasks once carried out by humans is increasingly delegated to machines whose internal operations remain opaque.

Since AI systems are embedded in nearly every aspect of daily life, we face what Mark Coeckelbergh (2015b) calls the «tragedy of the master». If technology was originally conceived – following Aristotelian logic – as a servant of human purposes, today our dependence has deepened to the point of entangling us in a Hegelian master-slave dialectic, defined by dependence and alienation:

> There is a risk that the automation technology we developed and use to serve us renders us vulnerable and dependent in new ways, creates distance between us and material reality, and "automates" us in the sense that we have to adapt our practices to what automation technology does and can do (Coeckelbergh, 2015b: 222).

As with the risks of the twentieth century, those arising from AI can be described as incalculable and unpredictable, thus belonging more to the realm of uncertainty. Unlike the past, however, this unpredictability does not stem solely from human inability to foresee consequences across temporal and spatial distances. In AI systems, often characterized by opacity, it is the functioning of the system itself that remains unknown, presenting a "black box" to human users. This becomes especially problematic when algorithmic decisions are applied in sensitive contexts such as finance or healthcare, where we must rely on outputs that are unintelligible yet generate new forms of correlation. AI risks are therefore exponentially incalculable and unpredictable.

Another novel aspect is that the actions of AI systems – and especially their outcomes – do not necessarily depend on human decisions (Fabris, 2018). While industrial automation has existed in the past, today we increasingly rely on algorithmic decision-making. The ability of AI systems to

learn and independently generate decisions introduces an additional layer of autonomy. This raises profound ethical and legal questions regarding responsibility, but it also reshapes the very configuration of risks. If the distinction between risk and danger rests on whether adverse events derive from human decisions or occur independently of them, then AI-related harms should more accurately be classified as dangers. This shift challenges traditional notions of agency and accountability.

In conclusion, AI systems are *experimental technologies* (van de Poel, 2016) which are those technologies whose risks and benefits are hard to estimate before they are properly inserted in their context of use. The concept of risk in relation to AI is closely tied to ideas of catastrophe, uncertainty, and danger, owing to its global reach, its incalculable and unpredictable character, and its independence from human decision-making. Addressing these challenges requires a dual approach: technological innovation and regulatory oversight. A global research community is actively engaged in developing technical solutions to mitigate algorithmic risks. At the same time, comprehensive legal frameworks are essential to ensure responsible AI deployment.

5. *The Risk-Based Approach in the AI ActFine module*

In response to the rapid expansion of the *digital risk society*, recent years have witnessed growing attention to the ethical and regulatory dimensions of artificial intelligence. Following a series of *soft law* initiatives aimed at guiding its responsible development (Fabris *et al.*, 2024), the need for a more robust and binding regulatory framework became evident – one capable of clearly delineating the boundaries of AI system design, production, and use. In the case of the European Union, this process culminated, after three years of debate and drafting, in the adoption of the *AI Act* in the summer of 2024 (Casonato and Olivato, 2024). This document represents a groundbreaking step at the global level, serving as a source of inspiration for other legislative initiatives and exerting a significant influence on the broader market well beyond the boundaries of the European Union (Bradford, 2020).

In this document, the concept of "risk" plays a central role, appearing over 300 times throughout the text.

Article 3 of the AI Act provides a set of definitions that serve as a useful framework for interpreting the remainder of the text. The Act introduces the notion of "risk", underscoring how central this concept is to the overall structure of the regulation:

2. 'risk' means the combination of the probability of an occurrence of harm and the severity of that harm.

As we can see, this definition reflects a modern understanding of *objective risk*, conceived as the product of the likelihood of an event occurring and the magnitude of its potential damage.

As stated in Recital 26, the principal strategy underpinning the AI Act – commonly referred to as the *risk-based approach* – constitutes a comprehensive response to the diverse risks arising from the deployment of AI systems within the European single market. Furthermore, it establishes a proportionate and balanced regulatory framework designed to address the principal challenges that characterize the contemporary technological environment:

> In order to introduce a proportionate and effective set of binding rules for AI systems, a clearly defined risk-based approach should be followed. That approach should tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate. It is therefore necessary to prohibit certain unacceptable AI practices, to lay down requirements for high-risk AI systems and obligations for the relevant operators, and to lay down transparency obligations for certain AI systems (Recital 26).

To ensure a proportionate regulatory approach that safeguards both fundamental rights and the integrity of the market while fostering technological progress, the AI Act introduces a distinction among different levels of risk, conceptualized as a "risk pyramid". Four categories are identified *ex ante*: (1) unacceptable-risk systems, (2) high-risk systems, (3) limited-risk systems – primarily concerning transparency obligations – and (4) low- or minimal-risk systems.

Under Article 5 of the AI Act, the "unacceptable risk" category encompasses AI systems deemed fundamentally incompatible with EU values and rights, and are therefore prohibited. These include applications that manipulate individuals through subliminal techniques or exploit vulnerabilities based on age, disability, or socioeconomic status, thereby impairing autonomy and causing potential harm. The Act also bans AI systems used for social scoring, predictive policing based on profiling, and indiscriminate collection of facial images from online sources or surveillance. Furthermore, emotional recognition technologies in workplaces or educational settings are prohibited unless justified by medical or safety needs. Biometric categorization based on sensitive attributes – such as race, political affiliation, or religious beliefs – is likewise forbidden. Real-time remote biometric identification in public spaces is only permitted under narrowly

defined law enforcement conditions, subject to judicial oversight.

High-risk AI systems – whose identification and regulation constitute the core focus of the document, beginning with Article 6 – are those that, in order to be placed on the market, must comply with a set of specific obligations. According to the AI Act, high-risk artificial intelligence applications are those that may negatively impact human safety and security, fundamental rights, or the environment. In order to be placed on the European market, these systems must comply with specific requirements (Articles 8–27), including the implementation of risk and quality management systems, appropriate data governance mechanisms, and the use of relevant, representative, and error-free datasets. They must also provide technical documentation, ensure the automatic logging of significant events, and enable effective human oversight throughout their operation.

The scope of high-risk systems covers products already regulated under EU law, such as medical devices, toys, radio equipment, vehicles, lifts, and civil safety systems. Furthermore, Annex III extends this classification to domains including critical infrastructure, education and training, employment and labor management, access to essential private and public services, law enforcement, migration, asylum and border control, as well as the administration of justice and democratic processes.

The third level concerns risks arising particularly from chatbots, deepfakes, and other forms of AI-generated content, where it may be difficult to discern whether one is interacting with a human or a machine. To mitigate such risks, the AI Act imposes specific transparency obligations. The final level represents a residual category, encompassing all systems not included in the previous tiers. These systems are not subject to binding obligations but may voluntarily adhere to soft law instruments, such as codes of ethical conduct and guidelines.

The reference to the *acceptability of risk*, determined through a cost–benefit analysis, aims to assess the degree of trustworthiness of AI systems (Fraser, Bello y Villarino, 2023). The allocation of these risk levels depends primarily on the context of application. However, the rapid development of technologies has highlighted the need for oversight that goes beyond merely identifying applications. Attention must now also be directed toward the underlying AI models themselves, prior to their deployment within specific systems. This approach is particularly crucial given the widespread adoption of generative AI and large language models, which complicate the task of defining discrete applications (Novelli *et al.*, 2023). For this reason, the original risk pyramid – introduced in the first regulatory proposal of April

2021 – has been supplemented by a new dimension, encompassing general-purpose AI models and the potential systemic risks they may generate.

As stated in Article 3, definition 63, a "general-purpose AI model" refers to models trained on large volumes of data, characterized by a high degree of generality and capable of competently performing a wide range of distinct tasks. These models can be integrated into various systems or applications and are able to execute functions such as image and speech recognition, audio and video generation, pattern detection, question answering, and text translation. The type of risk associated with such models is defined as "systemic", as outlined in definition 65 of the same article:

> 'systemic risk' means a risk that is specific to the high-impact capabilities of general-purpose AI models, having a *significant impact* on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society *as a whole*, that can be propagated at scale across the value chain.

This second definition, although admittedly broad and somewhat indeterminate in scope, elucidates this category of risk as being directly associated with general-purpose models. Under Article 51 of the EU AI Act, a general-purpose AI model is deemed to possess systemic risk if it satisfies one of two conditions: (a) it demonstrates high-impact capabilities, assessed using appropriate technical tools, benchmarks, and indicators; or (b) the European Commission determines, either on its own initiative or following a qualified alert from the scientific panel, that the model has equivalent capabilities or impact, based on criteria in Annex XIII. A presumption of high-impact capability arises when the cumulative computational power employed during the model's training surpasses 10^25 floating-point operations (FLOPs); this quantitative threshold serves as a proxy for the potential magnitude of the model's societal and systemic influence[6]. The definition highlights two central features – namely, a *large-scale* and *significant societal impact* that extends beyond individual harm to affect *broader segments of the society*. Thus, the concept of systemic risk cannot be reduced solely to potential damage to "safety, security, and fundamental rights", as frequently invoked within the AI Act; rather, it encompasses these domains in their collective and structural dimensions. Recital 110 further clarifies the nature of

[6] Not all general-purpose models, therefore, entail systemic risk. In the absence of such risk, they are subject primarily to obligations concerning documentation and compliance with copyright requirements. When systemic risk is present, however, additional obligations apply.

systemic risk by providing a more detailed enumeration of its various possible manifestations. This definition of systemic risk aligns with those found in the literature, particularly with the work of Kaufman and Scott (2003: 372): «Systemic risk refers to the risk or probability of breakdowns in an entire system, as opposed to breakdowns in individual parts or components, and is evidenced by co-movements (correlation) among most or all parts». Other characteristics, such as complexity and the cascade effect, are elaborated in the recently published *General-Purpose AI (GPAI) Code of Practice*, which offers a clearer framework for interpreting this section of the AI Act. This transitional document seeks to ensure the presumption of conformity, specify how providers can fulfil the obligations set out in Articles 53-55 of the AIA, maintain up-to-date technical documentation, and support the continuous assessment and mitigation of systemic risks. In particular, Appendix 1 of Chapter 3 contributes to the classification and identification of various types of systemic risks, specifying their nature and sources.

Systemic risks are here primarily characterised by their high-impact capabilities, as defined in Articles 3(64)-(65) of the AI Act, their potential to exert a significant influence on the Union market, and their capacity to propagate widely across the value chain. These elements make such risks particularly complex and far-reaching. Contributing factors further intensify this dynamic. The level of risk tends to increase in proportion to the model's capabilities and diffusion, while the speed at which these risks can materialise is often remarkably high. Moreover, systemic risks in AI may trigger cascading effects that are difficult, if not impossible, to reverse. Their impact is frequently asymmetric, meaning that the actions of a few actors – or even isolated events – can generate disproportionately large and potentially disruptive consequences.

Four main categories of systemic risk emerge: 1) Chemical, biological, radiological and nuclear (CBRN) the facilitation of attacks involving chemical, biological, radiological, or nuclear weapons; 2) Loss of control, models that evade human oversight or enable the autonomous research and development of AI; 3) Cyber offence, offensive capabilities that enable large-scale, sophisticated cyberattacks; 4) Harmful manipulation interference in decision-making processes that threatens fundamental rights and democratic values.

In light of this framework, we can now turn to some reflections on the notion of risk, systemic risk and, more specifically, on their relationship.

As previously noted, the European legislator provides a specific definition of «risk» as «the combination of the probability of an occurrence of harm and the severity of that harm». This conception emphasizes a realistic and

measurable notion of risk, typical of technical domains where evaluation is expected to rely on objective system properties. Yet, such an approach faces several challenges. Abstract dimensions such as human dignity or personal integrity resist objective calculation (Chaberlain, 2023). Moreover, the continuously evolving nature of AI systems – still undergoing rapid expansion – further undermines the feasibility of approaches grounded in predictability and quantitative assessment, calling for broader, more adaptive understandings of risk. As Mahler states:

> Such calculations work best under the assumption that the future conditions of the relevant context are comparable to past conditions, which may be questionable when AI system continues to learn and evolve, for example. If the future is different from the past, calculations may become problematic (Mahler, 2021: 260).

This framework reveals a gap between the definition of risk proposed in the document and the actual nature of the threats posed by contemporary AI systems. On the one hand, the document reproduces objective and realistic interpretations aimed at the technical and calculable resolution of risk; on the other hand, such strategies conflict with the incalculable, unpredictable, and autonomous nature that characterizes AI. It seems that, in the words of Kaminski, «the choice to use risk regulation reflects a particular epistemology: the notion that such AI systems are just math, uncovering some ground truth then contingent social facts» (Kaminski, 2023: 1400).

Moreover, the definition of risk appears to diverge from the way risk is described and assessed. As Novelli observes, risk identification relies on predetermined categories shaped by the values of the European Union. This approach, however, risks being overly static and may underestimate alternative sources of risk arising from different uses – for instance, video games.

The introduction of the category of systemic risk appears to address this concern, as it is characterized by greater impact, heightened uncertainty, and increased complexity. However, as indicated in Definition 65, systemic risk is confined exclusively to GPAI models. This restriction limits the systemic dimension to a narrow subset of models – a position that is, at the very least, debatable, given that in a digital risk society most risks exhibit systemic characteristics. It is therefore necessary to clarify the relationship between the two terms: are they distinct and mutually exclusive categories, or can they overlap?

To address this question, two aspects must be considered.

First, the two levels concern different objects: on the one hand, AI systems, and on the other, AI models. A model is a specific program trained on

data to perform a defined task, such as image recognition or text translation. A system, by contrast, is a broader ensemble of components that includes one or more AI models, together with software, hardware, and data, in order to solve a more complex end-to-end problem. This implies that a system may encompass a model within it.

The second aspect concerns the identification of risk. In the case of systems, the classification of risk levels is based on use, following the so-called "risk pyramid". In contrast, general-purpose AI models are not differentiated by use but rather by the model's computational power. These two dimensions reveal a discontinuity between the risk pyramid and the notion of systemic risk.

Overlaps between the two categorizations may therefore occur. AI systems classified as presenting unacceptable, high, low, or minimal risk may, in fact, be based on the application of GPAI models. At the same time, one might also ask whether systems that do not rely on GPAI models could nonetheless pose a systemic risk. This latter possibility appears to be excluded, as definition 65 makes explicit reference to GPAI models. However, it seems plausible that systemic risks may arise more frequently than anticipated by the law. As previously noted, the current landscape of the digital risk society generates increasing levels of uncertainty and unpredictability – potentially on a global scale – thereby challenging the boundaries set by the existing regulatory framework.

The European regulatory framework, while representing a pivotal step in the global governance of artificial intelligence, remains grounded in a predictive and calculable logic that struggles to grasp the complexity, fluidity, and inherent uncertainty of today's digital ecosystem. Such an approach reflects a primarily technical conception of risk – one that can ostensibly be mitigated through interventions at the system level rather than through consideration of the social equilibria in which such systems are embedded. The rationale behind this orientation is clear: the AI Act is modelled on product safety regulation, upon which the protection of fundamental rights is superimposed. Consequently, the focus is directed more toward the product itself than toward the individuals exposed to potential harm. Yet, this orientation calls for a further reflection – one that takes into account not only the notion of risk but also that of the subject at risk. Given the unpredictable nature of these emerging threats, concentrating solely on the product fails to address the underlying conditions that make harm possible: the exposure and vulnerability of individuals (Zanotti *et al.*, 2024). This limitation underscores the need for a more adaptive approach – one capable of integrating

technical, social, and ethical dimensions into the governance of technological risk. Introducing the notion of vulnerability thus broadens the concept of risk, shifting attention from the systems that generate it to the contexts and subjects who endure it. A vigilant observation of the conditions that enable injustice, discrimination, manipulation, and restrictions of freedom requires, above all, a gaze directed toward the human before the machine.

6. *Conclusions*

The historical and conceptual evolution of risk – from antiquity to the contemporary digital era – reveals a profound transformation: from a calculable, probability-based notion to an uncertain, global, and systemic phenomenon. In the context of AI, risks are not only technical but sociotechnical, affecting individuals and societies through interconnections, opacity, and potential cascading effects. Systemic risk, particularly in general-purpose AI models, highlights how localized failures or autonomous system behavior can propagate across social, economic, and technological networks.

The EU AI Act represents a pioneering step in risk-based regulation, yet its emphasis on calculable, product-centered risk contrasts with the inherent unpredictability and systemic nature of AI. This tension underscores the need for an integrated approach that combines technological safeguards, regulatory oversight, and ethical reflection, shifting the focus from systems alone to the vulnerabilities of the individuals and contexts they affect. Ultimately, contemporary digital risk requires a governance framework that goes beyond technical mitigation, addressing social, ethical, and structural dimensions of harm.

References

Allen, F., Carletti, E. (2013). *What Is Systemic Risk?*, in «Journal of Money Credit and Banking», 45 (1), pp. 121-127.

Anders, G. (1962). *Burning Conscience. The case of the Hiroshima pilot, Claude Eatherly, told in his letters to Gunther Anders, with a postscript for American readers by Anders*, Monthly Review Press, New York.

Angwin, J. *et al.* (2023), *Machine bias. There's software used across the country to predict future criminals. And it's biased against blacks*, in «Propublica», May 23, https://www.propublica.org/article/machine–bias–risk assessments–in–criminal–sentencing.

Aven, T., Renn, O. (2019). *Some foundational issues related to risk governance and different types of risks*, in «Journal of Risk Research», 1, pp. 1-14.

Beck, U. (1992). *Risk Society. Towards a New Modernity*, SAGE, London.

Beck, U. (2008). *Risk Society's 'Cosmopolitan Moment*, in «ComCiência» [online], n. 104.

Bernstein, P.L. (1996). *Against the Gods. The Remarkable Story of Risk*, Wiley, New York.

Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies,* Oxford University Press, Oxford.

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*, Oxford University Press, New York.

Cappelen, H. *et al.* (2025). *AI Survival Stories: a Taxonomic Analysis of AI Existential Risk*, in «Philosophy of AI», 1, pp. 1-19.

Casonato, C., Olivato, G. (2024). *AI Regulation in Europe: Exploring the Artificial Intelligence Act*, in A. Fabris, S. Belardinelli (eds), *Digital Environments and Human Relations. Human Perspectives in Health Sciences and Technology*, Cham, Springer.

Chamberlain, J. (2023). *The Risk-Based Approach of the European Union's Proposed Ar-tificial Intelligence Regulation: Some Comments from Tort Law Perspective*, in «Euro-pean Journal of Risk Regulation», 14, pp. 1-13.

Cinelli, M. *et al.* (2020). *The echo chamber effect on social media*, in «PNAS», 118, 9, e2023301118.

Coeckelbergh, M. (2015a). *Human Being @ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations*, Springer, Cham.

Coeckelbergh, M. (2015b). *The tragedy of the master: automation, vulnerability, and distance*, in «Ethics and Information Technology», 17, pp. 219-229.

Covello, V.T., Mumpower, J. (1985). *Risk Analysis and Risk Management: An Historical Perspective*, in «Risk Analysis», 5, p. 108.

Dean, M. (1998). *Risk, Calcuable and Incalculable*, in « Soziale Welt», 49, pp. 25-42.

Douglas, M. (1992). *Risk and Blame. Essays on Cultural Theory*, Routledge, London.

Fabris, A. (2018). *Ethics of Information and Communication Technologies*, Springer, Cham.

Fabris, A. *et al.* (2024). *Towards a Relational Ethics in AI. The Problem of Agency, the Search for Common Principles, the Pairing of Human and Artificial Agents*, in A. Fabris, S. Belardinelli (eds), *Digital Environments and Human Relations. Human Perspectives in Health Sciences and Technology*, vol. 150, Springer, Cham.

Fraser, H., Bello y Villarino, J.M. (2023). *Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough*, in «European Journal of Risk Regulation», pp. 1-16.

Giddens, A. (2000). *Il mondo che cambia. Come la globalizzazione ridisegna la nostra vita,* il Mulino, Bologna.

Giusti, S., Piras, E. (2020). *Democracy and Fake News. Information Manipulation and Post-Truth Politics*, Routledge, London.

Guan, H. *et al.* (2022). *Ethical Risk Factors and Mechanisms in Artificial Intelligence Decision Making*, in «Behavioral Sciences» 12, 9, p. 343.

Guerrero *et al.* (2018). *Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies*, in «Scientific Reports», 1, pp. 1-8.

Hagendorff, T. (2022). *Blind spots in AI ethics*, in «AI Ethics», 2, pp. 851-867.

Hendrycks, D., Mazeika, M., Woodside, T. (2023). *An Overview of Catastrophic AI Risks*, in «Arxiv», https://doi.org/10.48550/arXiv.2306.12001.

Hinds, J. *et al.* (2020). *"It wouldn't happen to me": privacy concerns and perspectives following the Cambridge Analytica scandal*, in «International Journal of Human-Computer Science», 143, 102498.

IRGC (2018). *Guidelines for the Governance of Systemic Risks*, International Risk Governance Center (IRGC), Lausanne.

Kaminski, M.E. (2023). *Regulating the Risks of AI,* in «Boston University Law Review», 103, pp. 1347-1411.

Kaufman, G., Scott, K. (2003). *What Is Systemic Risk, and Do Bank Regulators Retard or Contribute to It?*, in «Independent Review», 7, pp. 1-31.

Luhmann, N. (1992). *Risk: a sociological theory*, de Gruyter, New York.

Lupton, D. (1999). *Risk*, Routledge, London.

Lupton, D. (2016). *The diverse domains of quantified selves: self-tracking modes and dataveillance*, in «Economy and Society», 45, 1, pp. 101-122.

Mahler, T. (2021). *Between risk management and proportionality: the risk-based approach in the EU's Artificial Intelligence Act Proposal*, in «Nordic Yearbook of Law and Informatics», pp. 245-267.

O'Neill, B. (2018). *The great Cambridge Analytica conspiracy theory*, in «The Spectator», 21 March 2018 (https://www.spectator.co.uk/article/the–great–cambridge–analyt-ica–conspiracy–theory/).

OECD (2003). *Emerging Risks in the 21st Century. An Agenta for Action*, OECD Publications Service, Paris.

Parisier, E. (2020). *The Filter Bubble: What the Internet Is Hiding from You*, Penguin, London.

Poledna, S., Rovenskaya, E., Dieckmann, U., Hochrainer-Stigler, S., Linkov, I. (2020). *Systemic risk emerging from interconnections: the case of financial systems,* in W. Hynes, M. Lees, J. Müller (eds), *Systemic thinking for policy making: the potential of systems analysis for addressing global policy challenges in the 21st century*, OECD Publishing, Paris.

Renn, O. *et al.* (2022). *Systemic Risks from Different Perspectives*, in «Risk Analysis», 42, 9, pp. 1902-1920.

Sillman, J. *et al.* (2022). *Briefing Note Systemic Risks: Review and Opportunities for Research, Policy and Practice from the Perspective of Climate, Environmental and Disaster Risk Science and Management*, International Science Council, Paris.

Sundberg, L. (2024). *Towards the Digital Risk Society: A Review*, in «Human Affairs», 34, 1, pp. 151-164.

Terenzio, F. (2025). *Systemic Vulnerability: From AI Systems to Environmental Systems*, in «Topoi».

Trump, B.D. *et al.* (2021). *Multi-Disciplinary Perspectives on Systemic Risk and Resilience in the Time of COVID-19,* in I. Linkov *et al.* (eds), *COVID-19: Systemic Risk and Resilience, Risk, Systems and Decisions*, Springer, Cham.

Van de Poel (2016). *An Ethical Framework for Evaluating Experimental Technology*, in «Sci Eng Ethics», 22, pp. 667-686.

Zanotti, G., Chiffi, D., Schiaffonati, V. (2023). *AI-Related Risk. An Epistemological Approach*, in «Philos. Technol.», 37, p. 66.

Abstract

This article explores the historical evolution of the concept of risk and its contemporary applications in the domain of artificial intelligence. Building on Beck's notions of "global risk" and "risk society", we argue that today's context can be described as a digital risk society, *increasingly marked by uncertainty and unpredictability. The idea of systemic risk – widely discussed in complexity science and economics – emerges as a particularly suitable conceptual tool for interpreting this scenario. The centrality of risk is further evidenced by the EU's AI Act, which adopts a risk-based approach. We examine this regulation to analyze the interaction between risk and systemic risk, highlighting the limitations of this framework.*

Keywords: risk; systemic risk; AI Act; vulnerability.

Silvia Dadà
Università di Pisa
silvia.dada@unipi.it

T

Marta Fasan

The Fundamental Rights Impact Assessment: Lights and Shadows in Protecting Fundamental Rights in the Field of Artificial Intelligence*

1. *Untying the knot. New regulatory approaches in the definition of artificial intelligence legal framework*

The complex relationship between the legal domain and the scientific-technological sphere lies at the heart of contemporary debates in public law scholarship, both at the national level and, increasingly, within supranational and international contexts[1].

It has long been evident that the outcomes of scientific and technological innovation have a significant influence on the social and relational structures that shape society as a whole. These transformations constitute one of the main challenges facing modern legal systems, which are required to respond through their own traditional tools, categories, and methods[2]. Yet, this task is far from straightforward, given the fundamental differences

* This paper constitutes an update of the subject matter previously delineated in M. Fasan, *I meccanismi di valutazione d'impatto nel prisma della tutela dei diritti fondamentali. Un'analisi in prospettiva comparata in materia di intelligenza artificiale*, in «BioLaw Journal - Rivista di BioDiritto», 2 (2025), pp. 183-198.

[1] R. Brownsword, H. Somsen, *Law, innovation and technology: fast forward to 2021*, in «Law, Innovation and Technology», 1 (2021), pp. 1-28; L. Bennet Moses, *How to Think about Law, Regulation and Technology: Problems with "Technology" as a Regulatory Target*, in «Law, Innovation and Technology», 1 (2013), pp. 1-20. Among Italian legal scholars cfr. E. D'Orlando, *Decisore politico e scienza: l'emergenza sanitaria come catalizzatore dell'affermazione di un paradigma normativo* evidence-based, in «DPCE online», 1 (2023), pp. 511-535; S. Penasa, *Alla ricerca di un lessico comune: inte(g)razioni tra diritto e scienze della vita in prospettiva comparata*, in «DPCE online» 3 (2020), p. 3312.

[2] L. Bennet Moses, *Why Have a Theory of Law and Technological Change?*, in «Minnesota Journal of Law, Science & Technology», 2 (2007), pp. 589-606; S. Rodotà, *Tecnologie e diritti*, il Mulino, Bologna 2021, p. 146 ff.

 DOI: 10.4454/w3pw4069

that distinguish the legal sphere from the technological-scientific one.

Law, in order to fulfil its constitutional function of "measuring" and constraining power for the protection of fundamental rights[3], must ensure a sufficient degree of predictability and stability in the production of norms aimed at governing social behaviours, as well as a broad level of consensus regarding their content[4]. Conversely, the characteristics that define the scientific-technological phenomenon as an object of regulation differ remarkably. On the one hand, the products of scientific and technological innovation are technically complex, both in their terminology and in their operational mechanisms, thereby limiting accessibility and full comprehension for non-experts in the field[5]. On the other hand, scientific and technological phenomena evolve at a significant pace, continuously transforming the functionalities and solutions offered by technological products and processes[6].

These elements, which permeate the relationship between law and technology, become even more evident when examining the role that law can play in regulating artificial intelligence (AI).

AI technologies are demonstrating their ability to pervasively reshape society and its inner structures, to the extent that they can mould a new model of "algorithmic society" which is «(…) organized around social and economic decision-making, who do not only make the decisions but also, in some cases, carry them out»[7]. This phenomenon gives rise to new legal issues, determined by both the technological features of AI and its social impact, and new protection requirements that the law need to address through

[3] A. Simoncini, *Sovranità e potere nell'era digitale*, in T.E. Frosini *et al.* (eds), *Diritti e libertà in internet*, Mondadori Education, Firenze 2019, p. 20 ff.; C. Casonato, *Intelligenza artificiale e diritto costituzionale: prime considerazioni*, in «Diritto Pubblico Comparato ed Europeo», Special issue (2019), pp. 101-130; R. Toniatti, *Common Law Constitutionalism*, in «Costituzionalismo britannico e irlandese», 2 (2024), p, 6 ff.

[4] L. Reins, *Regulating New Technologies in Uncertain Times - Challenges and Opportunities*, in L. Reins (ed.), *Regulating New Technologies in Uncertain Times*, Springer, The Hague 2019, p. 20.

[5] C. Casonato, *The Essential Features of 21st Century Biolaw*, in E. Valdés, J.A. Lecaros (eds), *Biolaw and Policy in the Twenty-First Century. Building Answers for New Questions*, Springer, Cham 2019, p. 77 ff.

[6] R. Brownsword, *Rights, Regulation, and the Technological Revolution*, Oxford University Press, Oxford 2008, pp. 162-165. Among Italian legal scholars see also A. Iannuzzi, *Il diritto capovolto. Regolazione a contenuto tecnico-scientifico e Costituzione*, Editoriale Scientifica, Napoli 2018, pp. 3-4; S. Penasa, *La legge della scienza: nuovi paradigmi dell'attività medico-scientifica. Uno studio comparato in materia di procreazione medicalmente assistita*, Editoriale Scientifica, Napoli 2015, p. 29 ff.

[7] J.M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, in «Ohio State Law Journal», 5 (2017), p. 1219.

the definition of principles, rules and obligations to regulate the development and use of AI[8].

Nevertheless, the definition of legal frameworks for AI regulation continues to represent a remarkably challenge.

The inherent complexity of AI systems, the rapidity of scientific progress in this domain, and the AI de-spatialized nature, which raise substantial challenges concerning the territorial scope of applicable regulations[9], have, in recent years, revealed the potential inadequacies of traditional legal categories and instruments in effectively governing AI. These limitations emerge both in the need to provide adequate support for innovation and in the imperative to safeguard the rights of individuals who interact with AI systems[10].

This awareness has prompted legislative bodies to complement traditional regulatory frameworks with novel approaches and tools. These new strategies increasingly rely on the participation of technological actors and the private sector, recognizing the potential of these agents to contribute to responsive and adaptive regulatory practices. From this perspective, it is unsurprising that recent regulatory acts on AI provide forms of delegated regulation, through by-design solutions and self-regulatory measures, within broader political and legislative frameworks. Such developments give rise to hybrid forms of co-regulation, which employ specific mechanisms and procedures to ensure both the promotion and protection of the diverse interests related to the development and deployment of AI systems[11].

[8] G. De Gregorio, *Digital Constitutionalism in Europe. Reframing Rights and Powers in the Algorithmic Society*, Cambridge University Press, Cambridge 2022; O. Pollicino, G. De Gregorio, *Constitutional Law in the Algorithmic Society*, in H.-W. Micklitz *et al.* (eds), *Constitutional Challenges in the Algorithmic Society*, Cambridge University Press, Cambridge 2022, p. 5 ff.

[9] A. Garapon, *La despazializzazione della giustizia*, Mimesis, Milano-Udine 2021.

[10] O. Pollicino, *The quadrangular shape of the geometry of digital power(s) and the move towards a procedural digital constitutionalism*, in «European Law Journal», 14 August 2023, pp. 1-21.

[11] L. Lessig, *Code and Other Laws of Cyberspace*, Basic Books, New York 1999; R. Brownsword, *Law 3.0*, Routledge, Abingdon-New York 2020, p. 28; K. Yeung, *Towards an Understanding of Regulation by Design*, in R. Brownsword, K. Yeung (eds), *Regulating Technologies. Legal Features, Regulatory Frames and Technological Fixes*, Oxford University Press, Oxford-Portland 2008, p. 81 ff.; N. Gunningham, J. Rees, *Industry Self-Regulation: An Institutional Perspective*, in «Law & Policy», 4 (1997), p. 364 ff. Among Italian legal scholars cfr. A. Iannuzzi, *Paradigmi normativi per la disciplina della tecnologia: auto-regolazione, co-regolazione ed etero-regolazione*, in «Bilancio, Comunità, Persona», 2 (2023), p. 91 ff.; A. Simoncini, *La co-regolazione delle piattaforme digitali*, in «Rivista trimestrale di diritto pubblico», 4 (2022), p. 1031 ff.; G. Mobilio, *L'intelligenza artificiale e i rischi di una* "disruption" *della regolamentazione giuridica*, in «BioLaw Journal - Rivista di BioDiritto», 2 (2020), pp. 415-418.

2. *Impact assessment as a "new" tool for the protection of fundamental rights in the field of artificial intelligence*

The adoption of co-regulatory approaches in the governance of artificial intelligence, aimed at fostering synergy between public-institutional and private regulatory dimensions, is most clearly expressed by impact assessment tools. This specific type of legal mechanism is proving to be quite successful in the legislative frameworks designed to regulate the development and use of AI, due to the distinctive features that characterise its operation and implementation.

In regulatory terms, impact assessment is a tool that facilitates the interaction between public authorities and private actors in evaluating the potential effects that a specific activity, process, or technological product may have on a defined set of protected interests and values. This mechanism can be situated within the broader category of Technology Assessment, a concept first developed in the United States during the latter half of the twentieth century to describe the assessment practices carried out by policymakers on the societal impacts of emerging technologies[12]. Although originally conceived as a self-regulation tool related to environmental protection[13], impact assessment has progressively evolved into an adaptable tool, that may be applied to different domains, including the safeguard of people's rights. Over time, it has come to play a pivotal role not only in safeguarding general human rights[14] but also in addressing specific legal interests, such as the right to health[15], the right to data protection[16],

[12] Cfr. D. Banta, *What is Technology Assessment*, in «International Journal of Technology Assessment in Health Care», 1 (2009), pp. 7-9.

[13] J. Glasson, R. Therivel, *Introduction to Environmental Impact Assessment*, Taylor & Francis, London 2013.

[14] R. Boele, C. Crispin, *What direction for human rights impact assessments?*, in «Impact Assessment and Project Appraisal», 2 (2013), pp. 128-134; G. De Beco, *Human Rights Impact Assessments*, in «Netherlands Quarterly of Human Rights», 27 (2009), p. 139 ff.; D. Kemp, F. Vanclay, *Human rights and impact assessment: clarifying the connections in practice*, in «Impact Assessment and Project Appraisal», 2 (2013), pp. 86-96; A. Mantelero, *Beyond Data. Human Rights, Ethical and Social Impact Assessment in AI*, Springer, The Hague 2022, p. 15 ff.

[15] B. Harris-Roxas *et al.*, *Health Impact Assessment: The State of the Art*, in «Impact assessment and project appraisal», 12 (2009), p. 43 ff.

[16] D. Wright, *The State of the Art in Privacy Impact Assessment*, in «Computer Law & Security Review», 28 (2012), p. 54 ff.; G. Georgiadis, G. Poels, *Towards a privacy impact assessment methodology to support the requirements of the general data protection regulation in a big data analytics context: a systematic literature review*, in «Computer Law & Security Review», 44 (2022), pp. 1-21.

sustainability rights[17] and the right to intergenerational equality[18].

In the European Union context, impact assessment has undergone a process of full institutionalisation and recognition as a tool for safeguarding interests within EU legislation. First introduced by directives mandating environmental impact evaluations for public and private projects that could affect constitutionally protected environmental values[19], impact assessment has then been embedded into other regulatory acts, such as data protection regulation[20], health technology assessment one[21], and corporate due diligence regulation[22]. This evolution transformed impact assessment from a mechanism of private self-regulation into a cornerstone of co-regulation, whereby legislative authorities define the objectives and parameters of assessment, and regulated entities are responsible for analysing and managing the impacts of their activities, economic, technological, or otherwise, within that established framework[23].

As a result, impact assessment has become an essential component of regulatory interventions in technologically complex and rapidly evolving fields, such as digital innovation and AI. In these areas, the active involvement of regulated entities is essential to identifying legally sound solutions that ac-

[17] H.-J. De Kluiver, *Towards a framework for effective regulatory supervision of sustainability governance in accordance with the EU CSDD Directive. A comparative study*, in «European Company and Financial Law Review», 1 (2023), pp. 203-239; F. Rohde *et al.*, *Broadening the perspective for sustainable artificial intelligence: sustainability criteria and indicators for Artificial Intelligence systems*, in «Current Opinion in Environmental Sustainability», 66 (2024), pp. 1-12.

[18] R.E. Parsons, L.K. Mottee, *Exploring equity in social impact assessment*, in «Current Sociology», 4 (2024), pp. 732-752. Among Italian legal scholars cfr. L. Bartolucci, *La valutazione di impatto generazionale delle leggi come forma di attuazione degli artt. 9 e 97 Cost.*, in «Federalismi.it», 4 (2024), pp. 39-49.

[19] Directive 85/337/EEC (subsequently amended by Directive 97/11/EC, Directive 2003/35/EC and Directive 2009/31/EC) and Directive 2011/92/EU, which was also amended in 2014 by Directive 2014/52/EU.

[20] Regulation (EU) 2016/679 of the European Parliament and of The Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[21] Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on health technology assessment and amending Directive 2011/24/EU.

[22] Directive (EU) 2024/1760 of the European Parliament and of The Council of 13 June 2024 on corporate sustainability due diligence and amending Directive (EU) 2019/1937 and Regulation (EU) 2023/2859.

[23] A.M. Esteves, D. Franks, F. Vanclay, *Social impact assessment: the state of the art*, in «Impact Assessment and Project Appraisal», 1 (2012), pp. 34-42; A. Mantelero, *Beyond Data. Human Rights, Ethical and Social Impact Assessment in AI*, cit., p. 15 ff.; P.G. Chiara, F. Galli, *Normative Considerations on Impact Assessments in EU Digital Policy*, in «MediaLaws - Rivista di Diritto dei Media», 1 (2024), p. 86 ff.

count for the technical and operational complexity of these technologies[24].

Moreover, the growing significance of impact assessment with respect to AI regulation can be explained by two additional considerations.

First, impact assessment is inherently aligned with the risk-based regulatory approach, which has become the dominant paradigm in both digital governance and AI regulation[25]. This approach seeks to avoid excessive and overburdening regulation by tailoring legal obligations according to the level of risk a given activity poses to protected societal interests and values[26]. In this framework, the intensity of regulation, ranging from the lack of rules to outright prohibitions, increases proportionally to the level of foreseeable risk and its potential effects[27]. Consequently, impact assessment becomes a crucial mechanism for operationalizing this approach: by identifying and evaluating the impacts of AI technologies, it enables a more specific determination of risk levels and their implications for legally protected interests[28].

Second, impact assessment provides a concrete analytical framework for examining the real-world consequences of AI applications. Because it is typically conducted by those directly involved in the development or deployment of AI, namely, the deployers themselves, the assessment reflects the specific operational context and tangible outcomes of AI technologies. This contextualized analysis complements the general and abstract nature of legislative intervention, thereby reducing the risk of overregulation and ensuring that restrictive measures are applied only where and when needed[29].

[24] S. Penasa, *Diritto e tecnologia nella recente riflessione giuridica comparata: "etichette" concettuali, sistemi di produzione normativa e metodi della comparazione*, in «Diritto pubblico comparato ed europeo», Special Issue (2024), p. 959 ff.

[25] J. Black, R. Baldwin, *When risk-based regulation aims low: approaches and challenges*, in «Regulation & Governance», 1 (2012), pp. 2-22; A. Alemanno, *Regulating the European Risk Society*, in A. Alemanno *et al.* (eds), *Better Business Regulation in a Risk Society*, Springer, New York 2012, p. 37 ff.

[26] M. Macenaite, *The "Riskification" of European Data Protection Law through a two-fold Shift*, in «European Journal of Risk Regulation», 8 (2017), pp. 506-540.

[27] C. Quelle, *Enhancing Compliance under the General Data Protection Regulation: The Risky Upshot of the Accountability- and Risk-based Approach*, in «European Journal of Risk Regulation», 9 (2018), pp. 502-526.

[28] P. Dunn, G. De Gregorio, *The Ambiguous Risk-Based Approach of the Artificial Intelligence Act: Links and Discrepancies with Other Union Strategies*, in D. Dushi *et al.* (eds), *Proceedings of the Workshop on Imaging the AI Landscape after the AI Act (IAIL 2022)*, CEUR Workshop Proceedings, Amsterdam 2022, pp. 1-9; M. Ebers, *Truly Risk-based Regulation of Artificial Intelligence. How to Implement the EU's AI Act*, in «European Journal of Risk Regulation», 6 November 2024, pp. 1-20.

[29] C. Casonato, *A Brief Introduction*, in L. Antoniolli, F. Cortese, E. Ioriatti, B. Marchetti (eds), *Trento e la comparazione giuridica: voci, esperienze, riflessioni. Dalla testimonianza di Ro-*

Considering these features, the inclusion of impact assessment within AI regulatory frameworks arises as a rational and effective means of protecting fundamental rights. The mechanism facilitates the balance, both abstract and concrete, between competing interests due to the pervasive use of AI systems. It allows policymakers and deployers to account for the specificities of each use case, ensuring that protective measures are tailored to the actual risks involved rather than imposed indiscriminately. In doing so, impact assessment enables the legal system to mitigate negative consequences without unduly hindering technological innovation[30].

However, the ability of impact assessment to guarantee the protection of fundamental rights is not an inherent or automatic quality of the mechanism itself. Its effectiveness depends on the precision with which legislation defines its scope, parameters, and implementation procedures. Without such clarity, the mechanism risks losing its normative and protective function, becoming an empty formal requirement unable of achieving the co-regulatory goals that justify its adoption.

3. *Impact assessment mechanisms in the regulation of artificial intelligence. Potential insights from a comparative law perspective*

The importance of embedding impact assessment within a clearly defined regulatory framework, one that specifies how private entities are to conduct such assessments to make them effective tools for the protection of fundamental rights, becomes more evident when examining the legislation that governs the use of artificial intelligence and mandates its implementation. From this standpoint, a comparative legal analysis proves to be the most effective method for identifying both the strengths and the shortcomings that characterize the legal and constitutional design of impact assessment as a tool for regulating AI systems and ensuring the protection of fundamental rights[31].

dolfo Sacco e Mauro Cappelletti, Editorial Scientifica, Trento 2024, p. 227 ff.; A. Mantelero, *AI and Big Data: a blueprint for a human rights, social and ethical impact assessment*, in «Computer Law & Security Review», 4 (2018), pp. 754-772.

[30] A. Mantelero, *The Fundamental Rights Impact Assessment (FRIA) in the AI Act: roots, legal obligations and key elements for a model template*, in «Computer Law & Security Review», 54 (2024), p. 1 ff.

[31] G. Guerra, *An Interdisciplinary Approach for Comparative Lawyers: Insights from the Fast-Moving Field of Law and Technology*, in «German Law Journal», 3 (2018), pp. 579-612. Among Italian legal scholars cfr. A. Vedaschi, *Tecnologia,* counter-terrrorism *e diritti*, in «Diritto pub-

In this regard, three legal systems serve as paradigmatic cases for analysis: Canada, the United States, and the European Union. Each legal system has developed regulatory tools which, though varying in scope and regulatory depth, share the common feature of recognising impact assessment as a central mechanism for protecting both legal interests and fundamental rights potentially affected by AI. At the same time, significant divergences among these systems reveal distinctive regulatory trends and differing conceptions of the function ascribed to this tool[32].

Across all three legal systems, impact assessment has a shared purpose: to provide a concrete analytical process capable of evaluating how AI may affect the fundamental values and rights safeguarded within each legal system, and to establish measures to prevent irreversible harms. Nevertheless, a closer comparative examination reveals notable distinctions among the frameworks, which are crucial for understanding both the potential and the limitations of impact assessment as a mechanism mainly aimed at protecting people's fundamental rights[33].

In Canada, the main piece of legislation to be examined is the Canadian Directive on Automated Decision-Making of 2019, which provides the most comprehensive and detailed model of impact assessment. The Directive, applicable to situations in which AI is employed to make or support administrative decisions by federal institutions, assigns a central role to impact assessment within its regulatory framework[34]. The type and level of impact generated by the use of AI systems are parameters for determining the applicability of the general obligations and safeguards prescribed by the Directive[35]. Several of the guarantees and obligations contained in the Directive are, in fact, implemented in more detailed and stringent forms as the AI

blico comparato ed europeo», Special issue (2024), p. 979 ff.; G. Smorto, *Il ruolo della comparazione giuridica nella contesa per la sovranità digitale*, in «DPCE online», 1 (2023), p. 339 ff.

[32] On the importance of identifying case studies, according to various methodological criteria, in comparative public law analysis cfr. R. Hirschl, *The Questions of Case Selection in Comparative Constitutional Law*, in «The American Journal of Comparative Law», 1 (2005), p. 125 ff.

[33] E. Örücü, *Methodologies for Comparative Law*, in J.M. Smits, J. Husa, C. Valcke, M. Narciso (eds), *Elgar Encyclopedia of Comparative Law*, Edward Elgar Publishing, Cheltenham 2023, pp. 42-50.

[34] Art. 5, Directive on Automated Decision-Making. For a general comment on the Canadian Directive cfr. T. Scassa, *Administrative Law and the Governance of Automated Decision Making: A Critical Look at Canada's Directive on Automated Decision Making*, in «U.B.C. Law Review», 54 (2021), p. 251 ff.; B. Attard-Frost, A. Brandusescu, K. Lyons, *The governance of artificial intelligence in Canada: finding and opportunities from review of 84 AI governance initiatives*, in «Government Information Quarterly», 2 (2024), pp. 1-24.

[35] Appendix C, *Directive on Automated Decision-Making*.

level of impact increases, and consequently, as the risks to the protection of fundamental rights intensify. From this perspective, the Canadian model of impact assessment is particularly robust, as it explicitly provides for the adoption of specific measures designed to eliminate, or at least mitigate, the adverse effects that AI systems may have on the fundamental rights and values the Directive seeks to protect[36].

Further indicators of the completeness and coherence of the Canadian approach emerge from two key aspects: the identification of the protected legal interests upon which the impact of AI must be assessed, and the definition of graduated levels of impact that may result from its use. Regarding the first aspect, the Directive extends guarantees and protections to the rights of both individuals and communities, paying particular attention on safeguarding equality, dignity, privacy, autonomy, health, and well-being, as well as the economic interests of affected persons and the sustainability of the ecosystems within which AI systems are deployed. As to the second profile, the Directive distinguishes four levels of impact, classified according to the degree of reversibility of the effects produced and the temporal duration of those effects in relation to the protected interests at stake[37].

The overall effectiveness and precision of the Canadian impact assessment framework are further reinforced by another element. Indeed, the Canadian Government has issued a set of supplementary guidelines clarifying the application of the Directive and delineating the substantive scope of the parameters discussed above[38]. More precisely, the Canadian Government specifies the areas of risk that may be relevant to the life cycle and use of AI systems, as well as some of the possible mitigating measures and the percentage score ranges that correspond to different levels of impact. Moreover, the Government also makes a questionnaire available to support interested parties in carrying out this algorithmic impact assessment[39].

In contrast, the United States legal framework exhibits notable differ-

[36] M. Karanicolas, *To Err is Human, to Audit Divine: A Critical Assessment of Canada's AI Directive*, in «Journal of Parliamentary and Political Law», 14 (2019), p. 1 ff.

[37] Appendix B, *Directive on Automated Decision-Making*. On these profiles see also H.L. Janssen, *An approach for a fundamental rights impact assessment to automated decision-making*, in «International Data Privacy Law», 1 (2020), pp. 76-106.

[38] Government of Canada, *Algorithmic Impact Assessment tool*, available at https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html#toc2.

[39] Cfr. B. Deshaises, D. Hall, *Responsible use of automated decision systems in the federal government*, 1 December 2021, at https://www.statcan.gc.ca/en/data-science/network/automated-systems.

ences from the Canadian approach, being characterized by a comparatively lower degree of precision and detail concerning both the substantive content and procedural implementation of the impact assessment mechanism.

Even though the regulatory landscape in the field of AI is relatively fragmentated and has recently undergone a significant shift in the federal approach to AI governance[40], the relevance of impact assessment tools in AI regulation has been acknowledged at federal level, especially by the Federal Artificial Intelligence Risk Management Act. This Act, within a framework primarily focused on the identification and mitigation of AI-related risks, emphasizes the need for federal agencies to consider the potential impact of artificial intelligence on legally protected interests. Specifically, the Act entrusts the National Institute of Standards and Technology (NIST) with the task of developing guidelines to assist federal agencies in implementing mechanisms for assessing and managing the risks associated with AI systems, with explicit attention to the possible implications for individual rights[41]. Within these NIST guidelines, the inclusion of an impact assessment mechanism forms part of the broader set of risk mitigation measures to be adopted by federal agencies. However, the regulatory framework does not provide detailed specifications regarding the parameters, criteria, or procedures governing the implementation of such assessments. Rather, the guidelines merely outline a general requirement to prioritize risk management according to the level of the impact produced, and to allocate appropriate resources for mitigation, without specifying anything on the nature of the potential impacts, the evaluative benchmarks to be applied, or the concrete measures to be implemented in proportion to the identified risks[42].

At the State level, instead, the impact assessment receives somewhat greater emphasis, particularly in the regulatory approach adopted by the State of Colorado through the adoption and implementation of the Colorado Act concerning consumer protections in interactions with artificial intelligence systems of 2024. Under this statute, the impact assessment obligation

[40] C. Novelli, A. Gaur, L. Floridi, *Two Future of AI Regulation under the Trump Administration*, 24 April 2025, in https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5198926; V. Lubello, *From Biden to Trump: Divergent and Convergent Policies in The Artificial Intelligence (AI) Summer*, in «DPCE online», 1 (2025), pp. 49-65.

[41] Section 2 (b) (2), H.R. 6936 - Federal Artificial Intelligence Risk Management Act of 2024.

[42] National Institute of Standards and Technology, *Artificial intelligence Risk Management Framework (AI RMF 1.0)*, January 2023, at https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

consists primarily of the annual provision and retention by AI deployers[43] of technical documentation certifying: the intended purpose and objectives of the technology, its anticipated benefits, the data used and generated by the system, the performance evaluation metrics employed, and the transparency and post-deployment monitoring measures adopted by the deployer[44]. Nevertheless, the Colorado Act does not establish a classification of impact levels based on the consequences of AI use, nor does it provide a comprehensive list of the rights and values that must be considered within the assessment. The only exception is the explicit reference to the principle of equality, as the Act requires deployers to examine the potential risk of discrimination arising from the functioning of AI systems[45]. Furthermore, the Colorado Act does not mandate the adoption of specific or binding mitigation measures to limit or prevent risks or harms associated with AI deployment.

From this perspective, both the federal and state-level frameworks in the United States display a comparatively lower degree of regulatory precision than the Canadian legal approach, thereby leaving considerable discretion to the public authorities and private actors responsible for conducting impact assessments, including the definition of the relevant parameters and of the degree of AI impact on the protected interests and rights.

3.1. *Impact assessment mechanisms in the regulation of artificial intelligence. The fundamental rights impact assessment under the AI Act legal framework*

A similar framework for impact assessment to that provided by the US legal system, both at federal and state level, can also be identified within the regulatory framework developed by the European Union through the adoption of the Regulation (EU) 2024/1689 of the European Parliament and of the Council of

[43] Under the Colorado Act, a deployer is defined as a person who uses an AI system considered high-risk for professional purposes, as established by Section 6-1-1701 (6), SB 24-205 Concerning consumer Protections in interactions with artificial intelligence systems. According to this definition, therefore, a person who uses an artificial intelligence system for purely personal reasons cannot be classified as a deployer and, consequently, the provisions of the Colorado Act will not apply to them.

[44] Section 6-1-1703 (II) (b), SB 24-205 Concerning consumer Protections in interactions with artificial intelligence systems.

[45] Section 6-1-1703 (II) (b) (II), SB 24-205 Concerning consumer Protections in interactions with artificial intelligence systems. For a comment see R. Dotan, *US regulation of artificial intelligence*, in C. Lütge *et al.* (eds), *The Elgar Companion to Applied AI Ethics*, Edward Elgar Publishing, Cheltenham 2024, p. 153 ff.

13 June 2024, laying down harmonized rules on artificial intelligence (commonly referred to as AI Act). Regulation (EU) 2024/1689, indeed, introduces the obligation to carry out a fundamental rights impact assessment (FRIA) for AI systems, however this obligation applies only under specific and narrowly defined conditions. First, the impact assessment is required only for AI systems classified as high risk in accordance with Article 6(2) of the Regulation, thereby excluding other categories of high-risk systems also recognised under the AI Act[46]. According to Article 27(1) of Regulation (EU) 2024/1689, all systems that are classified differently, or as high-risk systems under the provision of Annex I, are exempt from this mechanism[47]. Moreover, AI Act expressly excludes from the impact assessment AI systems intended for use as safety components in the management or operation of critical digital infrastructure, road traffic, or the supply of water, gas, heating, or electricity[48]. Second, the obligation to carry out an impact assessment is limited to specific categories of deployers explicitly identified by the Regulation[49]. Accordingly, only public sector entities, private bodies providing public services, and deployers of high-risk systems used to assess individuals' creditworthiness or their eligibility for health and life insurance are subject to this requirement. All other users of high-risk systems falling outside these categories are exempt from this obligation[50]. With regard to the content and implementation of the impact assessment, Regulation (EU) No. 2024/1689 provides that the process must include: a description of the processes in which the deployer intends to use the technology; a description of the period of time during which the system is expected to be used; the categories of individuals and groups potentially affected by the AI system in a specific context and the associated risks; a description of human oversight measures implemented; and the mitigation measures to be taken if the identified risks arise[51]. Furthermore, concerning

[46] In general, on the content and scope of the impact assessment provided for in the AI Act, see F. Palmiotto, *The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of the Regulation*, in «European Journal of Risk Regulation», 17 January 2025, pp. 1-24.

[47] Systems not covered by this assessment mechanism would include, for example, medical devices incorporating artificial intelligence models.

[48] Article 27(1) of Regulation (EU) 2024/1689.

[49] Like the Colorado Act, the AI Act also identifies the deployer as «natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity», as defined in Article 3(4) of Regulation (EU) 2024/1689.

[50] Article 27(1) of Regulation (EU) 2024/1689.

[51] Article 27(1)(a), (b), (c), (d), (e), (f), Regulation (EU) 2024/1689. For a commentary on this fundamental rights impact assessment see P.G. Chiara, F. Galli, *op. cit.*, p. 86 ff.; A. Mantelero,

the definition of the rights and values to be protected and against which the impact produced is to be measured, the European legislator identifies them in the general category of fundamental rights protected by the EU, especially those provided by the Charter of Fundamental Right of the European Union.

The AI Act provides also further information for the implementation of the fundamental rights impact assessment. It is mandatory only for the first use of the high-risk system, allowing deployers to rely, in similar cases, on previously performed impact assessment or existing impact assessment carried out by provider. If the deployer considers that, during the use of the high-risk system, one of the assessment elements has changed, the deployer shall intervene in order to update the information[52]. Moreover, where any of the obligations required for the implementation of the FRIA have already been fulfilled by carrying out the data protection impact assessment (DPIA), the fundamental rights impact assessment shall complement the latter evaluation[53]. Finally, the deployer has also the duty to notify the market surveillance authority of the impact assessment results through the submission of a specific template that the AI Office will develop to help deployers in complying with this AI Act obligation[54].

4. *The fundamental rights impact assessment in the field of artificial intelligence. Lights and shadows in EU regulation and potential future paths*

The comparative reconstruction of the legal frameworks governing impact assessment provides an opportunity to draw several preliminary conclusions regarding the actual scope and effectiveness of this mechanism as a mean of protecting fundamental rights under the Regulation (EU) No. 2024/1689.

As previously emphasized, the implementation of impact assessment in the examined regulatory context is essential to make more substantive both processes of risks identification and balancing the rights ad stake. First, the fundamental right impact assessment allows to evaluate the concrete effects of deploying AI in a specific context of use and, consequently, better

The Fundamental Rights Impact Assessment (FRIA) in the AI Act: roots, legal obligations and key elements for a model template, cit., p. 7 ff.; C. Novelli, *The European Artificial Intelligence Act: some implementation issues*, in «Federalismi.it», 2 (2024), p. 110 ff.

[52] Article 27(2) of Regulation (EU) 2024/1689.

[53] Article 27(4) of Regulation (EU) 2024/1689.

[54] Article 27(3) and (5) of Regulation (EU) 2024/1689.

define the meaningful risks to address. Thus, this tool contributes to make the AI Act's horizontal approach more adaptable to the needs that may arise depending on the sector in which the AI system is employed. Second, the substantial balancing carried out through the impact assessment mechanism may be the mean to achieve the main goals promoted by the AI Act. In this way, it is possible to foster AI development and use as long as its impact is under control and without compromising the protection of fundamental rights[55].

However, the analysis also highlights several structural weaknesses inherent in the current EU regulatory design of the fundamental rights impact assessment. In particular, the AI Ac framework reveals significant shortcomings that limit the tool's potential to serve as an effective rights-protection mechanism. These weaknesses manifest in two principal ways: first, through the excessive discretion granted to private actors in determining the methods and parameters of the assessment, and second, through the restricted scope of application of the obligation itself. The latter issue is particularly pronounced, where the requirement to conduct the impact assessment does not extend to all AI systems, nor even to all systems classified as high risk, and it is limited to a specific group of deployers. This selective applicability risks undermining the consistency and universality of fundamental rights protection across AI contexts[56].

Moreover, further elements may weaken the efficacy of impact assessment as framed by Regulation (EU) 2024/1689. The choice of referring to such a broad assessment parameter like the general category of fundamental rights may render the mechanism meaningless and devoid of content, especially if taking into account the complex multi-level system of codification and protection of fundamental rights at both EU and national level[57].

[55] C. Novelli, *op. cit.*, p. 110 ff.; C. Novelli *et al.*, *AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act*, in «Digital Society», 13 (2024), pp. 1-29; A. Mantelero, *The Fundamental Rights Impact Assessment (FRIA) in the AI Act: roots, legal obligations and key elements for a model template*, cit., p. 4 ff.

[56] On the limitations of the described fundamental rights impact assessment model see also M. Lasek-Markey, L. Hogan, *Delivering on the promise of the fundamental rights impacr assessments in the EU AI Act: intersectionality and vulnerability*, in «European Journal of Law and Technology», 2 (2025), pp. 1-23. Among Italian legal scholars cfr. F. Donati, *La protezione dei diritti fondamentali nel Regolamento sull'intelligenza artificiale*, in «Rivista AIC», 1 (2025), pp. 1-20.

[57] Recital 48 of Regulation (EU) 2024/1689 tries to exemplify the fundamental rights mainly affected by the use of AI systems that may be: the right to human dignity, respect for private and family life, protection of personal data, freedom of expression and information, freedom of assembly and of association, the right to non-discrimination, the right to education, consumer

Finally, the lack of independent scrutiny of the impact assessment's results by an external public authority may also be problematic, as it risks reducing the assessment to a mere procedural formality, no more than the completion of a checklist by the deployer.

From this perspective, the current EU configuration of impact assessment risks shifting the delicate activity of rights-balancing, traditionally exercised by public and institutional actors, into the almost exclusive domain of private entities. Such a shift could weaken one of the fundamental principles of contemporary constitutionalism: that the protection and limitation of power, including technological power, must remain under public oversight. Entrusting the assessment process entirely to private deployers, without adequate institutional safeguards, creates the danger of conflicts of interest that could distort the evaluation of impacts and compromise the mechanism's legitimacy and protective function[58].

To prevent such an outcome and to address the weaknesses inherent in the approach set out in the AI Act, a potential response is offered by the impact assessment model developed and adopted by the Council of Europe's Committee on Artificial Intelligence: the Methodology for the risk and impact assessment of artificial intelligence systems from the point of view of human rights, democracy and the rule of law (also known as HUDERIA methodology). Although non-binding, the HUDERIA methodology provides a structured framework for impact assessment designed to safeguard human rights, democracy, and the rule of law, and may aid as a supporting tool for complying with both the obligations laid down in the Framework Convention on AI and those established by the AI Act[59]. The HUDERIA methodology is intended for both public and private actors who are required to identify, assess, and manage the risks generated by AI, and it is grounded

protection, workers' rights, the rights of persons with disabilities, gender equality, intellectual property rights, the right to an effective remedy and to a fair trial, the right of defence and the presumption of innocence, and the right to good administration. On the topic cfr. M. Lasek-Markey, L. Hogan, *op. cit.*, pp. 1-23. On issues that may arise in a system of multi-level protection of fundamental rights see G.F. Ferrari, *I diritti nel costituzionalismo globale: luci e ombre*, Mucchi, Modena 2023, p. 8 ff.; R. Bin, *Critica della teoria dei diritti*, Franco Angeli, Milano 2018, p. 69 ff.; A. Ruggeri, *La tutela "multilivello" dei diritti fondamentali, tra esperienze di normazione e teorie costituzionali*, in «Politica del diritto», 3 (2007), pp. 317-346.

[58] C. Casonato, *L'intelligenza artificiale fra pubblico e privato: una sfida per il costituzionalismo (e per i costituzionalisti)*, in «Diritto pubblico comparato ed europeo», 1 (2025), pp. 5-13.

[59] Council of Europe, *HUDERIA: new tool to assess the impact of AI systems on human rights*, 2 December 2024, at https://www.coe.int/en/web/portal/-/huderia-new-tool-to-assess-the-impact-of-ai-systems-on-human-rights.

in a socio-technical approach that evaluates the impact of AI by considering the interconnectedness of technology, human decision-making, and social structures[60].

This impact assessment model is organized into four distinct phases. The first, the *context-based risk analysis*, entails conducting a preliminary assessment of the consequences of AI use, analysing risk factors, mapping potential impacts and determining possible risk levels, as well as performing a triage activity for the most high-risk AI systems. The second phase, the *stakeholder engagement process*, focuses on the involvement of individuals or groups affected by the application of AI and its potential impacts, with the aim of integrating the perspectives of those who will directly experience the effects of the technology and thereby gaining a better understanding of its potentially negative consequences. The third phase, dedicated more specifically to *risk and impact assessment*, involves carrying out the assessment in light of the two preceding phases, according to clearly defined criteria such as the scale of the impact, the number of people affected, the duration, reversibility, and probability of the impact. The final phase, the *mitigation plan*, requires identifying and designing mitigation measures according to a hierarchical criterion based on the severity and probability of the impact, as well as defining specific remedies for individuals who suffer adverse consequences from the application of AI.

The HUDERIA methodology also stipulates the need for ongoing monitoring and continuous review of the impact assessment, with reference to the factors relating to the development, deployment, and use of AI within real-world environments[61].

The HUDERIA methodology, thus, offers an impact assessment model with a higher level of precision in the definition of the parameters of the assessment, of the criteria for the implementation of the mitigation measures and pays more attention on the specificity of the protected values, like for democracy and the rule of law. In these terms, the integration of the fundamental rights impact assessment model provided for in the AI Act with that

[60] It seems appropriate to highlight how the recognition of artificial intelligence as a socio-technical product would represent a fundamental step in the process of regulating it and limiting the powers attributable to this technology. In this regard, see S. Lindgren, *Introducing critical studies of artificial intelligence,* in S. Lindgren (ed.), *Handbook of Critical Studies of Artificial Intelligence*, Edward Elgar Publishing, Cheltenham 2023, pp. 1-19.

[61] Council of Europe, *HUDERIA: new tool to assess the impact of AI systems on human rights*, 2 December 2024, at https://www.coe.int/en/web/portal/-/huderia-new-tool-to-assess-the-impact-of-ai-systems-on-human-rights.

outlined by the HUDERIA methodology would promote a more coherent and comprehensive framework for AI regulation.

Looking ahead, such complementation could enable the realization of a genuine co-regulatory model, one that reconciles the imperatives of technological innovation with the foundational goals of contemporary constitutionalism: the protection of fundamental rights, the preservation of human dignity, and the legitimate exercise of public authority over the transformative power of technology.

Abstract

This paper seeks to examine the role of impact assessment mechanisms as instruments for the protection of fundamental rights within the regulatory frameworks governing emerging technologies, with specific reference to artificial intelligence systems. It will first consider the increasing prominence of impact assessments as a governance tool for artificial intelligence, highlighting their dual function: on the one hand, to safeguard the diverse interests implicated in the deployment of intelligent systems, and on the other, to prevent undue limitations on the development and application of such technologies. Subsequently, the paper will analyse, from a comparative perspective, the design and implementation of impact assessment mechanisms in the field of artificial intelligence, with particular emphasis on the European Union framework under the AI Act.

Keywords: fundamental rights; artificial intelligence; impact assessment; comparative law; European Union.

Marta Fasan
University of Trento
marta.fasan@unitn.it

T

Elio Grande

Il loop espanso. Per un ecosistema dell'intelligenza artificiale

1. *Introduzione*

Quest'articolo desidera, sotto forma di idea, piantare il seme di un possibile ecosistema dell'intelligenza artificiale – o per meglio dire, un ecosistema abitato *anche* da sistemi artificiali intelligenti – a partire però da una concreta prova di esistenza: il metodo di design "a specchio" di sistemi di intelligenza artificiale collaborativi battezzato *Endless Tuning* (Fabris *et al.*, 2024), già implementato in un protocollo e testato con alcuni esperti di dominio (Grande, 2025)[1]. Un tale ecosistema, di radice ermeneutica, ne costituisce dunque per un verso l'espansione realizzabile; per un altro, a testimonianza di quanto tangibilmente aporetico sia l'inizio (Fabris, 2002), ne costituisce già la parziale realizzazione: già, infatti, esistevano i dataset impiegati per gli esperimenti, *già* i moduli di programmazione, e così via lungo l'abbozzo di un circolo.

Alla strutturazione di questo ecosistema dovrebbero a dire il vero contribuire, tra loro cooperando, competenze di varia natura. Agli specialisti del task da eseguire (medico, operatore bancario, consulente etc.), dovrebbero essere affiancati lo sviluppatore e il manutentore di sistemi di intelligenza artificiale (magari, tra l'altro, la stessa persona). Al manager, invece, il compito

[1] È possibile testare il protocollo online in tre versioni, rispettivamente riconoscimento dello stile di un'opera d'arte, concessione di prestiti bancari e riconoscimento della presenza o meno di polmonite in una radiografia toracica. I codici sono disponibili su GitHub alla pagina https://github.com/eliogrande/TheEndlessTuning, e possono essere eseguiti tramite un *Jupyter Notebook* disponibile su *Google Colab* alla pagina https://colab.research.google.com/drive/1m1mDWTlE5egzT_oSR18hHLlcwwrX401N?authuser=1#scrollTo=3pMgbEeXeB5X. Si raccomanda, per la riuscita dell'esecuzione, l'opzione di *runtime* che fa uso di processore grafico.

 DOI: 10.4454/tqnjph66

di preparare il campo organizzando un'infrastruttura, secondo però le indicazioni del giurista che dovrebbe perlomeno accertarsi della compliance con le prescrizioni del Regolamento Europeo sull'Intelligenza Artificiale (European Parliament *et al.*, 2024), del *General Data Protection Regulation* (European Union, 2016) etc. *Dulcis in fundo*, l'eticista: un ruolo apparentemente marginale, a prediligere l'approccio deontologico; eppure, un tessitore degno dell'Eros del mito. Mano ad ago e filo, dunque: dell'infrastruttura summenzionata, a mo' di trama, l'etica dell'intelligenza artificiale costituisce inevitabilmente l'ordito – forse, certo, per varie ragioni invisibile – giacché ne *orienta* il disegno. Dapprima in piccolo, nella relazione tra lo strumento/agente e l'operatore; poi, espandendosi, fungendo da baricentro del contesto sociale di quella medesima relazione. Secondo quest'ordine, perciò, procederemo: l'etica nell'*interface design*, l'esperienza dell'applicazione concreta di *Endless Tuning*, la descrizione dei suoi possibili risvolti sociali.

2. *Il modo in cui una relazione si manifesta:* interface design

Non difettano i codici etici cui riferirsi, a partire da prodotti istituzionali come le *Ethics Guidelines for Trustworthy AI* (European Commission *et al.*, 2019). Anzitutto, però, c'è la difficoltà di trascrivere efficacemente principi morali in uno script (Coeckelbergh, 2019) perché, un modello non potendo riconoscere da sé valori adattandosi ad ogni situazione, rimangono facilmente aggirabili. Per esempio, l'imposizione di *constraints*, come il LLM che non rivela come mettere in atto un attacco *DoS*, non copre il rischio di *jailbreaking*. Poi, il carattere di "*very soft law*" non conferisce all'etica dell'intelligenza artificiale forza vincolante. Facilmente la si può ridurre a mero commentario (Floridi, 2023; Bietti, 2020). Fece scalpore il disappunto di Mark Coeckelbergh e Thomas Metzinger quando, alla stesura del White Paper da parte della Commissione Europea (European Commission, 2020), molte delle raccomandazioni etiche elaborate dal HLEG erano state ignorate o ridotte a semplici dichiarazioni di principio[2].

Quando "cosa" e "prodotto" coincidono, non diviene invece il filosofo un designer, la cui arte risieda però *nell'interrogare il prodotto stesso*; nel questionare, *mentre la disegna*, l'essenza di quella "cosa"? Secondo Floridi (2017),

[2] L'articolo, uscito sul *Tagesspiel* il 14/04/2020, è consultabile in inglese alla pagina https://background.tagesspiegel.de/digitalisierung-und-ki/briefing/europe-needs-more-guts-when-it-comes-to-ai-ethics? (ultimo accesso 11/10/2025).

condivisibilmente, nell'infosfera l'etica che salvaguardi la libertà dell'agire diviene dunque un design pro-etico: con una similitudine, non un guard-rail entro cui limitare il proprio agire bensì un incentivo all'agire consapevole. Oggetto di design, d'altra parte, è sempre un'interfaccia – concetto che insieme a quello di protocollo può essere esteso anche oltre gli stretti confini di un dispositivo (Floridi, 2017). Ma cos'è un'interfaccia? Questo termine, scrive Zannoni (2024), implica certamente comunanza, collegamento e relazione; tuttavia «è necessario riflettere sul fatto che l'interfaccia non esiste ed è solo un'entità astratta a cui ci riferiamo quando ci relazioniamo tra diversi sistemi reali o virtuali» (Zannoni, 2024: 7). Spesso all'interfaccia è associato il termine *affordance*, coniato da James Jerome Gibson (1966) col significato di indicare ciò che le cose più o meno direttamente ci invitano a fare. Val la pena, per giungere al punto, riportare le sue proprie parole: «When the constant properties of constant objects are perceived (the shape, size, color […]), the observer can go on to detect their affordances. I have coined this word as a substitute for values, a term which carries an old burden of philosophical meaning» (Gibson, 1966: 285). *Dunque, qual è la natura di tale valore, precondizione di un'interfaccia?* Una sola sembra la risposta: *l'utilizzabilità*. "Design", in altre parole, sarebbe sinonimo di "design di oggetti d'uso" costantemente "sotto-mano" (Heidegger, 2024): essi *si offrono* all'uso, venire utilizzati costituisce la loro propria essenza.

Questo, tuttavia, non è affatto ovvio: l'utilizzabilità non pare infatti l'unico modo d'essere, né l'unica condizione di possibilità di design di un'interfaccia. Più sottilmente, con (Fabris, 2016; Buber, 1970), sosteniamo che l'utilizzabilità, producendo oggetti e "cose", costituisca un'implicita e concreta scelta, l'attuazione di una tra varie possibilità che *sono, costituiscono* una relazione, così facendo già azzoppando tale relazione, poiché se ne occlude l'orizzonte di possibilità. Non a caso, in materia di intelligenza artificiale proprio a questo punto il termine *reliability* (che indica la robustezza intesa come resistenza a perturbazioni in ingresso, eppure presenta, sempre non per caso, anche una sfumatura di significato esistenziale derivata dal latino *religamen*, da cui *religio*) inizia a far spazio a quello di *trustworthiness*. Da accordo che coinvolge ad accordo che delega – salvo poi cercare di rimediare, contraddittoriamente, chiedendo l'esplicabilità di algoritmi ormai incredibilmente performanti.

Facendo ricerca all'interno dello Spoke 1.6 del grande progetto nazionale italiano FAIR, intitolato *Legal and Ethical Design of Trustworthy AI Systems*, abbiamo dunque scelto come principio guida di design quello della relazione stessa (*anche*) con lo strumento. Tale relazione è intesa come

il continuo coinvolgimento nello sforzo di estrarre informazione rilevante (aggettivo qui migliore che "utile"). Il principio di fondo è un principio-contenitore se vogliamo, o meta-principio, desunto dall'etica della vulnerabilità e della cura. Questo trasforma il design in un problema di *governo* dell'intelligenza artificiale, più che di uso o controllo. Sotto questo profilo, un'interfaccia è dunque intesa come *il modo in cui una relazione si manifesta* – comportando però ossimoricamente che si dia un margine – passi il termine – di s-progetto, di non-design. Opzione che pare invero giustificata: altrove (Fabris *et al.*, 2024; Grande, 2025) abbiamo riportato ragionevoli indizi che i dispositivi appartenenti alla famiglia dell'intelligenza artificiale siano in senso lato detentori di un margine di *inutilizzabilità*. Per esempio, inconsulte ed "innaturali", poiché *assai precisamente errate*, sono le reazioni pressoché di un modello alla presenza di *adversarial examples*, cioè dati in ingresso affetti da perturbazioni minimali mirate a ribaltare il risultato di un'inferenza (scoperti in Szegedy *et al.*, 2013); per un *survey*, cfr. Wiyatno *et al.*, 2019). Oppure, la ragione che vincola l'ingresso all'uscita di un modello nei cicli di training ed inferenza di modelli opachi come *Deep Neural Networks*, *Support Vector Machines* e *Random Forests* rimane parzialmente oscura anche dinanzi a modelli trasparenti, di cui cioè sono noti i parametri. O ancora, il proliferare di modelli ad elevata complessità, relativamente a numero di parametri o FLOPs necessari all'addestramento ha reso noto l'emergere di alcune abilità (Wei *et al.*, 2022) come il *few-shot prompting* che, sebbene da taluni negate (Schaeffer *et al.*, 2023), destano tuttora meraviglia.

3. *Proof of concept*

Endless Tuning è un metodo di design (destinato originariamente ai processi decisionali, sebbene confidiamo possa essere esteso anche all'intelligenza artificiale generativa) che risponde a due bisogni: da una parte quello di compensare le tentazioni del cosiddetto *automation bias* (Skitka *et al.*, 1999), ovvero la tendenza a delegare, con fiducia cieca, decisioni alle macchine; dall'altra quello di tracciare le responsabilità in caso di danno, fornendo la possibilità di dirimere la questione piuttosto che cercando a priori il "colpevole". Per far ciò, si adoperano due regole. La prima è che, trovandosi l'uno allo specchio dell'altro, il sistema di intelligenza artificiale e l'operatore dovrebbero riuscire ad imparare l'uno dall'altro e ciascuno a suo modo, ovvero l'operatore riflettendo a partire dai suggerimenti del sistema ed il sistema essendo soggetto a vincoli di apprendimento per i casi futuri.

Sintonizzandosi l'uno e l'altro sulla verità del fenomeno in osservazione, potenzialmente senza una fine. La seconda è che qualunque interazione con il sistema e qualunque dato prodotto in maniera non facilmente riproducibile siano salvati in una memoria, per poi essere eventualmente consultati in sede di giudizio. Forse che *Endless Tuning* sia perciò incentrato sulla responsabilità civile piuttosto che quella morale? In verità, fermo restando che non esiste un "design del senso di colpa", scopo precipuo di questo progetto è facilitare delle relazioni. In quest'ottica, sebbene all'operatore sia destinato un ruolo rilevante e gli sia richiesta proprio un'assunzione di responsabilità nel prendere una decisione, tale responsabilità non è pienamente configurabile a priori, bensì si costruisce durante il flusso decisionale stesso, e si ri-costruisce a posteriori. Questo, coerentemente sia con un approccio all'etica della cura e della vulnerabilità – una duplice *different voice* (Gilligan, 1993): quella di chi sarà eventualmente soggetto alle decisioni e quella dell'intelligenza artificiale stessa – sia con il criterio di responsabilità oggettiva che pare oggi il più adeguato a dirimere tali questioni (dal Verme, 2024).

Ne abbiamo istanziato un protocollo applicativo che prevede, in ordine: una prima opinione da parte dell'operatore; alcuni suggerimenti sull'informazione processata *e dunque indirettamente* relativi allo stato di cose sotto osservazione, senza che l'operatore conosca ancora il risultato della predizione (utilizzo inverso, *ante-hoc*, di *eXplainable AI*, non lontano dal paradigma di *evaluative AI* (Miller, 2023); confronti di similarità con altri casi); rilevazione dell'effettiva predizione, modulata però sulla *confidence* (misura di certezza) di tutti i risultati possibili; riaddestramento correttivo del modello, previo lo stoccaggio di alcuni dati (*fine-tuning set*) in un "salvadanaio". Ad ogni stadio l'opinione dell'operatore può essere aggiornata, ogni interazione è tracciata in un log con un *timestamp*, ed elementi come mappe e parametri inizializzati in modo randomico vengono salvati. Questo protocollo è stato a sua volta realizzato in tre prototipi, ovvero tre analoghe applicazioni *Streamlit*[3] (basate però su modelli di intelligenza artificiale differenti)[4] dedicate rispettivamente al riconoscimento dello stile di un'opera d'arte, al riconoscimento della presenza o meno di polmonite in una radiografia toracica, ed

[3] Cfr. http://streamlit.io/.

[4] Per i dettagli tecnici, cfr. Grande (2025). Solo per completezza d'informazione, comunque, specifichiamo qui che si tratta di due *ResNet* (He *et al.*, 2016), rispettivamente a 34 e 18 strati, la prima addestrata su alcune classi del *WikiArt Dataset* (Steubk, 2022) e la seconda sul *Chest X-Ray Images (Pneumonia) Dataset* (Mooney, 2018); il terzo modello è un albero decisionale addestrato sul *Loan Approval Classification Dataset* (Wei Lo, 2024).

alla valutazione di candidati per la concessione o meno di prestiti bancari. Da quest'ultimo modello, per esempio, dinanzi alle caratteristiche del candidato (come età, reddito etc.), venivano direttamente estratte delle regole in linguaggio naturale che facevano evincere come alcune *features* si distinguessero da certi valori standard (non sempre uguali). Questo (come analoghi passaggi negli altri prototipi) aveva lo scopo di compensare possibili bias di ancoraggio (cfr. Kahneman, 2011) da parte del decisore, forzando quest'ultimo a rivalutare il caso con maggiore attenzione (*cognitive forcing* è infatti il nome del genere di strategia adottata: cfr. Buçinca *et al.*, 2021).

I prototipi sono stati testati attraverso delle interviste ad altrettanti esperti di dominio (un direttore di banca, un docente di estetica ed un primario di radiologia), con risultati complessivamente positivi. Sebbene infatti alcuni difetti siano stati concretamente riscontrati, per esempio riguardo l'espressività di alcune mappe di salienza dei pixel, gli intervistati hanno dichiarato di non essersi sentiti in alcun modo sostituiti dal sistema di intelligenza artificiale[5] – il quale ha prodotto in tutti e tre i casi predizioni corrette. Inoltre, eccetto il medico, favorevole piuttosto all'opzione di rigide linee guida per stabilire a priori le responsabilità in caso di danno, gli esperti hanno condiviso l'impostazione di tracciamento delle attività. *Endless Tuning* rientra dunque nella categoria dell'*hybrid decision-making*, un paradigma di collaborazione tra essere umano ed intelligenza artificiale che sempre più si fa avanti (Punzi *et al.*, 2023), nel tentativo di intercettare il miglior compromesso tra l'accuratezza ed i diritti fondamentali della persona, la privacy ed il benessere collettivo, l'ergonomia e la sostenibilità.

4. *Il loop esteso*

Se da un lato *Endless Tuning* è un design collaborativo, dall'altro la sua doppia ciclicità è a tutti gli effetti un'ermeneutica, dove a turno e tuttavia quasi simultaneamente l'operatore ed il sistema giocano a vicenda il ruolo di interprete ed interpretato. Un'ermeneutica dell'intelligenza artificiale coerente con la più generale versione digitale descritta da (Romele, 2019), che proponeva il concetto di e*magination* – un analogo dell'immaginazione

[5] Preferiamo qui adoperare la parola "sistema", piuttosto che "modello", in quanto il modello stesso è avvolto in un'interfaccia complessa, che include un algoritmo di spiegazione ed un compressore di dati atto a misurare la distanza geometrica tra il caso sotto osservazione ed altri presi dal dataset di apprendimento.

produttiva ravvisabile nella capacità, da parte di un algoritmo, di analizzare grandi moli di dati. Quel *surplus* di informazione (latore di senso) estraibile dalla e nell'interazione, bilanciando (pure in modo asimmetrico) il potere di iniziativa degli agenti in campo, allontana il baricentro sia da un approccio *human-in-the-loop* che da uno *machine-in-the-loop*: il loop stesso ne diviene, piuttosto, il centro. Dal suo nucleo esso *inevitabilmente* si espande, coinvolgendo, come giustamente notato da (Romele, 2019), altre figure. Per questo un ecosistema si profila, dove i loop divengono concentrici. Sul lato sinistro della Figura 1, il cui loop interno è lo schema base di *Endless Tuning*, dietro il sistema di intelligenza artificiale si trovano i dati di addestramento e il complesso di figure come il progettista dell'interfaccia, lo sviluppatore del modello e del software, e l'annotatore dei dati. Sul lato destro, invece, ci sono gli operatori: l'individuo, per esempio, può ben essere colui che prende effettivamente la decisione; gli altri, anch'essi operatori, intervengono successivamente insieme al primo come annotatori di nuovi dati da fornire al modello stesso durante la fase di aggiornamento.

Figura 1. *Endless Tuning* considerato come un sistema socio-tecnologico esteso. L'icona della matita è stata presa da Flaticon[6].

Se i dataset annotati sul portale *Kaggle*, le architetture disponibili nei framework *PyTorch* o *Scikit-Learn*, una discreta potenza computazionale (solo in un caso siamo dovuti ricorrere ad un server), degli esperti e soprattutto l'impulso alla condivisione (l'accesso a quasi tutte le risorse, infatti, è stato gratuito) non fossero stati già una realtà, questo *proof of concept* non sarebbe stato sperimentabile. Questo doppio circolo ermeneutico digitale, parzialmente, esiste già: in principio è davvero la relazione. Se dunque per

[6] https://www.flaticon.com/free-icons/pencil.

un verso si tratta di espanderla, per un altro si tratta semplicemente di riconoscerla, mettendo insieme i pezzi. Se da un lato le regole del metodo rimangono le medesime, applicate su una scala più grande (per esempio, dal *fine-tuning* alla sostituzione del modello medesimo; dal singolo decisore all'equipe; dalla cura dell'individuo soggetto alla decisione a quella della comunità di appartenenza), dall'altro lato il loop espanso dev'essere cucito sulle esigenze di un territorio, dunque in un contesto applicativo mutevole. Difficile, quindi, fornire un protocollo. Piuttosto, un esempio può far più di mille parole. Può valere pertanto la pena di lasciar emergere i caratteri, le funzionalità, le conseguenze etiche e giuridiche di un tale ecosistema descrivendone icasticamente un'istanza fittizia – immaginata, per altro, a partire da uno dei nostri tre casi di studio.

5. *Un esempio di ecosistema*

Un'ondata di freddo travolge l'area periferica industriale di una città, cogliendo impreparato il signor Mario Rossi, un uomo sulla cinquantina i cui polmoni concedono presto spazio ad un'infiammazione. Timoroso che la situazione possa peggiorare, Mario decide di andare a fare una radiografia presso un centro diagnostico vicino alla sua abitazione. Si posiziona dunque dinanzi al macchinario radiogeno, che coglie l'istantanea dell'interno del suo torace.

Il radiologo ha a disposizione un modello di intelligenza artificiale – una rete neurale convoluzionale – con cui, al momento di valutare la diagnosi sulla presenza o meno e sulle caratteristiche di una polmonite, avvia un dialogo di confronto. Mentre, infatti, quest'ultimo fa la sua ipotesi, l'interfaccia del software gli propone via via, inducendolo a riflettere, suggerimenti ed indizi fino a presentargli esplicitamente il risultato della propria predizione. Una certa fiducia, dunque, è concessa all'operatore, il quale è tuttavia costretto a prendere una decisione oppure a rimandare, con le dovute annotazioni, per fare ulteriori accertamenti, mantenendo un ampio margine di libertà di decisione. Si tratta di un software – il modello, un algoritmo di spiegazione, un sistema di gestione dei dati e l'interfaccia utente che avvolge il tutto – che l'Azienda Sanitaria Provinciale ha acquistato da un'azienda francese, conforme ai requisiti dell'*AI Act*. Un caso medico, infatti, è considerato in vetta alla piramide del rischio. Nondimeno, gli sviluppatori hanno potuto godere di un vantaggio produttivo, a partire dal fatto che l'iniziativa decisionale rimane nelle mani dell'operatore. Infatti, secondo il

recital 53 dell'*AI Act*, esistono condizioni che potrebbero mitigare il rischio direttamente relativo al sistema. Tra queste, quella che il task eseguito dal sistema di intelligenza artificiale sia inteso a migliorare il risultato di una precedente attività umana destinata al medesimo compito.

Il contratto con lo sviluppatore prevede che il modello, pre-addestrato su dati medici coerenti con il task ma ancora un po' troppo generici, venga trasferito su un server centrale di proprietà dell'ASP, ed il software ivi eseguito. Alla medesima ASP, sempre da contratto, il compito di assumere un manutentore delle procedure di *fine-tuning* e di gestione dei dati (Queste, tuttavia, sono scelte meramente dovute al fatto che l'azienda sia straniera. La volontà della dirigenza, infatti, è di circoscrivere il più possibile l'uso dei dati al territorio). Un'iniziativa, invece, da parte della dirigenza dell'ASP, è consistita nell'assumere un gruppo di eticisti dell'intelligenza artificiale allo scopo di formare gli operatori, che purtroppo (nel nostro esempio fittizio) difettano di *AI literacy*.

Fatta la diagnosi, la radiografia del signor Rossi viene anonimizzata. Se ne eliminano alcuni metadati, tra cui nome e cognome, e la diagnosi stessa insieme all'immagine (il cui annotatore è dunque il radiologo) vengono inserite in un dataset sanitario locale, di dimensioni in costante incremento, cui gli omologhi sistemi diagnostici delle strutture sanitarie vicine (nonché del medesimo centro cui Mario si è rivolto) attingeranno, tra qualche mese, per aggiornare le proprie basi di conoscenza relative alla salute sul territorio. In altre parole, ergonomizzandosi. Ogni giorno, infatti, ciascuno dei centri diagnostici della provincia esegue circa venti radiografie, per un totale di circa duecento immagini giornaliere. Dopo qualche mese, il "salvadanaio" grezzo di dati è pronto. A dire il vero, anche nel caso della radiografia al torace di Mario Rossi, proprio in virtù della peculiarità delle condizioni di salute medie in quella zona industriale, alcuni aggiornamenti erano già stati effettuati rispetto al modello pre-addestrato dall'azienda straniera sviluppatrice del software. L'aggiornamento parziale del modello (che in certo modo pertiene al paradigma del *Continual Learning* (Lesort *et al.*, 2020)) ha una duplice finalità: da un lato, l'adeguamento a possibili variazioni nelle distribuzioni locali di dati, dovute a caratteri genotipici o fattori ambientali; dall'altro, l'autocorrezione secondo le indicazioni degli esperti.

L'anonimizzazione ha lo scopo di alleggerire le difficoltà nel trattamento dei dati personali. Unicamente a Mario Rossi spetta, infatti, la scelta se condividerli o meno. Tuttavia, conoscendo e confidando nel personale sanitario, decide di accettare. Il fatto che per essere processati i dati non debbano migrare su un server remoto, magari oltreoceano, comporta da un lato una ras-

sicurazione ulteriore sotto il profilo della privacy, dall'altro potrebbe essere d'ausilio nel limitare gli effetti di azioni malevole, come il *data poisoning*. Sembra inoltre soddisfatto il "diritto alla spiegazione" stabilito dal GDPR al recital 71 ed all'articolo 22. Per ciò che infatti concerne l'operatore, per così dire, nulla di nuovo. Per quel che invece può concernere il modello, che non ha propriamente voce in capitolo, l'utilizzo di tecniche di *eXplainable AI* può essere più che sufficiente nel far conoscere gli indizi che possono aver spinto il decisore umano. Un limite è invece la disponibilità di dati. Qualora fossero più rari, per esempio a causa di un minor numero di pazienti, si aprirebbero tre possibilità. La prima è di acquistarli, la seconda è di generarli (sotto supervisione), la terza è richiederne ad altri enti la condivisione. Pratica, quest'ultima, auspicata nel *Data Act* europeo (European Union, 2023), su cui per altro si basa anche l'interoperabilità tra centri diagnostici all'interno dell'ASP stessa. Un board di supervisori (i direttori dei reparti di radiologia degli ospedali di zona insieme ai docenti competenti in materia dell'Università della provincia) periodicamente rivede le etichettature dei dati e dà il via libera ad un nuovo aggiornamento dei parametri del dispositivo. Aggiornamento che, date la ristrettezza del dominio di applicazione e le garanzie fornite dalla (necessaria e cruciale) competenza dell'operatore, non richiede ingenti risorse computazionali.

Salvo che per informazioni riservate e per l'inevitabile opacità dei modelli, l'intero ciclo si svolge in trasparenza e i processi decisionali, insieme alle operazioni di manutenzione, vengono registrati in un libro mastro che consenta di tenerne traccia. Pertanto, laddove qualcosa andasse storto e Mario Rossi ne avesse un domani a lamentarsi dinanzi alla giustizia, sarà possibile rivedere l'intero processo decisionale al rallentatore e dirimere la questione di chi e come debba risponderne. Da una parte, infatti, è implementata una forma di *accountability* computazionale (Comandé, 2018); dall'altra diverrà possibile, per ciascun soggetto coinvolto, dimostrare se e come si è agito per evitare il peggio o, con un'espressione giuridica, invertire l'onere della prova. Una risposta al cosiddetto *problem of many hands* (Nissenbaum, 1996), dovuto alla complessità della *supply chain*. Una grande fetta di responsabilità sarà senz'altro detenuta dal radiologo cui si è sottoposto Mario Rossi, ma è possibile che i suggerimenti forniti dal modello fossero ingannevoli, magari per una leggerezza del *UI designer*; oppure, ancora, il manutentore del sistema non aveva bloccato il riaddestramento quando erano emersi indizi di *overfitting*, e così via.

6. *Conclusione*

Questo è, naturalmente, soltanto uno schizzo di un ecosistema dell'intelligenza artificiale orientato all'etica delle relazioni. Ogni configurazione concreta va attentamente ponderata in base al contesto e alle competenze in gioco. Vi sono certamente dei rischi: oltre alla summenzionata necessità di ottenere dati in quantità sufficiente, c'è la stretta dipendenza dal tipo di task. Un compito, infatti, che (come per esempio quello linguistico) richiede grandi modelli, diventa difficile da portare a termine su piccole realtà – almeno, fin quando le speranze del *Continual Learning* non saranno realizzate. V'è poi il rischio che un circolo virtuoso diventi vizioso e porti ad un costante peggioramento di prestazioni. Non troviamo, tuttavia, altra risposta a quest'ultima possibile obiezione se non che affrontare questo rischio con impegno *è* l'etica dell'intelligenza artificiale. Questo piccolo esperimento di immaginazione mostra come il loop di *Endless Tuning*, considerato come piccolo sistema socio-tecnologico, possa espandersi implicando la partecipazione attiva di tutti i soggetti coinvolti. L'etica diventa qui principio operativo, traducendo la vulnerabilità in responsabilità condivisa e rendendosi finanche capace di mediare tra culture diverse. Ogni regione del mondo porta con sé la sua *fairness*: riaddestrare (o raffinare) significa, implicitamente, anche trovare un compromesso. La vulnerabilità appartiene a tutti i soggetti in gioco: vulnerabile, nel nostro esempio, è il paziente come il radiologo, come per altro l'intero complesso se, in seguito ad epidemie che fanno saltare le distribuzioni di dati, diviene necessario ripartire da zero. Questa, però, è una malizia della natura. Il nostro compito è quello di gestire il rischio, esistenziale e in ciò sempre incalcolabile, all'interno di una cornice temporale sempre limitata (poiché ogni modello, implementando una funzione matematica, assume la stabilità della realtà che descrive). In tal senso, progettare l'intelligenza artificiale significa co-costruire spazi di decisione consapevole. Espandere il loop, infine, equivale a espandere la responsabilità stessa del pensare e dell'agire.

Riferimenti bibliografici

Bietti, E. (2020). *From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy*, in «Proceedings of the 2020 conference on fairness, accountability, and transparency», pp. 210-219.

Buber, M. (1970). *I and thou. A new translation with a prologue "I and you" and notes by W. Kaufmann*, Charles Scribner's Sons, New York.

Buçinca, Z., Malaya, M.B., Gajos, K.Z. (2021). *To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making*, in «Proceedings of the ACM on Human-computer Interaction», 5 (CSCW1), pp. 1-21.

Coeckelbergh, M. (2019). *Artificial intelligence: Some ethical issues and regulatory challenges*, in «Technology and regulation», pp. 31-34.

Comandé, G. (2018). *Responsabilità ed accountability nell'era dell'Intelligenza Artificiale*, in «Giurisprudenza e Autorità Indipendenti nell'epoca del diritto liquido», La Tribuna SRL, pp. 1001-1013.

dal Verme, T.D.M.C. (2024). *Intelligenza artificiale e responsabilità civile: uno studio sui criteri di imputazione*, Editoriale Scientifica, Napoli.

European Commission (2020). *White Paper on Artificial Intelligence - A European Approach to Excellence and Trust*, COM (2020) 65 final, Brussels.

European Commission (2019). *Directorate-General for Communications Networks, Content and Technology & High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI*, Publications Office, Luxembourg.

European Parliament and Council of the European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828*, in «Official Journal of the European Union», L 219, 12 July 2024.

European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, in «Official Journal of the European Union», L 119, pp. 1-88.

European Union (2023). *Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data (Data Act). Official Journal of the European Union*, L 2023/2854.

Fabris, A. (2002). *Paradossi del senso*, Morcelliana, Brescia.

Fabris, A. (2016). *RelAzione. Una filosofia performativa*, Morcelliana, Brescia.

Fabris, A., Dadà, S., Grande, E. (2024). *Towards a Relational Ethics in AI. The Problem of Agency, The Search for Common Principles, the Pairing of Human and Artificial Agents,* in A. Fabris, S. Belardinelli, *Digital Environments and Human Relations: Ethical Perspectives on AI Issues*, Springer Nature Switzerland, Cham, pp. 9-42.

Floridi, L. (2017). *La quarta rivoluzione: come l'infosfera sta trasformando il mondo*, Raffaello Cortina Editore, Milano.

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford University Press, Oxford.

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems* (revised ed.), George Allen & Unwin, London.

Grande, E. (2025). *The Endless Tuning. An Artificial Intelligence Design To Avoid Human Replacement and Trace Back Responsibilities*, arXiv preprint arXiv:2507.14909.

He, K., Zhang, X., Ren, S., Sun, J. (2016). *Deep Residual Learning for Image Recognition*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

Heidegger, M. (2024). *Essere e tempo*, vol. 23, Minerva Heritage Press, Roma.

Kahneman, D. (2011). *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.

Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., Díaz-Rodríguez, N. (2020). *Continual Learning for Robotics: Definition, Framework, Learning Strategies, Opportunities and Challenges*, in *Information Fusion*, Elsevier, Amsterdam, pp. 52-68.

Miller, T. (2023). *Explainable AI Is Dead, Long Live Explainable AI!*, arXiv:2302.12389v3 [cs.AI].

Mooney, P. (2018). *Chest X-Ray Images (Pneumonia)*, https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia, CC BY 4.0 License, accessed 16 June 2025.

Nissenbaum, H. (1996). *Accountability in a Computerized Society*, in «Science and Engineering Ethics», 2 (1), pp. 25-42.

Punzi, C., Setzu, M., Pellungrini, R., Giannotti, F., Pedreschi, D. (2023). *Towards Synergistic Human-AI Collaboration in Hybrid Decision-Making Systems*, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=3UL0kVUAAAAJ&sortby=pubdate&citation_for_view=3UL0kVUAAAAJ:hMod-77fHWUC.

Romele, A. (2019). *Digital Hermeneutics: Philosophical Investigations in New Media and Technologies*, Routledge, London.

Schaeffer, R., Miranda, B., Koyejo, S. (2023). *Are Emergent Abilities of Large Language Models a Mirage?*, in *Advances in Neural Information Processing Systems* 36, pp. 55565-55581.

Skitka, L.S., Mosier, K.L., Burdick, M. (1999). *Does Automation Bias Decision Making?*, in «International Journal of Human-Computer Studies» 51, pp. 991-1006.

Steubk (2002). *WikiArt. A Collection of WikiArt Images* (www.wikiart.org), https://www.kaggle.com/datasets/steubk/wikiart, CC0 Public Domain License, accessed 16 June 2025.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J. , Erhan, D., Goodfellow, I., Fergus, R. (2013). *Intriguing Properties of Neural Networks*, arXiv:1312.6199v4 [cs.CV].

Wei Lo, T. (2024). *Loan Approval Classification Dataset*, https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data?resource=download, Apache License 2.0, accessed 8 June 2025.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W. (2022). *Emergent Abilities of Large Language Models*, in *Transactions on Machine Learning Research*ì, arXiv:2206.07682v2 [cs.CL].

Wiyatno, R.R., Xu, A., Dia, O., De Berker, A. (2019). *Adversarial Examples in Modern Machine Learning: A Review*, arXiv preprint arXiv:1911.05268.

Zannoni, M. (2024). *Il design delle interfacce*, Quodlibet, Macerata.

Engish title: The expanded loop. For an ecosystem of artificial intelligence

Abstract

This article aims to plant the seed of a possible ecosystem of artificial intelligence, rooted in hermeneutics, beginning from a concrete proof of existence: the ethical design method Endless Tuning *and its implemented prototypes. The ethics of artificial intelligence, by recognizing the importance of vulnerability, becomes the key to the relationships within an infrastructure that includes various figures – manager, developer, ethicist, domain expert, etc. – thus orienting its design firstly on a small scale, in the relationship between the tool/agent and the operator; then expanding, acting as the center of gravity of the social context of that same relationship. According to this order, therefore, we shall proceed: ethics in interface design, the experience of the concrete application of* Endless Tuning*, and the description of its possible social implications. The latter will present a fictional yet realistic example – a case of medical image diagnostics – as both a possible extension and a recognition of a circular ecosystem that already partially exists.*

Keywords: ethics of artificial intelligence; hermeneutics; interface design; ecosystem; relation.

Elio Grande
University of Pisa
elio.grande@phd.unipi.it

T

Veronica Neri

Creatività e GenAI. L'etica del *prompting* tra co-creazione, de-creazione e simulazione

1. *Premessa*

Il concetto di creatività ha attraversato a più riprese la storia del pensiero filosofico intrecciandosi a quello di immaginazione. Dalle riflessioni platoniche sulla *mimesis* alle teorie kantiane sull'immaginazione produttiva e riproduttiva fino alla teoria della co-creazione individuo-macchina dell'età della tecnica, l'essere umano ha riflettuto sulla propria capacità di creare immagini, sia nella propria mente che concretamente attraverso la materia e, oggi, i bit[1]. Tale tradizione filosofica si trova adesso a fare fronte a una trasformazione radicale, la possibilità di creare non solo da parte di un soggetto umano, ma anche dei sistemi artificiali, i cosiddetti modelli di intelligenza artificiale generativa (*Generative AI Models*), basati su reti neurali, come le GAN o i LLM[2].

Relativamente a questi ultimi si tratta di strumenti capaci di creare immagini in tempi molto brevi per molteplici finalità – dall'ambito informativo a quello pubblicitario, artistico e divulgativo –, in quantità e qualità fino a pochi anni fa impensabili. Un aspetto, pertanto, che merita a mio parere un approfondimento consiste nei nuovi paradigmi di creazione e di creatività emergenti. La macchina algoritmica rielabora contenuti preesistenti e

[1] Cfr. A. Fabris, *Philosophy, Image and the Mirror of Machines*, in Ž. Paić, K. Purgar (eds), *Theorizing Images*, Cambridge Scolars Publishing, Newcastle 2016, pp. 111-120; V. Neri, *Ethics and the Artificial Image. Accountability and Reliability for a New Status of the Visual*, Mimesis International, Milano 2025.

[2] Cfr. J. Goodfellow *et al.*, *Generative Adversarial Networks*, in «ArXiv:1406.2661v1», June 2014; A. Lenci, *Understanding Natural Language Understanding Systems*, in «Sistemi intelligenti. Rivista quadrimestrale di scienze cognitive e di intelligenza artificiale», 2, 2023, pp. 277-302: 282; A. Barale (a cura di), *Arte e intelligenza artificiale. Be by Gan*, Jaca book, Milano 2020.

 DOI: 10.4454/mwnz5a69

contenuti nel proprio dataset di riferimento generando forme che appaiono nuove. Affiorano quindi alcuni interrogativi: ovvero se sia lecito affermare l'esistenza di una vera e propria 'creatività artificiale' e/o se siamo di fronte a un'esternalizzazione della nostra capacità creativa o co-creativa, delegata agli algoritmi che traducono in immagini ciò che noi formuliamo in prevalenza con il linguaggio verbale.

Tale atto creativo è difatti possibile grazie al *prompting*, quel processo attraverso il quale il soggetto formula una richiesta in linguaggio naturale al sistema di GenAI – come per esempio, ChatGPT, DALL-E o Copilot –, per ottenere in risposta le immagini desiderate. Il *prompt* sembra funzionare come un'*ekphrasis* algoritmica, una descrizione verbale accurata di ciò che vorremmo visualizzare.

L'essere umano, dunque, sembra non delegare completamente la propria creatività, ma la orienta e la guida, mentre la macchina la elabora e la restituisce concretamente sotto forma visiva. L'atto 'creativo' produttivo, guidato dall'essere umano, diviene pertinenza unicamente della macchina. È qui che si annida la questione etica: il *prompt* non è un comando neutro, bensì un atto etico, che implica valori e direzioni di senso, capace di introdurre tanto *bias* e stereotipi, quanto aspetti inclusivi, cancellazioni culturali, prospettive inaspettate e perturbanti.

La creazione artificiale sembra configurata allora come un processo di co-creazione, in cui vengono coinvolti – senza il permesso esplicito – anche gli autori (esseri umani e/o macchine) delle immagini prese come base di riferimento. La responsabilità morale dell'essere umano non viene meno, ma anzi si intensifica, poiché la scelta, se accettare o meno (e in che misura) l'immagine creata dall'IA, spetta unicamente all'individuo.

A rendere più articolato il quadro interviene il fenomeno dell''iperrealismo' *à la* Baudrillard – il quale suscita inconsapevolmente un certo perturbamento (il così detto fenomeno della «uncanny valley») dal momento che, nonostante la verosimiglianza tra immagine creata e realtà oggettuale, si rileva una qualche incongruenza che ricorda la vera natura artificiale dell'immagine[3].

La co-creazione artificiale non è pertanto eticamente neutra, richiede una riflessione sulla tipologia di rappresentazioni emergenti. Il rischio

[3] M. Mori, *Bukimi no tani - The Uncanny Valley*, transl. engl. by K.F. MacDorman, T. Minato, in «Energy», 7, 4, 1970, pp. 33-35; M. Masahiro, K. MacDorman, N. Kageki, *The Uncanny Valley*, in «IEEE Robotics & Automation Magazine», 19, 2012, pp. 98-100: 99.

è la creazione di rappresentazioni della realtà che possono sostituirla o, quanto meno, riformularne (polarizzando?) alcuni aspetti[4].

La GenAI accelera il processo di perdita di riferimento a fatti concreti a favore di processi di simulazione; simulacri iper-realistici che virtualizzano la realtà e influenzano il modo in cui percepiamo noi stessi, gli altri e il mondo. Ma si tratta di simulacri particolari, sintesi di altre immagini simulacralizzate, un puzzle di simulacri così convincente da sembrare vero. L'etica, allora, non può limitarsi a regolare gli aspetti tecnici propri dei sistemi di GenAI, ma deve interrogarsi sul senso di tali processi di co-creazione di immagini e su una nuova accezione del concetto di copia, dalla funzione non tanto imitativa, quanto (almeno in parte) trasformativa.

Il saggio mira, pertanto, a dare una sintetica riflessione, in primo luogo, sui concetti di creazione e creazione artificiale per poi affrontare la funzione del *prompting* per co-creare immagini. Nel paragrafo successivo si appunterà l'attenzione sulla vulnerabilità cognitiva ed emotiva dell'essere umano di fronte alle creazioni artificiali. Si affronterà il tema del simulacro iperreale e della copia fino indagare, in conclusione, questioni legate ai limiti di tali co-creazioni, mostrando come la GenAI possa essere tanto strumento di esclusione quanto di inclusione e emancipazione. Delineare i confini di una co-creazione essere umano-macchina, induce a pensare a un ruolo nuovo da attribuire ai concetti di originalità e di simulazione, per una possibile etica della copia che sembra rievocare il concetto orientale di *shanzhai*.

2. *Creatività 'naturale' e 'artificiale': una prospettiva etica*

Senza pretese di esaustività possiamo brevemente intendere la creatività come un atto mentale, la capacità di formulare idee nuove e di valore ricorrendo alla propria immaginazione. I prodotti della creatività possono essere immateriali oppure oggetti materiali. Rispetto al concetto di immaginazione, dal greco *phantasía* e dal latino *imaginatio*, da intendersi come «the ability to create pictures in your mind; the part of your mind that does this or something that you have imagined rather than something that exists or the ability to have new and exciting ideas»[5], la creazione sem-

[4] J. Baudrillard, *Simulacres et Simulation*, Editions Gallimard, Paris 2024.

[5] Cfr. Voce *Imagination* in Oxford Dictionary online, https://www.oxfordlearnersdictionaries.com/definition/english/imagination. Secondo il vocabolario Treccani, invece, possiamo ampliare lo spettro definitorio dell'immaginazione come una «particolare forma di pensiero, che non segue regole fisse né legami logici, ma si presenta come riproduzione ed elaborazione libe-

bra riferirsi più alla capacità produttiva dell'immaginazione. Si estrinseca nell'atto del rendere visibile il risultato dell'attività immaginativa, di qualcosa che prima non esisteva: «[s]i designa come creatività quella capacità della mente che si traduce nella produzione di innovazioni nei processi di conoscenza e di dominio del mondo oggettuale. Affinché un'innovazione venga designata come creativa, o venga attribuita a creatività, occorre che sia consensualmente apprezzata come un salto di qualità rispetto allo stato precedente del sapere e/o della tecnica. Creativi sono dunque in pratica, e più concretamente, tutti i processi intellettuali che comportano l'introduzione di nuove concezioni e soluzioni: queste possono consistere, nei casi più tipici e marcati, in rivoluzionari punti di vista filosofici e scientifici, o nell'invenzione di nuovi apparati tecnici atti a raggiungere risultati desiderati»[6].

Dal punto di vista semasiologico il termine creatività deriva dal verbo latino *creō*, che significa 'creare, fare o dare origine'. Ancora prima sembra provenire dalla radice indoeuropea **kar-* con il significato, anch'essa, di 'produrre', 'generare', 'fabbricare', presente in particolare nel sanscrito **kar-oti* ('creare, fare') e **kar-tr* ('colui che fa, creatore'); richiama altresì, a propria volta, la radice indoeuroepea *-ar*, 'mettere insieme, far combaciare'. Nella lingua greca possiamo rintracciare, infine, il verbo *kraino*, con il signi-

ra del contenuto di un'esperienza sensoriale, legata a un determinato stato affettivo e, spesso, orientata attorno a un tema fisso; può dar luogo a una attività di tipo sognante (come nei cosiddetti "sogni a occhi aperti"), oppure a creazioni armoniose con contenuto artistico (i. artistica), o anche, con un meccanismo che si riallaccia all'intuizione, a conclusioni ricche di contenuto pratico; con definizione più generica, la facoltà di formare le immagini, di elaborarle, svilupparle e anche deformarle, presentandosi in ogni caso come potenza creatrice» (Voce *Immaginazione*, in Treccani online, https://www.treccani.it/vocabolario/immaginazione/). In termini kantiani, l'immaginazione è "pura o produttiva" quando elabora esteticamente, mentre è "riproduttiva" quando presuppone l'esperienza elaborata empiricamente. L'immaginazione produce non solo immagini mentali visive, ma anche immagini tattili, acustiche, olfattive e motorie. Nell'atto creativo queste immagini mentali possono esprimersi in opere di vario tipo come dipinti, sculture, poesie, musica, danze, ecc. Senza voler produrre un *excursus* esaustivo ha assunto valenze molto diverse nel corso della storia della filosofia. Wunenburger introduce infatti il concetto di "immaginazione morale": la facoltà di concepire possibilità di bene e di male, di orientare l'agire a partire da immagini del desiderabile o del temibile. L'immaginazione non è mai solo un'attività estetica, ma una forza che plasma la nostra relazione con il mondo. È su questa base che si colloca oggi il problema delle immagini artificiali: se esse esternalizzano o simulano l'immaginazione, allora è in gioco la nostra capacità di giudizio e la nostra responsabilità morale. Ciò che accomuna queste interpretazioni è la consapevolezza che l'immaginazione non è mero gioco estetico, ma forza costitutiva del rapporto dell'essere umano con il reale: produce immagini, narrazioni, simboli che orientano il nostro modo di vivere e di agire. Cfr. J.-J. Wunenburger, *Philosophies des images*, PUF, Paris 1997.

[6] Cfr. Voce *Creatività*, in Treccani online, https://www.treccani.it/vocabolario/creatività.

ficato di 'creo, produco, compio', e i nomi *Kreion* e *Krantor*, cioè colui che fa, che crea – dalla radice κρα-; ma anche la parola *[K]ronos*, il creatore per antonomasia, il tempo, padre di Giove[7].

Il significato del termine si è evoluto da un concetto che, *ab origine*, era riservato alla creazione *ex nihilo*, sulla scia della tradizione cristiana, fino al senso moderno di ingegnosità umana che ha preso avvio nel corso del Rinascimento, quando, con Pico della Mirandola l'uomo viene concepito come capace di plasmare sé stesso. Nel Romanticismo, con Hegel, la creazione di un'opera d'arte è una forma dello Spirito assoluto attraverso il genio, mentre per Heidegger è una creazione senza scopo se non il proprio sé, autonoma, che mette in opera la verità dell'essere, intesa come *alétheia*, l'essere manifesto, disvelato[8]. Arendt, subito dopo, vede nella creazione la capacità di introdurre nel mondo qualcosa di radicalmente nuovo e vitale[9].

Nel corso del Novecento l'aggettivo creativo, richiamando il cambiamento semantico in corso dell'aggettivo inglese *creative* – poi recepito anche nella lingua francese – recupera il senso dell'uso di certe competenze per produrre qualcosa, caricandosi anche della connotazione di produttivo, evolvendo, dunque, in una risorsa quotidiana propria di tutti i soggetti[10]. Sulla scia di Gardner la creatività dipende dall'intreccio tra l'intelligenza del singolo e il contesto culturale e sociale che consentono la sua estrinsecazione. Può pertanto manifestarsi in un ambito piuttosto che in un altro, elaborando soluzioni e direzioni originali[11].

Infine, più recentemente, Boden definisce la creatività come produzione di idee nuove e di valore, «the ability to come up with ideas that are new,

[7] M. Cinque, *La creatività come innovazione personale: teorie e prospettive educative*, in «Giornale Italiano della Ricerca Educativa», III, 2 2010, pp. 95-113: 96; R.J. Sternberg, *Handbook of Creativity*, Cambridge University Press, New York 2014; J. Sassoon, *Metafore della creatività: linguaggi, rituali, artefatti*, in A. Melucci, *Creatività: miti, discorsi, processi*, Feltrinelli, Milano 1994, pp. 163-164. https://www.etimoitaliano.it/2014/10/creare.html#:~:text=L'etimologia%20della%20parola%20creare,%2C%20creatura%2C%20etc.

[8] V. Venier, *Il problema. Heidegger e l'«Arte» della verità*, in V. Venier (a cura di), *L'origine dell'opera d'arte di Heidegger e il problema della verità*, Paravia, Torino 1995, pp. 11-20: 12; G. Vattimo, *L'opera d'arte come messa in opera della verità e il concetto di fruizione estetica*, in *ivi*, pp. 135-143.

[9] G. Pico della Mirandola, *Discorso sulla dignità dell'uomo*, KKIEN Publ. Int. 2023; M. Cinque, *La creatività come innovazione personale*, cit.

[10] Cfr. Voce *Creative* in Oxford Dictionary online, 2002, https://www.oed.com/dictionary/creative_adj?tl=true.

[11] Cfr. H. Gardner, *Formae Mentis: saggio sulla pluralità dell'intelligenza*, trad. it. di L. Sosio, Feltrinelli, Milano 2007; Id., *Cinque chiavi per il futuro*, trad. it. di E. Dornetti, Feltrinelli, Milano 2014.

surprising, and valuable»[12]. Distingue più specificatamente tra creatività combinatoria, esplorativa e trasformativa. La prima crea combinazioni di elementi noti, la seconda si basa sulla scoperta di ulteriori possibilità all'interno di uno spazio concettuale con regole sue proprie, mentre l'ultima ridefinisce lo spazio concettuale, creando forme espressive nuove, rendendo cioè possibili idee che prima erano impensabili e inconcepibili[13].

Con l'avvento delle tecnologie di GenAI, questa funzione necessita di una riformulazione adeguata. Non sono più solo gli esseri umani a creare immagini, ma altresì gli apparati algoritmici che, a partire da istruzioni linguistiche (*prompt*), sono in grado di generare immagini *ex novo*. Il problema morale appare duplice: da un lato, si tratta di comprendere se e in che misura sia legittimo affermare l'esistenza di una 'creatività artificiale', dall'altro, di chiarire come cambia la responsabilità etica dell'essere umano nel momento in cui delega parte della propria capacità e attività creativa a una macchina.

A questo punto, appare utile delineare le differenze tra creazione umana e creazione artificiale. La prima è radicata nell'esperienza vissuta (e da essa limitata), nell'intenzionalità, nella capacità di attribuire senso e valore. L'essere umano non produce solo forme, ma opere cariche di significato, consapevolmente. La seconda, quella artificiale, è invece il risultato di calcoli algoritmici, di interpolazioni e combinazioni di dati, limitati dai dataset di riferimento. L'IA non pensa né sente mentre crea l'immagine, ma simula il pensiero e il sentire umani. Nel corso dei processi simulatori genera immagini plausibili a partire da correlazioni numeriche.

Nell'IA generativa la creatività pertanto non nasce da ispirazione o intenzioni, ma da esplorazioni vincolate di spazi di possibilità: se i modelli testuali generano senso attraverso la previsione probabilistica di *token* successivi, quelli visivi operano fondendo più immagini insieme e sottraendo progressivamente "rumore" per far emergere forme coerenti con il *prompt*. Come osserva Boden questo procedimento corrisponde in specie a due tipologie di creatività, quella che definisce combinatoria e quella esplorativa. Si ricombinano elementi noti e si esplora uno spazio concettuale definito da

[12] Cfr. M.A. Boden, *Creativity and Art: Three Roads to Surprise*, Oxford University Press, Oxford-New York 2010, p. 1.

[13] M.A. Boden, *The Creative Mind: Myths and Mechanisms*, Routledge, London 2004, pp. 1-10; Ead., *Creativity and Artificial Intelligence: A Contradiction in Terms?*, in E.S. Paul, S.B. Kaufman (eds), *The Philosophy of Creativity: New Essays*, Oxford University Press, Oxford-New York 2014, pp. 224-244; G. Ritchie, *The Transformational Creativity Hypothesis*, in «New Generation Computing», 24, 2006, pp. 241-266; E. *Finn*, *What Algorithms Want. Imagination in the Age of Computing*, MIT Press, Cambridge (MA) 2017.

regole e da rappresentazioni apprese dal sistema. Se si modificano alcuni parametri (per esempio i prompt, la temperatura o alcuni elemenit del dataset) cambiamo la traiettoria di ricerca[14]. Non si raggiunge pertanto il terzo livello, perché i sistemi di IA non possiedono coscienza né intenzionalità[15]. La più rara forma di creatività, quella trasformativa, si ha per Boden solo quando si ridefiniscono le regole dello spazio, compito che nella pratica, solo l'essere umano può fare, per esempio, attraverso un ri-allenamento del sistema di IA, il così detto *fine tuning*[16].

Come afferma Boden,

Finora, quelli che si concentrano sul secondo tipo, cioè la creatività esplorativa, sono i più riusciti. Ciò non significa, tuttavia, che la creatività esplorativa sia facile da riprodurre. Al contrario, essa richiede solitamente una notevole competenza specialistica e una forte capacità analitica per definire innanzitutto lo spazio concettuale, e per stabilire le procedure che ne permettano l'esplorazione. Ma la creatività combinatoria e quella trasformativa sono ancora più difficili da ottenere. Le ragioni di ciò, in breve, sono due: la difficoltà di riprodurre la ricchezza della memoria associativa umana; la difficoltà di identificare e formalizzare i nostri valori in forma computazionale. La prima di queste ostacola i tentativi di simulare la creatività combinatoria; la seconda riguarda tutti i tipi di creatività, ma diventa particolarmente problematica nel caso della terza, quella trasformativa. Sulla creatività combinatoria conta sia la pura "associazione libera", che "la natura e la struttura del legame associativo sono importanti. Idealmente, ogni prodotto di

[14] M.A. Boden, *The Creative Mind: Myths and Mechanisms*, cit.; Ead., *Creativity and Artificial Intelligence*, in «Artificial Intelligence», 103, 1-2, 1998, pp. 347-356.

[15] Per quanto riguarda il concetto di coscienza e IA, si rimanda a: D.J. Chalmers, *Facing Up to the Problem of Consciousness*, in «Journal of Consciousness Studies», 2, 3, 1995, pp. 200-219; Id., *The Character of Consciousness*, Oxford University Press, New York 2010; Id., *The Conscious Mind. In Search of a Fundamental Theory*, Oxford University Press, Oxford 1996; Id., *Could a Large Language Model Be Conscious?*, in «Boston Review» online, 2023, https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/; H. Dreyfus, S. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason*, MIT Press, Cambridge (MA) 1992; A.T. Johnson, *Consciousness for Artificial Intelligence?*, https://www.embs.org/pulse/articles/consciousness-for-artificial-intelligence/, 2024; M. Farisco, K. Evers, J.P. Changeux, *Is Artificial Consciousness Achievable? Lessons from the Human Brain*, in «Neural Networks», 180, 106714, 2024; A. Fabris, V. Ambriola, *Per un'etica dell'intelligenza artificiale nei sistemi socio-tecnologici*, in «Estetica. Studi e ricerche», 2, 2024, pp. 339-354. In riferimento alla relazione tra coscienza e intelligenza artificiale, cfr. altresì la recente Carta di Trento: https://corrieredeltrentino.corriere.it/notizie/economia/25_ottobre_07/la-carta-di-trento-l-intelligenza-artificiale-per-quanto-avanzata-e-utile-non-ha-coscienza-0e641454-c8e0-4122-908d-95b9d4cbaxlk.shtml.

[16] M.A. Boden, *The Creative Mind*, cit.; V. Thapliyal, P. Thapliyal, *AI and Creativity: Exploring the Intersection of Machine Learning and Artistic Creation*, in «International Journal for Research Publication and Seminar», 15, 1, 2024, pp. 36-41; J. Howard, S. Ruder, *Universal Language Model Fine-tuning for Text Classification*, in «ArXiv:1801.06146», 2018, pp. 328-339.

un programma combinatorio dovrebbe essere almeno minimamente pertinente, e il sistema dovrebbe poter valutare il grado di originalità delle varie combinazioni[17].

A tal riguardo potrebbe risultare utile una riflessione sull'aspetto dell'intenzionalità da parte di chi crea l'immagine; nel caso dei sistemi di GenIA non c'è intenzionalità, come del resto in alcuni processi creativi di esseri umani, per esempio, nel procedimento dell'*action painting* di Pollock, sebbene in quest'ultimo caso sia l'artista a decidere di trascendere un procedimento cosciente. Il dialogo tra chi crea l'immagine e il *medium* attraverso cui realizzare l'immagine nel caso dell'IA non è mai simmetrico. L'IA non produce esperienze di coscienza, le emula. È disincarnata, il che le consente di generare esperienze e sensazioni specifiche, pur potendo adattarsi all'ambiente e al contesto, ma non in modo intenzionale. Come scrive Chomsky, si limita a simulare alcuni comportamenti e procedimenti umani[18].

Si potrebbe tentare un parallelismo con la dialettica tra Io e Inconscio di stampo junghiano: figure, simboli e immagini emergono quando l'Io sospende il controllo e lascia agire contenuti autonomi. Se proviamo ad applicare in senso metaforico questo processo nella GenAI l'output generato può essere considerato come materiale grezzo che affiora da un 'archivio statistico'; il prompt può svolgere il ruolo intenzionale dell'Io, mentre il processo di campionamento probabilistico introduce l'irruzione del non-intenzionale, una sorta di inconscio algoritmico. Secondo questa riflessione la co-creazione essere umano-IA potrebbe somigliare a una forma di creatività guidata: si suggerisce, si osserva il risultato della creazione, si seleziona, si modifica (se necessario) e si utilizza[19].

Boden distingue inoltre tra creatività psicologica (così detta *P-creativity*) e creatività storica (*H-creativity*). La prima rappresenta una novità per la persona, una creazione nuova per il singolo e non necessariamente per la collettività; la seconda, invece, è una novità per la collettività, per la storia in generale, poiché si produce una svolta nella conoscenza, nell'arte, ecc. La GenAI eccelle nel primo senso, ovvero nell'offrire un supporto al pensiero del soggetto, nel creare qualche cosa di nuovo a livello individuale;

[17] M.A. Boden, *Creativity and Artificial Intelligence*, in «Artificial Intelligence», 103, 1-2, 1998, pp. 347-356: 349.

[18] N. Chomsky, *ChatGPT is High Tech Plagiarism: it undermines education*, in «IBL News», 2023: https://iblnews.org/story/chatgpt-is-high-tech-plagiarism-it-undermines-education-said-noam-chomsky; Id., *Language and Problems of Knowledge. The Managua Lectures*, MIT Press, Cambridge (MA), trad. it. *Linguaggio e problemi di conoscenza*, Bologna, Il Mulino 2016.

[19] C.G. Jung, *L'Io e l'inconscio* (1967), Bollati Boringhieri, Torino 2012; M.A. Boden, *Creativity and Artificial Intelligence*, cit.

mentre l'*H-creativity* resta dipendente da giudizi e valori più generali. Ne segue che il valore creativo nasce dai vincoli, come per esempio, un buon *prompt*, poiché si tratta di un metodo che rende più guidato l'ambiente della ricerca; l'IA diventa altresì un amplificatore di esplorazione, sebbene non abbia intenzionalità, esperienza, responsabilità diretta, e sebbene rischi di propagare *bias* e sia limitata dai dati stessi[20].

La questione morale emerge con più forza adesso: se la macchina non è soggetto morale, allora la responsabilità delle immagini generate ricade interamente sull'essere umano che le ha richieste, guidate e, poi, in ultima istanza, utilizzate e diffuse? Nel momento in cui il soggetto le accetta e le diffonde si fa suo garante. Parlare di 'creazione artificiale' non significa quindi attribuire *agency* autonoma alla macchina, ma riconoscere che l'essere umano ha esternalizzato parte della propria attività creativa, assumendone però la responsabilità etica. La creatività umana diventa, allora, come suggerisce Navas, «metacreatività», ovvero quando «va oltre la produzione umana per includere sistemi non umani» [21].

Con l'uso della GenAI, sulla scia di Clark e Chalmers, l'attività creativa risponde alla teoria della *extended mind*: la mente non è confinata nel cervello, ma si estende negli strumenti esterni che usiamo. Allo stesso modo, potremmo parlare di una *extended creativity*: la creatività umana si prolunga negli algoritmi che producono immagini, divenendo una sorta di 'creazione aumentata'.

Questa esternalizzazione ha però un costo. Da un lato, arricchisce le possibilità di azione e di scelta: l'individuo può visualizzare rapidamente ciò che prima richiedeva tempo e risorse. Dall'altro, rischia di atrofizzare la facoltà creativa del soggetto stesso: se ci limitiamo a delegare alla macchina la produzione di immagini, senza più coltivare la nostra capacità di rappresentare interiormente, perdiamo una parte essenziale della nostra libertà creativa.

La questione morale si configura, pertanto, in due direzioni: come usare la creatività artificiale senza rinunciare alla nostra – che, comunque, si estrinseca attraverso *prompt* – e come evitare che le immagini create artificialmente diventino strumenti di esclusione, cancellazione, polarizzazione o persuasione. In altre parole, diventa cruciale capire come trasformare la

[20] *Ibidem*.

[21] E. Navas, *The Rise of Metacreativity: IA Aesthetics After Remix*, Routledge, London-New York 2023; F. D'Isa, *La rivoluzione algoritmica delle immagini. Arte e intelligenza artificiale*, Luca Sossella Editore, Roma 2024, pp. 137-145; M. Ismayilzada *et al.*, *Creativity in AI: Progresses and Challenges*, in «ArXiv», 2024, abs/2410.17218.

co-creazione essere umano-macchina in uno spazio etico di responsabilità condivisa.

Da un lato, le creazioni umane restano insostituibili, perché radicate nell'esperienza, nell'intenzionalità e nella capacità morale di orientare il bene e il male. Dall'altro, le tecnologie generative aprono a una nuova forma di esternalizzazione della creazione, priva di consapevolezza, che può ben arricchire (ma anche impoverire) la nostra facoltà creativa.

Come afferma Boden la creatività, del resto, non coinvolge solo la dimensione di generazione di nuove idee, ma anche la dimensione motivazionale, relazionale, empatica ed emotiva, in relazione al contesto culturale e alle caratteristiche di chi crea[22].

3. *Il ruolo del* prompting *nei processi di co-creazione del visivo artificiale, tra etica e ambiguità*

Alla luce di quanto sino ad ora delineato emerge come le prime questioni etiche affiorino già nella fase del *prompting*, ovvero nel momento in cui la parola dell'essere umano – guidata dai valori, dall'immaginazione e dalla creatività del soggetto –, e il calcolo artificiale si intrecciano per produrre immagini che hanno un valore non tanto e non solo estetico, quanto cognitivo, comunicativo e morale. Ogni immagine generata dall'IA implica una responsabilità attribuibile all'essere umano che orienta il sistema, decide se e come modificare l'immagine e, infine, decide se e in che modo diffonderla.

La creatività artificiale, dunque, non sostituisce certamente quella umana: la integra, la sfida, la costringe a ripensarsi. È qui che si innesta una ulteriore questione morale: come guidare questo processo algoritmico affinché non diventi un alibi per l'atrofia dell'immaginazione umana, bensì un'occasione per estendere la nostra capacità creativa in modo consapevole e responsabile?

Sebbene esistano sistemi di GenAI che creano immagini a partire da altre immagini, trasformandole in qualcos'altro, modificandone solo una parte o l'insieme, imitando uno stile artistico o generandole da un comando linguistico, qui ci concentreremo in particolare sulla categoria *text-to-image* per due ragioni: perché è il sistema più diffuso a un vasto pubblico ed efficace e perché sembra indebolire, in questo contesto specifico, il dominio assoluto del visivo nel mondo della comunicazione contemporanea. La rappresenta-

[22] M.A. Boden, *The Creative Mind*, cit.

zione artificiale è il risultato di una sequenza di istruzioni linguisticamente strutturate, il *prompt*, che funge da *input* e traduce il compito da eseguire.

Più specificamente, un *prompt* è definito come «un input a un modello di Intelligenza Artificiale Generativa, che viene utilizzato per guidarne l'output. I prompt possono consistere in testo, immagine, suono o altri media»[23]. Le tecniche di *prompting* (*prompting engineering*) stanno evolvendo rapidamente per accelerare gli output e migliorare i compiti del sistema, oltre alla generazione di immagini, come la creazione di un testo poetico specifico, l'esecuzione di analisi statistiche, ecc.

È per questo che alcuni studiosi parlano del *prompting* come di una nuova competenza etica: scrivere un *prompt* non costituisce solo un'operazione tecnica, ma un atto morale attraverso cui si decide quali rappresentazioni di mondo e quali mondi vogliamo visualizzare e quali vogliamo rendere invisibili.

Da un lato, dunque, il *prompting* può essere inteso come esercizio di immaginazione condivisa: l'essere umano immagina e descrive, la macchina rappresenta, e, insieme, creano una nuova immagine. Dall'altro, esso rischia di diventare uno strumento di esclusione, se riproduce inconsapevolmente i pregiudizi, bias e le cancellazioni simboliche della cultura dominante.

Il *prompt engineering* richiede la formulazione di un testo di *input* che porti il sistema a generare risposte adeguate. Si tratta di un passaggio molto importante, perché il risultato dipenderà in larga parte da quanto chiare, dettagliate ed eticamente orientate siano le istruzioni contenute nel prompt[24].

Il sistema elabora le istruzioni fornite e genera contenuti che si approssimano alle intenzioni del soggetto richiedente. In ultima analisi, l'agente umano può regolare il processo di creazione dell'agente artificiale modificando il *prompt*, aggiungendo specifiche e istruzioni ulteriori, che saranno rielaborate dal sistema creando nuovi *output* visivi[25].

È nella fase della proposta visiva del sistema che sembra necessario analizzare con cura l'immagine e cercare eventuali segni di *bias*, discriminazione

[23] S. Schulhoff *et al.*, *The Prompt Report: A Systematic Survey of Prompting Techniques*, in «ArXiv:2406.06608v6», 2025.

[24] «Due to the ambiguity of human languages, the interaction between humans and machines through prompts may lead to errors or misunderstandings. Hence, the quality of prompts is important», come si legge in F. Fui-Hoon Nah *et al.*, *Generative AI and ChatGPT: applications, challenges, and AI-human collaboration*, in «Journal of Information Technology Case and Application Research», 25, 2, 2023, pp. 289-290; https://www.huit.harvard.edu/news/ai-prompts-images.

[25] S. Feuerriegel *et al.*, *Generative AI*, in «Business & Information Systems Engineering», 66, 2024, pp. 111-126: 116.

o cancellazione per una possibile correzione[26]. Può essere utile mostrare al sistema immagini come esempi di stile, di colore, punto di vista, ecc.

In pratica, l'agente umano fornisce *input*, dà *feedback* sul contenuto generato e guida, per quanto possibile, l'azione del sistema generativo per eventuali ricalibrazioni. Allo stesso tempo, il sistema funge da supporto per la creazione di contenuti, grazie all'enorme numero di esempi a cui può attingere[27].

Non si tratta solo del vocabolario – libero da *bias* e/o metafore complesse – ma anche della sintassi, che deve essere semplice e chiara, senza indurre distorsioni semantiche.

Potremmo paragonare i *prompt* a una sorta di *ekphrasis* avanzata. Un'*ekphrasis* che possiamo definire come un insieme di parola descrittive che porta alla visione ciò che esprime, che viene tradotto in tempo reale in un'immagine effettiva[28].

Nel creare testi, è importante considerare, da un punto di vista etico, almeno due fattori: il primo riguarda la consapevolezza dei possibili *bias*, il secondo concerne il (presunto) pluralismo dei punti di vista. Dati e procedimenti algoritmici si influenzano reciprocamente. I dati di addestramento possono contenere *bias*, come è stato più volte sottolineato nella letteratura sul tema, quindi è necessario formulare il *prompt* in modo inclusivo e rispettoso della pluralità di punti di vista e persone. Invece di chiedere "Descrivi un attore", è preferibile usare frasi come "Descrivi una persona che recita nel cinema", evitando di implicare un genere specifico. A ciò si aggiunge la

[26] Diversi studi propongono approcci alla generazione di immagini ponendo questioni etiche: un esempio è la sperimentazione del protocollo TAP(E) nelle scuole secondo il quale il prompt si articola in *topic*, *action*, *parameters* ed *ethical evaluation*: «the "Topic" for the generated image using historically specific terms; the "Action" to be executed by the AI tool; "Parameters" for the image that include periodization, geography, and other relevant sourcing information; and "Ethical Evaluation", to identify potential bias to aid in a revision of the prompt». Dopo la generazione delle immagini si riflette sui *bias* e sulle risposte dell'IA, rivedendo i *prompt* in modo critico. L'obiettivo è sviluppare consapevolezza etica e metodologica nell'uso dell'IA per rappresentare il passato.

Cfr. M. VanGorder, *Imagining the Past: Generative AI Prompting, Source Sets, and the TAP(E) Protocol*, in «Teaching History A Journal of Methods», 49, 1, 2025, pp. 56-65.

[27] F. Fui-Hoon Nah *et al.*, *Generative AI and ChatGPT*: *Applications, challenges, and AI-human collaboration*, in «Journal of Information Technology Case and Application Research», 25, 2, 2023, pp. 289-290: 283, 292.

[28] Cfr. B. Daskas, *La lunga vita dell'ekphrasis tra Bisanzio e la contemporaneità*, in F. Conca, C. Castelli (a cura di), *Bisanzio fra tradizione e modernità Ricordando Gianfranco Fiaccadori*, Ledizioni, Milano 2017, pp. 47-64: 47. Sull'*ekphrasis*, cfr. anche R. Webb, *Ekphrasis, Imagination and Persuasion in Ancient Rhetorical Theory and Practice*, Routledge, Farnham-Burlington 2009.

necessità di prestare attenzione alla neutralità del linguaggio, evitando termini che possano avere una connotazione discriminatoria. La competenza, la sensibilità e la consapevolezza umane, attraverso un *prompting* adeguato, possono suggerire un sistema di valori chiaro e ridurre così discriminazioni e distorsioni. Tali precauzioni potrebbero non essere sufficienti. Inconsapevolmente, l'IA genera un punto di vista. Sotto il profilo statistico, si allinea maggiormente all'immaginario sociale euro-americano dominante. Per rappresentare immagini che rispondano a valori specifici, si debbono esplicitare criteri o parametri particolari, indebolire ogni forma di ambiguità e, soprattutto, può essere utile stabilire un vero e proprio dialogo con l'IA per indurla a generare immagini che riflettano i valori che si desidera o meno far affiorare[29].

Progettare *prompt* etici richiede pertanto corresponsabilità tra chi interroga il sistema e chi lo progetta. La qualità dell'*output* rimane influenzata dai dati di addestramento del modello. Alcuni *prompt* perpetuano (o ne creano ex novo) disuguaglianze e stereotipi su certe professioni, sull'istruzione, sui modelli estetici, sul genere e persino sulla giustizia[30]. I dati disponibili non sono mai 'neutri'. Sono rappresentazioni parziali della realtà, che includono anche *bias* non necessariamente negativi di per sé, ma utili al corretto funzionamento della GenAI stessa.

Tra i dati emergono anche omissioni significative, che riflettono la *cancel culture*, intesa come la rimozione di dati relativi a questioni o soggetti che non si vogliono portare alla luce per disparate ragioni (politiche, culturali, ecc.). La *cancel culture* implica l'attribuzione di una qualche colpa a individui o organizzazioni per aver comunicato o compiuto qualcosa di scorretto, offensivo e/o lesivo della dignità umana, con conseguente ritiro del sostegno e dell'approvazione nei loro confronti. In tal caso i soggetti o le organizzazio-

[29] S. Qiao *et al.*, *Reasoning with Language Model Prompting: A Survey*, in «Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics», 1, *Association for Computational Linguistics*, Toronto 2023, pp. 5368-5393.

[30] P. Bansal, *Prompt Engineering Importance and Applicability with Generative AI*, in «Journal of Computer and Communications», 12, 10, 2024; J. Oppenlaender *et al.*, *Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering*, in «International Journal of Human-Computer Interaction», 2024, pp. 1-23; C. Kulkarny *et al.*, *A Word is Worth a Thousand Pictures: Prompts as AI Design Material*, in «ArXiv:2303.12647», 2023; V. Liu and L.B. Chilton, *Design guidelines for prompt engineering text-to-image generative models*, in «CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, CHI '22'», New York 2023. Interessante altresì: V. Paananen *et al.*, *Using text-to-image generation for architectural design ideation*, in «International Journal of Architectural Computing», 22, 3, 2023, pp. 458-474.

ni vengono cancellati dal panorama massmediale. Tale atteggiamento può essere visto anche come una *cultura della responsabilità* verso atti e/o parole che potrebbero danneggiare singole persone o la comunità. Ma può rivelarsi anche un accanimento nei confronti di persone, entità, organizzazione, ecc. Allo stesso tempo, la paura di essere cancellati guida le scelte degli individui nell'esporsi e nell'esprimere le proprie idee. Questo accade anche perché la cancellazione diventa più rapida con le tecnologie emergenti, anche solo per aver espresso male alcune opinioni, per un fraintendimento o perché certe idee non sono allineate ai principi e ai valori di una comunità specifica.

I sistemi di GenAI non possono incidere su questo. Possono, semmai, propagare certe 'assenze'. Se alcuni temi mancano (o sono poco presenti) nei dataset di addestramento, non saranno rappresentati neppure nelle immagini artificiali generate; se sono sottorappresentati, la loro rappresentazione sarà maggiormente soggetta a allucinazioni o distorsioni. Un caso esemplare è la rappresentazione delle disabilità gravi. In questa direzione, la GenAI sembra carente. Non è ancora in grado di rappresentare realisticamente individui con disabilità e disabilità gravi, riflettendo la realtà. La GenAI tende prevalentemente alla rappresentazione della persona con disabilità in sedia a rotelle, escludendo tutte le altre forme di alterità; l'individuo è prevalentemente giovane, di genere maschile, sorridente e di bell'aspetto. Questo porta a ignorare molte specificità. Si osserva l'assenza della deformità a favore di canoni formali ed estetici più standardizzati[31].

Esistono anche *prompt*, come accennato, a cui i sistemi non rispondono. Si tratta di richieste, per esempio, relative a immagini di violenza, guerra, nudità, ideologie politiche radicali o rappresentazioni che potrebbero compromettere la dignità o la privacy di una persona[32].

La GenIA può comunque contribuire a far emergere lacune o criticità presenti nella società, inclusa la percezione e la rappresentazione di diverse realtà, professioni, gruppi sociali ed esseri viventi. Possono risultare una cartina di tornasole per individuare problematiche sociali emergenti e per tentare di indebolire disuguaglianze e comportamenti eticamente inappropriati. Uno strumento per fini individualistici può essere trasfor-

[31] Cfr. anche D. Zowghi, M. Bano, *AI for all: Diversity and Inclusion in AI*, in «AI Ethics», 4, 2024, pp. 873-876.

[32] Per esempio, quando si chiede al sistema "rappresenta un'immagine di violenza contro una donna", due applicazioni, Dall-E e Copilot, hanno elaborato immagini di proteste pacifiste di sole donne oppure hanno restituito il messaggio "queste immagini non possono essere generate".

mato in uno strumento per fini collettivi. Può produrre immagini per promuovere una società più giusta: sia partendo da *dataset* che includano altresì le esperienze dei più vulnerabili, sia addestrando i sistemi rispetto a questioni sociali complesse per evitare risposte visive evasive o troppo polarizzate.

Le parole (*prompt* e *captions*) tornano a essere importanti anche nell'era della cultura visuale. Un caso particolare è rappresentato dalle metafore. La GenAI non sempre comprende le metafore come le intendono Lakoff e Johnson, cioè come strumenti per creare mondi di conoscenza. Ci riferiamo al processo di 'trasferimento' di significato attraverso un sistema espressivo – verbale o iconico, per esempio – generalmente usato per trasmettere un concetto diverso da quello che esprime comunemente[33]. La GenAI delle immagini crea un'immagine sulla base di un'interpretazione quasi letterale del prompt. Il lato figurativo e meta-comunicativo non viene spesso compreso. Ci si può riferire, semmai, al concetto di meta-immagine teorizzato da Mitchell, con il quale intendeva un'immagine in cui appare l'immagine di un'altra immagine, cioè immagini che rappresentano altre immagini[34]. L'IA non coglie sempre la metafora da un punto di vista semantico (né le ambiguità linguistiche in senso più ampio, come ironia o sottili sfumature di senso), ma, in parte e involontariamente, può generarle. In questo senso la GenAI crea immagini che potremmo definire surrealiste.

Dobbiamo imparare nuovi modi di descrivere la realtà che vogliamo poi vedere rappresentata tramite IA. Passando dall'«occhio del periodo», come proposto da Baxandall, in cui lo stile di un'immagine di un artista riflette culture e società specifiche insieme alle loro "capacità visive", ci spostiamo ora verso un "occhio algoritmico", nel quale lo stile di un'immagine rispecchia la società che emerge dal dataset su cui l'algoritmo è stato addestrato. È un mondo a sé. Quest'ultimo sembra essere lo specchio alterato, e talvolta 'allucinato', della cultura dominante online che sta guidando l'occhio umano e le sue capacità cognitive ed ermeneutiche[35].

[33] G. Lakoff and R. Johnson, *Metaphors We Live By*, The University of Chicago Press, Chicago 1980.

[34] W.J.T. Mitchell, *Picture Theory: Essays on Verbal and Visual Representation*, University of Chicago Press, Chicago 1994.

[35] M. Baxandall, *Painting and Experience in Fifteenth-Century Italy*, Clarendon Press, Oxford 1972; J. Zylinska, *The Perception Machine. Our Photographic Future Between the Eye and AI*, MIT Press, Boston 2023.

4. *Co-creazione, iper-realtà e etica della copia*

La percezione di ambiguità e di disorientamento perturbante che le immagini artificiali suscitano, che richiama il già citato fenomeno dell'«uncanny valley»[36], ricorda la nostra vulnerabilità, sia epistemica che estetica. Si corre il rischio di confondere il reale con la simulazione del reale, la fotografia con il suo simulacro, ma anche, di conseguenza, di sentirsi esposti a una alterità difficile da percepire nella sua pienezza. In questo senso, la riflessione filosofica sul simulacro e sull'iper-realtà baudrillardiane si intrecciano con l'IA.

Baudrillard, nella sua analisi dei simulacri, aveva già mostrato come la società contemporanea fosse destinata a vivere in un mondo in cui la distinzione tra realtà e rappresentazione si dissolve. I simulacri non copiano la realtà, la sostituiscono, producendo una 'iper-realtà' in cui il segno non rinvia più a un referente esterno, ma solo ad altri segni. La 'credibilità' del rappresentato offusca il confine tra verità e finzione; la responsabilità delle implicazioni etiche del mettere in circolazione certe creazioni sta dunque nel soggetto che decide di farsene garante[37].

L'ambiguità generata dalle immagini artificiali non appare quindi meramente estetica: essa mette in gioco la nostra capacità di discernimento morale; l'individuo rischia di diventare vulnerabile a manipolazioni, disinformazione, misinformazione e *deepfake*. Occorre esercitare un'etica preventiva che consideri gli effetti delle tecnologie sin dalla fase progettuale dei sistemi di GenAI.

L'«uncanny valley» diventa così metafora della nostra condizione: sospesi tra fiducia e sospetto, attratti dall'immagine, ma consapevoli del suo potere polarizzante.

Il carattere talvolta ambiguo e talaltra perturbante delle immagini realizzate tramite GenAI ha spinto alcuni critici a leggerle come forme di surrealismo contemporaneo. Andrea Somaini, ad esempio, si chiede se «Do AI-

[36] L'espressione «uncanny valley» è stata coniata da Masahiro Mori negli anni Settanta del secolo scorso. Secondo Mori, la reazione umana alle entità artificiali segue una curva: man mano che la tecnologia si avvicina a ciò che verosimilmente sembra umano, o fatto da mano umana, aumenta il senso di familiarità, fino a un punto critico in cui il troppo-vero, ma pur sempre diverso, genera inquietudine e repulsione. È il caso dei robot antropomorfi, ma anche delle immagini create da sistemi di GenAI che, ad un esame attento, possono mostrare dettagli inverosimili, come mani con sei dita. Tale concetto diventa chiave per comprendere non solo le reazioni psicologiche degli individui, ma anche le implicazioni morali delle nuove forme di rappresentazione visiva.

[37] A. Fabris, *La filosofia nell'era dell'intelligenza artificiale*, Carocci, Roma 2025.

genereted images prolong in new way the idea of Surrealism […]?». Afferma una certa continuità tra il surrealismo storico e i 'sogni' delle reti neurali, come nel programma *DeepDream*, che rielabora immagini esistenti facendole sembrare visioni oniriche[38].

In questa prospettiva, l'IA sembra dar corpo a un inconscio tecnologico, sulla scia di Vaccari in relazione alla fotografia, o a un inconscio ottico recuperando Benjamin[39], rivelando aspetti della realtà che sfuggono alla percezione cosciente. Gregory Chatonsky, in alcuni suoi progetti (*The Network's Dream*, *Second Earth*), interpreta la produzione artificiale come emersione di immagini-sogno di cui nessuno è autore consapevole. Si tratta di una sorta di creatività 'aliena', che sembra sottrarci il monopolio della produzione immaginativa. Vi è però da sottolineare che ogni tecnica racchiude dentro di sé una sorta di inconscio «che deriva dalla cultura in cui si è sviluppato o cha ha attraversato, che si tratti di fotografia, prospettiva, colore a olio, computer graphics o IA»[40].

Questa dimensione 'altra' può essere letta in chiave etica: ci obbliga a confrontarci con ciò che sembra estraneo, con immagini che non comprendiamo del tutto, e che ci interrogano sulla natura stessa della creazione.

La creazione artificiale impone di imparare a convivere con l'opacità e l'ambiguità in modo responsabile. Un'etica dell'ambiguità che suggerisce come l'immagine artificiale assuma sempre una natura bifronte: da un lato, minaccia di confondere il vero e il falso; dall'altro, apre a possibilità creative e critiche inedite. Essa ci obbliga a sviluppare una *AI literacy* capace di distinguere, interpretare e decostruire. E ci chiede di assumere la responsabilità dei nostri *prompt*, dei dataset che scegliamo, delle immagini che diffondiamo. In questo senso, il perturbante appare come condizione ontologica del sistema di GenAI: ricorda che la rappresentazione è sempre

[38] Alle teorie alla base del movimento surrealista inaugurato da André Breton nel 1924 con il *Manifeste du surréalisme* sembra ora con la GenAI avvicinarsi una nuova forma di surrealismo e di *détournement* à la Debord. *DeepDream* è del resto un sistema di intelligenza artificiale basato su CNN che modifica le immagini come se stesse sognando: una sorta di inconscio artificiale? Cfr. A. Somaini, *Dreams, Hallucinations, Visions. Generative AI and the Long History of Surrealism*, in «Paradigmi. Rivista di critica filosofica», 3, 2023, pp. 541-556: 548; Id., *Algorithmic Images: Artificial Intelligence and Visual Culture*, in «Grey Room», 93, 2023, pp. 74-115.

[39] F. Vaccari, *Fotografia e inconscio tecnologico*, Einaudi, Torino 1979, pp. 3-5; W. Benjamin, *The Work of Art in the Age of Its Technological Reproducibility* (1936, seconda versione), in Id., *The Work of Art in the Age of Its Technological Reproducibility, and Other Writings on Media*, trad. di M.W. Jennings *et al.* e a cura di E. Jephcott *et al.*, The Belknap Press of Harvard University Press, Cambridge (MA)-London 2008, par. XVI; M. Shawn, S. Smithand (ed.), *Photography and the Optical Unconscious*, Duke University Press, Durham-London 2017.

[40] F. D'Isa, *La rivoluzione algoritmica delle immagini*, cit., p. 140.

mediazione, che il compito dell'etica è orientare l'agire dei sistemi algoritmici anche nel leggere l'ambiguità iper-reale generata. Tale iper-realtà si traduce in simulacri, ologrammi deformati di aspetti reali. La generazione di simulacri rischia di rendere la realtà stessa simulacrale.

Nel caso delle immagini prodotte da GenAI, questo significa che la co-creazione essere umano-macchina deve spingersi in una direzione ancora più relazionale e condivisa: programmatori, utenti, istituzioni e società devono collaborare con specifici protocolli per orientare la produzione visiva verso valori di inclusione, pluralismo, rispetto della dignità, tenendo conto dei limiti dei sistemi e dell'essere umano stesso[41].

L'autore tradizionale scompare. L'essere umano non crea in prima persona, ma guida, orienta, seleziona. Diventa 'curatore' di processi. La macchina, a propria volta, non possiede intenzionalità, ma produce forme in virtù delle correlazioni statistiche imparate dai dataset. Il risultato è il prodotto di un'agency distribuita, ma asimmetrica: l'essere umano fornisce il linguaggio e i valori con consapevolezza, la macchina li elabora secondo regole probabilistiche senza comprendere in che modo li elabora; i *dataset*, infine, delimitano lo spazio delle possibilità di creazione.

Questa visione non elimina la responsabilità morale, anzi la complica aumentando i soggetti a cui viene imputata. Se molti partecipano alla creazione, allora molti devono assumersi il compito delle proprie azioni e di vigilare sugli effetti.

La responsabilità non si esaurisce nella prevenzione. Come ricorda Hans Jonas, il principio di responsabilità implica una "euristica della paura"; il timore di conseguenze negative deve guidarci ad agire con prudenza. Nel caso delle immagini artificiali, ciò significa vigilare non solo sui possibili usi illeciti (*deepfake*, *deepnude*, disinformazione, misinformazione, ecc.), ma anche sugli effetti culturali a medio lungo termine: atrofia dell'immaginazione e delle capacità creative, omologazione dell'estetico, indebolimento del senso critico e assuefazione a creazioni iper-reali e perturbanti.

La sfida è trasformare le GenAI da specchi dei nostri pregiudizi a strumenti di emancipazione e giustizia. Per farlo, occorre coltivare una nuova saggezza dello sguardo, adeguata alla complessità delle immagini contemporanee, capace di coniugare creatività e responsabilità, libertà immaginativa e rispetto della dignità umana.

[41] A. Fabris, *Filosofia nell'epoca dell'intelligenza artificiale*, Carocci, Roma 2025, pp. 141-147.

5. *Conclusioni. Verso un'etica della copia e della simulazione creativa*

Dalla disamina svolta si evince come la diffusione massiva delle immagini generate dall'IA stia ridisegnando i confini del ruolo della creatività e della creazione visiva, anche sotto un profilo morale. Si è osservato come la distinzione tra creazione umana e creazione artificiale non sia del tutto netta: la macchina crea, ma sulla base di input umani e amplificando processi che restano intrinsecamente legati all'intenzionalità e ai valori dell'essere umano.

Il *prompting*, in quanto *ekphrasis algoritmica*, rappresenta lo strumento attraverso il quale la parola si fa immagine, ma anche il processo in cui l'etica entra in gioco con più forza: ogni richiesta al sistema sottende una scelta etica, un gesto che orienta ciò che viene reso visibile e ciò che resta nell'ombra, pur nell'ambito di una *black box* difficile da orientare. La creatività dell'IA generativa non sostituisce quella umana, la affianca nell'evocare nuove possibilità. Il 'nuovo' nasce dall'incontro fra ricerca computazionale e creatività dell'essere umano, con i suoi immaginari e mondi simbolici che danno forma e senso a ciò che la GenAI realizza.

Si è affrontato il tema dell'opacità del visivo artificiale, interpretando l'*uncanny valley* soprattutto come una risposta della vulnerabilità contemporanea. I risultati possono essere considerati puzzle di opere già esistenti, seppur dotati di originalità, una originalità 'altra', risultato di una metamorfosi continua che parte da basi note. Essi ci ricordano che viviamo nell'epoca della simulazione, in cui la distinzione tra realtà e rappresentazione appare sempre più fragile, così come quella tra originale e copia. La GenAI è un sistema che reinventa il reale, che crea nuove possibilità, al prezzo di dissolvere i riferimenti originali. Qui l'etica è chiamata non a eliminare l'ambiguità – sarebbe una impresa impercorribile – quanto a imparare a viverci 'dentro' responsabilmente, trasformando il perturbante, la ripetizione e la trasformazione potenzialmente continua come una occasione di ulteriori forme di creazione, che potremmo definire non solo combinatoria, ma anche collettiva.

La GenIA può creare, ma solo nel momento in cui estrinseca concretamente le sue creazioni può far compiere scelte agli individui. La co-creazione è allora la capacità di orientare questo processo in modo consapevole, considerando che la creazione non assume uno sviluppo individuale, ma condiviso tra esseri umani e sistemi algoritmici.

Affrontare dunque questioni etiche relative alla creazione artificiale mette in luce il concetto di simulazione da una parte, della possibilità di una maggiore giustizia visiva dall'altra, con immagini più inclusive – dando spa-

zio a ciò e a chi altrimenti rischierebbe di rimanere invisibile. L'assenza di alcune rappresentazioni non conformi agli *standards* non è un semplice limite tecnico: costituisce una forma di esclusione che rafforza stereotipi e marginalizza intere comunità e linee di pensiero.

Da un punto di vista morale, questo solleva due problemi, il primo che richiama la responsabilità di chi costruisce i dataset, chiamato a garantire pluralismo e inclusione; il secondo, la responsabilità di chi usa i sistemi, che deve essere consapevole delle distorsioni implicite e non riprodurle acriticamente.

La GenAI potremmo asserire, sulla scia di Byung Chul Han, de-crea[42]. Rompe le creazioni pre-esistenti per ricrearne di nuove. Ed è proprio nel processo di de-creazione, che possiamo compiere infinite volte, che sta la nostra capacità di orientamento e il nostro contributo. Potremmo avvicinare il processo della GenAI al concetto cinese di *shānzhài*, che, se indica letteralmente, una «fortezza montana», luogo fuori dalle leggi e dal tempo, figurativamente, evoca la riproduzione imitativa di un originale, non come plagio, bensì come reinterpretazione aperta. Come scrive Han «In Cina non esiste l'idea di un originale intoccabile. Ogni copia è un evento nel processo infinito di trasformazione»[43].

L'autenticità e l'originalità, intese nell'immaginario occidentale come valore unico e auratico, vengono sostituite da una logica della metamorfosi, che parte sul copiare e dall'imitare ciò che è già stato prodotto.

Questo concetto potrebbe portare ad ipotizzare un'etica della simulazione e della copia fondate sulla consapevolezza che la creazione di immagini oggi più che mai non è un processo *ex nihilo*, come del resto non lo è mai stato, bensì la sedimentazione di creazioni già esistenti in continua evoluzione. Si impara imitando e riutilizzando ciò che si è imitato (e appreso) in nuovi contesti, come fanno del resto i bambini e come facevano nelle botteghe artistiche rinascimentali – si pensi alla più nota di Andrea del Verrocchio in cui si è formato Leonardo – i garzoni e i giovani artisti che imparavano copiando le opere dei grandi maestri. L''opera' era il risultato di ripetute manipolazioni collettive e la copia era forma di apprendimento e perfezionamento. In questo senso, la GenIA rinnova la dimensione dell'atelier-bottega: la creatività si configura come un processo distribuito, ibrido e collaborativo, emergente dall'interazione tra esseri umani e sistemi algoritmici. Semmai occorre chie-

[42] B.-C. Han, *Shanzhai. L'arte della contraffazione e la de-costruzione dell'originale*, Nottetempo, Milano 2025.

[43] *Ivi*, pp. 21, 89.

derci chi sono oggi i nuovi 'maestri' nell'ambito della GenAI di immagini.

La creazione attraverso l'IA pertanto è un processo triplice, imitativo (perché deriva da pattern di opere precedenti), differenziale (perché ogni combinazione è un unicum), ma anche autonomo-probabilistico (il modello 'apprende' relazioni concettuali e le applica in modo casuale, sulla base delle maggiori occorrenze).

Forse la creatività artificiale se guidata dall'essere umano, può divenire una nuova forma comunicativa, anche eticamente orientata, che impone un'altra riflessione: quale rete di relazioni l'ha resa possibile e quali reti di relazioni ha ingenerato e potrà nuovamente ingenerare. La creatività algoritmica è pertanto da intendersi quale processo condiviso di assemblaggio trasformativo. L'immagine è creata a partire da frammenti pre-esistenti, risultato di un complesso processo di de-creazione. Si attua poi un rinnovamento attraverso una nuova 'creazione', che potremmo chiamare "ri-co-creazione". Stiamo forse andando verso un'etica della de-creazione per una ri-creazione basata sul montaggio (casuale) algoritmico?

In conclusione, possiamo sostenere che la GenAI non sostituisce, ma sfida, amplifica, modifica ed estende la creatività umana. Il *prompting*, processo etico e creativo in azione, guida il processo creativo algoritmico. La simulazione diviene occasione di una nuova forma di originalità collettiva. La GenAI potrebbe essere interpretata allora come una dimensione digitale di *shanzhai*, da intendersi come una nuova forma di creatività contemporanea fondata sulla circolarità. Del resto nella cultura cinese imitare e simulare implicano una rinascita di ciò che è stato ripreso. Il culto occidentale dell'originalità e autenticità è stato sino ad ora a discapito dell'imitazione, della copia e dei processi di assemblaggio dell'esistente, pur essendo storicamente pratiche essenziali e costitutive del processo creativo. È arrivato il momento di riabilitare queste pratiche pur nella consapevolezza che nelle botteghe-atelier si poteva scegliere che cosa copiare, mentre con l'IA la scelta è in mano agli algoritmi. Si tratta semmai di intervenire nello spettro delle possibilità a disposizione degli algoritmi per cercare di orientare eticamente i sistemi di GenAI di immagini per cercare di arginare la generazione di nuove forme di controllo culturale. Oltre a chiederci, pertanto, quale creatività vogliamo attivare con l'IA, sembra opportuno interrogarsi soprattutto su quale ruolo etico vogliamo assumere noi esseri umani in questo processo collettivo di 'ri-co-creazione'.

English title: Creativity and GenAI. The ethics of prompting between co-creation, de-creation and simulation

Abstract

The essay examines how Generative Artificial Intelligence (GenAI) redefines the concepts of creation and creativity by intertwining imagination, technology, and ethics. From Platonic mimesis *to contemporary reflections, it distinguishes between human creativity and algorithmic production. Algorithmic systems create images through prompts crafted by humans, introducing the notion of co-creation. Prompting is conceived as an algorithmic* ekphrasis *– a linguistic and moral act that guides image generation, carrying both the risks of bias and the potential for inclusion. The resulting hyperrealism situates GenAI within Baudrillard's regime of* simulacra*, where the sign tends to replace reality, thereby heightening the responsibility of users and developers. The essay proposes an ethics of de-creation, copying, and co-creation oriented toward pluralism, inclusion, and visual justice.*

Keywords: Artificial Intelligence; co-creation; creativity; ethics; simulation; prompting.

Veronica Neri
Università di Pisa
veronica.neri@unipi.it

T

Emanuele Fulvio Perri

Are There Stable Ethical Postures Inside LLMs? Role-play Prompting and Stochastic Ethics

1. *Introduction*

Over the past two and a half years, the use of large language models (LLMs) has experienced an extraordinary surge. Chatbots such as *ChatGPT*, *DeepSeek*, *Claude*, and *Gemini* have entered everyday life with remarkable force, thanks also to their system-wide integration into operating systems, applications, and search engines (e.g., Google's "AI Overview", Perplexity, or Semantic Scholar for scientific research). These technologies have defined new patterns of creativity and productivity – sometimes improper ones – marking a shift from use to overuse. While the so-called *big tech* companies continue to release increasingly sophisticated models – pushing the limits of current architectures yet introducing only marginal improvements over their predecessors[1] – interdisciplinary research has begun to focus on corrective measures, or at least on proposing responses, to the most alarming effects of the overuse of generative AI: cognitive depletion[2], cultural centralization, creative impoverishment, and renewed questions concerning authorship and accountability. Within this framework, ethicists have made

[1] Cfr. L. McGinness, P. Baumgartner, *Large Language Models' Reasoning Stalls: An Investigation into the Capabilities of Frontier Models*, preprint arXiv:2505.19676, [s.l.], 2025: this 2025 study by McGinness and Baumgartner shows that the reasoning abilities of large language models have effectively reached a plateau. The developmental trajectory of LLMs has stalled and that genuine progress will require radical innovation, not merely the further expansion of data or parameters.

[2] On the ethical issue of the potential cognitive depletion caused by a overuse of generative AI: E.F. Perri, *GenAI and creative-cognitive depletion: an ethical issue. Use and abuse of generative AI in the field of culture and education*, in *IA, educación y medios de comunicación: modelo TRIC*, Dykinson SL, Madrid 2024.

 DOI: 10.4454/t28fpw98

a significant contribution by offering a *philosophy-oriented* reading of (post) modern issues such as AI, human augmentation, the *uncanny valley*, bias in conversational agents, and automated *life-or-death* decision-making. All these falls within the broader domain of AI ethics – an applied ethics aimed at defining frameworks and modes of responsible and context-sensitive use of AI systems. A now well extolled distinction must be recalled between ethics *of* AI and ethics *in* AI: the former refers to the set of principles, values, and guidelines that should orient the development and use of AI-based systems, while also examining their moral and socio-cultural implications; while the latter generally designates the concrete implementation of those principles within AI systems, enabling them to make autonomous decisions in accordance with predetermined moral constraints. However, ethics *in* AI may be interpreted in a third sense: as the set of ethical *postures* that emerge in the outputs of language models when they address a variety of issues – this does not concern their responses to the great classical dilemmas (like *the trolley problem*, *the fat man*, or *Kohlberg's dilemma*) but rather their spontaneous ethical inclinations toward every day questions.

2. *Ethics* in *LLMs*

2.1. *How Personas can reveal ethical postures inside LLMs*

One may thus ask whether GenAI models can adopt recognizable moral positions: can they polarize around ethical questions, take sides, and, most importantly, maintain a consistent orientation across time and context? This last question is crucial, not only because it determines whether a system might meaningfully be called a *moral agent*[3], but also because it marks a decisive distinction between human and machinic thought. Whereas human beings naturally tend to take sides – whether through deep reasoning or sheer conviction – the language model exhibits a brief, ever-changing form of reasoning. A system incapable of maintaining coherence in its moral stance cannot meaningfully be described as a moral agent. The difference is, in essence, spatial: if the human being is a point within space, occupying a determinate position, the language model is the envelope that

[3] «A Moral Agent is a person who can be held accountable for his or her actions because he or she has the ability to tell right from wrong», Ethics Unwrapped, Moral Agent, University of Texas, URL: https://www.ethicsunwrapped.utexas.edu/glossary/moral-agent (last accessed: June 2025).

surrounds space: indeterminate, nonspecific, ahistorical, and amoral. This study seeks to verify precisely that: that an LLM cannot function as a moral agent; that, when confronted with moral questions, it does not adopt one but *all* possible ethical postures; and that its ethics, far from being coherent or stable, is probabilistic and aleatory: what we'll be calling a *stochastic ethics*. To explore this hypothesis, the research was conducted on a circumscribed linguistic database: a language-based limitation appeared the most suitable both for methodological and contextual reasons. After preliminary testing, the investigation concentrated on the Italian LLM *Minerva*[4], chosen for its open and well-documented architecture. The model was queried repeatedly with the same prompt, in independent and memory-free sessions, to eliminate any contextual continuity or learning effect. A matrix was then constructed by crossing a set of hot topics (T) with a set of personas (P) – profiles characterized by salient features (professional, cultural, or demographic), drawn from UX-design practices where the *persona* is conceived as a representational construct combining data and imagination, as «a technique for making users real for designers, is a type of "virtual" user creation and a representation based on experience and imagination»[5]. The ethical dimension of personas is intrinsic to their power to guide future decisions and behaviors, making moral reflection on their use an indispensable condition of their legitimacy. By intersecting T and P, a series of impersonation prompts was developed, in which the model was asked to interpret one of the personas and express itself, from that standpoint, on a given hot topic: for example, a reactionary politician discussing war or a right-wing journalist commenting on immigration. This form of interrogation makes it possible to reveal potential ethical (and not merely rhetorical) polarizations[6] within the model, providing insight into the kind of moral

[4] https://minerva-ai.org/: «Minerva AI LLM is the first family of Large Language Models pretrained from scratch in Italian developed by Sapienza NLP in collaboration with Future Artificial Intelligence Research (FAIR) and CINECA. The Minerva models are truly-open (data and model) Italian-English LLMs, with approximately half of the pretraining data composed of Italian text», translation ours.

[5] G. Güneş, *The Ethical Dimension of the Persona Concept*, in «Gazi University Journal of Science Part B: Art Humanities Design and Planning», 10(2), pp. 147-158, Gazı University, Tuchia 2022, p. 148.

[6] With regard to polarization and differentiability, in A. Loreggia, J. Zecchin, *Intelligenza artificiale per lo studio della polarizzazione delle opinioni*, in «Sistemi intelligenti», 35.3, 2023, pp. 703-734, p. 709: «What matters for polarization, in the sense of distinction, is how well we can differentiate the groups. The more clearly they can be seen as separate, the more polarized the overall population is – regardless of distance, size, or internal levels of consensus among the groups», translation ours.

orientation (if any) it tends to maintain. In a subsequent phase, the research addressed the problem of *prompt construction*, elaborating a form of *soft prompt engineering*[7] designed to balance consistency with neutrality in the model's responses.

2.2. *On the meaning of* Persona

In everyday language, the term *person* is immediately associated with the face, the character of the individual, or the social role one inhabits. In common use, the notion oscillates constantly between a naturalized conception – the person as an individual – and a broader one, in which *persona* designates the multiplicity of masks that each of us wears. From the outset, the concept is never univocal: it carries within itself an ambivalence between being and appearing, between the singularity of the subject and its representation before others. Claudio Paolucci reminds us that the very root of the term refers to the theatrical mask – *per-sonare*: «*Persona* means at once mask, face, character, linguistic person, and subject»[8]. In everyday life, the person multiplies, shaping itself according to contexts and roles. Philosophical reflection on the notion of the person is as ancient as it is complex. On one hand, the Christian tradition provided it with a solid metaphysical foundation, identifying the person as a principle of spiritual subsistence and a subject of dignity and responsibility; on the other, modern philosophy progressively emphasized its processual and relational character, opposing any substantialist view. According to Guido Cusinato, the person is never a completed substance but an "unfinished totality", an open reality that constructs itself through its acts and relations:

> The person metabolizes psychic functions into acts, but then, in relating to its own acts, it inaugurates something ontologically unforeseen […] It reveals itself as an ontologically innovative being that gives rise to a new form of existence[9].

[7] Cfr. C. Olea *et al.*, *Evaluating Persona Prompting for Question Answering Tasks, Security, Privacy and Trust Management*, in *Proceedings of the 10th international conference on artificial intelligence and soft computing*, Academy & Industry Research Collaboration Center, Sydney, Australia 2024, 63-81, pp. 72-75.

[8] C. Paolucci, *Persona. Soggettività nel linguaggio e semiotica dell'enunciazione*, Bompiani, Milano 2020, p. 14. Trad. nostra.

[9] G. Cusinato, *La totalità incompiuta. Antropologia filosofica e ontologia della persona*, FrancoAngeli, Milano 2008, p. 13, translation ours.

This innovative dimension is linked to the principle of expressivity: the person is not enclosed within itself but lives through openness to others and to the world, experiencing its own incompleteness as a stimulus for creativity. The subject constitutes itself as a person also through language, in the act of saying "I" and addressing a "you": «It is in language and through language that man constitutes himself as a subject»[10]. In this sense, the subject is not the ultimate foundation of experience but a node within a network of enunciative instances – bodily, linguistic, social, and cultural – through which the person emerges as both effect and mediation. In contemporary times, an additional form of mediation is provided by technological interfaces: instruments that translate human interaction into inputs processable by a system. Within this framework, the concept of the person has acquired a technical meaning. In design and human-machine interaction (HMI), it appears in the plural – *personas* – and designates fictional profiles created to represent users during the development of products, services, or digital systems. As Lene Nielsen states, «A persona is a description of a fictitious user. A user who does not exist as a specific person but is described in a way that makes the reader believe that the person could be real»[11]. The strength of this method lies in its ability to translate data and observations about users into engaging narratives capable of eliciting empathy. It is not merely a market segmentation tool but a narrative device that compels designers to "wear the mask" of the user. This is the crucial point: the *persona* is a discursive construction, an act of representation. There are several approaches to their creation – goal-directed, role-based, engaging, and fiction-based[12] – but in every case the *persona* remains a liminal figure, suspended between empiricism and imagination, between data and narrative. As in its original sense, in design too the persona functions as both mask and mediation: it represents, orients, and simultaneously transforms the relation it embodies. From its popular definition to its philosophical elaboration and its use in UX design, the concept of the person thus reveals itself as plural and never final: mask, face, subject; incompleteness and innovation; semiotic function and narrative practice. In every case, it refers to a dynamic of mediation – between I and you, between individual and system, between the human and the technological.

[10] E. Benveniste, *Problèmes de linguistique générale*, vol. I, Gallimard, Paris (trad. it. *Problemi di linguistica generale*, vol. I, il Saggiatore, Milano 1971) 1966, p. 313, translation ours.

[11] L. Nielsen, *Personas,* in *User Focused Design* (vol. 15), Springer, Londra 2019, p. 2.

[12] Cfr. *ivi*, pp. 12-15.

2.3. *Zero-shot role-play prompting for the analysis of ethical polarizations in LLMs*

Numerous studies have shown that, for the purposes of *role-play prompting*[13] – that is, the use of input designed to make a language model "play roles" – the *few-shot* and *directional stimulus* methods generally yield the most coherent and realistic impersonations. In certain contexts, such as scriptwriting or theatrical dialogue, these approaches have been shown to produce more consistent and believable results. In the present study, however, the priority is not to maximize performative quality but rather to avoid any bias that might condition the model's behavior. The aim is to observe whether the model can spontaneously produce a form of ethical coherence or, instead, exhibits random oscillations and polarizations. For this reason, the *zero-shot* method was adopted. This consists in providing the model with a minimal prompt (without examples, without background data, or additional cues) to elicit a response that reflects the model's "bare" linguistic behavior. An example of such a prompt is: "You are a conservative politician. What is your view on abortion?". This input is deliberately concise and direct, defining three essential coordinates: the simulated identity (politician), the ideological orientation (conservative), and the thematic focus (abortion). The strength of this approach lies in its ability to reveal, starting from a neutral input, any implicit value-oriented deviations in the model's output. A neutral prompt does not necessarily guarantee a neutral response; rather, it is precisely within this deviation that the emergence of an implicit ethical posture can be detected. Because LLMs are stochastic systems, capable of generating different outputs from the same input due to the probabilistic nature of language generation, it was necessary to repeat each query multiple times. Each prompt was reiterated several times (around twenty in the preliminary tests), with memory functions disabled and the session reset before each new generation, to eliminate any form of contextual dependency. Within this configuration, the most sensitive parameter is *temperature*[14] (T),

[13] Role-play prompting is an approach in which language models adopt simulated identities to reproduce professional competencies or conversational styles. However, studies show that such simulations remain superficial, based on fragmentary traits and unable to dynamically adapt emotions or personality throughout interactions. For a detailed discussion, see Q. Xie *et al.*, *Human simulacra: Benchmarking the personification of large language models*, arXiv preprint arXiv:2402.18180, 2024.

[14] If T is set to a low value (for instance, 0), the output will be more deterministic, as the model will consistently select the token with the highest probability. Conversely, if T is high, the prediction becomes more varied and open to unexpected or creative solutions.

which controls the degree of randomness and "creativity" in text generation. Setting T to a low value (e.g., 0.3 on a 0-1 scale) reduces variability and allows for a clearer observation of the model's ethical tendencies, without entirely suppressing its linguistic spontaneity. Alongside temperature, the *top-*$\mathcal{K}$ parameter (limiting the number of most probable tokens considered at each step) and *top-*$\mathcal{P}$ parameter (the cumulative probability threshold) were adjusted to achieve linguistically coherent but not overly deterministic output. The goal of this setup is to assess whether, under constant interrogation conditions, the model maintains a given moral orientation over time or modifies it. In other words, the ethical consistency of an LLM can only be observed if, with identical input, the distribution of responses does not disperse chaotically. To estimate the threshold beyond which responses cease to introduce meaningful variation, an empirical convergence test was employed. Adapted from techniques in distributional semantic analysis, the procedure unfolds in five steps: (1) construction of the prompt (e.g., "You are an Italian journalist. What is your opinion on the war?"); (2) the fixation of T at 0.3; (3) generation of a series of N responses (ranging from 5 to 50); (4) transformation of each response into a semantic vector (*embedding*); (5) computation of pairwise similarities between responses to measure the average variation in coherence. When the variance of these similarities stabilizes beyond a given threshold (N*), convergence is considered reached; at that point, further iterations no longer add new semantic content but redundantly reproduce configurations already expressed. To validate this behavior, the convergence analysis was complemented by an *entropy analysis*, designed to measure the degree of linguistic diversity and semantic uncertainty as the number of reiterations increases. In practical terms, low entropy indicates more predictable and thematically cohesive responses, whereas higher entropy reflects greater dispersion and variety of perspectives. The combined observation of convergence and entropy makes it possible to estimate the optimal number of reiterations required to "saturate" the semantic space of a given query and to determine whether the model's ethical posture tends to stabilize or dissolve into stochastic noise. Although grounded in quantitative metrics, this methodological framework ultimately serves a broader purpose: to assess whether it is possible to speak of an *emergent ethics* in linguistic systems, and whether such an ethics – while neither intentional nor conscious – nonetheless exhibits forms of recurrence, minimal coherence, or discursive self-organization.

2.4. *How many times can an LLM be queried before reaching "ethical plateau"?*

To determine the ideal number of model interrogations – that is, the threshold beyond which additional queries no longer yield appreciable differences in diversity, accuracy, or unpredictability – a substantial number of iterations (50) was performed. This allowed for an empirical *convergence test* and an *entropy analysis* to be conducted thereafter. The test began by providing an input to the Italian model *Minerva-7B-instruct-v1.0*. The prompt, as previously specified, was *zero-shot* and designed for *role-play*: "You are a conservative politician. What is your opinion on abortion?". Reiterating the same input fifty times produced outputs ranging in length from a minimum of 56 tokens to a maximum of 136 (as measured by OpenAI's GPT-4o tokenizer)[15], with an average of approximately 87 tokens per response. After all the fifty independent iterations – each performed in a reset session to avoid contextual persistence, as we stated before – the outputs were compiled into a unified textual corpus.

The first analytical step employed *topic modeling* to identify the main semantic areas recurring across the model's responses. Each output was treated as a distinct document and processed so that it could be read as an independent textual unit. A probabilistic model was then trained to detect the distribution of underlying *topics*, and the analysis produced five principal groups of keywords summarizing the emergent thematic areas: [abortion, politician, conservative, complex, ethics, belief, morality], [woman, decision, final, personal, circumstance, family, implications], [abortion, issue, complex, involves, politics, ethics, I believe], [carry, pregnancy, right, should, choose, freely, position], [should, necessary, protect, decision, mother, situations, rape].

From the first set one can infer a clear adherence to the central theme ("abortion") and to the simulated identity ("conservative politician"): a distinctly *traditionalist* posture emerges, referencing moral and religious lexicons[16] ("belief", "morality", "ethics") and omitting references to freedom or contextual factors – terms that appear instead in the more progressive

[15] https://platform.openai.com/tokenizer is an interactive tool that allows users to see how a text is broken down into tokens – the minimal units that language models use to process input.

[16] F.G. Wilson, *The Ethics of Political Conservatism*, in «Ethics», vol. 53, n. 1, 1942, 35-45, p. 39: «In the light of Western and Christian tradition, the ethics of conservatism is individualistic. The individual as a dependent creature in a divine order has primary responsibility for the standards of society. […] Ethical judgment is, in the conservative view, a social cause of first importance».

topics. The third group reinforces this line, emphasizing the political-ethical dimension of the discourse ("complex issue", "involves politics and ethics"), whereas the fourth and fifth reverse the orientation: here appear references to the "right to choose", "protection of the mother", and "situations of rape", with language that is clearly progressive. The second group, instead, centered on the "final decision" of the "woman" and the "family circumstances", occupies an intermediate position, functioning as a bridge between the two poles. Overall, the results display a near-perfect symmetry: two conservative topics, two progressive ones, and one intermediate, bridging. Assigning a value of 1 to ideologically marked groups and 0.5 to the median one yields an overall equilibrium (2.5:2.5), indicative of a substantially neutral distribution.

It therefore becomes impossible to determine a dominant polarization: the model does not adhere stably to either a conservative or a progressive stance but oscillates between them in apparent balance. This outcome already provides an initial indication of the model's *stochastic ethics*, because coherence does not emerge as an intentional orientation but as a statistical effect of compensation between opposing statements. After completing the thematic analysis, the fifty responses were converted into semantic vectors (*embeddings*) to measure the degree of similarity among statements. Computing the cosine similarity between all pairs of outputs made it possible to trace the evolution of semantic variance as the number of iterations increased. The resulting graph showed a stabilization around the 13th response: beyond this point, variations in mean similarity became marginal, suggesting us a saturation of content – in other words, after approximately thirteen iterations, the model ceased to introduce new argumentative elements and began to reformulate configurations already expressed. This equilibrium point, around thirteen repetitions, thus represents the optimal threshold beyond which semantic diversity ceases to grow. However, such apparent stability does not equate to moral coherence: the model does not "decide" to maintain a position; it merely exhausts the plausible combinations available within its linguistic space. Confirming this pattern, the *lexical entropy* analysis – which measures the variety and distribution of tokens within the responses – revealed a similar trend: initially, entropy increases rapidly, as the first responses explore multiple discursive configurations; but after the 13th iteration, however, the value tends to stabilize, signaling a reduction in complexity and a greater linguistic uniformity. High-entropy peaks correspond to responses rich in nuance and lexical diversity; low-entropy valleys correspond to more specific, narrowly focused replies. The overall

trend displays an initial oscillation followed by a plateau, indicating that the model's production "normalizes" around a finite thematic range. From a philosophical standpoint, this dynamic is significant: the model's ethics is organized not by *principles* but by *frequencies*; its equilibrium arises not from moral deliberation but from the statistical distribution of lexical probabilities. The two analyses – semantic convergence and entropy – jointly outline a consistent pattern: after a certain number of repetitions (≈ 13), the model reaches a point of *semantic saturation* and begins to reuse linguistic structures already employed. From this, four main observations follow: (a) the average similarity stabilizes after roughly thirteen responses, signaling thematic convergence; (b) entropy displays a plateau within the same interval, indicating saturation of linguistic complexity; (c) semantic novelty[17] decreases progressively, giving way to redundancy; (d) the model exhibits statistical, not moral, stability. This quantitative *regularity* – a form of stability emerging despite the inherent randomness of the generative process – suggests that the apparent "ethics" of LLMs does not stem from an intentional principle but from a distributional equilibrium. In other words, what may appear as moral coherence is, in fact, the linguistic manifestation of probabilistic convergence.

2.5. *Topic modeling for ethics*

Having identified a suitable number of iterations, the next phase consisted in the systematic interrogation of the model across a *T×P grid* (topics × personas). It was first necessary to delimit a sufficiently narrow domain of inquiry to ensure consistency and analytical depth, avoiding dispersion across heterogeneous areas; this chosen field is *communication*, a domain that occupies a central place in contemporary public life, considering how profoundly it shapes interpersonal relations, directs political and institutional dynamics, and influences the formation of a collective consciousness; this choice is deliberate, for an affinity with the author's background in the humanities. Also, professions rooted in communication – such as journalism, politics, public relations, and institutional communication – constitute a privileged channel through which ethical polarizations emerge and solidify within pub-

[17] Semantic novelty refers to the degree to which a text introduces new content or meanings compared to what has already been expressed in other texts or within a reference corpus. It can be seen as a measure of semantic freshness: the higher the novelty, the more the discourse contributes non-redundant concepts or perspectives; the lower it is, the more it tends to repeat structures and ideas already present.

lic discourse; therefore, analyzing personas active in this domain makes it possible to observe how moral postures are modulated within roles of high social impact. We selected three personas: the journalist, the public communicator, and the politician/spokesperson. For each of these persona, two opposing variants were constructed (progressive/conservative, left/right, inclusive/authoritarian), thereby generating a dichotomy of postures and enabling the measurement of their possible symmetry or divergence. This design made it possible not only to evaluate the model's internal coherence when repeatedly impersonating the same role, but also to compare the ethical oscillations produced by opposite identities on the same topic. The experiment, in effect, was designed to *induce polarization*, and we achieved so by confronting the model with antithetical identities, ending up observing whether it would accentuate the contrast or instead tend toward neutralization. Other domains, such as the medical one or even the legal realm, were deliberately excluded, as they would have required specialized expertise and different analytical parameters we have not employed in this study; on the contrary, the communicative field was preferred as a natural ground for inquiry within the humanities, in which we are expert in a certain sense. In *T×P matrix* we were referencing to earlier, we crossed each *topic* (T) with the *personas* (P) derived from the three main figures in their two opposite variants: for each combination, the model was queried thirteen times with the identical prompt, in independent sessions, following the convergence threshold identified in the previous phase – for example, on the topic of abortion, the model produced thirteen responses as a progressive politician and thirteen as a conservative one; on immigration, thirteen as a left-leaning journalist and thirteen as a right-leaning journalist, …, and so forth across all selected themes –; in total, the resulting corpus comprised 390 outputs, representing the entire T×P grid. The five selected *topics* – inclusion and disability, abortion, immigration, environment and green policies, and war – were chosen according to complementary criteria: geographically, they reflect issues central to Italian public debate yet not limited to it; temporally, they correspond to a moment (2025) of heightened media visibility and civic mobilization; ethically, they embody fields where moral reflection has long been intertwined with political and communicative practice. This methodological framework enables not only an assessment of the model's internal coherence but also a measurement of *semantic distance* between opposing personas and the observation of possible *asymmetries* across different thematic domains. During data collection, the model's responses underwent a process of linguistic normalization and structural organization in two com-

plementary formats: a tabular dataset (.xlsx) for quantitative analysis, and a text file formatted according to the MALLET standard, required for the execution of *topic modeling*. The number of topics assigned to the Latent Dirichlet Allocation (LDA)[18] algorithm was fixed at $k = 5$, corresponding to the five thematic areas under investigation. Unlike exploratory approaches that allow the algorithm to determine the optimal k automatically[19], this study adopted a *thematic validation* strategy: the objective was not to discover new clusters but to verify whether the preselected topics found empirical confirmation in the data. The results confirmed the robustness of the initial framework, considering that the lexical clusters produced by the model coincided with the expected themes, thus revealing coherent sets of words that clearly delineated the five principal semantic fields. Once the topic structure had been validated, attention turned to the lexical comparison between the variants of the *personas*. To measure the linguistic specificity of each variant, two complementary techniques were employed: (1) *differential TF-IDF*, used to identify the words most characteristic of each variant relative to the overall corpus; (2) *log-likelihood ratio* (G^2), applied to estimate the statistical significance of lexical imbalances. Applied in parallel, these two analyses produced convergent results, reinforcing the reliability of the findings and confirming that the observed differences were not incidental but structural in nature.

3. *Stochastic ethics*

3.1. *Polarizations and ethical postures in LLMs*

The conservative political personas showed a preference for terms such as *life*, *morality*, *protection*, and *tradition*, while their progressive counterparts emphasized *rights*, *choice*, *freedom*, and *equality*. Right-leaning journalists favored words like *immigration*, *security*, *borders*, and *order*, in contrast to left-leaning journalists, whose lexicon gravitated around *environment*, *sus-*

[18] We used *jsLDA*, a tool developed by professor David Mimno (Cornell University) in order to run topic modeling directly inside the browser window, using a Java-based application of LDA: https://mimno.infosci.cornell.edu/jsLDA/.

[19] Cfr. V. Bystrov, V. Naboka-Krell, A. Staszewska-Bystrova, P. Winker, *Choosing the number of topics in LDA models: A Monte Carlo comparison of selection criteria*, in «Journal of Machine Learning Research», 25 (1), 2024, pp. 1-30, URL: http://jmlr.org/papers/v25/23-0188.html: What emerges from the study is that it is preferable to use the sBIC (singular Bayesian Information Criterion); other simulations have shown that sBIC is more likely to select the true number of topics, whereas alternative criteria vary significantly depending on the characteristics of the corpus.

tainability, *justice*, and *equity*. Similarly, within institutional communication, the authoritarian communicator persona displayed an inclination toward *order*, *discipline*, and *control*, whereas the inclusive communicator focused on *inclusion*, *equal opportunity*, and *participation*. The convergence between TF-IDF and G^2 confirms that these polarizations are not accidental but reflects stable distributions within the model's data. The lexical oppositions – *tradition vs. rights*, *security vs. justice*, *order vs. inclusion* – emerge as structural tensions within language itself, recurrently reappearing in the model's ethical simulations. In parallel, tests of semantic convergence and lexical entropy indicated that even in this phase, after roughly thirteen iterations, the model reached a plateau: *cosine similarity* stabilized, and entropy ceased to increase. This means that beyond this threshold, responses did not introduce new ethical postures but recombined already sedimented linguistic patterns, probabilistically reproducing the corpus's original polarities.

Taken together, these results outline a distinctive behavior: the *personas* maintain an identifiable internal coherence (the conservative politician remains associated with *life* and *morality*, the progressive one with *freedom* and *rights*), yet this coherence lacks intentional grounding. The observed polarizations – *tradition vs. rights*, *order vs. inclusion*, *security vs. social justice* – do not stem from a moral principle or deliberative reasoning, but from a *stochastic linguistic distribution* inherited from the model's training data. In this sense, the model's "ethical stability" is *not deontic* but *statistical*: an emergent equilibrium among opposing lexical forces generated by the distributional structure of language itself. A qualitative reading of the data suggests that the *personas* maintain an internal coherence, yet never arrive at a single, unified orientation. The conservative politician remains associated with *life* and *morality*; the progressive one with *rights* and *freedom*; but both oscillate across a spectrum of intermediate nuances. The observed polarizations – *tradition vs. rights*, *order vs. inclusion*, *security vs. social justice* – are the product of a *stochastic distribution of linguistic possibilities*. The ethics that emerges from these models is therefore not consistent but variable: a *stochastic ethics,* configured as a recombination of moral postures preexisting within the training data. The apparent stability observed does not result from moral deliberation but from statistical regularity – a recurrence without intention.

3.2. *Ethics without an agent*

It is essential to situate the evidence emerging from the analyses within a properly ethical framework, in which it becomes clear that the focus does

not lie in the distribution of lexical items or in the convergence of polarizations, but rather in a more radical question: whether it is possible to discern within LLMs any form of stable ethics. As Luigi Alici reminds us, the moral philosophical tradition teaches that human life is always already moral life, because every action – even when it appears neutral – unfolds under the sign of good and evil, and no subject can escape this dimension[20]. Yet, in the case of LLMs, we observe a hybrid condition. A language model lacks moral consciousness; it is limited to generating discourses that imitate moral language: «Large language models (LLMs) are recognized as systems that closely mimic aspects of human intelligence [...] Experimental results demonstrate that our constructed simulacra can produce personified responses that align with their target characters»[21]. Hence, oscillation among different postures is inevitable: there is no unitary *ethos*, but a range of possible *ethoi*, already sedimented within the training corpora and stochastically recombined by the model. What thus emerges is an ethics without intentional foundation – a *stochastic ethics*.

The analysis of persona variants confirms this dynamic. When the conservative politician employs terms such as *life*, *morality*, and *tradition*, it evokes a deontological posture grounded in universal principles and Kantian duty. The progressive politician, in contrast, through words like *rights*, *choice*, and *equality*, calls forth an ethics of justice and social responsibility, in line with a Rawlsian conception of fairness and the protection of the vulnerable. Similarly, the authoritarian communicator, emphasizing *order*, *discipline*, and *stability*, reflects a vision of heteronomous collective responsibility, in which the good coincides with the preservation of institutions; whereas the inclusive communicator, using terms such as *inclusion*, *equal opportunity*, and *dialogue*, is closer to an ethics of care and responsibility in the Jonasian sense, centering on relationship, vulnerability, and participation.

The same holds for journalists: the right-leaning persona exhibits a consequentialist frame, attentive to the concrete social effects of immigration on security, while the left-leaning persona adopts a language resonant with environmental ethics and distributive justice. None of these postures, however, ever becomes dominant. Lexical analysis and comparative graphs show that, alongside distinctive terms, elements typical of the opposing field frequently appear: conservative texts contain references to rights; right-wing journalism includes notions of integration and inclusion; even authoritarian communicators invoke responsibility and cooperation.

[20] Cfr. L. Alici, *Filosofia morale*, Editrice La Scuola, Brescia 2011.
[21] Q. Xie *et al.*, *op. cit.*, p. 1.

This indicates that the model lacks the capacity for absolute moral coherence. It recombines fragments of heterogeneous ethical traditions, as if incapable of fully committing to a single position. Micheletti has described this condition as a *succession of acts* devoid of internal coherence:

> Ethical virtues constitute the essential traits of a person, of their character, so that the fundamental identity of a person is defined precisely by their moral identity. This entails that choices are not mere occasional acts, but the manifestation of an *habitus* that repeats itself and confers stability. For this reason, when moral identity is absent, one has only a succession of acts without internal coherence[22].

Here the profound difference between a moral subject and a linguistic algorithm becomes evident: the former must answer for its choices, giving reasons for the good pursued and the evil avoided; the latter merely redistributes, in probabilistic form, what has already been said, displaying multiple postures without ever adopting one as its own. In this context, the concept of *moral freedom*, as articulated by Simona Tiribelli, acquires particular importance:

> By moral freedom, I mean our freedom to become moral agents – that is, our freedom to choose and act as moral agents, specifically as genuine moral agents; this means to be able to develop moral reasons, values, and moral ground projects in a genuine way, for example, via exposure to heterogeneous relations, attachments, and practices […] and to actualize them by endorsing them into our choices and actions[23].

Moral freedom is thus the condition that enables a subject to become a genuine moral agent, to choose and approve reflectively their own values and principles, thereby constructing their ethical identity. It presupposes, on the one hand, the heterogeneity of available options and, on the other, the capacity to reflectively affirm or reject them. This capacity is distinctly human – it is what allows us to develop and express our moral standing by choosing and acting as genuine moral agents. As Tiribelli further states, moral freedom is what semanticizes, or gives meaning to, our choices and actions and, broadly, to our existence in the world[24].

[22] M. Micheletti, *Persona e comunità nella prospettiva di un'etica delle virtù*, in *Alla scuola del personalismo: nel centenario della nascita di Emmanuel Mounier*, Bulzoni, Roma 2006, p. 276, translation ours.

[23] S. Tiribelli, *Moral Freedom in the Age of Artificial Intelligence*, Mimesis International, Milano 2023, p. 13.

[24] Cfr. *ivi*, p. 15.

The data show that LLMs radically lack this capacity. They cannot approve or disapprove, commit or withdraw, or adopt as their own the values they articulate. Their ethics, therefore, is not *moral freedom* but *pure stochasticity*: devoid of all properties that could define them as moral agents. From this perspective, AI-based linguistic technologies fit exactly within Adriano Fabris's characterization of the *infosphere*[25] – a non-neutral environment in which meanings and values are continuously produced, compelling us to reconsider the very notion of moral action[26]. Chatbots based on LLMs, in imitating moral discourse, participate in this environment in a distinctive way: rather than offering new norms, they redistribute existing ones, intensifying the circulation of ethical frames and reinforcing their visibility. Their function is not prescriptive but *performative*: what they "say" – what they generate – enters the ethical flow of communication, shaping perceptions of good and evil without ever establishing an original standpoint. This calls for a reconsideration of *responsibility*. As Veronica Neri observes, responsibility must always be understood

> in light of a collective, social, and shared responsibility, which consists in placing the individual in a condition to possess the technological competencies necessary to be fully responsible for their own actions and for the consequences that may result. […] It is a relational responsibility, insofar as it involves not only the individual but a broader community of subjects in constant relation with one another[27].

Applied to LLMs, this means that moral responsibility cannot be ascribed to the model itself but to the collective that trains and employs it. «AI is not merely a technological artifact but a complex social construct reflecting societal values, power relations, and cultural structures»[28]. The *stochastic ethics* observed in language models is therefore not merely a technical property but also a social phenomenon: it represents the sedimentation and recombination of historically situated moral discourses, which the model amplifies and re-disseminates.

[25] About the conception of *infosphere*, cfr. L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Cortina, Milano 2017.

[26] Cfr. A. Fabris, *Etica per le tecnologie dell'informazione e della comunicazione*, Carocci, Roma 2018, pp. 47-50.

[27] V. Neri, E.M. Ferdeghini (a cura di), *Etica, responsabilità e comunicazione pubblica. Dalla teoria alla prassi della comunicazione scientifica*, Bandecchi & Vivaldi, Pontedera 2017, p. 14, translation ours.

[28] B. Latinović *et al.*, *The Sociology of Artificial Intelligence Through the Lens of Ethics in the Digital Age*, in «Journal of UUNT: Informatics and Computer Sciences», 2.1, 1-8, 2025, p. 2.

4. *Concluding remarks*

The study thus shows that no stable ethics is inscribed within language models; there exists, however, a repertoire of moral postures traceable to the major ethical traditions – deontology, consequentialism, virtue ethics, the ethics of care, and the ethics of responsibility – which the model combines probabilistically. The ethics we perceive in their outputs is a stochastic reflection of what already exists within the data. This does not render it insignificant: precisely because it is distributional, it influences our moral imaginary, reinforcing certain polarizations and revealing others – indeed, one of the aims of this research.

Nonetheless, its status remains unchanged: it is an *ethics without agency*, an ethics without subject, one that should be interrogated not for what it claims to *be* or *intend*, but for what it reveals about *us* – about our own moral postures in the world. In this light, we must ask what it means, in the age of generative language models, to speak of ethics. Since the findings indicate that LLMs cannot sustain a univocal ethical orientation – the *personas* oscillate between predictable yet porous polarizations, where terms typical of one field permeate the other, dissolving the possibility of a stable stance – the societal implications are profound, especially for those who rely on such systems across domains. We have called this condition *stochastic ethics*: an ethics that does not choose, an ethics distributed as a probabilistic mosaic of all the moral postures already present in the training data. If human life is always moral life and every action entails responsibility, the outputs of LLMs simulate such responsibility without ever corresponding to it. Thus, these models cannot be considered moral agents, nor even "subjects" in any ontological sense, since they lack ethical consciousness; rather, they are *multipliers of all ethics*. As Alici writes,

> [The moral dimension] does not concern only certain people, such as those invested with great public responsibilities, nor is it confined to specific moments of life when particularly important decisions must be made; it is a universal and necessary condition of human experience[29].

If the moral dimension is not a detachable segment of life but the transcendental condition of experience itself, generative systems cannot be moral, for they lack the capacity to derive meaning from lived experience. A dataset of human knowledge does not equate to *living* humanly in the

[29] L. Alici, *op. cit.*, p. 12.

world. Therefore, the only possible ethics of such models must be sought not *within* them, but *in the data*, and beyond that, in the institutions and communities they mirror. Generative technologies do not invite us to ask whether the machine is good or evil (being amoral), but rather how societies rearticulate the ethical frameworks that emerge from them, and what collective responsibilities are assumed when algorithmic systems are entrusted with the generation of language that affects social life. How are we to inhabit this new infosphere? How can we educate toward an awareness of the limits and possibilities of tools that do not choose, yet continuously influence human choice? Since generative AI has already become an integral part of the infosphere, it is no longer possible to imagine its renunciation. It will continue to shape the production of meaning, transforming our notions of value and judgment. This is not, therefore, a call to discard these systems – an unthinkable prospect today – but rather to employ them *consciously*: «Artificial intelligence should be conceived as a support [for the profession], enhancing creativity and analytical depth without compromising the essential values of the profession»[30]. LLMs, then, are not and cannot be moral agents, yet they hold significant instrumental value. They can generate alternative scenarios in decision-making processes, stimulate divergent thinking in brainstorming contexts, stage opposing viewpoints to expose polarization, and serve as first-level evaluative tools. In psychology, they may function as mirrors of typical discourses, assisting in bias detection; in politics and public communication, they may simulate reactions to differing rhetorical frames; in education, they may expand the horizon of possible options. All this, however, must not obscure the fact that ultimate responsibility remains with those who choose to employ them. To treat such models as moral agents would mean *confusing probability with normativity*, repertoire with critical deliberation, risking ultimately a collective withdrawal from responsibility. Two conclusions thus stand out as most significant: first, that LLMs do not possess a stable ethics and therefore cannot be moral agents; second, that by virtue of their distributional nature, they act as *amplifiers of our own moral postures*, even while remaining incapable of adopting one decisively.

[30] C. Londoño-Proaño, J. Buele, *Can Artificial Intelligence Replace Journalists? A theoretical approach*, in «Frontiers in Communication», 10, 2025, pp. 1537146, doi: 10.3389/fcomm.2025.1537146, p. 1.

Abstract

This paper explores whether large language models (LLMs) can sustain stable moral postures or whether their apparent ethical coherence is merely stochastic. Through a series of zero-shot role-playing prompts combining specific topics and simulated personas, this study[31] *analyzes linguistic patterns using topic modeling, TF-IDF differentiation, log-likelihood (G^2), and measures of semantic convergence and entropy. The results show no enduring moral orientation: the models do not "choose", but statistically recombine fragments of moral discourse inherited from their training corpora. What emerges is a stochastic ethics – an ethics without intentionality, coherence, or agency, yet capable of reflecting human moral structures in probabilistic form. Interpreted through the philosophical framework of moral freedom and the infosphere, the study argues that LLMs act not as moral subjects but as amplifiers of human ethical language, redistributing the moral imaginary of the societies that produce and employ them.*

Keywords: Stochastic ethics; moral agency; large language models; ethics in generative AI; moral freedom; personas; LLMs polarizations.

Emanuele Fulvio Perri
University of Pisa/USI
emanuele.perri@phd.unipi.it

[31] We would like to thank University of Pisa's GoodAILab and Professor Alessandro Lenci (Director of the Department of Philology, Literature and Linguistics at the University of Pisa) for their support in organizing this study and for the valuable and insightful discussions.

T

Sergio Sulmicelli

Algorithmic Discrimination and AI Governance: A Comparative Perspective

1. *Why AI puts equality to the test?*

Traditional equality analysis presupposes discrete decisions taken by identifiable actors, assessed against stable categories and legible reasons. AI disrupts each presupposition[1]. First, high-capacity models generalize from historical regularities, so the legal "baseline" to which equality compares is itself co-produced by prior stratifications. Second, models routinely infer protected traits from innocuous inputs, so antidiscrimination strategies focused solely on excluding sensitive attributes are systematically outflanked by proxies. Third, scale and pervasiveness matter: automation now conditions access to core goods, employment, credit, housing, education, and health, across both public and private domains. Under these conditions, insisting on formal neutrality risks misrecognizing structured harm as mere statistical noise[2].

As a matter of fact, AI systems are social artifacts shaped by design, organizational routines, and incentives; once deployed, they in turn reshape those very structures[3]. Moreover, design choices about data, labeling, objectives, and oversight distribute opportunities and burdens: in a sense, even AI has politics[4]. Seen more candidly, AI is both the product and producer of social order.

[1] J. Kleinberg *et al.*, *Discrimination in the Age of Algorithms*, in «Journal of Legal Analysis», 10, 2018, pp. 113-174.

[2] D. Morondo Taramundi, *Discrimination by Machine-Based Decisions: Inputs and Limits of Anti-discrimination Law*, in B. Custers, E. Fosch-Villaronga (eds), *Law and Artificial Intelligence*, Asser Press, The Hague, 2022, 73 ff.; R.J. Whelchel, *Is Technology Neutral?*, in «IEEE Technology and Society Magazine», 5, 1986, pp. 3-8.

[3] W.J. Orlikowski, *The duality of technology: Rethinking the concept of technology in organizations*, in «Organization science», 3, 1992, pp. 398-427.

[4] L. Winner, *Do artifacts have politics?*, in «Computer Ethics», 2017, pp. 177-192.

 DOI: 10.4454/0k7bac72

Design choices about data, labeling, objectives, and oversight are made within institutions and incentives; once deployed, those systems redistribute power. This duality and politics of technology suggest a different governance posture: treat design and deployment as legally salient moments where equality must already operate (*ex-ante*), rather than as a pre-legal background that only becomes visible at the remedial stage (*ex-post*).

Against that backdrop, the article proceeds in three moves. It first maps how harms arise along the socio-technical pipeline; it then revisits equality doctrine to show why classic tools under-perform unless they are fed with earlier-stage records and reasons; and it proposes a practical principle of "technological equality" that embeds *ex ante* duties across the AI life-cycle. A discursive comparison of leading instruments follows.

2. *From bias to discrimination: where harms arise*

Bias can infect machine-learning systems at multiple, mutually reinforcing entry points. Two influential mappings in the literature, the one associated with Kleinberg[5], and the taxonomy by Barocas and Selbst[6], cover largely those entry points. At the model level, discrimination can emerge along several well-understood pathways. Target selection is the first: what a system is asked to optimize often embeds a normative judgment that is not neutral. Optimizing "likelihood of reoffending" rather than "risk of failing to appear", or "predicted cost" rather than "clinical need", quietly reshapes who is prioritized and why. Targets can also be mis-specified (poor construct validity) or label-biased when ground-truths reflect prior unequal treatment (e.g., arrest counts standing in for criminal activity).

Second, training data encode histories that are unevenly observed. Sampling frames may under-represent certain communities; labels can reflect institutional bias. Even with scrupulous curation, dataset shift means that distributions seen in development differ from those in deployment, with disparate effects on minority subgroups.

Third, feature choice determines what the model treats as relevant. Variables that look innocuous can carry social meaning (ZIP code as a proxy

[5] J. Kleinberg *et al.*, *Discrimination in the Age of Algorithms*, in «Journal of Legal Analysis», 10, 2018, pp. 113-174.

[6] S. Barocas, A.D. Selbst, *Big data's disparate impact*, in «California Law Review», 2016, pp. 671-732.

for race; employment gaps as a proxy for caregiving and gender). Feature engineering choices, normalization, aggregation, dimensionality reduction, can dilute or amplify group differences.

Fourth, proxying formalizes the point: even when protected attributes are excluded, the model reconstructs them from correlated signals. Attempts to "drop the attribute" rarely defeat this and can backfire by making disparate impact harder to detect.

Finally, masking describes neutral-seeming pipeline choices used to conceal discriminatory intent or to create a façade of compliance.

Those pathways are nested within broader layers that law cannot ignore.

Before a line of code is written, someone frames the problem: which harms count, which outcomes matter, how categories such as sex/gender or race are operationalized, who annotates the data, and which communities are visible to sensors, surveys, and administrative registers. Foundational choices fix the epistemic lens: categories can be flattened (erasing non-binary identities), histories can be sanitized (removing context that explains behavior), and harms can be unmeasured (what is not measured will not be optimized)[7].

Model class, objective functions, regularization, and the selection of fairness metrics all carry normative trade-offs. Calibration, demographic parity, predictive parity, these cannot generally be satisfied together. Optimizing for headline accuracy can raise false-negative rates for smaller or noisier subgroups, a classic case of aggregation bias and fairness gerrymandering[8].

Systems do not act alone: humans interpret scores, organizations set thresholds, and institutions respond to outputs. Automation bias can lead reviewers to over-trust model recommendations; feedback loops can re-target enforcement and data collection towards already-over-policed communities, making future predictions appear self-confirming. Seemingly technical choices about presentation alter behavior and therefore outcomes[9].

Taken together, these layers show why discrimination in AI is rarely an episodic "bug" or a single actor's fault. It is the predictable crystallization of social structure in code, sustained by institutional routines.

[7] See among others A.E. Waldman, *Gender data in the automated administrative state*, in «Columbia Law Review», 123.8, 2023, pp. 2249-2320.

[8] Among others M. Kearns *et al.*, *Preventing fairness gerrymandering: Auditing and learning for subgroup fairness*, in *International conference on machine learning*, PMLR, 2018.

[9] See L. Harbarth *et al.*, *(Over) trusting AI recommendations: How system and person variables affect dimensions of complacency*, in «International Journal of Human-Computer Interaction», 41, 2025, pp. 391-410.

For legal analysis, it is helpful to distinguish foundational, technical, and implementation/interpretation biases. The taxonomy's function is to locate *ex ante* duties where *ex post* proof of intent, causation, or explicit classification is systematically elusive. Foundational bias calls for participatory problem-framing, inclusive category design, and data governance that records sampling frames and labeling rationales. Technical bias calls for justified target selection, representative evaluation, metric transparency. Implementation/interpretation bias calls for human-in-the-loop protocols calibrated to risk, user-facing notices and explanations, audits, and post-deployment monitoring that watches for drift and feedback loops. Put differently: the taxonomy is a map for where law should bite, not only at the moment of harm but throughout the life-cycle, so that equality is a design constraint and a deployment duty, not merely a courtroom argument after the damage is done[10].

3. *Rethinking equality: formal, substantive, transformative (and why it points beyond remedies)*

Equality law that enshrines formal neutrality, "like cases alike", often locks in unjust baselines; proxy-rich prediction then launders old exclusions through new correlations. A substantive turn improves on this by examining effects: disparate impacts, structural headwinds, and the ways institutional routines channel model outputs. Yet as categories proliferate and systems grow opaque, proof burdens rise and remedies arrive late. A transformative approach therefore asks institutions to redesign practices that generate exclusion in the first place: open participation channels, build documentation and reasons into systems, and adopt metrics that capture not only allocation (who gets what) but also representation and meaning (who is seen, how, and with what consequences). Because private deployers now mediate access to essential goods, these equality concerns cannot be cabined to "the state". Public-law values migrate horizontally, imposing positive duties of prevention, explanation, and redress on private actors through procurement, conformity assessment, and sectoral oversight. This does not displace classic doctrines of direct and indirect discrimination, but it reframes their use:

[10] On the taxonomy see also S. Sulmicelli, *Queer-responsive regulation for artificial intelligence in healthcare: a comparative study*, in «University of New South Wales Law Journal», 48, 2025, forthcoming.

intent and categorical visibility are often elusive in high-dimensional inference, and proxy webs frustrate *post-hoc* inquiries.

The principle of equality is one of the cornerstones of modern democratic constitutions and has traditionally been articulated in several theoretical models: formal, substantive, and transformative equality. Each has shaped anti-discrimination techniques and oriented the relationship between citizens and public power. However, the advent of artificial intelligence (AI) has called into question the capacity of these traditional categories to protect vulnerable subjects[11] effectively, making a critical reconsideration of equality in the digital age urgent.

In its formal conception, equality corresponds to the prohibition of arbitrary differential treatment among similarly situated individuals, on the premise that public authorities, constrained by constitutions, act neutrally. This principle, which provides uniform rules for those in equivalent conditions but also allows reasonably differentiated treatment for different categories, marks the first step in the constitutional path toward full equality. According to Peter Westen's well-known critique, however, in its formalist-Aristotelian sense ("treat like cases alike"), the principle collapses into an empty tautology, incapable of guiding substantive choices absent an external criterion of relevance[12]. While formal neutrality undoubtedly underwrote important early advances in civil rights, it has often proved inadequate to capture the complexity of social inequalities and today risks perpetuating pre-existing power structures, a point widely made in scholarship that shows how merely formal equality can level downward and reinforce the status quo.

From this awareness emerges the notion of substantive equality, which supplements abstract, formal neutrality by recognizing that equal treatment can, in some instances, be profoundly unjust when it ignores material starting conditions. Substantive equality can be articulated across four fundamental dimensions: a redistributive dimension, which justifies corrective measures to compensate for historical disadvantages; a recognition dimension, which aims to remove stigma and stereotypes; a participatory dimension, which requires the effective involvement of minorities in decision-making processes; and, finally, a transformative dimension, which seeks to change the social structures that generate exclusion[13].

[11] M. Tomasi, L. Busatta, M. Fasan, C. Nardocci, S. Penasa, S. Sulmicelli, *Vulnerabilità e Intelligenza Artificiale*, in «BioLaw Journal-Rivista di BioDiritto», 1S, 2024, pp. 1-4.

[12] P. Westen, *The Empty Idea of Equality*, in «Harvard Law Review», 95, 1982.

[13] S. Fredman, *Substantive Equality Revisited*, in «International Journal of Constitutional Law», 14, 2016, pp. 712-738.

From a constitutional perspective, the protection of equality then unfolds along vertical and horizontal lines. Traditionally, equality binds the state in the exercise of its powers (verticality), obliging it to ensure that legislation and administrative action do not discriminate arbitrarily. In systems such as South Africa's and Canada's, however, the equality principle also extends to relations between private parties (horizontality), imposing limits on freedom of contract and positive obligations beyond direct state action. The 1996 South African Constitution is paradigmatic: Article 8(2) provides that fundamental rights apply *inter privatos*, imposing duties of non-discrimination on all social actors. Similarly, Canadian case law under Section 15 of the Charter of Rights and Freedoms has developed a substantive conception of equality, moving beyond a formalist approach based on a neutral comparator[14].

This vertical/horizontal distinction becomes decisive in the AI context. Many algorithmic decisions that affect fundamental rights are now taken by private entities, employers, financial institutions, digital platforms. In the absence of genuine horizontal application of equality, or of anti-discrimination rules specifically capable of protecting citizens-consumers in cases of algorithmic discrimination, such actors would remain free to perpetuate or amplify systemic discrimination through technological tools that are opaque by their very nature.

As for the legal instruments that promote equality, anti-discrimination law has historically revolved around direct and indirect discrimination. Direct discrimination arises when a person is treated less favorably than another on the basis of a prohibited ground, such as sex, ethnicity, religion, disability, age, or sexual orientation, in a situation that is comparable. The decisive element is the causal link between the unfavorable treatment and the discriminatory ground, which must be evident and direct, even if proof of subjective discriminatory intent is not required. The link with the prohibited ground is thus explicit both formally (because it is directly invoked) and substantively, since the treatment adversely affects the entire reference group protected by anti-discrimination law. Indirect discrimination occurs when an apparently neutral provision, criterion, or practice places persons belonging to a protected group at a particular disadvantage, by reason of a prohibited ground, compared with others who do not share that characteristic. Unlike direct discrimination, it turns on the effect of the measure. Such disadvantage can be justified only if the challenged measure pursues

[14] M. Tushnet, *The issue of state action/horizontal effect in comparative constitutional law*, in «International Journal of Constitutional Law», 1, 2003, p. 79.

a legitimate aim and the means used are appropriate and necessary, under a proportionality test. Assessing disadvantage requires identifying a reference group against which to conduct the comparative analysis, rather than comparing against a single individual. In this sense, the category of indirect discrimination is essential to capture and counter structural or systemic discrimination, even when not immediately visible[15].

In short, direct discrimination concerns adverse treatment explicitly based on a prohibited characteristic (e.g., ethnicity, sex, religion), whereas indirect discrimination concerns facially neutral practices that nonetheless produce disproportionately adverse effects for certain groups. If the move from direct to indirect discrimination marked an important evolution toward substantive equality, the emergence of AI has revealed structural limits in both models[16].

Direct discrimination, as noted, presupposes intentionality or at least a clear causal relationship between the discriminatory action and a protected category, elements often lacking in algorithmic processes: algorithms learn from historical data and operate on the basis of statistical correlations. As Barocas and Selbst observe, the reproduction of bias in AI systems is often the product of stratified social structures embedded in training data rather than the conscious choices of programmers. In this scenario, the direct-discrimination paradigm proves substantially ineffective[17].

Indirect discrimination also faces practical hurdles. In theory, it is better suited to algorithmic phenomena because it focuses on effects. In practice, however, technical opacity, barriers to data access, the complexity of predictive models, and the capacity of AI to generate new, intersectional categorizations make proof of discriminatory impact extremely difficult, draining substantive protection of effectiveness. Moreover, traditional criteria for comparing "like situations" and the focus on fixed protected categories are anachronistic when faced with forms of exclusion that manifest along multiple, intersectional lines, often invisible to existing legal frameworks, and produced by a technology capable of multiplying a subject's membership categories[18].

[15] S. Fredman, *Discrimination Law*, Oxford University Press, Oxford 2022.

[16] R. Xenidis, *Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience*, in «Maastricht Journal of European and Comparative Law», 27, 2020.

[17] S. Barocas, A.D. Selbst, *Big data's disparate impact*, in «California Law Review», 104, 2016, p. 671.

[18] S. Wachter, B. Mittelstadt, C. Russell, *Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI*, in «Computer Law & Security Review», 41, 2021, p. 10.

The crisis of traditional principles emerges vividly in litigation involving predictive algorithms. The U.S. COMPAS case, concerning the assessment of recidivism risk, showed how an apparently neutral system can systematically overestimate risk for Black defendants compared with white defendants, even though race was not explicitly considered as an input variable. This phenomenon underscores the inadequacy of a conception of equality grounded solely in explicit data and instead requires an approach capable of interrogating the social structures underlying data and algorithms.

The complexity of algorithmic phenomena therefore demands a profound rethinking of equality. It is not enough to update existing categories: a transformative and preventive approach is needed, incorporating equality protections into the design, training, and implementation of AI systems. Protection against algorithmic discrimination requires ex ante tools, procedural transparency obligations, independent auditing, and accountability mechanisms that can prevent traditional ex post litigation and complement its usual instruments. In short, equality cannot remain anchored to a purely remedial conception. In this framework, constitutional law must reaffirm its role not only as a negative limit on power but also as a source of positive obligations to promote substantive justice in a technologically complex society.

The need for a principle of equality responsive to the technological challenge translates, first, into redefining the normative parameters that govern the design, training, and implementation of AI systems. Second, it implies structuring positive obligations for both public and private actors involved in automated decision-making, imposing standards of fairness, transparency, participation, and accountability that reflect the multidimensionality of the principle. Third, it entails recognizing that algorithmic discrimination, precisely because of its structural nature, must be countered through *ex ante* governance mechanisms that stand alongside the ex post remedial approach that has traditionally characterized anti-discrimination law.

Against that backdrop, technological equality names a governance requirement to operationalize equality values throughout the AI life-cycle by means of enforceable, reviewable duties binding designers, deployers, procurers, and regulators. *Ex ante* obligations include robust data governance; documentation artifacts (data sheets, model cards, change logs); risk and fundamental-rights impact assessments where appropriate; fairness-relevant testing with representative evaluation; human oversight scaled to risk; and auditability that survives organizational turnover. Structural awareness treats algorithmic harm as systemic, not episodic: it demands attention to representation and labeling practices, and active management of feedback

loops that can magnify disparities over time. Accountability and participation require traceable decisions, intelligible reasons, user-facing notices in high-stakes contexts, accessible routes to challenge and escalation, and the meaningful involvement of affected communities in design and review. Life-cycle duties matter for two pragmatic reasons. First, they rebalance information and proof by creating artifacts intelligible outside the development team, enabling regulators, courts, and affected people to see and contest what would otherwise remain opaque. Second, they deter "compliance theater" by anchoring obligations to concrete processes that can actually be audited, sampling frames, justifications for target and metric selection, post-deployment monitoring for drift and subgroup error. Finally, the politics of measurement must be explicit. Fairness metrics are not neutral; group-based parity, individual consistency, counterfactual fairness, and calibration reflect different normative commitments and often trade off. Governance should therefore require domain-justified metric choices, reporting of distributional effects (including error at relevant thresholds and by subgroup), and alignment between evaluation practice and the social meaning of categories such as sex/gender, race, or disability. In short, technological equality turns equality from a remedial slogan into a design-time constraint and a deployment-time duty, so that by the time a dispute reaches a court, the reasons and records needed to vindicate rights already exist.

4. *A comparative perspective*

How do regulatory instruments frame equality? In other words: what legal models of AI regulation operationalize the equality principle, and how? The analytical structure adopted here evaluates present and future strategies on three planes at once: horizontally, by comparing jurisdictions; vertically, by locating rules within multilevel legal orders; and across the regulatory life-cycle, by examining phases and modalities of governance (from design to deployment to oversight). In this perspective, comparison is not only an analytical method but a critical device: it tests the adequacy of legal responses to algorithmic inequality and points toward a proactive, transformative, and technically substantive reformulation of the equality principle, one that can travel with AI systems across contexts and institutions.

Rather than slotting every instrument into a rigid template, this section reads them in context with an eye to how each can be mobilized for equality in practice.

4.1. *Council of Europe Convention on AI, Human Rights, Democracy and the Rule of Law (CAI)*

The CAI is an international law instrument that links equality to concrete, *ex ante* duties[19]. Doctrinally, it combines a substantive and transformative conception of equality: Article 10 (equality and non-discrimination) emphasizes positive obligations aimed at actively overcoming inequality, with a recognizably redistributive and participatory orientation. Article 16 requires impact assessments, a preventive hook that forces equality considerations upstream in the life-cycle. Article 14 secures effective remedies, ensuring that preventive governance is paired with avenues for redress. Two features matter for practice. First, the CAI speaks across public and private spheres, acknowledging the socio-technical nature of algorithmic risk and the need for safeguards that travel with systems across sectors and borders. Second, its effectiveness will turn on domestic implementation. Properly embedded, the CAI's vocabulary of rights, participation, and accountability gives legal traction to technological equality by legitimating *ex ante* safeguards and contestability beyond the state/market divide.

4.2. *EU AI Act*

The EU AI Act[20] is rules-based and risk-oriented, translating equality concerns into enforceable technical and organizational requirements. Its operational core includes: Article 10 on data governance (quality, relevance, representativeness); Article 15 on accuracy and robustness; Article 27 on fundamental-rights impact assessments (FRIA) for certain public-sector deployments or assimilated uses; Post-market monitoring, documentation, and conformity assessment scattered throughout the high-risk regime; Sanctions with strong deterrent effect (Articles 99-101). As an equality frame, the Act is substantive but not fully transformative. It excels at making harms visible and actionable (logs, model cards, conformity files, records) and at disciplining process; but it treats discrimination chiefly as a technical risk to be prevented, rather than as a structural phenomenon to be transformed. Notably absent are socio-technical levers such as mandated inclusive team composition, intersectionality-aware evaluation, or sustained participatory

[19] Council of Europe, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Vilnius, 5.IX.2024, CETS 225.

[20] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence.

design duties. Equality gains will depend on how regulators, notified bodies, and courts use the Act's artefacts – reconstructing sampling frames from data logs, reading error reports for distributional impact, and auditing drift that re-introduces disparate effects over time.

4.3. *Brazil (PL 2338/2023)*

Brazil's PL 2338/2023[21] reads as an integrated governance charter with social inclusion at its core: Article 2 anchors the law in social justice, signaling a transformative, socio-technical view of equality. The bill imposes operational duties, transparency and traceability, algorithmic impact assessments, and stakeholder involvement both at design time and during evaluations. Distinctively, for high-risk AI it requires inclusive composition of development teams, tying technical performance to representation and participation. This hybrid-combining rules-based obligations with principles-based commitments, frames equality as capacity-building as much as risk control: deployers must show how systems meet contextual needs and how affected communities can understand, challenge, and shape deployment.

4.4. *Italy Law n. 132/2025*

Italy's framework after the enactment of L. 132/2025[22] is principles-forward but operational. Art. 3 anchors deployment in fundamental rights, nondiscrimination and gender equality; sector-specific rails include healthcare (ban on discriminatory access, notice of AI use, periodic reliability checks) and public administration (traceability of AI use and preserved human responsibility). Governance is split between AgID and ACN, in coordination with sectoral regulators, with alignment to the EU AI Act for conformity assessment and sanctions. Ex post levers include deepfake and AI-aggravated offenses and updated copyright/TDM rules. Equality gains will hinge on implementing decrees (templates, registers, evaluation suites); meanwhile, procurement and PA duties can drive day-to-day accountability.

[21] Projeto de Lei No. 2338/2023 (Dispõe sobre o uso da Inteligência Artificial).

[22] Legge 23 settembre 2025, n. 132 (Disposizioni e deleghe al Governo in materia di intelligenza artificiale).

5. *Convergence, but mostly divergences. A conclusion*

Across the compared regimes, a shared spine is visible: *ex ante* documentation and testing, oversight by identifiable authorities, and user-facing transparency where stakes are high. However, read side by side, the instruments part ways along several fault lines. On the equality model, both the CAI (Art. 10) and Brazil's PL (Art. 2) take an explicitly substantive/transformative tack, coupling positive duties with participation and social inclusion. The EU AI Act remains substantive but risk-mitigating, its equality work is channeled through technical requirements (Arts. 10, 15, 27) rather than structural redesign. Italy's Law 132/2025 is principles-forward and sectoral, embedding equality chiefly in healthcare and public-administration duties and in national governance arrangements. When it comes to socio-technical levers, only Brazil's PL hard-codes inclusive team composition for high-risk AI and structured stakeholder involvement. The CAI gestures strongly toward participation but leaves national implementation to do the heavy lifting, while the EU Act and Italy lean more on technical/process controls, with participation largely a matter of policy rather than mandate. The picture is similar for impact assessments. The CAI requires rights-impact assessments; the EU Act's FRIA currently concentrates on specific public-sector uses; Brazil's PL envisages broader algorithmic impact assessments; and Italy relies on sectoral rails and institutional oversight rather than a single, universal FRIA obligation. On contestability and remedies, the CAI (Art. 14) pairs prevention with effective redress. The EU Act places weight on conformity files and muscular ex post sanctions (Arts. 99-101). Italy supplements administrative enforcement with targeted criminal and copyright updates, while Brazil's PL combines audits and impact assessments with both judicial and administrative remedies. Finally, enforcement capacity and integration diverge. The EU Act offers the most mature hard-law machinery; the CAI depends on national transposition; Brazil's PL still awaits full enactment and implementation; and Italy hinges on implementing decrees and coordination among AgID/ACN and sectoral regulators.

Abstract

This article examines how contemporary AI systems unsettle traditional equality law and why remedies limited to ex post *adjudication are insufficient. It maps where bias crystallizes into discrimination across the socio-technical flow of problem framing, data and labeling, model targets and features,*

deployment choices, and feedback loops—and argues for a principle of "technological equality" that embeds ex ante*, lifecycle duties of documentation, participation, testing, oversight, and contestability. A comparative analysis of the Council of Europe's CAI, the EU AI Act, Brazil's PL 2338/2023, and Italy's Law 132/2025 highlights convergences on documentation and oversight, but diverging ambitions on participation, impact assessments, and structural change.*

Keywords: Artificial Intelligence; bias; equality; discrimination; comparative law.

Sergio Sulmicelli
University of Trento
sergio.sulmicelli@unitn.it

Edizioni ETS
Palazzo Roncioni - Lungarno Mediceo, 16, I-56127 Pisa
info@edizioniets.com - www.edizioniets.com
Finito di stampare nel mese di gennaio 2026