

RIVISTA DI DIRITTI COMPARATI



Rivista Quadrimestrale
SPECIAL ISSUE VIII (2026)

Rivista di diritti comparati

Rivista quadrimestrale

Special Issue VIII/2026

DIREZIONE

Andrea Buratti – Università di Roma “Tor Vergata”
Giacomo Delledonne – Scuola Universitaria Superiore Sant’Anna di Pisa
Alessandra Di Martino – Sapienza Università di Roma
Cristina Fasone – LUISS Guido Carli
Giuseppe Martinico – Scuola Universitaria Superiore Sant’Anna di Pisa
Anna Mastromarino – Università di Torino
Oreste Pollicino – Università commerciale “Bocconi” di Milano
Giorgio Repetto – Università di Perugia
Francesco Saitto – Sapienza Università di Roma
Raffaele Torino – Università Roma Tre

COMITATO SCIENTIFICO

Richard Albert (Texas University, Austin), Vittoria Barsotti (Univ. Firenze), Francesco Bilancia (Univ. Sapienza Roma), Roberto Bin (Univ. Ferrara), Giuseppe Bronzini (Corte di cassazione), Ermanno Calzolaio (Univ. Macerata), Paolo Carrozza † (Scuola Sant’Anna, Pisa), Marta Cartabia (Univ. Bocconi), Ginevra Cerrina Feroni (Univ. Firenze), Francesco Cerrone (Univ. Perugia), Roberto Conti (Corte di cassazione), Diego Corapi (Univ. Sapienza, Roma), Barbara De Donno (Luis “Guido Carli”), Pasquale De Sena (Univ. Palermo), Giuseppe De Vergottini (Univ. Bologna), Giuseppe Franco Ferrari (Univ. Bocconi), Tommaso Edoardo Frosini (Univ. Suor Orsola Benincasa), Anna Gamper (Universität Innsbruck), Javier García Roca (Universidad Complutense de Madrid), Michele Graziadei (Univ. Torino), Peter Hay (Emory University), Nicola Lupo (Luis “Guido Carli”), Elena Malfatti (Univ. Pisa), Miguel Poiars Maduro (European University Institute), Giovanni Marini (Univ. Perugia), Francesco S. Marini (Univ. Roma Tor Vergata, Corte costituzionale), Roberto Mastroianni (Univ. Napoli Federico II), Petros Mavroidis (Columbia University, NY; Université de Neuchâtel), Antonello Miranda (Univ. Palermo), Luigi Moccia (Univ. Roma Tre), Laura Montanari, (Univ. Udine), Massimo Papa (Univ. Roma Tor Vergata), Ernst Ulrich Petersmann (European University Institute), Valeria Piccone (Corte di Cassazione), Cesare Pinelli (Univ. Sapienza, Roma), Giovanni Pitruzzella (Univ. Palermo, Corte costituzionale), Marie-Claire Ponthoreau (Université de Bordeaux), Patricia Popelier (University of Antwerp), Paolo Ridola (Univ. Sapienza, Roma), Roberto Romboli (Univ. Pisa), Antonio Ruggeri (Univ. Messina), Alejandro Saiz Arnaiz (Universitat Pompeu Fabra), Roberto Scarciglia (Univ. Trieste), Robert Schütze (Durham University, Luis “Guido Carli”), Francesco Viganò (Univ. Bocconi, Corte costituzionale)

REDAZIONE

Marco Bassini (Tilburg University) (Coordinatore), Silvia Filippi (Università di Perugia) (Coordinatrice), Nicola Cezzi (Sapienza Università di Roma), Giulia Chinaglia (Università di Torino), Giovanni De Gregorio (Católica Global School of Law), Claudio Di Maio (Università della Calabria), Alessandro Francescangeli (Università di Roma “Tor Vergata”), Alessia Fusco (Università del Piemonte Orientale), Giampiero Gioia (Sapienza Università di Roma), Umberto Lattanzi (Università Commerciale “L. Bocconi”), Matteo Monti (Università Commerciale “L. Bocconi”)

TABLE OF CONTENTS

Artificial Intelligence and Human Rights. In Search of Coherence across Fragmented Frameworks

edited by Marco Bassini

editorial assistance by Giovana Peluso Lopes

- MARCO BASSINI, *Artificial Intelligence and Human Rights. Connecting the Dots* [pp. 1-4]
- GIOVANA PELUSO LOPES, PRATIKSHA ASHOK, *Freedom of Expression and AI-Driven Content Moderation* [pp. 5-41]
- THOMAS MARGONI, LEONA KING, *Authority-Mediated Trade Secrecy in the AI Act: An Example of Functional Constitutionalisation of Transparency?* [pp. 42-70]
- FEDERICA PAOLUCCI, *The New Face of Privacy: AI, Power, and the Disappearing Private Sphere* [pp. 71-106]
- GLORIA GONZÁLEZ FUSTER, *Caught Between AI and the AI Hype: How the Right to Personal Data Protection was Ambushed* [pp. 107-124]
- AIMEN TAIMUR, *Neurorights in the Age of AI: Universalism, Cognitive Vulnerability, and the Limits of Legal Translation* [pp. 125-155]
- COSTANZA NARDOCCI, *Discrimination Revised. How AI Is Reshaping Anti-Discrimination Law* [pp. 156-192]



Contributions published in the *Rivista di Diritti Comparati* are made available free of charge and without any fees for authors. All contributions are published under a Creative Commons Attribution – NonCommercial 4.0 International (CC BY-NC 4.0) licence. This licence permits reproduction, distribution, communication to the public, and adaptation of the content for lawful, non-commercial purposes, provided that appropriate credit is given to the author and the original source is acknowledged in accordance with the terms of the licence.

For matters not expressly governed by the licence, the applicable provisions of copyright law shall apply.

The journal adheres to the *Code of Conduct and Best Practice Guidelines for Journal Editors* developed by the *Committee on Publication Ethics* (COPE). The quality and scientific rigour of contributions are ensured through differentiated blind peer-review procedures, depending on the type of submission, as specified in the journal's regulations. These procedures are carried out by a panel of experts selected on the basis of their expertise and subject to periodic renewal.

The regulations governing the review process for contributions published in the journal, as well as the editorial guidelines, are available at: www.diritticomparati.it/rivista.

I contributi pubblicati sulla *Rivista di Diritti Comparati* sono redatti dagli autori e concessi a titolo gratuito, senza oneri per gli stessi. Tutti i contributi sono pubblicati sotto licenza “Creative Commons Attribuzione – Non commerciale 4.0 International (CC BY-NC 4.0)”. La licenza consente la riproduzione, la distribuzione, la comunicazione al pubblico e la rielaborazione dei contenuti per scopi leciti e non commerciali, a condizione che sia adeguatamente attribuita la paternità dell'opera e indicata la fonte originale secondo i termini previsti dalla licenza. Per quanto non espressamente disciplinato dalla licenza, restano applicabili le normative vigenti in materia di diritto d'autore e dei diritti connessi. La rivista aderisce al *Code of Conduct and Best Practice Guidelines for Journal Editors* elaborato dal *Committee on Publication Ethics* (COPE). La qualità e il rigore scientifico dei contributi sono garantiti mediante procedure di valutazione differenziate in relazione alla tipologia di contributi, come specificato nel Regolamento, che sono affidate a un comitato di esperti selezionati secondo criteri di competenza, rotazione e aggiornamento periodico.

Il regolamento relativo alla procedura di valutazione dei contributi pubblicati nella *Rivista* e le *Norme editoriali* sono disponibili on line all'indirizzo www.diritticomparati.it/rivista.

Editore: Andrea Buratti, Giacomo Delledonne, Alessandra Di Martino, Cristina Fasone, Giuseppe Martinico, Anna Mastromarino, Oreste Pollicino, Giorgio Repetto, Francesco Saitto, Raffaele Torino

Coordinatore Editoriale: Serenella Quari

Sede: Via Roentgen, 1 – 20136 Milano / Via Cracovia, 50 – 00133 Roma

ISSN: 2532-6619

Artificial Intelligence and Human Rights. Connecting the Dots

Marco Bassini

The contributions in this special issue offer a genuinely diverse overview of the evolving relationship between artificial intelligence and the protection of human rights within contemporary legal systems, with a particular focus on the European Union. As artificial intelligence systems become increasingly pervasive in mediating access to information, processing large volumes of personal data, and structuring decision-making processes, they no longer merely function as objects of regulation; rather, they have become constitutive elements of the environments in which fundamental rights are exercised. The contributions investigate this transformation from a number of perspectives, including freedom of expression, cognitive freedom, privacy, data protection and non-discrimination, as well as transparency. In doing so, they address both the limitations and the opportunities inherent in the existing legal frameworks. Following this trajectory, the special issue aims to highlight the fragmentation across different legal frameworks, which may hinder the development of a coherent framework, while also mapping convergences and connections across domains – such as content and data – that may reveal underlying common patterns.

A first axis of inquiry concerns the impact of AI on freedom of expression and the (increasingly private) governance of online speech. Giovana Lopes and Pratiksha Ashok engage in a detailed analysis of AI-driven content moderation, highlighting its dual role as both an enabler and a constraint on free speech rights. On the one hand, automated moderation is indispensable for managing the scale of modern digital communication and for maintaining safe online environments. On the other hand, the authors demonstrate how such systems engender significant risks, including the over-blocking of lawful content, false positives and false negatives, and the opacity of decision-making processes. These dynamics not only affect the active dimension of freedom of expression – i.e., the right to speak – but also its passive dimension – i.e., the right to receive information – thereby reshaping the twofold nature of this right in the digital sphere. The contribution's key finding is that while frameworks such as the Digital Services Act and the AI Act begin to address these concerns, they must be

interpreted and applied in a manner that structurally integrates fundamental rights protection into platform governance.

The question of opacity and its implications for accountability is further developed by Thomas Margoni and Leona King, who investigate the transparency obligations established by the AI Act, particularly in relation to general-purpose AI models. Their contribution identifies a fundamental tension between transparency, understood as a precondition for the effective exercise of fundamental rights, and trade secrecy, conceived as a legitimate protection of commercial interests. Through the notion of “authority-mediated trade secrecy”, the authors propose a model in which these competing values are not treated as mutually exclusive but are reconciled through procedural mechanisms that subject claims of confidentiality to independent review. Central to their analysis is the concept of “functional constitutionalisation of transparency”, on the basis of which transparency obligations acquire constitutional force because they work as indispensable preconditions for the exercise of fundamental rights, including access to information, effective remedy, and freedom of expression. The key insight is that the future of AI regulation is best served by institutional designs that can effectively mediate between competing rights in a proportionate and reviewable manner.

Federica Paolucci provides a complementary perspective, investigating the transformation of privacy in environments increasingly permeated by AI-driven surveillance technologies, particularly biometric identification systems. Departing from the conventional understanding of privacy as a spatial or secrecy-based concept, her article proposes a reconceptualization of privacy as a relational and constitutional condition. This new conceptualisation enables individuals to engage in social and democratic life without being subjected to constant identification and monitoring. The erosion of this “intermediate space”, especially through technologies such as facial recognition, raises profound concerns not only for privacy but also for related freedoms, including freedom of assembly and expression. The analysis centres upon the landmark *Glukhin v. Russia* judgment of the European Court of Human Rights to illustrate how the use of biometric identification systems can have a chilling effect on democratic participation. This, in turn, extends the impact of surveillance beyond the individual to the collective sphere. The article ultimately underscores the need for legal frameworks that do not merely regulate the use of such technologies but actively preserve the conditions under which a right to (truly) private life remains possible.

Marco Bassini
*Artificial Intelligence and Human Rights.
Connecting the Dots*

In her article, Gloria González Fuster critically explores the evolving status of the right to personal data protection, questioning whether recent policy developments emerging in the Digital Omnibus and prioritising AI innovation risk undermining this cornerstone of EU digital constitutionalism. Her contribution revisits the right to the protection of personal data, traditionally conceived as a cornerstone of the EU's digital constitutionalism, and questions its current trajectory in light of policy developments that prioritise AI innovation. The article contends that data protection is increasingly regarded not as a safeguard but as an impediment to technological progress, resulting in pressures for its “simplification” or recalibration. This shift poses a considerable risk of subverting the fundamental rationale of the right, which lies in the framing of the parameters within which the processing of personal data can be conducted in a manner that is consistent with individual autonomy and democratic principles. The main takeaway is that while the AI Act introduces new regulatory layers, it cannot substitute for a robust commitment to data protection; rather, the two must operate in tandem if fundamental rights are to be effectively safeguarded.

Aimen Taimur's contribution extends the analysis into emerging domains, addressing the concept of neurorights and the challenges of regulating cognitive vulnerability in AI-mediated environments. As technologies increasingly enable the inference and potential manipulation of mental states, traditional legal categories such as privacy and freedom of thought are placed under strain. The article explores the efforts to articulate a set of universal principles, most notably through UNESCO's 2025 Recommendation on the Ethics of Neurotechnology, while also highlighting the difficulties of translating these principles into concrete obligations across diverse jurisdictions. A fundamental dichotomy underpinning this analysis concerns the interplay between universalism and contextualism. While the definition of global standards is imperative to address the transnational nature of artificial intelligence and neurotechnology, their implementation must be aligned with local legal and social conditions. The contribution situates neurorights within a broader continuum of fundamental rights, suggesting that the challenge lies less in the creation of new rights and more in the specification of how existing ones apply to novel forms of cognitive intrusion.

Finally, Costanza Nardocci discusses the implications of AI for anti-discrimination law, arguing that algorithmic systems give rise to forms of discrimination that cannot be fully captured by traditional legal categories. In contrast to human-driven discrimination, which is typically characterised

Marco Bassini
*Artificial Intelligence and Human Rights.
Connecting the Dots*

by identifiable intent or explicit reliance on protected characteristics, AI-based discrimination frequently operates through what the author terms “proxy discrimination” – a distinct legal category where discriminatory outcomes arise from factors that serve as predictive indicators of membership in a protected class. This complicates the establishment of causal relationships and the integration of such practices within the prevailing frameworks of direct or indirect discrimination. The result is a regulatory gap, in which cases of harm may not be subject to legal scrutiny or remedy. The contribution calls for a reconceptualisation of anti-discrimination law that takes into account the specificities of algorithmic decision-making. These include its opacity, scalability, and capacity to reproduce structural biases.

The content of these contributions is highly relevant to the current debates on AI regulation and the protection of human rights in Europe, which have been revamped after the Commission’s proposals for a Digital Omnibus and a Digital Omnibus on AI. The creation of a comprehensive digital rulebook within the EU legal system is indicative of an ambition to establish the EU as a global standard-setter in this field. However, as demonstrated by the contributions, regulation alone cannot resolve some deeper and underlying tensions. To address this gap, it is necessary to engage with the foundational principles of fundamental rights in the age of AI and the institutional mechanisms through which these rights are operationalised. By bringing together diverse perspectives, this special issue not only maps the challenges ahead but also aims to contribute to ongoing efforts to ensure that the development and deployment of AI technologies remain aligned with the core values of democratic societies.

This special issue is funded under the “RetrAIIn– Enforcing Constitutional Rights in the Age of Generative Artificial Intelligence” Dutch government-funded starter grant, for which Marco Bassini serves as principal investigator.

Marco Bassini – Assistant Professor of Fundamental Rights and Artificial Intelligence, Tilburg Institute for Law, Technology, and Society – Tilburg University, Tilburg, Netherlands (m.bassini@tilburguniversity.edu)

Freedom of Expression and AI-Driven Content Moderation*

Giovana Peluso Lopes, Pratiksha Ashok

TABLE OF CONTENTS: 1. Introduction. – 2. Content Moderation in the Digital Age. – 3. Risks of AI-Driven Content Moderation to Freedom of Expression. – 3.1. Technical Limitations of Content Moderation Systems. – 3.2. Risks to Freedom of Expression. – 3.2.1. Over-Removal of Content. – 3.2.2. False Negatives and False Positives. – 3.2.3. Proactive Moderation and Prior Restraint. – 3.2.4. Opaque Decisions and Lack of Transparency. – 4. From Algorithmic Harms to Regulatory Remedies. – 4.1. The Digital Services Act. – 4.2. The Artificial Intelligence Act. – 5. Conclusion.

1. Introduction

Human beings hold substantial interests in communicating their thoughts and in receiving the expressions of others. The significance of these interests renders coercive restrictions on communications difficult to justify, thereby providing a plausible foundation for a moral right to speak and to listen that merits legal protection. Freedom of expression is a central commitment of political liberalism, recognised as a fundamental human right across various legal frameworks, including international treaties, regional systems, and national constitutions¹, underscoring its universal dimension. It refers to the liberty to express ideas, thoughts, and opinions, as well as to seek, receive, and impart information through diverse means of communication, free from unjustified interference by public authorities.

In line with the Universal Declaration of Human Rights (UDHR), Article 19 of the International Covenant on Civil and Political Rights (ICCPR) affirms such a right as encompassing the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers and by any means of choice. Likewise, the right to freedom of expression is

* Giovana Peluso Lopes is funded by RetrAIIn (Dutch government-funded Starter Grant) and CONSENTIS (Grant Agreement No 101168011); Pratiksha Ashok is funded by AI4POL (Grant Agreement No 101177455). This article was subjected to double-blind peer review.

¹ A. Stone, *The Comparative Constitutional Law of Freedom of Expression* in T. Ginsburg and R. Dixon (eds), *Comparative Constitutional Law*, Cheltenham, 2011, p. 406 ff.

enshrined in Article 10 of the European Convention on Human Rights (ECHR), and reaffirmed in Article 11 of the Charter of Fundamental Rights of the European Union (CFREU), both of which protect the freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.

The freedom to hold opinions is often referred to as the active dimension of freedom of expression, whereas the individual freedom to form an opinion based on accessing diverse information constitutes the passive dimension of this right (also understood as freedom of information). Within the traditional European constitutional framework, freedom of expression thus encompasses at least two equivalent rights, namely, that of speakers to express themselves and that of audiences not to be deprived of valuable and useful information². In other words, freedom of expression is “double-sided”, providing for speakers’ and audiences’ rights³.

Moreover, in the European constitutional tradition, the centrality of freedom of expression does not exclude potential restrictions due to the need to prevent abuses or to balance its exercise with other rights, which are equally worthy of constitutional protection⁴. This is explicitly acknowledged, for instance, in the ECHR, according to which, since freedom of speech «carries with it special duties and responsibilities»⁵, it may be subject to certain restrictions provided by the law and necessary to guarantee the rights and reputation of others, and to protect relevant public interests such as national security, public order, or public health. This approach is not, however, reflected in other legal frameworks, such as that of the United States, where the First Amendment emphasises the active right of every individual to speak without restriction, and where interests considered legitimate counterweights to freedom of expression in other liberal democracies are instead subordinated to the overriding constitutional priorities of freedom of speech and the press⁶.

² G. De Gregorio and P. Dunn, *Artificial Intelligence and Freedom of Expression* in A. Quintavalla and J. Temperman (eds), *Artificial Intelligence and Human Rights*, Oxford, 2023, p. 77.

³ L. Woods, *Article 11* in S. Peers et al. (eds), *The EU Charter of Fundamental Rights: A Commentary*, Oxford, 2015, p. 323.

⁴ G. De Gregorio and O. Pollicino, *The European Constitutional Way to Address Disinformation in the Age of Artificial Intelligence*, in *German Law Journal*, vol. 26, 2025, p. 7.

⁵ Art. 10(2) ECHR.

⁶ F. Schauer, *The Exceptional First Amendment* in M. Ignatieff (ed), *American Exceptionalism and Human Rights*, Princeton, 2005, p. 29 ff.

While numerous justifications exist for granting freedom of expression robust legal protection in liberal societies, two approaches have, over time, gained special significance⁷. The first highlights the inherent value of freedom of expression for the individual. From this perspective, expression is regarded as a fundamental component of personal autonomy, self-development, and self-fulfilment. The capacity to articulate one's thoughts and to pursue lines of inquiry inspired by one's imagination is seen as a defining feature of a free society. On this account, freedom of expression merits protection even in the absence of wider social benefits, given its role in developing one's character and potentialities as a human being through the open exchange of ideas. The second justification underscores the instrumental value of freedom of expression for society. Here, its importance lies in, among other goods, sustaining a marketplace of ideas, enabling the pursuit of truth and, most importantly, advancing democratic self-government⁸. Indeed, the digital age has made salient one of the central purposes of freedom of speech, namely, to protect and foster a democratic culture, consisting of a culture in which individuals have a fair opportunity to participate in the forms of meaning-making and mutual influence that constitute them as individuals⁹.

The advent of digital technologies, however, has profoundly changed the ways in which individuals express themselves and enjoy freedom of expression in a democratic society¹⁰. In particular, Artificial Intelligence (AI) systems are increasingly shaping the production and dissemination of speech. This is taking place, for instance, through generative tools that expand opportunities for personal and artistic self-expression, but also, due to their low cost and accessibility, enable the mass production of persuasive and hyper-realistic content that is difficult to detect, thereby amplifying the risks of large-scale disinformation and eroding trust in authentic media¹¹.

AI is also defining the contours and limits of speech via content moderation practices. Broadly speaking, content moderation is defined as «the governance mechanisms that structure participation in a community to

⁷ E. Barendt, *Freedom of Speech*, Oxford, 2007, p. 1 ff.

⁸ M. Goswami, *Algorithms and Freedom of Expression* in W. Barfield (ed), *The Law of Algorithms*, Cambridge, 2020, p. 561.

⁹ J. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, in *UC Davis Law Review*, vol. 51, 2018, p. 1149 ff.

¹⁰ *Ibid.*, p. 1151.

¹¹ K. Bontcheva et al., *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*, 2024, available at verrai.eu.

facilitate cooperation and prevent abuse»¹². Moderation thus involves not only the administrators or moderators with power to remove content or exclude users, but also the design decisions that organise how the members of a community engage with one another¹³.

On the one hand, AI systems used for content moderation have become a necessary tool to defend freedom of expression and the values underlying it, creating the conditions for a robust and vibrant democratic exchange on online platforms. Due to the enormous amount of content created and circulated daily, which vastly exceeds the capacity of intermediaries to rely solely on human review prior to upload, these platforms are turning to automated processes to assist in the detection and analysis of problematic content¹⁴.

On the other hand, content moderation can influence both individuals' ability to express their views and their access to information on digital platforms, thereby impacting the active and passive dimensions of freedom of expression. For instance, the unjustified demotion of content incorrectly labelled as disinformation may restrict the freedom of expression of the person posting it, as well as limit the public's right to be informed. In contrast, curation of search results or newsfeeds primarily impacts the freedom to receive information – for example, when algorithmic curation filters out opposing political perspectives, preventing a supporter of one party from being exposed to alternative viewpoints¹⁵. Such risks have led the United Nations Special Rapporteur on Freedom of Expression to criticise content moderation systems for their vague operational rules, inconsistent enforcement practices, and excessive reliance on automation, which can result in over-blocking and prior censorship¹⁶.

¹² J. Grimmelmann, *The Virtues of Moderation*, in *Yale Journal of Law and Technology*, vol. 17, 2015, p. 42.

¹³ R. Gorwa, R. Binns and C. Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, in *Big Data & Society*, vol. 7, 2020, p. 3.

¹⁴ E. Llansó, J. van Hoboken, P. Leerssen and J. Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression*, in *Transatlantic Working Group*, 26 February 2020, available at ivir.nl.

¹⁵ M. Brkan, *Freedom of expression and Artificial Intelligence: on personalisation, disinformation and (lack of) horizontal effect of the Charter*, in *Maastricht Faculty of Law Working Papers*, 17 March 2019, available at papers.ssrn.com.

¹⁶ United Nations, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression* (U.N. Doc A/HRC/38/35), 6 April 2018, available at obchr.org.

Giovana Peluso Lopes, Pratiksha Ashok
Freedom of Expression and AI-Driven Content Moderation

In the European Union (EU), recent regulatory initiatives have begun to address the implications of AI-driven content moderation for freedom of expression from distinct yet complementary perspectives. The Digital Services Act (DSA)¹⁷ governs the responsibilities of online intermediaries and platforms in moderating and curating content, seeking to ensure that these private governance structures operate consistently with users' fundamental rights, including freedom of expression. The Artificial Intelligence Act (AI Act)¹⁸ by contrast, regulates AI systems as such, extending its scope to those employed in moderation and recommender functions that shape online discourse.

Based on the existing research, this article investigates the following research question: How does the increasing reliance on AI-driven content moderation affect the protection of freedom of expression in the European Union, and to what extent do the Digital Services Act (DSA) and the Artificial Intelligence Act (AI Act) provide an adequate regulatory response to these risks?

To answer this question, the article adopts a doctrinal and normative legal analysis, complemented by a structured mapping of technological risks. First, it conceptualises AI-driven content moderation and distinguishes between different forms of moderation and curation in the digital environment (Section 2). Second, it identifies and systematises the technical and structural risks that automated moderation systems pose to the active and passive dimensions of freedom of expression (Section 3). This risk-mapping exercise serves as the analytical baseline against which the regulatory framework is evaluated. Third, the article examines the DSA and the AI Act, analysing their scope, obligations, and underlying regulatory logic in light of the previously identified risks (Section 4). Finally, it assesses whether the combined operation of these instruments establishes a coherent and rights-sensitive governance architecture for AI-driven content moderation, and whether the classification of certain moderation systems as high-risk under the AI Act would strengthen fundamental rights protection (Sections 4.1–4.2).

¹⁷ Regulation (EU) 2022/2065, Digital Services Act.

¹⁸ Regulation (EU) 2024/1689, Artificial Intelligence Act.

2. *Content Moderation in the Digital Age*

Content moderation can be described as the processes and mechanisms through which online platforms monitor, filter, and regulate the material that users upload or circulate¹⁹. In other words, it is the attempt to balance the principles of freedom of expression with the need to maintain safe, lawful, and hospitable online environments. Yet the concept of “content moderation” is not monolithic; it encompasses different models, approaches, and philosophies²⁰.

Literature often distinguishes between “hard” and “soft” content moderation, as well as “content moderation *stricto sensu*,” and “content curation,” each reflecting different functions and degrees of intervention²¹. Hard content moderation refers to the direct removal, blocking, or filtering of user-generated content deemed unacceptable²². This can include takedowns of posts containing, for instance, hate speech, child sexual abuse material, or copyright-infringing works. Hard moderation operates in a binary mode in which content is either deemed permissible or impermissible, and platforms often enforce this approach under a mix of legal obligations and their own community standards. For example, EU law mandates the removal of terrorist content within one hour of receipt of a removal order issued by a Member State competent authority, compelling platforms to deploy rapid and decisive moderation strategies²³.

Soft content moderation, by contrast, involves less intrusive interventions that do not necessarily remove the content but alter its visibility or accessibility. This can include downranking posts in recommendation algorithms, placing warning labels, fact-checking notes, or limiting the sharing functions for disputed content. The “soft” approach has become especially significant in the context of misinformation and “lawful but awful” speech – the latter referring to speech that is offensive or morally repugnant to many people but is nevertheless protected by

¹⁹ S. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*, Yale, 2019, p. 33 ff.

²⁰ B. Rieder and Y. Skop, *The Fabrics of Machine Moderation: Studying the Technical, Normative, and Organizational Structure of Perspective API*, in *Big Data & Society*, vol. 8, 2021.

²¹ M. Klos, *Wrongful Moderation: Regulation of internet intermediary service provider liability and freedom of expression*, 21 September 2021, available at scholarlypublications.universiteitleiden.nl.

²² F. Stjernfelt and A. Lauritzen, *Nipples and the Digital Community* in F. Stjernfelt and A. Lauritzen (eds), *Your Post has been Removed: Tech Giants and Freedom of Speech*, Cham, 2020, p. 95 ff.

²³ Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online.

freedom of expression²⁴. Rather than outright eliminating these sorts of materials, platforms try to reduce their reach or add contextual information, thereby walking the line between respecting user autonomy and curbing harmful effects.

Content curation differs conceptually from moderation, even though the two are frequently conflated. Curation refers to the ways platforms organise, recommend, and prioritise content in feeds, timelines, or search results²⁵. Whereas moderation is primarily concerned with excluding undesirable material, curation is about amplifying or foregrounding certain kinds of content. Yet, the boundary is blurry. A decision to downrank extremist content in recommendations could be seen as both moderation (a soft form) and curation (an editorial choice about what users should see).

Content moderation *lato sensu* can be implemented through human judgment, automated systems, or a hybrid model combining both. Each approach has advantages and drawbacks, and in practice, most large platforms use a combination of these according to the type of content being moderated and the goals of content moderation practices.

Manual moderation relies on human moderators who review flagged content and make case-by-case decisions²⁶. In the early days of the internet, most moderation was manual, with community managers or volunteers deciding whether posts complied with forum rules. Even today, platforms like Reddit and Discord rely heavily on human moderators at the community level²⁷. Manual moderation has the benefit of contextual understanding: humans can detect sarcasm, cultural nuance, or ambiguous speech in ways algorithms struggle to replicate²⁸. However, it is resource-intensive and emotionally taxing. Reports have documented the severe psychological toll on commercial moderators who spend hours reviewing

²⁴ D. Keller, [Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users](#), in *The University of Chicago Law Review Online*, June 2022, available at lawreview.uchicago.edu.

²⁵ A. Higson, *Netflix – The Curation of Taste and the Business of Diversification*, in *Studia Humanistyczne AGH*, vol. 20, 2021, p. 7 ff.

²⁶ J. Wamai, M. Kalume, M. Gachuki and A. Mukami, *A New Social Contract for the Social Media Platforms: Prioritizing Rights and Working Conditions for Content Creators and Moderators*, in *International Journal of Labour Research*, vol. 12, 2023, p. 98 ff.

²⁷ T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale, 2018, p. 18.

²⁸ A. Demopoulos, [Free the Nipple: Facebook and Instagram Told to Overhaul Ban on Bare Breasts](#), in *The Guardian*, 18 January 2023, available at theguardian.com.

violent or abusive material²⁹. Moreover, manual moderation is often too slow to handle the sheer scale of modern social platforms, where millions of posts are uploaded every minute.

Automated moderation, by contrast, leverages algorithms and machine learning systems to detect and filter problematic content at scale. For instance, YouTube employs automated copyright detection tools like Content ID, while Facebook uses machine learning to proactively detect and remove nudity or terrorist propaganda³⁰. Automation allows for speed and efficiency, especially in contexts where rapid removal is legally required. Yet, these systems are prone to error. As will be discussed shortly, they often over-remove, censoring legitimate expression, or under-remove, failing to catch harmful content³¹.

Finally, hybrid moderation attempts to combine the strengths of human judgment with the efficiency of automated systems. Typically, AI tools pre-screen content and flag potentially problematic posts, which are then reviewed by human moderators. This tiered approach reduces the volume of material requiring human review while still ensuring contextual evaluation for borderline cases. Hybrid systems are increasingly common: for example, Meta reports that AI removes large volumes of policy-violating content automatically, but complex or ambiguous cases are escalated to human teams³².

These different approaches to content moderation aim to achieve various purposes, which extend beyond simple compliance with law, but rather encompass a wide range of objectives linked to safety, trust, and governance. At the most basic level, moderation ensures that platforms do not become havens for illegal content. This includes categories such as child sexual abuse material, terrorist propaganda, and copyright infringement. Legal obligations vary across jurisdictions, but platforms often operate under a “notice and takedown” model, where they must remove illegal content once notified.

Content moderation practices also seek to address harmful but not necessarily illegal content. This includes some forms of hate speech,

²⁹ O. Solon, [Facebook Is Hiring Moderators. But Is the Job Too Gruesome to Handle?](#), in *The Guardian*, 4 May 2017, available at [theguardian.com](#).

³⁰ T. Gillespie, *Custodians of the Internet*, cit., p. 3.

³¹ J. Cobbe, *Algorithmic Censorship by Social Platforms: Power and Resistance*, in *Philosophy & Technology*, vol. 34, 2021, p. 739.

³² S. Balendra, *Meta's AI Moderation and Free Speech: Ongoing Challenges in the Global South*, in *Cambridge Forum on AI: Law and Governance*, vol. 1, 2025, p. 21.

harassment, self-harm promotion, and graphic violence³³. In many jurisdictions, such content may not reach the legal threshold of illegality but still poses risks to users. For instance, the category of “lawful but awful” content captures this grey area of expression that is legal but undesirable. Examples include racist slurs that do not meet the legal definition of hate speech, or misinformation that is not outright illegal³⁴. While misinformation involves the unintended sharing of false data, disinformation reflects the deliberate spread of misleading content to further specific agendas³⁵. The COVID-19 pandemic underscored how misinformation can have serious public health consequences³⁶, while electoral disinformation threatens democratic processes³⁷. Platforms responded with new measures such as fact-checking labels, content demotion, and promotion of authoritative sources.

Platforms often remove or restrict materials that are not illegal under their own policies to protect user safety, mitigate reputational harm, and maintain advertiser-friendly environments. In fact, moderation also serves broader platform policy and business objectives. Advertisers often demand brand-safe environments, leading platforms to adopt stricter standards than those required by law. Moderation thus also functions as a tool of market self-regulation and enhances consumer trust³⁸.

This section has examined the principal dimensions of content moderation, distinguishing between strict moderation and content curation, the varying degrees of human and automated involvement in these processes, and the multiple purposes they serve – including the removal of illegal content, the mitigation of misinformation and disinformation, and the management of “lawful but awful” material under platform community standards. The following section will consider how freedom of expression is impacted by content moderation practices increasingly driven by artificial intelligence technologies within digital platforms.

³³ E. Nave and L. Lane, *Countering Online Hate Speech: How Does Human Rights Due Diligence Impact Terms of Service?*, in *Computer Law & Security Review*, vol. 51, 2023, p. 1 ff.

³⁴ S. Chesterman, *Lawful but Awful: Evolving Legislative Responses to Address Online Misinformation, Disinformation, and Mal-Information in the Age of Generative AI*, in *The American Journal of Comparative Law*, vol. 72, 2024, p. 933 ff.

³⁵ K. Jaidka et al., *Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy*, in *Digital Government: Research and Practice*, vol. 6, 2025, p. 2.

³⁶ F. Caceres et al., *The impact of misinformation on the COVID-19 pandemic*, in *AIMS public health*, vol. 9, 2022, p. 262 ff.

³⁷ C. Marsden, T. Meyer and I. Brown, *Platform values and democratic elections: How can the law regulate digital disinformation?*, in *Computer Law & Security Review*, vol. 36, 2020.

³⁸ T. Gillespie, *Custodians of the Internet*, cit., p. 126.

3. *Risks of AI-Driven Content Moderation to Freedom of Expression*

Artificial intelligence now forms the backbone of modern content moderation, as the vast scale of online information surpasses the capacity of human reviewers to monitor uploaded material and identify content that is illegal or violates platform rules. Machine learning classifiers are widely used to identify categories of harmful content, such as hate speech, extremist material, nudity, or spam. For instance, natural language processing (NLP) systems scan text for offensive language, while computer vision models detect violent or pornographic imagery³⁹. These techniques are particularly effective at categorising undesirable content, yet their use also carries risks, as they may inadvertently exclude legitimate expressions or overlook illegal or harmful materials, thereby posing potential threats to freedom of speech. This section begins by identifying technical limitations of AI-driven content moderation tools and how they can negatively impact the right to freedom of speech. These limitations stem from their inability to fully account for evolving context and practice, lack of diversity in training datasets and the generation of biased outputs.

3.1. *Technical Limitations of Content Moderation Systems*

The effectiveness of content moderation systems largely depends on their ability to accurately distinguish between acceptable and unacceptable material. Developing such reliable systems requires minimising the number of errors they produce. To achieve this, the machine learning techniques used for moderating content must be trained on large datasets with specific qualities – most importantly, data that accurately represents the domain in question. If certain categories are either underrepresented or overrepresented in the training data, the system may fail to properly identify items belonging to those categories. This can compromise the tool's accuracy, resulting either in the removal of harmless content or in the unintended publication of material that should have been restricted⁴⁰.

³⁹ S. Govindankutty and A. Kumar, *Design and Implementation of Automated Content Moderation Systems in Social Media*, in *Integrated Journal for Research in Arts and Humanities*, vol. 4, 2024, p. 380 ff.

⁴⁰ G. Sartor and A. Loreggia, *The Impact of Algorithms for Online Content Filtering or Moderation*, European Parliament, September 2020, p. 45, available at [europarl.europa.eu](https://www.europarl.europa.eu).

One example is the lack of linguistic diversity in training datasets. Many moderation systems are trained primarily on high-resource languages, such as English and Spanish, leaving them less effective in detecting harmful content in other languages⁴¹. This creates global inequities: platforms may be more effective at moderating speech in the United States or Europe than in the Global South, where linguistic diversity and lack of resources reduce coverage. The lack of linguistic diversity can result in illegal or harmful speech being unidentified or misidentified, as highlighted by the 2017–2018 Rohingya crisis in Myanmar, as Facebook was slow to act against hate speech in Burmese⁴².

Another challenge lies in the lack of contextual understanding. AI systems cannot account for the contextual factors, intentions, linguistic subtleties, cultural variations, and their evolving dynamics that are necessary for effective content evaluation. Certain formats of content, such as audiovisual material, also present greater challenges for accurate identification and moderation. Moreover, AI often struggles to interpret elements such as misspellings, variations in syntax, and the use of images, GIFs, or memes to convey meaning. Consequently, this can lead to the unintended removal of lawful content, including reports on extremist incidents or expressions of political dissent, raising concerns regarding accuracy, potential censorship, and the broader implications for freedom of expression⁴³.

AI also struggles to interpret the subtle dimensions of human communication, especially when it comes to identifying speaker intent or motivation – both of which are key elements in moderating content, particularly hate speech. For instance, it fails to distinguish between harmful and benign uses of the same language⁴⁴, or to interpret the context specific language used by marginalised communities, such as forms of ‘mock

⁴¹ G. Nicholas and A. Bhatia, *Toward Better Automated Content Moderation in Low-Resource Languages*, in *Journal of Online Trust and Safety*, vol. 2, 2023, p. 1 ff.

⁴² M. Kettemann, *How Platforms Respond to Human Rights Conflicts Online: Best Practices in Weighing Rights and Obligations in Hybrid Online Orders*, Verlag Hans-Bredow-Institut, 2022, available at soar.info; F. Shahid and A. Vashista, *Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?*, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, article no. 391, p. 1 ff.

⁴³ S. Udupa et al., *Artificial Intelligence, Extreme Speech and the Challenges of Online Content Moderation*, AI4Dignity Project, 2021, available at disinfoobservatory.org.

⁴⁴ M. Alawamleh, N. Shammam, K. Alawamleh and L. Ismail, *Examining the Limitations of AI in Business and the Need for Human Insights Using Interpretive Structural Modelling*, in *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, 2023.

impoliteness' and the use of reclaimed slurs within the LGBTQ community⁴⁵. Similarly, research indicates that tweets written in African American English are twice as likely to be flagged as offensive, underscoring the presence of racial bias within algorithmic moderation systems⁴⁶. This highlights concerns about potentially discriminatory content moderation practices and the unjust removal of legitimate material, especially in relation to expressions from minoritised or historically marginalised communities.

As previously mentioned, a lot of these issues stem from biased, inaccurate or limited training data. Machine learning systems acquire their capacity to recognise and differentiate between various types of content through the datasets on which they are trained. Many of these systems rely on publicly available labelled datasets; however, when such datasets lack examples of language use from diverse linguistic or social groups, the resulting tools are unable to accurately interpret those communities' forms of expression. Furthermore, natural language processing models tend to perform optimally in contexts that closely resemble their training data. Developing tools that function effectively across different platforms, languages, cultures, interest groups, and subject areas remains challenging. When applied to domains or speaker groups not well represented in the training data, these systems risk producing misclassifications that disproportionately impact underrepresented groups⁴⁷. Algorithmic bias in moderation thus reflects and amplifies existing social inequalities⁴⁸.

In addition, content moderation systems face adversarial challenges, as users seeking to disseminate harmful content often develop strategies to evade detection, such as altering spellings, embedding messages within images, or using coded language. Moderation systems must continually adapt, creating a cat-and-mouse dynamic⁴⁹. Consequently, there is a need for flexible and adaptive models, as human communication patterns change rapidly and users restricted by automated filters are especially motivated to find ways to bypass them. Static machine learning models, therefore, risk

⁴⁵ T. Dias, D. Antonialli and A. Gomes, *Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risk to LGBTQ Voices Online*, in *Sexuality and Culture*, vol. 25, 2020, p. 700 ff.

⁴⁶ M. Sap, *The Risk of Racial Bias in Hate Speech Detection*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, p. 1668 ff.

⁴⁷ E. Llansó, J. van Hoboken, P. Leerssen and J. Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression*, cit., p. 8.

⁴⁸ F. Shahid and A. Vashistha, *Decolonizing Content Moderation*, cit., p. 42; B. Rieder and Y. Skop, *The Fabrics of Machine Moderation*, cit., p. 20.

⁴⁹ A. Arora et al., *Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go*, in *ACM Computing Surveys*, vol. 56, 2023, p. 1 ff.

becoming obsolete and ineffective at accurately classifying user content over time⁵⁰.

In order to overcome some of the technical challenges presented thus far, Large Language Models (LLMs) are increasingly integrated into automated content moderation, particularly for detecting and mitigating hate speech online. For instance, Google's Perspective API, a transformer model, detects toxic content in 12 languages, whereas Meta AI developed and employs tools like RoBERTa and XLM-R, which are transformer-based and trained on text in 100 languages. Compared to conventional moderation techniques, they offer notable advantages in contextual reasoning, adaptability, and the ability to automatically generate large datasets to improve classifier performance⁵¹. However, significant shortcomings remain. Studies have shown that while LLMs provide valuable reasoning support in moderation decisions, they are best suited as tools within hybrid moderation frameworks, where automated outputs are complemented by human oversight⁵². Moreover, concerns persist regarding their stability and reliability, as moderation outcomes can vary unpredictably between models, with larger models not necessarily performing better than smaller ones. Traditional moderation tools often continue to outperform LLMs in consistency, while issues of high computational cost, limited interpretability, and dependence on contextual prompting highlight the need for further research into achieving a balanced and sustainable integration of LLMs into content moderation ecosystems⁵³.

⁵⁰ E. Llansó, J. van Hoboken, P. Leerssen and J. Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression*, cit., p. 8.

⁵¹ E. Penagos, *ChatGPT, Can you solve the content moderation dilemma?*, in *International Journal of Law and Information Technology*, vol. 32, 2024, p. 9.

⁵² *Ibid.*, p. 1 ff.; F. Gilardi, M. Alizadeh and M. Kubli, *ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks*, in *Proceedings of the National Academy of Sciences: Political Sciences*, vol. 120, 2023, p. 1 ff.; L. Li, L. Fan, S. Atreja and L. Hemphill, "HOT" ChatGPT: *The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media*, in *ACM Transactions on the Web*, vol. 18, 2024, p. 1 ff.; D. Kumar, Y. Abuhashem and Z. Durumeric, *Watch Your Language: Investigating Content Moderation with Large Language Models*, in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, p. 865 ff.; K. Caramancion, *Harnessing the Power of ChatGPT to Decimate Mis/Disinformation: Using ChatGPT for Fake News Detection*, in *IEEE World AI IoT Congress (AIoT)*, 2023, p. 42 ff.

⁵³ E. Penagos, *ChatGPT, Can you solve the content moderation dilemma?*, cit., p. 9.

3.2. *Risks to Freedom of Expression*

Having outlined the main technical limitations of AI-driven content moderation systems, the following section turns to examine how the integration of automated systems in content moderation introduces unique risks for freedom of expression. It analyses key issues arising from the deployment of automated moderation tools, including the risks of over-removal of content, the persistence of false positives and false negatives, the emergence of proactive filtering as a form of prior restraint, and the opacity of algorithmic decision-making.

3.2.1 *Over-Removal of Content*

Content moderation requires carefully balancing two often conflicting objectives: minimising the individual and societal harms caused by illegal or harmful content, while also promoting free speech and enabling open, civil interaction online. Automated systems, including AI-driven moderation tools, inevitably operate with some level of error because they rely on statistical and probabilistic models⁵⁴. As Sartor and Loreggia observe, at any given level of technical performance, lowering the rate of false negatives (that is, reducing the chances that harmful content will be overlooked) typically comes at the cost of increasing the rate of false positives (the rejection of legitimate content). In other words, enhancing a system's sensitivity or recall inevitably entails a corresponding reduction in its specificity or precision, meaning platforms must navigate a trade-off between under-blocking harmful material and over-blocking legitimate expression.

This trade-off is further compounded by the regulatory and economic incentives that shape platform behaviour. In practice, the risk of liability and financial penalties can encourage platforms to err on the side of caution, prioritising the removal of potentially unlawful or controversial material rather than risking non-compliance⁵⁵. Dias highlights how the volume increase of content to be moderated and regulatory requirements to swiftly act have caused companies to «act proactively in order to avoid liability (...)

⁵⁴ G. Sartor and A. Loreggia, *The Impact of Algorithms for Online Content Filtering or Moderation*, cit., p. 45.

⁵⁵ N. Alkiviadou, *Moderating Hate or Moderating Rights? The Paradox of the European Approach to Online Hate Speech and Platform Liability*, CELE Research Paper no. 69, 2025, p. 12, available at papers.ssrn.com.

in an attempt to protect their business models»⁵⁶. In this context, the Council of Europe has cautioned that increasing dependence on AI-based moderation tools may result in excessive content removal, thereby endangering freedom of expression⁵⁷. Moreover, given the lack of harmonisation between national legal frameworks – for example, in how hate speech and other forms of illegal content are defined – platforms may choose to follow the most restrictive regulatory standards across jurisdictions to ensure overall compliance, rather than aligning with the more permissive ones. This dynamic further amplifies the tendency toward over-blocking and can have a chilling effect on legitimate expression online.

3.2.2 *False Negatives and False Positives*

Errors arising from AI-driven content moderation can affect both the active and passive dimensions of freedom of expression, by either restricting the right to speak and the right to receive information. On one hand, false negatives – which we have seen are instances where harmful or inappropriate content is not detected – can produce chilling effects on free speech. When abusive, misleading, or otherwise harmful material remains online, it can discourage individuals, particularly those belonging to vulnerable or marginalised groups, from participating in public discussions for fear of harassment or hostility. For instance, exposure to online hate speech can cause fear, emotional distress and, importantly, self-censorship, particularly among members of minority groups⁵⁸. This silencing effect not only limits the voices of those directly affected (i.e., the active dimension of freedom of expression) but also deprives others of the opportunity to hear diverse perspectives (i.e., the passive dimension of freedom of speech), ultimately weakening public discourse.

Moreover, false negatives can have a disproportionate effect on different social groups participating in public discourse, particularly when the AI systems are trained on biased or unrepresentative data. Dietrich illustrates this with the example of two opposing groups that regularly

⁵⁶ T. Dias, *Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression*, in *Human Rights Law Review*, vol. 20, 2020, p. 608.

⁵⁷ Council of Europe, [*Algorithms and Human Rights: Study On the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications*](#), 2018, available at [edoc.coe.int](#).

⁵⁸ N. Alkiviadou, *Moderating Hate or Moderating Rights? The Paradox of the European Approach to Online Hate Speech and Platform Liability*, cit.

exchange hostile and insulting comments toward each other⁵⁹. Each group uses distinct linguistic styles or “speech codes,” which may not be equally represented in the training data used to develop the AI moderation system. If the dataset includes mainly examples of harmful speech from one group, the system will become more effective at detecting violations by that group while overlooking similar behaviour from the other. This imbalance results in biased moderation outcomes: one group faces more frequent content removal, while the other is able to disseminate harmful speech with little interference. Such disparities risk silencing or intimidating certain communities, undermining the principle of equal access to freedom of expression. At the same time, the perception that one group is unfairly targeted while another is spared may erode trust in moderation systems and weaken public commitment to platform rules and broader norms against harmful speech⁶⁰.

False positives, in their turn, which happen when legitimate content is mistakenly flagged or removed, pose a more direct restriction on freedom of expression. Automated deletion of legitimate speech prevents users from communicating their ideas and simultaneously denies audiences access to that information. This problem extends beyond explicit removal to more subtle forms of suppression through content curation, with representatives of minority groups and of discriminated or marginalised communities having repeatedly claimed that the recommendation systems of most platforms often tend to significantly diminish the visibility of the counter-speech content they post, sometimes even resulting in forms of “shadowbanning”⁶¹. The latter refers to the practice of limiting the broader visibility of a user’s content without their knowledge, whilst still maintaining the original publication on the poster’s profile page⁶². By amplifying certain voices and downplaying others, AI-driven ranking systems shape the visibility of content, indirectly moderating what users encounter. This has profound implications for democratic discourse and information diversity⁶³.

⁵⁹ F. Dietrich, *AI-Based removal of hate speech from digital social networks: chances and risks for freedom of expression*, in *AI and Ethics*, vol. 5, 2025, p. 2943 ff.

⁶⁰ *Ibid.*, p. 2947.

⁶¹ G. Nicholas, *Shedding Light on Shadowbanning*, Center for Democracy and Technology, April 2022, available at cdt.org; C. Are, *The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram*, in *Feminist Media Studies*, vol. 22, 2022.

⁶² K. Gilbert, *How Shadow Banning Can Silently Shift Opinion Online*, in *Yale Insights*, 9 May 2024, available at insights.som.yale.edu.

⁶³ L. Bojic, *AI Alignment: Assessing the Global Impact of Recommender Systems*, in *Futures*, vol. 160, 2024, 103383.

As highlighted in a recent study by Zaman and Chen, by selectively muting and amplifying posts, online platforms can shift users' positions or increase overall polarisation, often in ways that appear neutral to an outside observer⁶⁴.

3.2.3 Proactive Moderation and Prior Restraint

The use of AI to proactively moderate content, by filtering or screening all user-provided content at the moment of upload to determine whether it contains illegal or harmful elements, can also result in prior restraint or prior censorship on speech. Prior restraint occurs when a speaker is required to seek approval from some empowered third party – most typically a public official –, before being allowed to speak or to publish her views. Under international human rights law, there is a strong presumption against the validity of prior censorship, guided by the need to ensure that all individuals, groups, ideas, and means of expression have the opportunity to participate freely in public discourse, and only later face subsequent punishment for any law or rule which may have been violated⁶⁵.

According to Llansó, there are three characteristics of systems of prior censorship that are replicated by proactive content moderation⁶⁶. The first relates to exposing more speech to evaluation and approval, in the sense that «prior restraints subject a much greater breadth and variety of content to government scrutiny and surveillance than a system of subsequent prosecution and punishment»⁶⁷. Whether imposed by government mandate or corporate practice, filters inherently exert this kind of expanded scrutiny, by treating every piece of uploaded content as a potential rule violation.

The second characteristic is that of removing procedural hurdles to censorship. These procedural safeguards create friction within systems for determining the illegality of speech, ensuring that more than a simple administrative action is required to suppress speech, and allowing people to defend themselves before an independent adjudicator before the decision is

⁶⁴ Y. Chen and T. Zaman, *Shaping Opinions in Social Networks with Shadow Banning*, in *PLoS ONE*, vol. 19, 2024, 1 ff.

⁶⁵ E. Llansó, *No amount of "AI" in content moderation will solve filtering's prior-restraint problem*, in *Big Data & Society*, vol. 7, 2020, p. 1 ff.

⁶⁶ *Ibid.*, p. 3.

⁶⁷ J. Balkin, *Old-school/new-school speech regulation*, in *Harvard Law Review*, vol. 27, 2014, p. 2296 ff.

enforced. The deliberate introduction of friction stands in fundamental tension with the drive toward upload filtering and other automated content moderation tools aimed at blocking or removing speech more rapidly⁶⁸.

Finally, systems of prior restraint function as «low-visibility systems of control»⁶⁹, in the sense that they can operate outside of public scrutiny, going against the requirement that limitations to freedom of expression must be clearly defined in advance so that people can regulate their conduct accordingly. Regarding this characteristic, filtering can greatly hinder people's capacity to recognise, interpret, and evaluate the censorship mechanisms shaping their online information environment. Content moderation systems may implement upload filters opaquely, and the use of machine-learning tools can further obstruct public scrutiny if they are not designed to provide transparency or explainability. For instance, when a machine-learning model creates a classifier to differentiate between categories of speech (such as hate speech and non-hate speech), the features it relies on may not correspond to concepts intelligible to humans⁷⁰.

Proactive moderation also raises questions under the DSA's liability regime. Article 6 grants hosting providers a conditional exemption based on their role as neutral intermediaries, yet systematic upload filtering blurs the line between passive hosting and active control over user speech. Although Article 7 confirms that voluntary own-initiative measures do not in themselves remove the exemption, extensive *ex ante* filtering intensifies the tension between neutrality and editorial intervention, reinforcing concerns about platforms' expanding gatekeeping power over online expression.

3.2.4 Opaque Decisions and Lack of Algorithmic Transparency

The lack of transparency regarding content moderation practices and the opacity of AI systems involved aggravates the issues discussed thus far. A first concern relates to the opacity of algorithmic decision-making. Many AI-driven moderation systems operate as “black boxes,” offering little to no insight into how they are coded, what data they are trained on, or how they reach their conclusions. This opacity is often justified by platforms on the grounds of protecting intellectual property or preventing manipulation

⁶⁸ E. Llansó, *No amount of “AI” in content moderation will solve filtering’s prior-restraint problem*, cit., p. 3.

⁶⁹ J. Balkin, *Old-school/new-school speech regulation*, cit., p. 2296.

⁷⁰ E. Llansó, *No amount of “AI” in content moderation will solve filtering’s prior-restraint problem*, cit., p. 3.

of their systems, yet it ultimately shields moderation decisions from public scrutiny and accountability⁷¹.

A second issue concerns the lack of notice provided to users when automated tools are employed to make or influence moderation decisions. Often, individuals whose content is removed, downranked, or otherwise limited are not informed about whether AI was involved in that decision, nor are they given sufficient information to challenge it effectively. Without clear notice on when and how AI is being used in decisions pertaining to content limits the ability for users to appeal the decision⁷².

Conceptual Map of Risks in AI-Driven Content Moderation

Level	Type of Harm	Core Mechanism / Source	Primary Impact / Dimension	Illustrative Examples
1. Technical / Data Level	Accuracy and Bias Errors	Inaccurate, unrepresentative, or biased training data; linguistic imbalance; limited contextual comprehension	Misclassification, unequal treatment of linguistic cultural groups	Over-flagging African American English; under-detection of hate speech during Rohingya crisis
	Contextual Deficiency	Inability to process nuance, satire, intent, or cultural variations	Loss of meaning; misinterpretation of expression	Mistaking reclaimed slurs or political dissent for hate speech
	Adversarial Adaptation	Evasive tactics by users (coded language, obfuscation)	Reduced model efficacy, ongoing arms race	Users are modifying spelling or embedding harmful content in memes

⁷¹ S. Singh, *Everything in Moderation: An analysis of how internet platforms are using artificial intelligence to moderate user-generated content*, Open Technology Institute, 22 July 2019, available at newamerica.org.

⁷² S. Udupa et al., *Artificial Intelligence, Extreme Speech and the Challenges of Online Content Moderation*, AI4Dignity Project, 2021, available at disinfoobservatory.org; S. Udupa, A. Maronikolakis, and A. Wisiolek A., *Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence*, in *Big Data & Society*, 10, 2023, p. 1 ff.

Level	Type of Harm	Core Mechanism / Source	Primary Impact / Dimension	Illustrative Examples
2. Systemic / Operational Level	Trade-off Errors (False Positives vs. False Negatives)	Probabilistic nature of ML systems—precision/recall tension	Over-blocking legitimate content; under-blocking harmful content	Removal of lawful speech; exposure to hate content
	Stability and Reliability of Models (LLMs)	Model inconsistency, dependence on prompting, computational cost	Unpredictable moderation outcomes; uneven enforcement	LLMs producing inconsistent toxicity scores
	Opacity and Lack of Explainability	Proprietary algorithms; lack of notification or transparency	Accountability deficit; reduced contestability of moderation decisions	No user notice when content removed; opaque classifier logic
3. Social / Distributional Level	Discriminatory Impacts	Embedded bias in datasets and models	Unequal restriction and protection of expression across groups	Minority or Global South speech disproportionately censored
	Inequitable Global Reach	Resource language asymmetries	Regional and inequality in content moderation quality	Better moderation in high-resource languages like English-speaking countries
4. Normative / Legal Level	Chilling Effects (False Negatives)	Persistence of harmful content leading to harassment	Self-censorship; reduced participation of vulnerable groups	Marginalised users withdrawing from discourse
	Censorship (False Positives)	Removal or downranking of legitimate content	Direct violation of freedom of expression; diminished information diversity	Shadowbanning, visibility reduction

Level	Type of Harm	Core Mechanism / Source	Primary Impact / Dimension	Illustrative Examples
	Prior Restraint	Proactive, upload-level filtering	Preventive censorship; lack of due process	Automated pre-screening before publication
	Opacity as Structural Control	Invisibility of algorithmic decision-making	Undermines rule of law, procedural fairness, public scrutiny	No notice or appeal; black-box and moderation
5. Democratic / Epistemic Level	Distortion of Public Discourse	Algorithmic amplification and suppression	Polarisation; loss of pluralism; manipulation of public debate	Selective muting or amplification of shaping perceptions
	Erosion of Trust and Legitimacy	Perceived unfairness and opacity	Reduced public confidence in moderation systems and rules	Unequal moderation practices between groups

Recognising the challenges previously discussed, the European Union has introduced two major regulatory instruments that, while distinct in scope and purpose, share an overarching commitment to safeguarding fundamental rights in the digital environment: the DSA and the AI Act. Each instrument targets a different stage of the content moderation pipeline. The former focuses on the procedures and outcomes of moderation and how platforms make, justify, and communicate moderation decisions, while the latter regulates the design and deployment of the underlying AI systems that perform those functions. Taken together, these frameworks have the potential to form a comprehensive governance architecture for protecting freedom of expression online. The following section examines how the DSA and the AI Act interact, where their scopes overlap or diverge, and discusses the possibility of categorising AI-based content moderation systems as high-risk under the AI Act.

4. *From Algorithmic Harms to Regulatory Remedies*

The challenges outlined above reveal deep structural vulnerabilities of AI-driven content moderation systems. These tools, while indispensable for managing the scale of online communication, produce systematic errors that strike at the core of freedom of expression. They frequently misclassify lawful speech, suppress marginalised voices, and erode the procedural and substantive guarantees that underpin democratic discourse. The opacity and inaccessibility of these systems – both at the technical and organizational level – further exacerbate these harms, leaving users without meaningful recourse or understanding of how automated moderation decisions are made.

In effect, automated moderation systems have become powerful arbiters of visibility and silence, shaping who can speak and who is heard. As Dietrich argues, they now function as private infrastructures of speech governance⁷³, operating at the intersection of technological determinism and corporate discretion. The outcome is a form of algorithmic governance that often escapes public accountability, producing not only individual injustices but also systemic distortions of the online information environment. Galli similarly observe that platforms' hybrid status – as both market actors and quasi-regulators – has blurred the line between public and private ordering of speech, creating new regulatory asymmetries and normative ambiguities⁷⁴.

Against this backdrop, a dual-layered governance approach is required: one that constrains how moderation systems are designed and deployed (a technical, *ex ante* dimension), and another that guarantees procedural fairness and user rights when those systems are applied (an organisational, *ex post* dimension). This is precisely the gap that the EU seeks to close through the DSA and the AI Act. Together, these instruments represent a concerted regulatory response to the algorithmic governance of expression. The DSA focuses on the procedural side of content moderation – ensuring transparency, contestability, and accountability in platform decisions – while the AI Act targets the technical conditions under which AI systems are built and placed on the market. The former is directed at platform operators and aims to secure users' procedural rights; the latter

⁷³ F. Dietrich, *AI-Based removal of hate speech from digital social networks*, cit., p. 2944 ff.

⁷⁴ F. Galli, A. Loreggia and G. Sartor, *The Regulation of Content Moderation* in D. Vicente, S. Casimiro and C. Chen (eds) *The Legal Challenges of the Fourth Industrial Revolution: The European Union's Digital Strategy*, Cham, 2023, p. 63 ff.

addresses developers and deployers, setting out harmonised, risk-based obligations to protect fundamental rights at the design stage.

4.1. *The Digital Services Act*

The Digital Services Act represents the most significant overhaul of the EU's regulatory framework for online platforms since the adoption of the e-Commerce Directive in 2000⁷⁵. It establishes horizontal obligations for all intermediaries that provide services in the EU, with a particular focus on ensuring that content moderation practices respect fundamental rights, notably freedom of expression and information under Article 11 CFREU. The DSA's approach to freedom of expression is primarily procedural: it does not dictate what content must or must not be removed but instead sets out obligations to ensure that platforms act with transparency, accountability, and fairness when moderating user-generated content, whether through human or automated means.

At the core of the DSA's content moderation regime are the obligations imposed on "hosting services" and, more specifically, "online platforms" as defined in Article 3. Article 14(1) requires platforms to establish "clear and unambiguous" terms and conditions specifying any restrictions they impose on the use of their services, including restrictions based on the platform's own policies or applicable law. Crucially, under Article 14(4), such terms must be applied «in a diligent, objective, and proportionate manner,» taking into account the fundamental rights of users as enshrined in the CFREU⁷⁶. This provision aims to embed respect for freedom of expression directly into platforms' governance structures, ensuring that private rules governing speech do not arbitrarily restrict lawful expression.

The DSA introduces a structured due process framework for content moderation decisions. Under Article 16, online platforms must inform users when they take restrictive measures against content or accounts such as removal, suspension, or demotion and must indicate whether automated tools were used in the decision. Article 17 further requires that platforms provide users with a "statement of reasons," specifying whether the restriction was based on legal obligations or on the platform's own terms

⁷⁵ Directive 2000/31/EC, Directive on electronic commerce.

⁷⁶ M. Senftleben, *Human Rights Outsourcing and Reliance on User Activism in the DSA*, in *Verfassungsblog*, 21 February 2024, available at [verfassungsblog.de](https://www.verfassungsblog.de).

and conditions. The statement must include the legal or contractual basis for the decision, the facts and circumstances relied upon, and, if applicable, information on the use of automated means⁷⁷. These obligations represent an unprecedented step toward transparency in content moderation, reflecting the EU's commitment to procedural fairness and to enabling users to understand and contest decisions that affect their expressive rights.

To give these procedural safeguards practical effect, the DSA mandates accessible redress mechanisms. Article 20 provides users with the right to an internal complaint-handling system, allowing them to contest moderation decisions within six months. Platforms must handle complaints «in a timely, diligent, and objective manner» and must reinstate content where the removal or restriction was unjustified. Beyond internal remedies, Article 21 establishes the right to engage in out-of-court dispute settlement through certified bodies, which must operate independently and free of conflicts of interest⁷⁸. These procedural rights collectively transform the governance of online speech by providing users with avenues to challenge opaque or arbitrary moderation decisions that would previously have gone unreviewed.

For very large online platforms (VLOPs) and very large online search engines (VLOSEs) – services with more than 45 million monthly active users in the EU – the DSA imposes enhanced systemic obligations that link content moderation to the protection of fundamental rights. Under Articles 34 to 37, VLOPs must identify, assess, and mitigate systemic risks arising from the design, functioning, and use of their services, including risks to freedom of expression and information. They must conduct annual risk assessments, implement proportionate mitigation measures, and undergo independent audits to evaluate their compliance⁷⁹. These obligations recognise that the impact of platforms on public discourse extends beyond individual takedown decisions to the broader architecture of algorithmic amplification, ranking, and recommendation. By requiring VLOPs to assess systemic risks, the DSA seeks to ensure that freedom of expression is

⁷⁷ P. Leerssen, [The DSA's First Shadow Banning Case](#), in *DSA Observatory*, 6 August 2024, available at [dsa-observatory.eu](#).

⁷⁸ T. Hughes, [Practical Considerations for Out-of-Court Dispute Settlement \(ODS\) under Article 21 of the EU Digital Services Act \(DSA\)](#), in *DSA Observatory*, 8 February 2024, available at [dsa-observatory.eu](#).

⁷⁹ B. Zeybek, [The DSA and the Risk-Based Approach to Content Regulation: Are We Being Pulled into More Advanced Automation?](#), in *DSA Observatory*, 1 October 2021, available at [dsa-observatory.eu](#).

protected not only at the level of individual user interactions but also at the structural level of information flows within the digital public sphere.

Despite these advances, several limitations have been identified in the DSA's approach to content moderation. First, the statements of reasons provided under Article 17 often risk being generic and formulaic, especially when platforms operate at scale. Empirical studies of pre-DSA transparency reports already showed that explanations for content removals are frequently reduced to broad categories such as "hate speech" or "misinformation," offering little insight into the specific rationale or the standards applied⁸⁰. While the DSA requires platforms to disclose whether automated tools were used, it does not compel them to provide detailed information about how such tools function, how they are trained, or the error rates associated with their deployment. As a result, users are informed of automation only in the abstract, without the technical transparency necessary to evaluate or contest algorithmic decisions effectively.

Second, the DSA's transparency obligations under Articles 15, 24, and 42, requiring platforms to publish periodic transparency reports on their content moderation practices, suffer from a lack of harmonisation. The regulation mandates the publication of data on content removals, complaints, and automated tools, yet it leaves significant discretion to platforms regarding the format and granularity of reporting⁸¹. This has led to concerns that transparency reports will remain inconsistent and incomparable across services, undermining the goal of creating a coherent and accountable moderation ecosystem. Scholars and civil society groups have warned that without standardised metrics or oversight the transparency obligations may evolve into a "box-ticking exercise" that offers little substantive accountability.

Third, the DSA does not prescribe substantive requirements for the design or functioning of automated content moderation systems. While Article 16 obliges platforms to disclose their use, and Article 35 encourages VLOPs to mitigate systemic risks, there are no explicit obligations to evaluate accuracy, bias, or proportionality in algorithmic moderation

⁸⁰ R. Kaushal et al., *Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database*, in FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, p. 1121 ff.

⁸¹ European Commission, *Commission Harmonises Transparency Reporting Rules under the Digital Services Act*, in *Shaping Europe's Digital Future*, 4 November 2024, available at digital-strategy.ec.europa.eu/; J. Ohnesorge, *Counting Without Accountability? An analysis of the DSA's Transparency Reports*, Humboldt Institute for Internet and Society, 25 September 2025, available at hiig.de.

decisions⁸². Consequently, even though the DSA seeks to safeguard users' expressive rights procedurally, it leaves unaddressed the material question of whether the systems performing moderation are capable of doing so reliably and without discriminatory outcomes.

It could be argued that Articles 34 and 35 DSA are capable of encompassing risks related to the accuracy, bias, or proportionality of algorithmic moderation systems, given that content moderation processes are explicitly identified as potential sources of systemic risks and that testing and adaptation of algorithmic systems are listed as mitigation measures. However, the systemic risk framework remains structurally limited in several respects. First, it applies exclusively to VLOPs and VLOSEs, leaving other deployers of AI-driven moderation systems outside its enhanced governance regime. Second, Article 35 does not impose harmonised or technically specified obligations regarding dataset quality, bias mitigation, or accuracy thresholds, nor does it mandate lifecycle risk management comparable to that required for high-risk AI systems under the AI Act. Third, the concept of "systemic risk" remains open-textured and operationalised with considerable discretion by platforms, resulting in divergent methodological approaches. While Articles 34-35 may indirectly incentivise improvements in algorithmic moderation, they do not guarantee the structured ex ante safeguards that a high-risk classification under the AI Act would entail. In this sense, the AI Act would not replace the DSA's systemic governance model but rather complement it by embedding fundamental-rights protection at the design and development stage of AI systems.

Finally, the DSA's enforcement landscape remains uncertain. National Digital Services Coordinators (DSCs) are responsible for supervising intermediary services, but their capacities vary widely among Member States⁸³. The European Commission directly oversees VLOPs, yet the mechanisms for consistent enforcement, cross-border cooperation, and independent scrutiny are still evolving. Without robust oversight, the procedural guarantees of the DSA risk remaining aspirational rather than transformative.

The DSA articulates a procedural vision of freedom of expression that prioritises transparency, fairness, and accountability in content

⁸² European Data Protection Board, *Guidelines 3/2025 on the interplay between the DSA and the GDPR*, 11 September 2025, p. 4, available at edpb.europa.eu.

⁸³ European Commission, *Digital Services Coordinators*, in *Shaping Europe's Digital Future*, 17 October 2025, available at digital-strategy.ec.europa.eu; J. Jaursch, *Overview: Digital Services Coordinators in Europe*, in *Interface*, 8 February 2024, available at interface-eu.org.

moderation. It represents a major step toward subjecting platforms to rule-of-law principles, ensuring that decisions affecting users' speech are reasoned, contestable, and open to scrutiny. Yet its effectiveness ultimately depends on how these obligations are implemented in practice, particularly with regard to automation, where procedural transparency has not yet translated into substantive accountability. This underscores the need for complementary legislation, such as the AI Act, to address the technical design and risk profile of the automated systems that increasingly govern online expression.

4.2. *The Artificial Intelligence Act*

While the DSA grants users necessary procedural safeguards in relation to content moderation on online platforms, regardless of whether an automated system is involved in the process, the AI Act specifically targets AI systems and sets regulatory standards for their deployment in the EU market⁸⁴, seeking to ensure that they are safe and respect fundamental rights. For the purposes of the regulation, an AI system is defined as «a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments» (Article 3(1) AIA). Most of the AI systems currently deployed for content moderation, such as automated flagging tools, recommendation algorithms, comment-ranking systems, and image- or text-recognition classifiers, fall under this definition.

The AI Act adopts a risk-based framework, classifying AI systems into four categories, namely, unacceptable risk, high-risk, limited risk, and minimal risk, each with corresponding legal obligations. AI systems deemed to pose unacceptable risks, such as those manipulating users, assigning social scores, or performing remote biometric identification for law enforcement, are strictly prohibited (Article 5 AIA). High-risk systems face stringent regulatory requirements to ensure safety and protection of fundamental rights. Systems in the limited-risk category, including chatbots, emotion recognition tools, and deepfake generators, are subject primarily to transparency obligations, requiring users to be informed of the AI's

⁸⁴ F. Galli, A. Loreggia and G. Sartor, *The Regulation of Content Moderation*, cit., p. 85.

involvement (Article 50 AIA). All other AI applications are considered minimal risk and are largely unregulated, though voluntary codes of conduct may guide their development to align with high-risk standards (Article 95 AIA).

Article 6 AI Act sets out the methodology for classifying high-risk AI systems. A system may qualify as high-risk either because it is a product or safety component under the New Legislative Framework (such as autonomous vehicles or surgical robots), or because it falls within the use cases listed in Annex III. Under Article 6(2), AI systems covered by Annex III are automatically deemed high-risk and must comply with the requirements laid out in Chapter III. The list in Annex III contains 8 areas where the use of AI systems is considered to present a high risk to the health, safety or protection of fundamental rights of natural persons, as well as specific use cases within each area⁸⁵. The broad application areas encompass biometrics, critical infrastructure, education and vocational training, employment, essential services, migration and administration of justice and democratic processes. Notably absent from Annex III is the use of AI systems for content moderation, despite a proposal by the EU Parliament to include recommender systems used by social media platforms in Annex III, which was not agreed with in the trilogue process⁸⁶. Only AI systems intended to be used for «influencing the outcome of an election or referendum or the voting behaviour of natural persons» are considered to be high-risk under point 8(b) of Annex III, therefore failing to encompass other problematic uses of content moderation systems.

This has prompted calls to classify AI systems used for content moderation as high-risk under the AI Act, via its inclusion in Annex III⁸⁷. As discussed in Section 3, these systems carry significant risks for fundamental rights, particularly freedom of expression. Due to technical choices and limitations, AI moderation tools are prone to misclassification: false positives result in the removal or downranking of lawful content, effectively censoring users, while false negatives allow harmful or discriminatory material to remain, creating toxic environments that silence vulnerable groups. Beyond technical errors, these systems can also be exploited for illicit or unethical purposes, such as suppressing political

⁸⁵ Recital 52 AI Act.

⁸⁶ European Parliament, [Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence \(Artificial Intelligence Act\) and amending certain Union legislative acts](#), 21 January 2024, available at [artificialintelligenceact.eu](#).

⁸⁷ F. Galli, A. Loreggia, and G. Sartor, *The Regulation of Content Moderation*, cit., p. 85; F. Dietrich, *AI-Based removal of hate speech from digital social networks*, cit., p. 2948-2949.

debate or manipulating public opinion. Such outcomes directly affect the right to freedom of expression, highlighting the potential for significant adverse impacts on fundamental rights. As stated in Recital 48, the extent of the adverse impact caused by the AI system on the fundamental rights protected by the CFREU is of particular relevance when classifying an AI system as high-risk, among which is the right to freedom of expression and information.

Under Article 7(1) AI Act, the European Commission can adopt delegated acts to amend the list of high-risk use cases in Annex III. This provision aims to enable the AI Act to evolve over time, considering the rapid pace of technological progress in this field and the possible unforeseen developments in the use of AI systems⁸⁸. The amendment via delegated acts would require two cumulative conditions, namely: the relevant AI system would have to fall within a high-risk area already listed in Annex III; and it would need to present a potential risk to health, safety, or fundamental rights equivalent to or exceeding those of existing high-risk systems (Article 7(1) AIA). Considering how the area of content moderation is not included in Annex III, the intervention of the EU legislator would be required in order to classify AI-based content moderation systems as high-risk under the AI Act.

By including AI systems used for content moderation under the high-risk classification, the AI Act could have added another layer of protection additional to that of the DSA. Requirements applicable to high-risk systems as regards risk management, the quality and relevance of data sets used, technical documentation and record-keeping, transparency and the provision of information to deployers, human oversight, and robustness, accuracy and cybersecurity could help mitigate some of the risks previously identified.

For instance, Article 9 AI Act contains an obligation directed at providers to establish, implement, document and maintain a risk management system for high-risk AI systems. This must be understood as a continuous iterative process, planned and run throughout the entire lifecycle of the AI system. The risk management system must encompass the identification and analysis of the known and reasonably foreseeable risks to health, safety or fundamental rights, both when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse. It must also include the evaluation of other

⁸⁸ G. Couneson, *Article 7: Amendments to Annex III* in C. Pehlivan, N. Forgó and P. Valcke (eds) *The EU Artificial Intelligence (AI) Act: A Commentary*, Alphen aan den Rijn, 2024.

risks possibly arising, based on the analysis of data gathered from the post-market monitoring system. Following risk identification, suitable and proportionate risk management measures shall be implemented to mitigate risks to a level at which both the residual risks linked to individual hazards and the overall residual risk of the high-risk AI system are deemed acceptable.

The second part of the risk management system is the testing procedure, which serves the purpose of identifying the most appropriate and targeted risk management measures to ensure that high-risk AI systems perform consistently for their intended purpose and in compliance with other requirements laid out in the AI Act. Testing can help evaluate which measure to reduce risk is appropriate, if the system is likely to perform poorly in the environment it is intended for (perhaps due to a difference between the training and the actual environment), and whether certain obligations such as sufficient transparency (Article 13 AIA) or an appropriate level of accuracy (Article 15 AIA) are fulfilled⁸⁹.

While the obligation to implement risk management applies solely to high-risk AI systems, some scholars advocate for its voluntary extension to lower-risk systems. Schuett, for example, maintains that in a holistic risk management approach, all risks should initially be considered, since AI systems not classified as high-risk can still pose potential threats⁹⁰. He further notes that the costs associated with establishing a risk management framework are predominantly fixed, and thus expanding its scope to include additional systems would involve only marginal cost increases.

Both the AI Act and the DSA impose risk management obligations on the entities they govern. Consequently, there is a potential for practical overlap between the risk management obligations under the two instruments. It is important to note, however, that under the AI Act, these obligations apply exclusively to providers of AI systems. When VLOPs or VLOSEs act solely as deployers (such as when they use an AI system supplied by another entity), their obligations remain separate; while related and potentially burdensome, they do not overlap from a regulatory perspective. In contrast, when a VLOP or VLOSE also functions as a provider of the AI system or model used on its platform or search engine (e.g., by developing it, placing it on the market, or operating it under its own

⁸⁹ D. Schneeberger, W. Hotzendorfer and C. Tschohl, *Article 9: Risk Management System* in C. Pehlivan, N. Forgó and P. Valcke (eds), *The EU Artificial Intelligence (AI) Act: A Commentary*, Alphen aan den Rijn, 2024.

⁹⁰ J. Schuett, *Risk Management in the Artificial Intelligence Act*, in *European Journal of Risk Regulation*, vol. 14, 2023, p. 367 ff.

name), a factual overlap may occur, particularly if the system qualifies as high-risk⁹¹. While major platforms such as Meta and Google have developed proprietary AI-driven systems for moderating content, a range of companies also offer third-party digital content moderation solutions to other clients. These include Amazon, Microsoft, Accenture, and Appen⁹².

Should overlapping risk management obligations arise, the AI Act explicitly accommodates this possibility. Article 9(10) permits the combination of AI Act requirements with existing frameworks established to meet DSA compliance. Moreover, Recital 118 offers interpretive guidance by stating that, in such cases, the AI Act's obligations should be considered fulfilled through DSA-based risk management frameworks, unless significant systemic risks not covered by the DSA emerge and are identified in relevant AI models.

Beyond the establishment of a risk management system, other requirements directed at high-risk AI systems could play a significant role in addressing several of the shortcomings identified in AI-driven content moderation systems, such as the data and data governance requirements set out in Article 10. As discussed in Section 3, biased and unrepresentative training datasets often lead to discriminatory over-removal of content and under-detection of harmful material, particularly across different linguistic and cultural contexts. By requiring that training, validation, and testing datasets be relevant, representative, complete, and of high quality, Article 10 directly targets these issues. Its emphasis on robust data governance practices – including continuous monitoring, bias detection, and mitigation – would help ensure that moderation algorithms are trained on more diverse and accurate data. In turn, this would promote fairer and more consistent moderation outcomes.

The accuracy requirement set out in Article 15 AI Act could also help address persistent issues in AI-driven content moderation, particularly false positives and false negatives. Accuracy, as the name suggests, seeks to obtain accurate predictions in machine learning models, and is measured by how accurately the model predicts the value of test data sets. This testing for accuracy in machine learning models may be conducted through precision tests to assess the performance by measuring the accuracy or

⁹¹ H. Graux et al., *Interplay between the AI Act and the EU digital legislative framework*, Policy Department for Transformation, Innovation and Health Directorate-General for Economy, Transformation and Industry, October 2025, available at europarl.europa.eu.

⁹² F. Galli, A. Loreggia, and G. Sartor, *The Regulation of Content Moderation*, cit., p. 66-67.

precision of the AI model⁹³. By mandating that AI systems perform as accurately as possible and are empirically validated against test datasets, Article 15 promotes the development of models with proportionate and well-calibrated classification thresholds, which could reduce the occurrence of misclassifications of content.

The human oversight obligations established under Article 14 AI Act, in its turn, could help counter the opacity of automated moderation systems by ensuring the possibility of meaningful human review, particularly in “hard cases” that require contextual or cultural understanding. As Dietrich observes⁹⁴, *ex post* monitoring currently represents the only feasible form of oversight for large-scale content moderation; however, this approach remains insufficient without *ex ante* design obligations. The AI Act’s framework has the potential to institutionalise oversight mechanisms at the system level, rather than limiting them to individual complaint handling.

Complementing this, the transparency obligations outlined in Article 13 further strengthen accountability. The AI Act underscores that transparency involves designing and deploying AI systems in ways that enable traceability and explainability, ensuring that humans are aware when interacting with an AI system, and that deployers are informed about the system’s capabilities, limitations, and appropriate use. Recital 27 confirms transparency as a guiding principle applicable across all AI systems, while Article 13 specifically tailors these requirements to high-risk systems, ensuring that deployers can understand how the system functions, why it produces specific outputs, and how to interpret them responsibly.

Whether integrated into a core service or employed for content moderation, AI systems used by intermediary services are subject to extensive transparency obligations. Under the AI Act, prospective providers and deployers must, for instance, inform individuals when they are directly interacting with an AI system. In turn, the DSA obliges intermediary services to publish annual, publicly accessible reports outlining their content moderation practices, including details on the use of automated tools – such as AI systems applied for content filtering. While these transparency duties arise from distinct legal frameworks and do not generally overlap, practical intersections may occur when AI systems used for moderation are classified

⁹³ J. Constantino, *Article 15: Accuracy, Robustness and Cybersecurity* in C. Pehlivan, N. Forgó and P. Valcke (eds) *The EU Artificial Intelligence (AI) Act: A Commentary*, Alphen aan den Rijn, 2024.

⁹⁴ F. Dietrich, *AI-Based removal of hate speech from digital social networks*, cit., p. 2949 ff.

as high-risk under the AI Act or are based on general-purpose AI models presenting systemic risks⁹⁵.

Finally, it should also be acknowledged that certain LLMs used in content moderation may fall within the regime applicable to providers of general-purpose AI models with systemic risk under the AI Act. Where an LLM qualifies as a general-purpose AI model (GPAIM) with systemic risk, its provider is subject to the systemic risk management obligations set out in Article 55, including requirements to identify and mitigate systemic risks, conduct model evaluations, and implement appropriate technical and organisational safeguards. To the extent that content moderation tools rely on such models, these obligations may contribute to addressing some of the concerns discussed above, particularly those relating to bias, uneven performance across languages, and model robustness. However, automated moderation is still predominantly carried out through more specialised machine-learning classifiers rather than frontier LLMs. Accordingly, while the GPAIM systemic risk framework is relevant and may provide partial safeguards where applicable, it does not exhaust the regulatory questions raised by AI-driven content moderation systems that rely on other types of AI architectures.

In conclusion, the AI Act has the potential to complement the DSA by ensuring that the AI systems underpinning content moderation processes are developed and deployed in ways that are safe, transparent, and respectful of fundamental rights. Many of the obligations imposed on high-risk AI systems – such as those relating to data quality, risk management, human oversight, and transparency – could meaningfully mitigate the risks to freedom of expression and other fundamental rights identified earlier. However, this additional layer of protection is not yet available under the current framework. AI systems used for content moderation are not explicitly listed in Annex III of the AI Act (except where they are intended to influence democratic processes), and their inclusion would require careful and nuanced deliberation. Not all content moderation systems pose comparable risks. There is a clear difference, for instance, between algorithmic recommendation tools used for entertainment platforms such as Netflix or Spotify and large-scale content filtering or recommender systems deployed by VLOPs or VLOSEs that may shape political discourse or amplify filter bubbles.

⁹⁵ H. Graux et al., *Interplay between the AI Act and the EU digital legislative framework*, cit., p. 48 ff.

Because such systems are not included among the limited-risk use cases either, they currently fall outside the scope of the AI Act's binding obligations, relying instead on voluntary adherence to codes of conduct under Article 95. While this places them primarily within the realm of self-regulation, providers could still adopt certain high-risk requirements voluntarily to mitigate rights-based risks.

Importantly, the AI Act's technical governance mechanisms and the DSA's procedural safeguards are not redundant but complementary. The DSA regulates how platforms exercise and communicate moderation decisions, while the AI Act governs how the underlying AI systems are designed, tested, and validated. Together, they can establish a comprehensive framework that secures both the technical integrity and the procedural fairness of algorithmic content governance – addressing the entire lifecycle of moderation, from system design to user redress, and reinforcing accountability across both the technological and human dimensions of content moderation.

5. Conclusion

The pervasive reliance on artificial intelligence in content moderation reflects a structural necessity of the contemporary digital environment. AI tools are now deployed not only for content moderation *stricto sensu* – the identification, blocking, or removal of illegal or harmful material, but also for content curation, shaping the visibility, ranking, and dissemination of lawful speech within online platforms. Their use is driven by the sheer volume and velocity of content uploaded every second, which far exceeds the capacity of human moderators to review, assess, or contextualise. Platforms are therefore compelled to automate significant parts of their governance processes to meet legal obligations, manage operational burdens, and maintain safe online environments. This necessity, however, sits uneasily with the complex, contextual, and culturally embedded nature of human communication that these systems must evaluate.

As this article has shown, the use of AI for content moderation creates significant risks for fundamental rights, particularly the right to freedom of expression. Through a systematic mapping of risks, it identified how technical limitations, operational dynamics, and the broader sociotechnical environment in which AI systems function can interfere with both the active and the passive dimensions of freedom of expression. Errors arising from biased, incomplete, or unrepresentative training data can lead

to discriminatory outcomes, disproportionately silencing certain groups while failing to protect others. False positives directly suppress lawful speech, sometimes through opaque demotion or shadowbanning, depriving speakers of their ability to communicate and audiences of their ability to receive information. False negatives, in turn, allow harmful material to persist online, producing chilling effects that deter vulnerable groups from participating in public discourse. Proactive moderation can amount to a form of prior restraint, subjecting all users' contributions to pre-publication filtering without adequate procedural safeguards. Finally, the opacity of AI systems, combined with the limited information provided to users about automated decisions, undermines transparency, accountability, and the ability to contest moderation measures. Together, these risks highlight the structural ways in which AI-driven moderation can distort the online information environment.

Considering these risks, this article analysed the existing EU regulatory framework governing content moderation, showing how the Digital Services Act and the Artificial Intelligence Act approach the problem from different yet related angles. The DSA focuses on the organisational and procedural aspects of content moderation, establishing obligations concerning transparency, statements of reasons, user notice, internal complaints procedures, and systemic risk assessment. These mechanisms seek to ensure that moderation (whether carried out by human reviewers or automated systems) is exercised in a diligent, proportionate, and rights-respecting manner. At the same time, some shortcomings remain: the explanation duties risk becoming generic and uninformative at scale; transparency obligations remain insufficiently harmonised; and the DSA does not impose substantive requirements on how automated moderation tools are developed, trained, or evaluated. As a result, while the DSA strengthens procedural guarantees, it does not directly address the technical sources of risk inherent in AI-driven moderation.

Against this backdrop, the article argued that the AI Act's requirements for high-risk AI systems could provide an important complementary layer of protection for freedom of expression if certain AI-based content moderation systems were to be included within its scope. Obligations relating to risk management, data quality and governance, human oversight, transparency, and accuracy and robustness have the potential to mitigate some of the harms identified here, particularly those arising from technical and operational limitations of the AI systems deployed. These requirements could help mitigate misclassification, discriminatory outcomes, opacity, and lack of contextual understanding by

Giovana Peluso Lopes, Pratiksha Ashok
Freedom of Expression and AI-Driven Content Moderation

embedding safeguards at the design and development stage. Although the AI Act currently does not classify content moderation systems as high-risk, and their inclusion would likely require legislative amendment, such a classification would meaningfully contribute to addressing the structural vulnerabilities of AI-driven moderation. Meanwhile, voluntary observance of the relevant high-risk requirements remains possible under the current framework.

The DSA and the AI Act should not be understood as overlapping instruments, but rather as mutually reinforcing components of the EU's emerging governance architecture for online expression. The DSA regulates how platforms moderate, focusing on procedural fairness, transparency, and user rights. The AI Act regulates what they moderate with, targeting the design, training, and deployment of the underlying systems. Far from duplicating each other, the two instruments operate at different layers of the content moderation pipeline. If applied in tandem, they could form a coherent framework in which technical integrity and procedural safeguards work together to protect freedom of expression in the digital sphere. Although the AI Act's protective potential is not yet fully realised in relation to content moderation, its future alignment with the DSA could provide the additional protective layer necessary to ensure that AI-driven governance of speech remains compatible with the values of a democratic society.

Giovana Peluso Lopes, Pratiksha Ashok
Freedom of Expression and AI-Driven Content Moderation

ABSTRACT: Artificial intelligence has become integral to the governance of online speech, increasingly shaping how content is produced, disseminated, and restricted across digital platforms. As the volume of user-generated material continues to outpace the capacity of human reviewers, AI systems now play a central role in both the detection of potentially harmful or unlawful content and the curation of information flows through algorithmic ranking and recommendation. While such systems are indispensable for managing large-scale communication and maintaining safe online environments, their deployment carries significant implications for the fundamental right to freedom of expression. This article examines the relationship between AI-driven content moderation and freedom of expression, with the primary aim of identifying and mapping the risks that automated moderation practices pose to both the active and passive dimensions of this right. It also analyses the emerging European regulatory framework applicable to AI-based content moderation. In particular, it explores how the Digital Services Act governs the procedural and organisational aspects of moderation, and how the Artificial Intelligence Act regulates the design, development, and deployment of the underlying AI systems. The article further assesses the interplay between the two instruments, arguing that they approach the governance of online speech from distinct yet complementary perspectives. Finally, it investigates how certain provisions of the AI Act – especially those applicable to high-risk AI systems – could help mitigate some of the harms identified, if AI-based content moderation systems were to be brought within the scope of the high-risk classification.

KEYWORDS: content moderation – freedom of expression – artificial intelligence – AI Act – Digital Services Act.

Giovana Peluso Lopes – Postdoctoral Researcher, Tilburg Institute for Law, Technology, and Society – Tilburg University, Tilburg, Netherlands (g.lopes@tilburguniversity.edu)

Pratiksha Ashok – Postdoctoral Researcher, Tilburg Institute for Law, Technology, and Society – Tilburg University, Tilburg, Netherlands (p.ashok@tilburguniversity.edu)

Authority-Mediated Trade Secrecy in the AI Act: An Example of Functional Constitutionalisation of Transparency?

*Thomas Margoni, Leona King**

TABLE OF CONTENTS: 1. Introduction. – 2. A Targeted Transparency Architecture under Article 53 AI Act: Normative Context. – 3. Trade Secrets as a Conditional Constraint: The Legal Standard under the Trade Secrets Directive. – 4. Authority-Mediated Trade Secrecy: *CK v Dun & Bradstreet*. – 5. The Commission’s Template: Promise, Gaps, and Risks of Under Disclosure: From Principle to Practice. – 6. Conclusion: Towards a Principle of Transparency and Mediated-Authority Trade Secrecy

1. Introduction

Artificial intelligence (AI) systems increasingly structure access to information, goods, and public services across most aspects of modern life¹. Their operation often relies on large data-driven models, including large language models (LLMs), whose provenance and composition are often opaque to affected parties². This opacity is not merely technical; it raises questions of accountability, fairness, and the practical ability of individuals and institutions to exercise rights under Union law³.

* Leona King is funded by the Research Foundation – Flanders (FWO) [grant no. 11A0025N]. This article was subjected to double-blind peer review.

¹ Between October 2024 and March 2025, OpenAI’s ChatGPT search alone reached 41.3 million average monthly users in the EU – a sign of how deeply embedded these systems are becoming in everyday life, see OpenAI, [EU Digital Services Act \(DSA\) Monthly Active Recipients](#), October 2025, available at [help.openai.com](#).

² F. Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information*, Cambridge (MA), 2015, p. 193 and p. 217.

³ C. Novelli, M. Taddeo and L. Floridi, *Accountability in Artificial Intelligence: What It Is and How It Works*, in *AI & Society*, vol. 39, 2024, p. 1871 ff.; L. Grozdanovski, *In Search of Effectiveness and Fairness in Proving Algorithmic Discrimination in EU Law*, in *Common Market Law Review*, vol. 58, 2021, p. 99 ff.; P. Hacker, *Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination under EU Law*, in *Common Market Law Review*, vol. 55, 2018, p. 1143 ff.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

The EU Artificial Intelligence Act⁴ (AI Act) responds to this threat of opacity with a risk-based approach that embeds transparency into the governance of, among other forms of AI, general-purpose AI (GPAI). While styled as product-safety regulation, the Act also serves broader constitutional objectives grounded in the Charter of Fundamental Rights: notably the freedoms of expression and information, rights to privacy and data protection, good administration and effective remedy, and protections for property and the freedom to conduct a business⁵. These rights intersect each other and may very well collide, particularly when supporting opposite interests, such as in those cases where transparency and confidentiality are both seen, by different agents in the same transaction, as direct emanation of Charter prerogatives⁶. The Act therefore restages a familiar but pressing dilemma: how to reconcile transparency as a condition of accountability and trust with secrecy, or at least confidentiality, as a condition of innovation and competition⁷.

This article explores the reach and limits of trade secrecy in relation to the new obligations introduced by the AI Act, showing the sophisticated processes through which it operates as a conditional and reviewable legal entitlement mediating between the interests of confidentiality with those of transparency. As Aplin and others have argued, trade secrecy is inherently limited by competing rights and public interests⁸ and recent data governance instruments and case law increasingly subject trade secrecy claims to independent review⁹. On this trajectory, the article proposes that a comparable model of authority-mediated trade secrecy be adopted, or perhaps is already present in some form, in the AI context, enabling a proportionate reconciliation between transparency obligations and trade secret protection.

The analysis focuses on Article 53 AI Act and its disclosure framework for GPAI models. Article 53 establishes a tiered transparency architecture comprising: (a) technical documentation for competent

⁴ Regulation (EU) 2024/1689, “AI Act” or “AIA”.

⁵ Recital 48 AI Act.

⁶ C. Kuner, *Transborder Data Flows and Data Privacy Law*, Oxford, 2013.

⁷ V. Mayer-Schönberger and T. Ramge, *Reinventing Capitalism in the Age of Big Data*, London, 2018.

⁸ T. F. Aplin, *The Limits of EU Trade Secret Protection*, in S. Sandeen, C. Rademacher and A. Ohly (eds), *Research Handbook on Information Law and Governance*, Cheltenham, 2020.

⁹ E. De Noyette, L. Stähler and T. Margoni, *Data Secrets: The Data Act’s New Trade Secrets Framework*, in *IIC – International Review of Intellectual Property and Competition Law*, vol. 56, 2025, p. 984.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

authorities; (b) information for downstream integrators; (c) a copyright compliance policy; and (d) a public “sufficiently detailed summary” of training content. This differentiation reflects the Union’s embedding principles of transparency and proportionality¹⁰: sensitive information flows to narrower audiences under confidentiality; information essential to public accountability flows outward. Recital 107 clarifies the objective of the public summary: to enable «parties with legitimate interests, including copyright holders» to exercise and enforce their rights. Article 78, in turn, protects trade secrets but expressly subjects confidentiality to the Trade Secret Directive’s (TSD)¹¹ lawful disclosure and public interest boundaries. Read together, these provisions encode transparency and secrecy as mutually conditioning rights rather than hierarchical absolutes.

This reading is supported by doctrine¹² and recent case law¹³. The TSD protects information only where it is secret, derives commercial value from that secrecy, and is subject to reasonable protective measures, criteria that generally fit structured, bounded know-how rather than diffuse training corpora assembled from public sources. The Directive also recognises lawful disclosure where «required or allowed by Union law» and for the

¹⁰ K. Lenaerts, *Exploring the Limits of the EU Charter of Fundamental Rights*, in *ECLR*, vol. 8, 2012, p. 375; R. Schütze, *European Constitutional Law*, Cambridge, 2021, p. 223–227; G. de Búrca, *The Principle of Proportionality and Its Application in EC Law*, in *Yearbook of European Law*, vol. 13, 1993, p. 105 ff.; G. De Gregorio, *Digital Constitutionalism in Europe*, Cambridge, 2022, p. 262–272.

¹¹ Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (Trade Secrets Directive).

¹² T. F. Aplin, *The Limits of Trade Secret Protection*, cit., p. 1014, p. 1021; E. De Noyette, L. Stähler and T. Margoni, *Data Secrets*, cit., T. F. Aplin, [presentation at “\(Re\)evaluating Trade Secrets Protection in Light of AI”](#), Centre for Intellectual Property and Information Law Spring Conference, University of Cambridge, 2025, available at [cipil.law.com.uk](#); M. Leistner and L. Antoine, *IPR and the Use of Open Data and Data Sharing Initiatives by Public and Private Actors*, Brussels, 2022; T. F. Aplin, *Trading Data in the Digital Economy: A Trade Secrets Perspective*, in S. Lohsse, R. Schulze and D. Staudenmayer (eds), *Trading Data in the Digital Economy: Legal Concepts and Tools*, Baden-Baden, 2017, p. 59–72; T. F. Aplin, A. Radauer, M.-A. Bader et al., [The Role of EU Trade Secrets Law in the Data Economy: An Empirical Analysis](#), in *IIC – International Review of Intellectual Property and Competition Law*, vol. 54, 2023, p. 826.

¹³ CJEU, 4 May 2023, C-487/21, *Österreichische Datenschutzbehörde and CRIF*, §§ 33, 34 and 39; CJEU, 26 October 2023, C-307/22, *FT (Copies of Medical Records)*, § 73; Opinion of Advocate General Richard de la Tour, 12 September 2024, in Case C-203/22, *CK v Dun & Bradstreet Austria GmbH*, § 45; Opinion of Advocate General Cruz Villalon, 9 July 2015, in Case C-201/14, *Bara*, § 74; Article 29 Working Party, *Guidelines on Transparency under Regulation 2016/679*, adopted 29 November 2017, last revised and adopted 11 April 2018, WP260 rev.01, p. 4.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

protection of legitimate interests such as textual gateways that accommodate the AI Act's transparency duties. Interestingly, albeit in the different context of personal data protection, in *CK v Dun & Bradstreet* (C-203/22) the Court of Justice rejected blanket invocations of trade secrecy to defeat access rights and required case-by-case, authority-led proportionality review where secrecy is claimed. While the case arises under the GDPR¹⁴, its logic travels: where fundamental rights are engaged, controllers cannot self-certify secrecy to displace transparency.

Against this backdrop, the Commission's 2025 Explanatory Notice and Template¹⁵ for Article 53(1)(d) aims to operationalise the public summary. While guidance and standardisation are welcome, the current design risks a degree of defensive opacity by over-delegating disclosure choices to data holders (providers in the AIA parlance), relying on narrative description, and limiting dataset level identifiability. The result may be formal transparency without functional verifiability, frustrating Recital 107's aim of enabling rightsholders and other parties with legitimate interests to act.

The article proceeds in four steps. Section 2 reconstructs Article 53's tiered transparency architecture and clarifies how the Act differentiates audiences (regulators, downstream integrators, the public) and purposes (oversight, interoperability, accountability). Section 3 supports and further substantiates the claim that trade secrets function as a conditional constraint, not a categorical bar¹⁶, and explains how Articles 53 and 78 incorporate the TSD's lawful disclosure and public interest exemptions in a coherent way. Section 4 develops the idea of authority-mediated trade secrecy as a procedural standard, borrowing from the framing of the CJEU in *CK v Dun & Bradstreet* as well as from scholar analysis in the Data Act¹⁷. It translates the secrecy-transparency dichotomy into a reviewable process:

¹⁴ Regulation (EU) 2016/679 (General Data Protection Regulation).

¹⁵ European Commission, [Explanatory Notice and Template for the Public Summary of Training Content for General-Purpose AI Models](#), Brussels, 2025, available at digital-strategy.ec.europa.eu.

¹⁶ U.-M. Mylly, *Transparent AI? Navigating Between Rules on Trade Secrets and Access to Information*, in *European Journal of Law and Technology*, vol. 14, 2023, p. 1014; T. F. Aplin, *The Limits of Trade Secret Protection*, cit.; A. Van Caenegem and L. Desautettes-Barbero, *Trade Secrets and Intellectual Property*, 2nd ed., Cheltenham, 2025.

¹⁷ 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 and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828, OJ L 2023/2854 (Data Act, DA); see T. Margoni and E. De Noyette, *Bedrijfsgeheimen in de Dataverordening: Een Administratieve Wending*, in *Intellectuele Eigendom en Reclamerecht*, vol. 41, no. 1, 2025, p. 1–4.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

the burden lies with providers to substantiate secrecy; competent authorities decide; reasons are recorded and reviewable; and disclosure must remain as full as possible. Section 5 applies this framework to the Commission's Template, identifying where its design under-delivers on Recital 107 requirements and proposes a practical blueprint that reconciles transparency and trade secrecy through a multilayered disclosure framework combining targeted transparency¹⁸ with authority-mediated trade secrecy. By introducing a model of authority-mediated trade secrecy, the proposed blueprint would successfully embed oversight and accountability into the regulatory design of the AI Act, thereby advancing what, in our views, is the Union's broader, yet perhaps implicit, project of *functional constitutionalisation of transparency*.

The conclusions argue that, read through the TSD and informed by the CJEU's approach, the AI Act supports a rebuttable presumption of transparency for information necessary to support legitimate interests and rights enforcement. Secrecy persists, but only when justified by proved necessity and verified by independent authority. This is not a model to dilute innovation incentives; rather, it is to ensure that transparency becomes auditable and effective, not merely declaratory. In that sense, the AI Act offers a template for a governance model in which transparency and trade secrecy are co-managed through proportionate procedures rather than asserted as absolutes.

In this article, "functional constitutionalisation of transparency" refers to the process by which transparency obligations, while not framed as an autonomous Charter right, acquire constitutional force because they function as indispensable preconditions for the exercise of other fundamental rights, including access to information, effective remedy, and freedom of expression. Where transparency serves this enabling function, secrecy cannot operate as a self-certified or categorical exception, but must be justified through proportionate, authority-mediated review.

¹⁸ P. Keller and T. F. Aplin, *Reconciling Trade Secrets and AI Public Transparency*, 4 December 2024, available at ssrn.com (developing the concept of targeted transparency in the context of AI and data law); F. Pasquale, *Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries*, in *Northwestern University Law Review*, vol. 104, 2010, p. 105 (introducing the notion of qualified transparency); M. E. Kaminski, *Understanding Transparency in Algorithmic Accountability*, in W. Barfield (ed.), *The Cambridge Handbook of the Law of Algorithms*, Cambridge, 2020, ch. 5, p. 121–138; M. Maroni, *Mediated Transparency: The Digital Services Act and the Legitimation of Platform Power*, in M. Hillebrandt, P. Leino-Sandberg and I. Koivisto (eds), *(In)visible European Government: Critical Approaches to Transparency as an Ideal and a Practice*, London, 2023, ch. 16.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

2. *A Targeted Transparency Architecture under Article 53 AI Act:
Normative Context*

Transparency occupies an increasingly central position in the EU fundamental rights discourse, yet its conceptual and normative foundations remain diffuse. It has been framed as a *principle*, a *right*, a *policy tool*, and a *technique of accountability*¹⁹. Each form expresses a slightly different logic – procedural fairness in administration, access to data and documents, algorithmic explainability, or market oversight²⁰. In this sense, transparency operates less as a single legal obligation than as a context-dependent framework through which competing rights and interests are mediated²¹.

Article 53 AI Act forms the structural core of the Regulation's transparency regime with regard to GPAI models. It operationalises several Charter rights, most notably the freedoms of expression and information (Article 11 CFR; Article 10 ECHR), and the rights to good administration and access to information (Articles 41–42 CFR), by imposing positive disclosure obligations on private actors whose activities have public interest effects. At the same time, it safeguards other fundamental rights and freedoms by ensuring that these disclosure obligations remain proportionate and do not extinguish legitimate commercial confidentiality.

Rather than imposing a uniform transparency requirement, Article 53 differentiates the intensity of disclosure according to the audience and purpose of information use. Information flows to regulators, downstream providers, and the public through distinct channels, each governed by a different degree of precision and confidentiality²². In parallel, the *General-Purpose AI Code of Practice*²³ – a voluntary instrument developed by independent experts at the request of the European Commission – offers interim guidance to support responsible development and deployment of

¹⁹ A. Buijze, *The Principle of Transparency in EU Law*, Utrecht, 2013, p. 47–53; A. Buijze, *Transparency: The Swiss Knife of EU Law*, in *European Review of Public Law*, vol. 26, no. 3, 2013, p. 1123 ff.

²⁰ J. Krook, et al., *A Systematic Literature Review of Artificial Intelligence (AI) Transparency Laws in the European Union (EU) and United Kingdom (UK): A Socio-Legal Approach to AI Transparency Governance*, in *AI and Ethics*, vol. 5, 2025, p. 4069; M. D. Cole, et al., *Algorithmic Transparency and Accountability of Digital Services*, Strasbourg, 2023, available at rm.coe.int; A. Koene, et al., *A Governance Framework for Algorithmic Accountability and Transparency*, Brussels, 2019, available at europarl.europa.eu.

²¹ A. Buijze, *Transparency: The Swiss Knife of EU Law*, cit., p. 1123.

²² U.-M. Mylly, *Transparent AI?*, cit., p. 1014, p. 1021.

²³ European Commission, *General-Purpose AI Code of Practice* (Digital Strategy, 2024), available at digital-strategy.ec.europa.eu.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

GPAI models in anticipation of the AI Act's entry into force, though its detailed implications fall outside the scope of this paper.

2.1. *Regulatory Disclosure*

The first tier concerns regulatory disclosure to the AI Office and national competent authorities. Under Article 53(1)(a) providers of GPAI models must supply upon request the technical documentation listed in Annex XI. Following a different language, but reaching an arguably equivalent obligation, Article 53(1)(c) requires providers of GPAI to provide their copyright compliance policy to the AI Office in the light of a general duty to collaborate and the monitoring powers of the Office (e.g., Arts. 53(1)(3), 56, 68, 88, 89). Access to these materials is intended to allow regulators to verify conformity *ex ante* and investigate compliance *ex post*. Regulatory transparency thus functions as oversight – enabling administrative review while remaining within a confidential channel protected by Article 78.

2.2. *Downstream Disclosure*

The second tier addresses downstream disclosure between GPAI providers and those integrating the model into AI systems. Article 53(1)(b) requires the provision of «the information and documentation necessary» for compliant integration. Although typically implemented through contractual terms, this duty originates in statute and therefore has a public law character within private relationships.

This horizontal transparency extends accountability through the value chain, allowing system developers to meet their own regulatory duties under the AI Act. It also exemplifies how the Regulation projects public law principles, such as fairness, proportionality, and reason giving, into the sphere of private governance. Yet the obligation is explicitly conditioned by Article 53(7), which reiterates that the sharing of information must not compromise trade secrets or confidential business information. The result is a form of targeted transparency regime: disclosure is mandated, but its scope must be limited to what is strictly necessary for downstream compliance.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

2.3. *Public Disclosure: The “Sufficiently Detailed Summary”*

The third and most visible tier introduces public disclosure through Article 53(1)(d), requiring GPAI providers to make *publicly available* a «sufficiently detailed summary about the content used for training the model». This summary is the only element of proactive transparency directed toward the general public. Its purpose is to render AI development traceable, to enable the exercise of rights, particularly copyright²⁴, (personal) data protection²⁵, and non-discrimination claims²⁶, and to foster trust in the regulatory system.

Recital 107 clarifies the intended equilibrium: to respect legitimate secrecy while ensuring that the summary is «generally comprehensive rather than technically detailed». Providers should identify the «main data collections or sets» used for training, including major public or private databases, and give a narrative account of other sources. This formulation establishes an intermediate standard: the summary must reveal the character and provenance of training data without exposing its specific configuration or internal processing. In other words, the legislator defines transparency by function – a disclosure sufficient to make rights exercisable – rather than by “substance”.

The Commission’s 2025 Template translates this open-textured obligation into a structured format, distinguishing between publicly available, privately licensed, scraped, and synthetic datasets. The model reflects proportional reasoning: the more public the dataset, the more specific the required disclosure; the more private or confidential, the more general the permitted description. Yet, as later sections argue, this calibration may err on the side of caution, allowing providers to default to generality and thereby diluting the summary’s value as an enabler of transparency and of the rights that rely on such transparency in order to be effectively exercised.

²⁴ On the relationship with copyright and AI act see A. Peukert, [Copyright in the Artificial Intelligence Act – A Primer](#), in *GRUR International*, vol. 73, no. 6, June 2024, p. 497 ff.

²⁵ M. Nisevic, A. Cuypers and J. De Bruyne, [Explainable AI: Can the AI Act and the GDPR Go Out for a Date?](#), in *Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, 2024, p. 1 ff.

²⁶ F. Lütz, [The AI Act, Gender Equality and Non-Discrimination: What Role for the AI Office?](#), in *ERA Forum*, vol. 25, 2024, p. 79 ff.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

2.4. *A Model of Tiered Transparency*

Taken together, these three layers compose a tiered or targeted transparency architecture²⁷. Each tier serves a different function:

Tier	Primary recipient	Function	Legal basis	Fundamental rights implicated
Regulatory	AI Office / competent authorities	Oversight and enforcement	Arts 53(1)(a), (c), 78	Good administration (Art 41); effective remedy (Art 47); freedom to conduct a business (Art 16); property (Art 17)
Downstream	Integrating providers	Interoperability and compliance	Arts 53(1)(b), (7)	Freedom of the arts and sciences (Art 13); freedom to conduct a business (Art 16).
Public	General public	Accountability and trust	Art 53(1)(d), Rec 107	Freedom of expression (Art 11); access to documents (Art 42); privacy & data protection (Arts 7–8); copyright property (Art 17(2))

Table 1 – Multi-tiered transparency structure under the AI Act (author’s own assessment)

As Keller and Aplin observe, contemporary transparency regimes have evolved beyond the undifferentiated “right to know” of Freedom-of-Information (FOI) towards targeted transparency, i.e., disclosures tailored to specific purposes, procedures, and parties, with calibrated interfaces to trade secret claims²⁸. Transparency functions meaningfully only when its form and audience are aligned with its underlying accountability objective: disclosure without contextual intelligibility does not ensure oversight or

²⁷ P. Keller and T. F. Aplin, *Reconciling Trade Secrets and AI*, cit. (arguing for meaningful transparency achieved through targeted and context-specific disclosure mechanisms).

²⁸ P. Keller and T. F. Aplin, *Reconciling Trade Secrets and AI*, cit.; See related scholarship conceptualising *layered* or *qualified transparency*. A. Fung, M. Graham and D. Weil, *Full Disclosure: The Perils and Promise of Transparency*, Cambridge, 2007; M. E. Kaminski, *Understanding Transparency*, cit.; F. Pasquale, *The Black Box Society*, cit.; G. De Gregorio, *Digital Constitutionalism in Europe*, Cambridge, 2022; E. Kosta, *Peeking into the Black Box: Transparency Rights under the GDPR and the AI Act*, in *Computer Law and Security Review*, vol. 52, 2023, article 105819.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

fairness²⁹. The literature has accordingly emphasised the need for the AI Act to supplement traditional FOI regimes, normally limited to public administrations, with structured, audience-sensitive forms of transparency³⁰. Read in this light, the AI Act's Article 53(1)(d)'s requirement of a "sufficiently detailed summary" can be understood as a targeted instrument that enables rights-bearing entities (data subjects, rightsholders, researchers, regulators) without compromising secrecy.

The EU's broader regulatory framework already operationalises targeted transparency through differentiated, context-sensitive access rights. As Keller and Aplin outline³¹, this approach manifests across several instruments including Article 15 (right of access) and Article 22 (right to explanation) GDPR³² as well as the access granted to vetted researchers under Art 40 DSA³³. The list could be further expanded with the numerous access and portability rights, and their relationship with TS, introduced by the Data Act³⁴.

However, without mechanisms for independent review of confidentiality claims, the risk is that tiered transparency collapses into self-certified opacity. These tensions between disclosure obligations and the protection of confidential information are not new³⁵, but the AI Act restages them in the novel context of AI systems and data governance³⁶. The following sections therefore examine how trade secrets operate as a conditional constraint and how authority-mediated procedures can ensure that secrecy remains the exception, not the rule.

²⁹ M. E. Kaminski, *Understanding Transparency*, cit.; P. Keller, [Participatory Accountability at the Dawn of Artificial Intelligence](#), Dickson Poon School of Law Legal Studies Research Paper Series, 2019, p. 15–16.

³⁰ H. P. Olsen, et al., *The Right to Transparency in Public Governance*, in *Digital Government Research*, vol. 5, 2024.

³¹ P. Keller and T. F. Aplin, *Reconciling Trade Secrets and AI*, cit.

³² L. Edwards and M. Veale, *Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For*, in *Duke Law and Technology Review*, vol. 16, 2017, p. 18 ff.

³³ M. Maroni, *Mediated Transparency: The Digital Services Act and the Legitimation of Platform Power*, in M. Hillebrandt, P. Leino-Sandberg and I. Koivisto (eds), *(In)visible European Government: Critical Approaches to Transparency as an Ideal and a Practice*, London, 2023.

³⁴ E. De Noyette, L. Stähler and T. Margoni, *Data Secrets*, cit.

³⁵ E. A. Rowe, *Striking a Balance: When Should Trade-Secret Law Shield Disclosures to the Government?*, in *Iowa Law Review*, vol. 96, 2010, p. 791.

³⁶ P. Keller and T. F. Aplin, *Reconciling Trade Secrets and AI*, cit.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
 An Example of Functional Constitutionalisation of Transparency?*

3. *Trade Secrets as a Conditional Constraint: The Legal Standard under the Trade Secrets Directive*

Trade secrets occupy an intermediate space in EU law. As Desauettes-Barbero observes, «it is challenging to link [trade secret] protection to the interests safeguarded by specific fundamental rights»³⁷. The Trade Secrets Directive³⁸ (TSD) itself is neutral on any fundamental-rights foundation: its Charter references chiefly ensure it does not restrict freedom of expression and information (Recital 19), rather than asserting a positive fundamental right to secrecy³⁹. This textual silence underscores the absence of a settled fundamental rights basis for trade secret protection and, correspondingly, its adaptability within the broader Charter framework.

Scholarly and judicial analysis have linked trade secrecy to different Charter rights depending on context: as a facet of *property* (Article 17 CFR)⁴⁰, as privacy-adjacent (Article 7, e.g., *Varec* (C-450/06)⁴¹, or as part of to the *freedom to conduct a business* (Article 16 CFR)⁴².

The absence of a clear statutorily defined fundamental-rights anchor makes the balancing of trade secrecy with other Charter rights, such as freedom of expression, access to information, or data protection, particularly complex and invites further research into how proportionality and institutional design might provide coherence across these intersecting

³⁷ L. Desauettes, *Trade Secrets Legal Protection: From a Comparative Analysis of US and EU Law to a New Model of Understanding*, in *Munich Studies on Innovation and Competition*, no. 19, Munich, 2023, p. 115.

³⁸ Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (Trade Secrets Directive).

³⁹ European Commission, *Proposal for a Directive of the European Parliament and of the Council on the Protection of Undisclosed Know-How and Business Information (Trade Secrets) Against Their Unlawful Acquisition, Use and Disclosure*, COM (2013) 813 final, 2013/0402 (COD), explanatory memorandum noting that the initiative promotes the rights to property and to conduct a business, incorporates safeguards for the rights of defence and to a fair trial, ensures respect for freedom of expression and information, and recognises the importance of safeguarding the rights to privacy and data protection.

⁴⁰ L. Desauettes, *New Model of Understanding*, cit., p. 115; T. F. Aplin, *The Limits of Trade Secret Protection*, cit.; T. F. Aplin, L. Bently, P. Johnson and S. Malynicz, *Gurry on Breach of Confidence*, 2nd ed., Oxford, 2012, ch. 5.

⁴¹ E. De Noyette, *Rights of Access and Trade Secrets: Conflicting Interests, Carefully Balanced*, 2025, available at stradalex-com.kuleuven.e-bronnen.be.

⁴² V. Cassiers, *La directive 2016/943/UE du 8 juin 2016 sur les secrets d'affaires*, in *Journal des Tribunaux*, no. 385, 2017, p. 396; V. Cassiers and A. Strowel, *La directive du 8 juin 2016 sur la protection des secrets d'affaires*, in V. Cassiers (ed.), *Le secret*, Brussels, 2017, p. 89 ff.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

regimes, especially in emerging contexts such as AI governance and data access regulation.

Against this broader fundamental rights framing, the TSD provides the positive law framework for determining what qualifies as a protectable secret. Article 2(1) identifies the familiar conditions of a trade secret as information that is: (i) not generally known; (ii) has commercial value because it is secret; and (iii) has been subject to reasonable steps to keep it secret⁴³.

These cumulative conditions embed proportionality within the concept itself: secrecy protection extends only as far as confidentiality is demonstrably maintained and commercially justified⁴⁴. The Directive also establishes a qualified rather than absolute privilege. Articles 1(2)(b), 3(2), and 5 make clear that lawful disclosure may occur where required or permitted by Union or national law, or to protect a legitimate interest, including whistleblowing, workers' representation, and public interest disclosure, all relevant elements in the case of AI⁴⁵. In effect, the TSD asserts trade secrecy as a conditional entitlement, constrained by higher order legal and societal values⁴⁶.

3.1. *Trade Secrets in the AI Act*

The AI Act incorporates this conditional logic directly. Articles 53(7) and 78(1)(a) require the AI Office, national authorities, and other competent bodies to protect «intellectual property rights and confidential business information or trade secrets, including source code». Article 78 AI Act further qualifies the obligation: it does not apply «in the cases referred to in Article 5 of Directive (EU) 2016/943». By explicitly recalling – *ad abundantiam* – the TSD's exceptions the AI Act reinstates, beyond any possible hermeneutic exercise, the nature of trade secrecy.

Read together, Articles 53 and 78 create a two-step test. First, information may be withheld only where it qualifies as a trade secret within the meaning of the TSD, as recognised in Article 78(1). Second, that

⁴³ Art. 2(1) Directive (EU) 2016/943 (Trade Secrets Directive).

⁴⁴ T. F. Aplin, *The Limits of Trade Secret Protection*, cit.

⁴⁵ U.-M. Mylly, *Transparent AI?* cit., p. 1014, p. 1021; T. F. Aplin, *The Limits of Trade Secret Protection*, cit.

⁴⁶ A. Ohly, *Jurisdiction and Choice of Law in Trade Secrets Cases: The EU Perspective*, in S. Sandeen, C. Rademacher and A. Ohly (eds), *Research Handbook on Information Law and Governance*, Cheltenham, 2021.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

confidentiality is not absolute: Article 53(1)(d) expressly requires providers of general-purpose AI models to «make publicly available a sufficiently detailed summary about the content used for training», while Article 78(1) permits disclosure of trade secrets in the situations referred to in Article 5 TSD, i.e. where required by Union law or for the purpose of protecting a legitimate interest. Recital 107 elaborates the rationale for this balance, clarifying that the public summary aims to «facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law» while still respecting trade secret protection. Confidentiality under the AI Act is therefore reactive: it restricts onward dissemination by authorities (Articles 53(1)(a) and 78(1)) but does not nullify the provider's primary duty to disclose in the first place when Union law so requires, including publicly in specific situations (Article 53(d)).

This design reflects the Union's established method of reconciling openness and secrecy through proportional reasoning, a common approach in access-to-documents jurisprudence⁴⁷. Secrecy is never a self-executing defence; it is an interest that must be balanced against competing rights and assessed in context.

3.2. *Definitional Scope and the Likelihood of Trade Secret Qualification*

Although the complexity of AI systems raises difficult questions, the prevailing literature suggests that only a limited subset of the underlying information is likely to meet the TSD's definitional threshold in practice⁴⁸. As Aplin observes, trade secrets law presupposes information that is structured, identifiable, and human generated⁴⁹.

Though not an explicit condition under Article 2 TSD, identifiability is emerging as a jurisprudential requirement closely linked to the Directive's "reasonable steps" criterion. The EUIPO's 2023 litigation survey⁵⁰

⁴⁷ See, e.g., GC, 12 October 2007, Case T-474/04, *Pergan v Commission*, CJEU, 18 July 2017, *Commission v Bavarian Lager*, Case C-213/15 P.

⁴⁸ S. K. Sandeen and T. F. Aplin, *Trade Secrecy, Factual Secrecy and the Hype Surrounding AI*, in R. Abott (ed.), *Research Handbook on Intellectual Property and Artificial Intelligence*, Cheltenham, 2022, p. 454–456.

⁴⁹ T. F. Aplin, [presentation at "\(Re\)evaluating Trade Secrets Protection in Light of AI"](#), *Centre for Intellectual Property and Information Law Spring Conference*, University of Cambridge, 2025, available at cipil.law.cam.ac.uk.

⁵⁰ European Union Intellectual Property Office, *Trade Secret Litigation Trends in the EU: Report on 2023 Case-Law Developments*, Alicante, 2023, available at euipo.europa.eu.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

highlights decisions across Member States applying this evidentiary discipline: claims framed at a high level of abstraction e.g. vague or blanket claims such as asserting that “all company information” or “everything discussed” constitutes a trade secret are routinely rejected for lack of precision⁵¹.

This specificity requirement matters for AI. Large training datasets used to train large scale AI models are rarely so bounded. Collections such as Common Crawl or LAION-5B aggregate heterogeneous, publicly available material, much of which lacks the confidentiality or internal coherence necessary to constitute a trade secret. Even where data are curated or (pre-)processed, they often remain too diffuse to qualify as an identifiable “body” of knowledge and are commonly drawn from sources that anyone in the field could access⁵². In these circumstances, the claim that disclosure of a “sufficiently detailed summary” would reveal a trade secret appears legally tenuous.

Perhaps, more than the dataset itself, what may form part of a potential trade secret are the parameters or selection logic developed by the AI developer to identify and extract particular data sources. The precise “recipe” to select those sources (e.g., a set of instructions or algorithms given to a crawler on how to identify data that is believed to be relevant for the purposes), could perhaps more logically represent secret information which has commercial value in reason of its confidentiality, thereby satisfying the criteria of Article 2(1) TSD. However, the template does not require disclosure of such instructions or algorithms, but only the resulting datasets. Although one might, in theory, attempt to infer the elements of the selection logic through reverse engineering of the disclosed data, this

⁵¹ Estonian Criminal Case Against K. M., J. K., M. K. and BloomEst OÜ (15 May 2020); German Regional Labour Court of Düsseldorf, docket no. 12 SaGa 4/20 (3 June 2020); Hungarian Supreme Court, Gfv.VII.30.179/2020/4 (21 January 2021), Italian Supreme Court of Cassation, no. 34337 (27 December 2019): as cited in *EUIPO, Trade Secret Litigation Trends in the EU: Report on 2023 Case-Law Developments* (n.); T. F. Aplin and L. Bently, *Gurry on Breach of Confidence*, 3rd ed., Oxford, 2020, ch. 5.

⁵² See, especially with respect to data generated by connected devices: J. Drexl, *Designing Competitive Markets for Industrial Data – Between Propertisation and Access*, in *Journal of Intellectual Property, Information Technology and E-Commerce Law*, vol. 8, 2017, p. 257–292 <https://doi.org/10.2139/ssrn.2862975>; J. Drexl, *The (Lack of) Coherence of Data Ownership with the Intellectual Property System*, in N. Bruun, G. B. Dinwoodie, M. Levin and A. Ohly (eds), *Transition and Coherence in Intellectual Property Law: Essays in Honour of Annette Kur*, Cambridge, 2021, p. 213–223; J. Drexl, *Data Access and Control in the Era of Connected Devices*, study on behalf of the European Consumer Organisation (BEUC), Brussels, 2018, available at benc.eu.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

would rarely meet the threshold for revealing protectable information. The relationship between dataset composition and the underlying selection parameters is neither linear nor transparent: a range of alternative filtering, weighting, or ranking methods could generate comparable outputs. Without insight into the provider's internal infrastructure, any attempt to reconstruct the "recipe" would rely on conjecture rather than demonstrable access. In legal terms, such speculative inference would in all likelihood not make the information "readily accessible" within the meaning of the Directive, nor deprive it of the quality of secrecy. Consequently, while the risk of reverse engineering merits contextual evaluation, the structure of the template, which requires only a high-level, categorical identification of training data rather than detailed work-by-work documentation, substantially limits the likelihood that disclosure under Article 53 would compromise a valid trade secret.

Nevertheless, trade secret law can have a *de facto* categorical effect, since the determination of protection typically occurs *ex post* and through litigation between usually unequally situated parties. The result is an asymmetry of information and power, where dominant actors may use trade secret claims strategically to preserve opacity⁵³. Furthermore, certain subsets of training data may satisfy the TSD criteria: proprietary, human-generated datasets compiled for specialised domains (for instance, annotated medical images, industrial schematics, or internal user interaction logs). Where such data have been maintained in confidence, possess commercial value by virtue of their secrecy, and are secured through reasonable measures (contractual restrictions, limited access, technical safeguards), they may legitimately attract trade secret protection. The extent to which this protection should constrain transparency obligations, and how the relative tension is addressed in practice, is examined below Section 4 and in Section 5.

3.3. *Confidential Business Information, Overclaiming, and Emerging "Data Secrets"*

The AI Act's reference to «confidential business information or trade secrets» raises potential definitional ambiguity. As Mylly notes, the two

⁵³ T. F. Aplin, A. Radauer, M. Bader and N. Searle, *The Role of EU Trade Secrets Law in the Data Economy: An Empirical Analysis*, in *IIC – International Review of Intellectual Property and Competition Law*, vol. 54, 2023, p. 826, p. 836.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

terms are likely used interchangeably, with “confidential business information” serving as a descriptive restatement of the TSD’s definition rather than a separate category⁵⁴. Nonetheless, the phrasing risks strategic overclaiming: providers may designate as “confidential” any dataset, annotation method, or data curation workflow, thereby undermining Article 53’s transparency function. To maintain coherence, “confidential business information” should be interpreted *ejusdem generis* with the TSD, restricted to information that is genuinely secret, commercially valuable because of that secrecy, and subject to reasonable protection measures.

This problem resonates with a broader conceptual evolution in EU information law: the rise of “data secrets”⁵⁵. Under instruments like the Data Act, trade secrecy increasingly coexists with duties of access, sharing, and oversight. Information may remain confidential yet still be subject to controlled disclosure, mediated by administrative procedures or technical protection measures⁵⁶. What is particularly interesting in the proposed category of data secrets is the role of (national) administrative authorities which must adjudicate cases where a data holder under a Data Act obligation to share data wants to elude this obligation. A prominent “defence” in this situation is the claim that the information under a sharing obligation contains, or constitutes, a trade secret. Whereas different types of sharing obligations lead to different types of assessment, a common element is the fact that the administrative authority will have to decide *prima-facie* the validity of the trade secret (counter-) claim by the data holder. That institutional posture maps neatly onto the authority-mediated model advanced here – and anticipated by Keller & Aplin’s recommendation to place independent bodies between private claims to secrecy and public claims to transparency⁵⁷.

4. *Authority-Mediated Trade Secrecy: CK v Dun & Bradstreet*

The Court of Justice’s recent judgment in *CK v Dun & Bradstreet Austria GmbH* (C-203/22)⁵⁸ provides a crucial reference point for understanding how Union law reconciles transparency with trade secret

⁵⁴ U.-M. Mylly, *Transparent AI?* cit., p. 1014.

⁵⁵ E. De Noyette, L. Stähler and T. Margoni, *Data Secrets*, cit.

⁵⁶ *Ibid.*

⁵⁷ P. Keller and T. F. Aplin, *Reconciling Trade Secrets and AI*, cit.

⁵⁸ CJEU, 27 February 2024, C-203/22, *CK v Dun & Bradstreet Austria GmbH* (*CK v Dun & Bradstreet*),

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

protection in data intensive contexts⁵⁹. The case concerned an individual's request for access under Article 15(1)(h) of the General Data Protection Regulation (GDPR) to personal data processed by a credit scoring company. The company refused full disclosure, invoking trade secret protection.

The Court, building on prior case law⁶⁰, and considering the transparency obligations under Article 12(1) GDPR held that where disclosure might affect trade secret interests, information must still be made available at least to a competent authority or court which then conducts a case-by-case balancing and determines the scope of access in light of the specific circumstances⁶¹. Importantly, the Court emphasised that such balancing cannot be predetermined by national law: Member States may not enact provisions that automatically give priority to trade secret protection⁶².

4.1. *Emerging Principles*

Three strands of reasoning in *CK v Dun & Bradstreet* are particularly relevant for interpreting Articles 53 and 78 AI Act. First, the rejection of blanket confidentiality. The Court affirmed that transparency cannot be displaced by abstract references to business secrecy, reaffirming that «That right [of access] should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property [...] However, the result of those considerations should not be a refusal to provide all information to the data subject»⁶³.

Second, the requirement for independent determination. Only supervisory authorities or courts may decide whether a trade secret claim is valid and whether nondisclosure is proportionate. Private actors may not define for themselves the limits of transparency. This point directly

⁵⁹ For review of the case see E. De Noyette, *Rights of Access and Trade Secrets*, cit.

⁶⁰ Opinion of Advocate General Richard de la Tour, 12 September 2024, in Case C-203/22, *CK v Dun & Bradstreet Austria GmbH*, cit.; see also ECJ, 7 December 2023, C-634/21, *SCHUFA Holding and Others (Scoring)*; CJEU, 26 October 2023, C-307/22, *FT (Copies from the Medical File)*, cit.; and CJEU, 4 May 2023, C-487/21, *Österreichische Datenschutzbehörde and CRIF*, cit.

⁶¹ This case by case recommendation echoes the proposal of G. Malgieri, *Trade Secrets v Personal Data: A Possible Solution for Balancing Rights*, in *International Data Privacy Law*, vol. 6, no. 2, May 2016, p. 102 ff.

⁶² See *SCHUFA Holding and Others (Scoring)*, C-634/21, ECLI:EU:C:2023:957, § 70 and the case law cited; see also § 75 of the judgment.

⁶³ Judgment, § 3, citing Recital 63 GDPR.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

undermines any reading of the AI Act that would allow providers to unilaterally withhold information under Article 53 on the basis of self-declared confidentiality. At the same time, it paves the way for a much more significant role for the AI Office (and other authorities) in precisely defining this dynamic – a role that these bodies must be willing to embrace⁶⁴.

Third, the obligation of intelligibility. Even where some information remains confidential, controllers must provide explanations that are «concise, transparent, intelligible and easily accessible». Complexity or proprietary character cannot nullify the duty to communicate in a form that enables understanding. This requirement of functional comprehensibility should arguably apply *mutatis mutandis* to AI model documentation: summaries must be understandable enough to allow external scrutiny, even if they do not reveal technical detail.

These elements together logically demand a clear procedural standard of proportionality: secrecy claims must be specific, verified and justified, and transparency must remain the rule and confidentiality the exception.

It is important to emphasise that this judicial approach is not confined to data protection law⁶⁵. A comparable model of authority-mediated transparency has long been articulated in EU public procurement case law when trade secrets are invoked. In *Antea Polska* (C-54/21)⁶⁶, the CJEU held that contracting authorities may not automatically accept a tenderer's assertion that information constitutes a trade secret. Instead, they must require the economic operator to substantiate the genuinely confidential nature of the information and must themselves conduct an active assessment of whether confidentiality is justified (§ 65). Even where information does qualify as a trade secret, authorities remain under an obligation to disclose the *essential content* of that information in a neutral form, to the extent possible, including through summaries or redacted versions (paras 66–67).

⁶⁴ T. Margoni and E. De Noyette, *Bedrijfsgeheimen in de Dataverordening: Een Administratieve Wendings*, cit.

⁶⁵ See also N. Lee, *Governing valuable confidential data in the EU: Transparency as fairness*, in *Improving Intellectual Property*, Cheltenham, 2023, arguing that EU data governance debates extend beyond data protection and reflect broader regulatory and administrative law logics.

⁶⁶ CJEU, 28 April 2022, C-54/21, *Antea Polska and Others v Państwowe Gospodarstwo Wodne Wody Polskie (Antea)*.

Thomas Margoni, Leona King

*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

This follows from earlier case law, including *Varec* (C-450/06)⁶⁷ and *Klaipėdos* (C-927/19)⁶⁸, which similarly requires review bodies and courts to balance the protection of trade secrets against the right to an effective remedy, confirming that confidentiality cannot be permitted to undermine transparency, fair competition, or effective judicial protection. Taken together, this line of jurisprudence reveals an increasingly proceduralised conception of transparency, in which claims to secrecy are subject to verification, balancing, and controlled disclosure under the supervision of public authorities and courts.

4.2. *Applying the Standard to the AI Act*

The right of access under Article 15 GDPR is instrumental to the effective exercise of other data subject rights, including rectification, erasure, and restriction of processing. Transparency in this context functions not as an end in itself but as a facilitative right, a procedural condition for the realisation of substantive rights⁶⁹. In the *CK* case, the CJEU confirmed that where access may interfere with trade secrets, disclosure must nonetheless occur, if not directly to the data subject, then through a competent authority capable of balancing the rights and interests involved on a case-by-case basis. A similar logic underpins the AI Act's disclosure regime in Article 53 and Recital 107, where transparency is likewise designed to enable the exercise of other Union law rights, such as – but not limited to – those of copyright holders. This elevates transparency to a quasi-constitutional level, one however that is functionally justified (and limited) by the enablement of other fundamental rights. A framework of authority-mediated disclosure provides the mechanism for achieving this balance, ensuring that confidentiality is protected only to the extent strictly necessary and without undermining the rights that transparency is intended to guarantee.

Accordingly, a functionally constitutionalised reading of the transparency requirements of Articles 53 and 78 AI Act in light of the TSD and CJEU case law, suggest a procedural model composed of the following three operational principles:

⁶⁷ CJEU, 8 November 2007, C-450/06, *Varec SA v Belgian State (Varec)*.

⁶⁸ CJEU, 7 September 2021, C-927/19, *Klaipėdos regiono atliekų tvarkymo centras v UAB (Klaipėdos)*.

⁶⁹ E. De Noyette, *Rights of Access and Trade Secrets*, cit., p. 144.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

- a. No automatic exemption – Providers cannot rely on Article 78 to refuse disclosure of the relevant information contained in the “sufficiently detailed summary”. They must demonstrate that disclosure would genuinely reveal information satisfying the TSD definition, and that secrecy is necessary. Drawing on Malgieri’s notion of decontextualisation developed in relation to personal data, trade secret holders could mitigate disclosure risks by providing context-stripped or abstracted information within the “sufficiently detailed summary”⁷⁰. Such decontextualisation would preserve the informational value necessary for transparency and rights enforcement, allowing verification of data provenance, representativeness, or compliance, while withholding the specific relational or operational details that confer commercial value and constitute the essence of the trade secret.
- b. Independent verification – The AI Office or competent national authorities should evaluate each confidentiality claim through a reasoned decision subject to administrative and judicial review.
- c. Minimum intelligibility – Even where trade secret protection is upheld, providers must still furnish high level, intelligible information (dataset categories, data sources, timeframes, linguistic coverage) sufficient to allow oversight and the exercise of legitimate interests under Recital 107.

4.3. *Constitutionalising Transparency*

The *CK* decision and AI Act may therefore reflect a deeper shift in the proper legal categorisation of transparency, namely towards its constitutionalisation. What began as a procedural condition for exercising other rights is increasingly taking on an autonomous constitutional function. This development should not surprise: as public governance, commercial transactions, and even daily life become mediated by data-driven systems, transparency itself becomes a prerequisite for meaningful participation and accountability in the digital order. Transparency has arguably always existed in modern legal frameworks but often only implicitly or embedded in other fundamental rights, such as expression,

⁷⁰ G. Malgieri, *Trade Secrets v Personal Data: A Possible Solution for Balancing Rights*, cit.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

information, and good administration⁷¹. Yet in an environment that is not merely digital but almost fully “datafied”, where the ability to access, understand, and verify information determines the effective enjoyment of many other rights, transparency can no longer remain a secondary or implicit norm. It must be disembedded from its parent rights and recognised as a structural principle of constitutional significance: a form of public accountability appropriate to algorithmic governance.

As a matter of fact, the legislative and judicial trends so far surveyed seem to suggest that transparency may even hold a higher hierarchical position when it conflicts with secrecy, at least when this is due to its function in enabling participation, accountability, and the exercise of other fundamental rights. This fundamental right dimension of transparency, either as an autonomous right or as a necessary enabler of other rights, should, or perhaps must, translate in a rebuttable presumption of transparency to comply with the applicable EU fundamental rights framework. Under this point of view, claims to trade secrecy should be subject to authority-mediated review and sustained only where secrecy is proved as necessary.

Read in light of Articles 53(1)(d) and Recital 107 AI Act, the same presumption should apply to AI training data summaries. Providers must publish information enabling the exercise of legitimate interests under Union law; secrecy may prevail only when independently verified as proportionate. This approach maintains doctrinal coherence with Article 3(2) TSD, which deems lawful any disclosure «required or allowed by Union law», and aligns with broader EU data law trends, particularly the Data Act and the European Health Data Space Regulation⁷² (EHDSR), where access to confidential data is permitted under controlled, authority-supervised conditions⁷³.

To implement this presumption, the AI Office could adopt procedures akin to those of competition and access-to-documents authorities⁷⁴. Accordingly, an argument in favour of the institutionalisation of authority-mediated trade secrecy can be advanced, one in which

⁷¹ P. Keller and T. F. Aplin, *Reconciling Trade Secrets and AI*, cit.; M. E. Kaminski, *Understanding Transparency*, cit.; A. Buijze, *Transparency: The Swiss Knife of EU Law*, cit.

⁷² Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847 (Text with EEA relevance) [2025] OJ L 2025/327 .

⁷³ E. De Noyette, *Rights of Access and Trade Secrets*, cit.

⁷⁴ See Arts. 16–18 of Regulation (EC) No 773/2004; Commission Notice on the Rules for Access to the Commission File in Competition Cases (2005/C 325/07).

Thomas Margoni, Leona King

*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

independent oversight bodies verify the legitimacy and proportionality of secrecy claims before they restrict disclosure. This proceduralisation would deter overclaiming, foster consistent interpretation, and could support building a *corpus* of soft law clarifying what constitutes a trade secret with respect to training data. Such institutionalisation is indispensable given the danger of an increasingly fragmented governance landscape created by the AI Act, Data Act, EHDSR and related instruments. Coordination among the AI Office, national competent authorities, and related regulators will be essential to ensure coherence and to prevent administrative capacity from becoming the bottleneck of transparency. Without such institutional investment, the Union’s promise of balanced openness could risk collapsing into regulatory inertia.

5. *The Commission’s Template: Promise, Gaps, and Risks of Under Disclosure: From Principle to Practice*

The Commission’s 24 July 2025 Explanatory Notice and Template for the Article 53(1)(d) “sufficiently detailed summary”⁷⁵ (hereinafter, referred to as the “Commission’s Template”) translates an open-textured duty into a standardised disclosure format. This represents a significant institutional advancement, as it aims to enhance comparability and mitigate compliance uncertainty. However, the Template also embodies structural ambivalence. While its tiered design reflects the proportionality logic of the AI Act by calibrating disclosure according to source and sensitivity, it may delegate too much discretion to providers, thereby risking the consolidation of defensive opacity. The central question is whether the mandated disclosures are (i) unlikely to reveal genuine trade secrets under Article 2(1) TSD, and (ii) sufficient to enable «parties with legitimate interests» to exercise their Union law rights (Recital 107). On both counts, the Template may underdeliver.

⁷⁵ European Commission, [Explanatory Notice and Template for the Public Summary of Training Content for General-Purpose AI Models](#), Brussels, 2025, available at [digital-strategy.ec.europa.eu](#).

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

5.1. *The Template as Institutionalised Transparency?*

The Commission's Template is an important step towards standardising disclosure obligations under the AI Act at an institutional level. Its tiered structure reflects the Regulation's proportionality logic by distinguishing between public, commercial, and proprietary data sources. Yet, while the Template advances comparability and provides greater regulatory certainty, it also delegates substantial discretion to providers, thereby risking the consolidation of defensive opacity, a model in which confidentiality claims expand unchecked, undermining the Act's transparency objective.

From the perspective of balancing trade secrecy and transparency, three dimensions are particularly salient. First, disclosures concerning public and licensed datasets are central to the exercise of legitimate interests, including those of copyright holders under Recital 107. The Template requires identification of the main public or licensed corpora used for training. This obligation should not expose proprietary know-how; rather, it provides provenance information enabling rightsholders to verify lawful text and data mining (TDM) and copyright compliance.

Second, trade secret concerns arise primarily in relation to bespoke or third-party proprietary datasets – namely corpora assembled through targeted collection, annotation, or preprocessing that embed technical or organisational know-how⁷⁶.

Third, transparency regarding data processing measures, such as respect for TDM opt-outs and the removal of illegal content, is indispensable for accountability and seems to entail negligible risk to trade secrets. In doing so, the Template focuses on compliance behaviour rather than proprietary techniques, enabling authorities and affected parties to assess whether providers have used data lawfully and mitigated unlawful or biased outputs.

As discussed earlier, what may more plausibly constitute a trade secret in this context are not the datasets themselves, but the selection parameters – the proprietary instructions, algorithms, or criteria used to identify and extract relevant data from public sources. These processes could represent confidential know-how with commercial value by virtue of their secrecy. Crucially, the Template does not require disclosure of such parameters, but

⁷⁶ A. Nordberg, *Trade Secrets, Big Data and Artificial Intelligence Innovation: A Legal Oxymoron?*, in J. Schovsbo, T. Minssen and T. Riis (eds), *The Harmonization and Protection of Trade Secrets in the EU: An Appraisal of the EU Directive*, Cheltenham, 2020.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

only of the resulting datasets, thereby already excluding the elements most likely to fall within the TSD threshold.

5.2. *Proportionality and Legitimate Interest Functionality*

While only time (and arguably litigation) will tell, it seems that the current template, while certainly representing a major advancement, probably a unique example of institutionalised transparency tool on a global level, still favours precautionary secrecy over transparency. Measured against Recital 107's functional test, right holders have more information than before to exercise their rights, however, the information contained in the Template is probably not by itself sufficient to enable the exercise and enforcement of all rights or legal entitlements. Building on the logic developed in Section 4, a workable settlement between transparency and confidentiality can be achieved through a targeted transparency framework complemented by authority-mediated trade secrecy. This approach translates the principle of confidentiality into a structured, multilayered disclosure model that aligns with Articles 53 and 78 AI Act. The framework should contain a:

- a. Public layer (mandatory): The public layer would contain dataset identifiers for major public and scraped sources, categorical identifiers for licensed corpora, annotated domain lists, language and timeframe metadata, opt-out compliance metrics, and high-level information on user and synthetic data fields. These disclosures pose minimal trade secret risk but are essential to enable rights-holders, data subjects, and regulators to exercise their legitimate interests under Recital 107.
- b. Confidential layer (to authorities under Article 78): At the confidential level, providers would submit declarations for each claimed secret, along with sensitive annexes, such as the precise composition of proprietary datasets, lodged for verification rather than publication. This could include controlled access mechanisms such as a secure processing environment (SPE), borrowing the concept from other EU data law, such as the Data Governance Act (DGA) and EHDSR, allowing enhanced inspection by authorities or vetted parties. A comparable logic already exists in trade secret litigation, where *confidentiality clubs* enable limited disclosure of sensitive material to

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

identified parties under strict non-disclosure conditions⁷⁷. Incorporating a similar procedural safeguard in the AI Act context would preserve confidentiality while ensuring that those with legitimate oversight functions retain effective access for verification and enforcement.

- c. Oversight and consistency: The creation of a public registry of anonymised trade secrecy determinations to deter over-claiming and promote convergence in national practice, an approach already proposed in the context of “data secrets”⁷⁸. A comparable registry could serve a similar function under the AI Act, fostering predictability, procedural fairness, and cross-border consistency in the treatment of confidentiality claims. This proportionate implementation model operationalises the presumption of rebuttable transparency discussed earlier. It ensures that trade secrets are protected where necessary yet remain subject to verification through independent authority review. In doing so, it gives concrete effect to the argued Union’s broader project of functional constitutionalisation of transparency, embedding oversight and accountability within the digital regulatory order.

6. *Conclusion: Towards a Principle of Transparency and Mediated-Authority Trade Secrecy*

This article has argued that the AI Act establishes a model of targeted transparency in which trade secret protection operates as a conditional and reviewable constraint rather than as an absolute bar to disclosure. Focusing on Article 53 and its tiered disclosure architecture for general-purpose AI models, and read together with Article 78 and Recital 107, the analysis has shown how the Regulation reconciles transparency obligations with trade secret protection through proportional differentiation of audiences, purposes, and levels of detail. Drawing on the TSD and case law, the article has developed the concept of authority-mediated trade secrecy, under which claims to confidentiality must be substantiated, assessed, and, where necessary, limited by independent authorities to ensure that transparency

⁷⁷ See, for example, Art. 9 of Directive (EU) 2016/943 (Trade Secrets Directive); E. De Noyette, L. Stähler and T. Margoni, *Data Secrets*, cit.; E. De Noyette, *Rights of Access and Trade Secrets*, cit., p. 139–140.

⁷⁸ E. De Noyette, L. Stähler and T. Margoni, *Data Secrets*, cit.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

remains effective for the exercise and enforcement of rights under Union law.

By “functional constitutionalisation of transparency”, this article does not suggest the emergence of a formally recognised autonomous right to transparency under the Charter. Rather, it describes a process through which transparency begins to operate as a constitutional norm in practice. This occurs where transparency structures relations between private and public actors, conditions the exercise and limitation of fundamental rights, and triggers heightened procedural safeguards. Transparency thus acquires constitutional relevance through function, not formal designation. In this sense, transparency under the AI Act mirrors developments in other areas of EU law, where procedural guarantees increasingly serve as the backbone of rights effectiveness in data-intensive and asymmetrical governance environments.

Within this framework, the AI Act mediates between the Charter’s information and access rights (e.g., Articles 7, 8, 11 and 42 CFR) and the freedoms that may underpin trade secrecy – most notably the freedom to conduct a business (Article 16 CFR) and the right to property (Article 17 CFR). The operative equilibrium lies in Articles 53 and 78, read together with Recital 107 AI Act, which translate transparency into a legal rule: disclose what is necessary to make accountability real, protect only what is necessary to preserve legitimate secrecy.

A broader question emerging from this analysis is whether transparency, long treated as an instrument for the exercise of rights, should now be recognised as a constitutional principle in its own right. The trajectory of privacy, from a “right to be let alone”⁷⁹ to a fully articulated human right to private life, illustrates how procedural values can evolve into substantive guarantees in line with technological evolution. In the digital age, a context specific “right to know” may be necessary to counterbalance the concentration of informational power within private infrastructures of algorithmic governance⁸⁰. Transparency may therefore be undergoing a

⁷⁹ S. D. Warren and L. D. Brandeis, *The Right to Privacy*, in *Harvard Law Review*, vol. 4, 1890, p. 193.

⁸⁰ The concept of a “right to know” has been more extensively developed in US scholarship: M. Schudson, *The Rise of the Right to Know: Politics and the Culture of Transparency, 1945–1975*, Cambridge (MA), 2015; L. Watson, *The Right to Know: Epistemic Rights and Why We Need Them*, London, 2021; A. Florini (ed.), *The Right to Know: Transparency for an Open World*, New York, 2007. For a recent UK v. EU comparison, see P. Keller and T. F. Aplin, *Reconciling Trade Secrets and AI*, cit. In the EU context, see U.-M. Mylly, *Transparent AI?*, cit.,

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

comparable transformation, one that warrants further conceptual and doctrinal exploration.

It may be argued that within a framework of digital constitutionalism⁸¹, transparency would operate both as a necessary mechanism of oversight as well as a mode of participation⁸². It would enable independent verification, informed public scrutiny, and accountability. Yet transparency must be substantive rather than merely symbolic. The economic and technological concentration that characterise the current AI landscape may very well confine transparency to a mere formal requirement – another box to tick. The authority-mediated architecture advanced in this article offers an argumentative, and partly interpretative, pathway for rendering transparency effective in practice rather than nominal in form.

For rights to have effective meaning, they must be capable of being exercised. Transparency often serves as a precondition for that exercise. This reasoning, which appears in both CJEU⁸³ and ECtHR⁸⁴ jurisprudence, positions transparency as a functional requirement for the enjoyment and enforcement of rights, rather than a right in itself. However, as digital

p. 1018, noting that the access-to-information dimension of freedom of expression has been recognised in ECtHR case law, e.g. ECtHR, 25 June 2013, app. 48135/06, *Youth Initiative for Human Rights v Serbia*, and ECtHR, 28 November 2013, app. 39534/07, *Österreichische Vereinigung zur Erhaltung, Stärkung und Schaffung eines wirtschaftlich gesunden land- und forstwirtschaftlichen Grundbesitzes v Austria*.

⁸¹ G. De Gregorio, *The Rise of Digital Constitutionalism in the European Union*, in *International Journal of Constitutional Law*, vol. 19, 2021, p. 41; O. Pollicino, *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?*, Oxford, 2021; E. Celeste, *Digital Constitutionalism, EU Digital Sovereignty Ambitions and the Role of the European Declaration on Digital Rights*, REBUILD Centre Working Paper No. 16, 2024.

⁸² M. E. Kaminski, *Understanding Transparency*, cit., p. 121 ff.

⁸³ CJEU, C-203/22, *CK v Dun & Bradstreet Austria GmbH*, cit.; CJEU, 4 September 2025, C-413/23 P, *Single Resolution Board*. These judgments clarified that pseudonymised data remains personal data from the controller's perspective, thereby triggering transparency obligations at the point of collection. This reinforces the principle that transparency is necessary for fundamental rights to be actionable.

⁸⁴ ECtHR, 8 November 2016, app. 18030/11, *Magyar Helsinki Bizottság v Hungary*. The European Court of Human Rights recognised a right of access to public information under Art. 10 ECHR, linking transparency directly to the exercise of freedom of expression. The Court identified criteria for assessing restrictions on access to information, framing transparency as essential for democratic participation; see also D. Voorhoof, [Freedom of Expression in the Digital Environment: How the European Court of Human Rights Has Contributed to the Protection of the Right to Freedom of Expression and Information on the Internet](#), in E. Psychogiopoulou and S. de la Sierra (eds), *Digital Media Governance and Supranational Courts: Selected Issues and Insights from the European Judiciary*, Cheltenham, 2022, p. 112 ff.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

technologies increasingly mediate access to information, goods, and public services, this instrumental conception may no longer be sufficient.

Two interpretive trajectories can be identified. A maximalist view would recognise a distinct right to transparency as part of a new generation of digital constitutional rights. A minimalist view would treat transparency as a derivative functional principle that gives effect to existing rights. The AI Act occupies a position between these poles. It translates freedom of information and accountability ideals into operational duties for private actors whose systems have public impact effects. In doing so, it contributes to what may be described as a *functional constitutionalisation of transparency*.

The Commission's Template for the "sufficiently detailed summary" represents an important step in operationalising this architecture, but its effectiveness will ultimately depend on whether confidentiality claims are meaningfully scrutinised rather than merely formalized. Finally, the future development of this framework will depend in no small part on judicial interpretation. Just as *CK v Dun & Bradstreet* construes GDPR transparency through a proportionate, case-specific assessment of competing interests, rather than allowing blanket reliance on trade secrecy, so too the CJEU can ensure that Article 78 AI Act is applied as a qualified constraint, not an automatic escape from disclosure. When the inevitable conflicts between disclosure obligations and trade secret claims reach CJEU, the Court should affirm that information disclosure required by Union law cannot constitute misappropriation within the meaning of the TSD. Such a holding would preserve coherence across the Union's data governance framework and cement authority-mediation as the procedural standard for reconciling transparency and confidentiality in data-driven societies.

ABSTRACT: This article examines how the EU Artificial Intelligence Act (AI Act) seeks to reconcile transparency obligations with the protection of trade secrets through a multi-layered, regulator-facing disclosure framework centred on Article 53. It argues that the AI Act does not merely accommodate competing interests but participates in an emerging model of *authority-mediated trade secrecy* that is reshaping EU data governance. Across recent legislation – including the Data Act and the European Health Data Space Regulation – trade secret protection is increasingly removed from purely private assertion and litigation, and instead assessed, operationalised, and constrained through administrative procedures and institutional oversight.

Thomas Margoni, Leona King
*Authority-Mediated Trade Secrecy in the AI Act:
An Example of Functional Constitutionalisation of Transparency?*

Drawing on the EU Trade Secrets Directive, the AI Act, and recent case law of the CJEU (e.g. *CK v Dun & Bradstreet* C-203/22), the article conceptualises authority-mediated trade secrecy as a governance model in which independent authorities are tasked with assessing confidentiality claims. The analysis shows that the AI Act establishes a form of targeted transparency that balances, rather than hierarchises, competing Charter interests, including the rights to information and good administration, and the freedoms of property and to conduct a business through differentiated audiences, purposes, and levels of detail.

The article further contends that, despite this design, the current configuration of Article 53(1)(d) and the Commission's 2025 disclosure template risks entrenching defensive opacity by over-delegating discretion to AI providers. The article concludes that, taken together, this trend can be read as establishing *a principle of functional constitutionalisation of transparency*, whereby transparency obligations, while not framed as an autonomous Charter right, acquire constitutional force because they function as indispensable preconditions for the exercise of other fundamental rights. Interpreted in this way, the AI Act consolidates an ongoing evolution in EU data governance, signalling a shift towards a rebuttable presumption of transparency grounded in procedural safeguards, authority oversight, and institutional accountability.

KEYWORDS: AI Act – trade secrets – transparency – fundamental Rights – AI Office template.

Thomas Margoni – Professor, Centre for IT & IP Law (CiTiP), Faculty of Law, KU Leuven, Leuven, Belgium (thomas.margoni@kuleuven.be)

Leona King – FWO fellow, Doctoral Researcher, Centre for IT & IP Law (CiTiP), KU Leuven, Leuven, Belgium (leona.king@kuleuven.be)

The New Face of Privacy: AI, Power, and the Disappearing Private Sphere

Federica Paolucci*

TABLE OF CONTENTS: 1. Introduction. – 2. A Starting Point: The Face of Privacy as Pictured by the *Glukhin* Case on the Use of Biometric Identification Systems. – 3. The Stress Test: The AI Act Regulation of Biometric Identification Systems. – 4. The Elephants in the Room. – 4.1 Authorisation. – 4.2 The Normative Design. – 4.3 The Italian Case. – 5. Conclusion.

1. Introduction

Zwischenraum is the term used by Warburg to describe the spatial-temporal separation between individuals, images, and the environments they inhabit¹. By *distance*, Warburg referred to the act of stepping back when observing a painting or sculpture in a museum: distance allows the object to be seen without immediately absorbing or overwhelming the viewer. It creates a margin in which perception is not yet action, and visibility is not yet use; in a nutshell, the historic formulation of the «right to be let alone»².

In contemporary environments increasingly mediated by digital infrastructures, this distance is progressively eroding. As this paper will argue, thus, the erosion of the intermediate space has a constitutional dimension³: in its true meaning, privacy cannot be reduced to a mere spatial opposition between what is public and what is private⁴. Even in public space, where visibility is inevitable, the law has long recognised a qualitative difference between *being seen* and *being known*⁵. The traditional sites of the

* This article was subjected to double-blind peer review.

¹ C. D. Johnson, *Warburg's Zwischenraum: Between Hieroglyph and Diagram*, in D. Freedberg and C. Wedepohl (eds.), *Alby Warburg 150: Work, Legacy, Promise*, Berlin, 2024, p. 75 ff.

² S. D. Warren and L. Brandeis, *The Right to Privacy*, in *Harvard Law Review*, vol. 4-5, 1890, p. 193 ff.

³ M. Bassini, *Il diritto costituzionale alla privacy nel prisma dell'evoluzione tecnologica*, in *Rivista di diritto costituzionale*, vol. 1, 2023, p. 83 ff.

⁴ On the evolution of the conceptualisation of privacy operated by the courts: O. Pollicino, *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?*, London, 2021.

⁵ B. van der Sloot, *The right to be let alone by oneself: narrative and identity in a data-driven environment*, in *Law, Innovation and Technology*, vol. 13, 1, 2021, p. 223 ff.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

private sphere – as in home, correspondence, intimate communication, but also the ordinary anonymity of urban life – are protected not because hidden, but because they consist in «[the] zone of interaction between a person and others which, even in a public context, may fall within the scope of “private life”»⁶. Therefore, what is ultimately at stake is not a marginal interest in secrecy, but the preservation of a sphere in which individuals can appear, speak, and interact, as extending protection not only to when the individual is alone in his private home, but also to those social formations, to use the language of Article 2 of the Italian Constitution, within which each person’s life unfolds. The protection of privacy in these social formations is thus not ancillary to democratic life; it is the precondition for pluralism, dissent and personal development to disclose without being continuously filtered through classificatory and predictive operations.

It is therefore from a constitutional vantage point, one attentive to the equilibrium of powers⁷, that this analysis proceeds. Nonetheless, a fundamental rights conception of privacy cannot be examined in isolation from the profound transformations introduced by recent technological innovation, all the more so in a landscape reshaped by the increasing reliance on artificial intelligence systems (hereafter, AI)⁸. In particular, in a context highly modified by the presence of technological tools, the relationship is no longer one-directional, since individuals may also unwittingly participate in their surveillance by providing data through daily interactions with AI, platforms, and technology in general, which then use that data to inform and influence future behaviours and patterns⁹. This

⁶ ECtHR (IV Sec.), 11 April 2006, app. 56550/00, *Mólka v Poland*.

⁷ «The end of law is not to abolish or restrain, but to preserve and enlarge freedom. For in all the states of created beings capable of laws, where there is no law, there is no freedom. For liberty is to be free from restraint and violence from others; which cannot be where there is no law: and is not, as we are told, a liberty for every man to do what he lists. But a liberty to dispose and order freely as he lists his person, actions, possessions, and his whole property within the allowance of those laws under which he is, and therein not to be subject to the arbitrary will of another, but freely follow his own», J. Locke, *Due trattati sul governo e altri scritti politici*, in the ed. by L. Pareyson, Turin, 1982 (aut. transl.).

⁸ In this sense, this paper relies on the definition of Artificial Intelligence systems as in Article 3(1), 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, hereafter AI Act.

⁹ This evolution has been already analysed by Lyon with respect to the intersections between society and technology. *Inter alia*, D. Lyon, *The Culture of Surveillance: Watching as a Way of Life*, London, 2018; J. Locke, *Due trattati sul governo e altri scritti politici*, cit.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

dynamic is particularly evident in the uses of technology that may be able to impact the ordinary public behaviour. For instance, the Venezuelan state-sponsored application, under Maduro's presidency, *VenApp* was an illustrative example: originally conceived as a utility-reporting tool, it now enables users to report "suspicious individuals" directly to authorities, effectively turning everyday visibility into a form of reciprocated monitoring¹⁰. This mutual entanglement of humans and technology complicates the traditional "watcher" and "watched" dynamic, making it less about observation alone and more about the co-creation of control and influence in the digital ecosystem¹¹.

Starting from these premises, the perspective adopted here seeks to investigate how the boundary between the public and the private is reconfigured when biometric identification technologies are deployed by law enforcement authorities. Such systems, like facial recognition technologies, «extend the ability of government authorities to monitor, identify, and track individuals in public areas, but their indiscriminate, biased, and opaque nature has the potential to deter protesters from exercising their right to assembly»¹². The guiding concern of this inquiry is therefore how, within this evolving landscape, constitutional law can still secure a robust conception of privacy: not as a fragile residuum of secrecy, but as a structuring principle that limits the ways in which biometric technologies may reconfigure what it means to appear in public. This specific context raises a set of questions not only concerning the actual use of such tools, but also regarding the normative environment within which

¹⁰ S. Pozzebon, *Venezuela's Maduro, fearing US attack, promotes app to report suspect behavior*, in CNN, November 5, 2025, available at [cnn.com](https://www.cnn.com). As documented by human-rights observers, such tools transform the simple act of appearing in public into material for interpretation and intervention by public power, eroding the intermediate space on which privacy and freedom of expression rely.

¹¹ This aspect has also been commented on as "an erosion of privacy", or, in the words of the Korean philosopher B. Han, *Psychopolitics: Neoliberalism and new technologies of power*, Milan, 2017 «the neoliberal regime transforms subjects into projects. As projects, individuals constantly police, manage, and enhance themselves. They exploit themselves on behalf of the system». Specifically, Han highlights that the digital era introduces a new form of dominance, where control is exerted not through physical force but through data and information. For a critique and an analysis of the computer-human relationship respectively: F. Pasquale, *Watching (and improving) the watchers*, in *The black box society. The secret algorithms that control money and information*, New York, 2015, p. 140 ff.; B. Friedman *et al.*, *The Watcher and the Watched: Social Judgments About Privacy in a Public Place*, in *Human-Computer Interaction*, vol. 21, 2, 2006, p. 235 ff.

¹² M. Warthon, *Artificial Intelligence and Freedom of Assembly*, in A. Quintavalla and J. Temperman (eds), *Artificial Intelligence and Human Rights*, Oxford, 2023.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

they are authorised and operated. At this juncture, marked by the slow and at times uneven¹³ transition of the AI Act from legislative text to applied law, the aim of this contribution is not to assess the desirability of using these technologies¹⁴. Rather, it is to take stock of the composite regulatory framework and to observe how delicate this moment is, insofar as it requires Member States to articulate their national legal architectures in ways that remain faithful to the constitutional commitments and, generally, to the rule of law¹⁵. In this context, the question should not only focus on how personal data are processed once collected, but also whether the legal order still preserves any meaningful space in which individuals may appear without being immediately translated into identifiable data suitable for action by public or private power.

The normative starting point of this research is that such a transformation cannot be assessed solely in terms of formal compliance. As several authors have reminded us by looking back to the twentieth century, grave injustices have often been committed *through* the law rather than in its absence, under frameworks that respected the rule of law in a narrow, procedural sense while departing from any substantive idea of justice¹⁶. This paper adopts that cautionary perspective to interrogate the EU regulatory model on AI, and its national adaptations, as possible sites where formally correct procedures may coexist with structural erosion of fundamental rights.

Against this background, the analysis proceeds in two steps. First, it treats the erosion of the intermediate space as a diagnostic lens, showing how biometric identification systems reconfigure the conditions of privacy as a constitutional guarantee. Second, it examines a set of concrete legal challenges – from the design of the *lex specialis* on facial recognition to the choice of authorising authorities and the role of fundamental rights bodies – to explore whether, and how, EU and national law can function as

¹³ At the time in which this paper is written, the European Commission is evaluating a belated entry into application of part of the AI Act, especially of high risk obligations, as in Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI) {SWD(2025) 836 final}, 19.11.2025.

¹⁴ It is recalled the scholarship, as in N. Menéndez González and G. Mobilio (eds), *Next Democratic Frontiers for Facial Recognition Technology (FRT): The Legal, Ethical and Democratic Implications of FRT*, Cham, 2025.

¹⁵ *Ex multis*, ECtHR (GC), 25 June 1997, app. 20605/92, *Halford v United Kingdom*.

¹⁶ M. Cartabia and N. Lupo, *The Constitutional Court*, in *The Constitution of Italy*, London, 2022.

Federica Paolucci

The New Face of Privacy:

AI, Power, and the Disappearing Private Sphere

methods of constitutional resistance to the forces of simplification and trivialisation that seems to be pervading the regulation and integration of AI systems in the humans' experience. Throughout this analysis, privacy will therefore be treated as a litmus test: a standard against which to measure whether legal frameworks surrounding biometric identification merely accommodate technological expansion, or whether they actively reconstruct the conditions under which a private life, in the sense given by Article 7 of the Charter of Fundamental Rights of the European Union (CFREU)¹⁷ and Article 8 of the European Convention of Human Rights (ECHR)¹⁸, remains possible.

2. *A Starting Point: The Face of Privacy as Pictured by the Glukhin Case on the Use of Biometric Identification Systems*

Any analysis of biometric identification technologies must begin from a renewed understanding of what is meant by “privacy”. Traditionally, this right has been treated and narrated as a “negative freedom”, hence «the right [of] the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures»¹⁹. Almost the same wording is indeed used in the Universal Declaration of Human Rights, adopted by the United Nations General Assembly on December 10, 1948, which states in its Article 12 that «none shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks». Therefore, aside from the identification of privacy as a negative freedom, it is possible to recall also a “positive stance”, as the right of the individuals, stated in the second part of Article 12, to be protected by the law in the exercise and enjoyment of such rights. In other words, if, on the one hand, the law protects, individuals should have the right to claim and exercise such protection.

Furthermore, privacy is not a monolith. As also underscored by the scholarship, it is indeed a complex activity to define “privacy” since it is almost viewed as a particular kind of harm²⁰. However, the concept

¹⁷ Charter of Fundamental Rights of the European Union OJ C 364/1 2000.

¹⁸ Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR).

¹⁹ Fourth Amendment (Amendment IV) to the United States Constitution.

²⁰ D. Solove, *A Taxonomy of Privacy*, in *University of Pennsylvania Law Review*, vol. 154, 3, 2006, p. 477.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

becomes intelligible only when one shifts from abstraction to the concrete activities that place it at risk and recognises the broad range of situations in which privacy might be triggered. What matters, essentially, is not only the use of data, even though the supremacy of the application of Article 8 CFREU in technology related problems prevails in the reasoning of the courts, but the conditions under which intrusions occur: e.g., the institutional architecture that enables them, the technological environment that amplifies them, and the legal safeguards, or their absence, that determine whether they remain within constitutional bounds.

Seen from this perspective, privacy emerges not simply as a defensive shield but as a relational, operational right, one that requires an ongoing assessment of how practices, technologies, and regulatory frameworks interact. A constitutional approach to privacy thus demands a specific, contextualised inquiry: not merely whether data are processed lawfully, but whether the surrounding legal system ensures that any interference is authorised, proportionate, reviewable, and embedded within a structure that enables individuals to react to abuses before and after they materialise. This is precisely where the positive dimension of privacy becomes indispensable.

In this sense, even though the negative conception of privacy seems to prevail, both Article 7 CFREU and Article 8 ECHR²¹ have been interpreted by their respective courts as encompassing both negative and positive obligations²². They certainly protect individuals against arbitrary intrusions, the classical negative obligation, but they also impose positive duties on legislatures and administrative systems. This positive dimension, visible in judgments such as *Digital Rights Ireland*, *Schrems*, and *La Quadrature du Net*²³, imposes on public authorities a duty to ensure that actions of “surveillance” are governed by precise rules, subject to independent oversight, and embedded within a broader system that guarantees transparency, intelligibility, and access to remedies.

²¹ G. Martinico, *Art. 7. Rispetto della vita privata e della vita familiare*, in R. Mastroianni, O. Pollicino et al. (eds.), in *Carta dei Diritti Fondamentali dell'Unione Europea*, Milan, 2017.

²² B. van der Sloot, *Privacy as Human Flourishing: Could a Shift towards Virtue Ethics Strengthen Privacy Protection in the Age of Big Data*, in *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, vol. 5, 3, 2014, p. 230 ff.; B. van der Sloot and E. Kosta, *Big Brother Watch and Others v UK: Lessons from the Latest Strasbourg Ruling on Bulk Surveillance*, in *European Data Protection Law Review (EDPL)*, 2019, p. 252 ff.

²³ CJEU (GC), 8 April 2014, Joined C-293/12 and C-594/12, *Digital Rights Ireland and Seitlinger and Others*; CJEU (GC), 6 October 2015, C-362/14, *Maximilian Schrems v Data Protection Commissioner*; CJEU (GC), 6 October 2020, Joined C-511/18, C-512/18, and C-520/18, *La Quadrature du Net and Others v Premier ministre and Others*.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

More clear in this sense is actually the approach under Article 8 ECHR and the interpretation of the court, which is explained in the following, looking at the judgment that the European Court of Human Rights (ECtHR) delivered in the Application n. 11519/20. In this case, the Court relies on a concept of privacy that is substantiated not only in the claim “to be let alone”, but as a demand that the legal order itself erect the procedural and institutional architecture necessary to make such a claim meaningful, reading together the values of Article 8 with that of Article 6, ECHR. Privacy, thus, under this light, becomes a right that lives, or fails, within the conditions under which the practices and the activities that restrict its application are recognised, addressed, analysed, and, if there are the conditions, authorised.

In this context, *Glukhin v Russia*²⁴ is not just the first case decided by an international court²⁵, namely the ECtHR, about the use of biometric identification systems, but especially it is a case about the use of face recognition technology (hereafter, also FRT)²⁶ by law enforcement that analysed the compatibility of such systems with the human rights enshrined in the ECHR²⁷.

Precisely, at the moment of the contested facts, the applicant, Nikolay Sergeyeovich Glukhin, was standing silently in a crowded Moscow metro, holding a banner advocating for human rights. Without their knowledge, surveillance cameras capture their image, which is then circulated in

²⁴ ECtHR (III Sec.), 4 July 2023, app. 11519/20, *Glukhin v Russia*.

²⁵ As a matter of fact, previous cases regarded the use of wiretapping systems or the use of biometric data by the police, as in the famous *S. and Marper v UK*, id. Indeed, in another case, the ECtHR, 3 June 2020, app. 45245/15, *Gaughran v The United Kingdom*, § 37, the Court found that the UK’s police ability to upload the facial images of the claimant “from forces” local custody IT systems onto the Police National Database (PND) from which they could be enrolled in «the facial recognition gallery making them searchable using facial recognition software» constituted a violation of Art. 8 ECHR. Nonetheless, such a use of the data is not at all a face recognition *per se*, as it is in the case at hand. See also for this distinction F. Palmiotto and N. Menéndez González, *Facial recognition technology, democracy and human rights*, in *Computer Law & Security Review*, vol. 50, 2023.

²⁶ Facial recognition falls under the umbrella of biometric identification systems: Z. Tan and G. Guo, *Face Recognition Research and Development*, in S. Z. Li, A. K. Jain and J. Deng (eds), *Handbook of Face Recognition*, Cheltenham, 2024; I. Badri and M. Sayyouri, *Face Recognition: A Mini-Review*, in S. Motahhir and B. Bossoufi (eds), *Digital Technologies and Applications*, Cheltenham, 2023. These systems include fingerprint, voice recognition, emotion recognition, and they share a common feature: they rely on unique physiological or behavioural characteristics to identify individuals. However, their applications and risks vary significantly.

²⁷ See § 59 of the *Glukhin v Russia*’s judgment on the admissibility.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

Telegram groups. Facial recognition technology silently scanned their face and matched it against a government database. Days later, this data was weaponised by authorities to locate and arrest the individual, accusing them of violating public assembly laws. In the ruling, it emerged²⁸ that the practice used by the police in the metro against the applicant was a violation of Articles 10 (freedom of expression) and 8 (right to respect for private life) of the ECHR,²⁹ however, the implications of this case and the practice hereby condemned can extend far beyond these specific rights, undermining a broader spectrum of human rights³⁰, and contemporary understanding of the right to privacy. For this reason, this section analyses this right as triggered by the judgment under scrutiny, with the purpose of unfolding the purpose of the right as well as the legitimacy of the interferences with it. Specifically, Mr. Glukhin's story represents just an example of a paramount assault on the right to protest, the right to freedom of assembly and broader principles of democratic participation. Moreover, highlights what profound dangers might be posed by a critical intersection of crucial elements: where the power is more inclined in the suppression of

²⁸ § 88: «The Court observes that the applicant was prosecuted for a minor offence consisting of holding a solo demonstration without prior notification – an offence classified as administrative rather than criminal under the domestic law. He was never accused of committing any reprehensible acts during his demonstration, such as the obstruction of traffic, damage to property or acts of violence. It was never claimed that his actions presented any danger to public order or transport safety. The Court has already found that the administrative-offence proceedings against the applicant breached his right to freedom of expression. It considers that the use of highly intrusive facial recognition technology to identify and arrest participants in peaceful protest actions could have a chilling effect in relation to the rights to freedom of expression and assembly».

²⁹ § 82: «the Court considers that it is essential in the context of implementing facial recognition technology to have detailed rules governing the scope and application of measures, as well as strong safeguards against the risk of abuse and arbitrariness. The need for safeguards will be all the greater where the use of live facial recognition technology is concerned».

³⁰ The case, being the first one at the international level on FRT, was commented by many scholars. This research especially relied on the analysis of G. Mobilio, *La Corte EDU condanna il ricorso alle tecnologie di riconoscimento facciale per reprimere il dissenso politico: Osservazioni a partire dal caso Glukhin c. Russia*, in *DPCE Online*, 2024, p. 695; F. Palmiotto and N. Menéndez González, *Facial Recognition Technology, Democracy and Human Rights*, cit.; M. Zalnieriute, *Glukhin v. Russia. App. No. 11519/20. Judgment*, in *American Journal of International Law*, vol. 117, 2023, p. 695; C. Nardocci, *Il riconoscimento facciale sul "banco" degli imputati. Riflessioni a partire, e oltre, Corte EDU Glukhin c. Russia*, in *BioLaw Journal - Rivista di BioDiritto*, 2024, p. 279.

civil liberties, advanced surveillance technologies and the curtailment of democratic rights³¹.

Coming to the merits of the case, first of all, the Court identified two distinct uses of FRT in this case. The authorities used images shared on Telegram and matched them with CCTV footage to retrospectively identify Glukhin as a “protest”³² participant³³. Secondly, after the police identified Mr. Glukhin, they searched him through the use of “real-time” biometric systems to identify the location of the applicant in Moscow. The applicant was arrested on the same day at an underground station. Even though the Russian Government never explicitly admitted its use of FRT in real-time³⁴, the Court accepted, in light of the particular circumstances of the case, that FRT was used as claimed by the applicant³⁵.

The first aspect that should be underlined is how tenuous the boundaries are between different forms of facial recognition and how, in practice, these modalities are often combined rather than deployed as mutually exclusive alternatives³⁶. As a consequence, the action sequence reconstructed by the Court – *i.e.*, remote identification followed by an immediate search across a real-time use – demonstrates that the distinction between *ex post* and live uses rests on a continuum of latency rather than on neatly separable categories³⁷. What appears as a technical difference in timing may, in reality, involve only a minimal temporal gap, one that becomes functionally irrelevant once the technology enables near-

³¹ By mean of example, it can be recalled the use of biometric systems in Latin America against protesters, Al Sur, *Facial Recognition in Latin America: Trends in the Implementation of a Perverse Technology*, 2021, p. 7; Derechos Digitales, *Gender Impacts and other Inequalities. Identity systems and social protection in Venezuela and Bolivia*, 2024, p. 1 ff.

³² The use of the commas is intended to underline the arbitrariness of the decision given that the protest was both silent and just composed by Mr. Glukhin. It was not a huge assembly: this data is crucial in order to understand the framework for the proportionality analysis conducted by the ECtHR, as this paper will further developed *infra*.

³³ §§ 71-73. However, as Mobilio has well noted, the Court remains on the surface of the distinction between “a posteriori” and “real-time” FRTs, without grasping the different legal implications arising therefrom, and, «nevertheless, the too general nature of these passages exposes the Court to a risk of abstractness and equivocality», G. Mobilio, *La Corte EDU condanna il ricorso alle tecnologie di riconoscimento facciale*, cit.

³⁴ §§ 60-63 of the judgment.

³⁵ §§ 37, 69, 76, 87 of the judgment.

³⁶ EDPB-EDPS, *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence*, 2021.

³⁷ See F. Paolucci, *Enhancing Oversight and Addressing Gaps: Assessing the Impact of the AI Act on Biometric Identification Systems*, in N. Menéndez González and G. Mobilio (eds.), *Next Democratic Frontiers for Facial Recognition Technology (FRT)*, cit., p. 71 ff.

instantaneous matching across vast datasets³⁸. As the analysis will show, this apparent subtlety carries significant legal consequences.

However, the court in delimiting its *thema decidendum* does not go so far as to consider more broadly the risks associated with this technology in the broader framework of the fight against crime, but merely «whether the processing of the applicant’s personal data was justified under Article 8 § 2 of the Convention in the present case»³⁹. In this sense, the ECtHR expands the scope of the right to private life, clarifying its contours when linked to the use of artificial intelligence technologies, such as biometric identification systems⁴⁰. Drawing on its established jurisprudence regarding the collection and retention of individual data and images⁴¹, the Strasbourg Court seamlessly applies these principles to the case at hand, ultimately finding the interference suffered by the applicant to be unlawful. Both the “versions” of biometric identification systems were evidently used both to identify the applicant during the protest and subsequently to locate him for the purpose of his arrest.

However, as required by Article 8 ECHR, determining whether the interference violates the Convention necessitates an assessment of whether it meets the criteria outlined in the second paragraph of the provision:⁴² «there shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others».

Hence, although it may appear almost self-evident to recall that the conditions under which interferences with the right to privacy are permissible must be determined by law, this reminder is far from

³⁸ It is recalled the definition of post remote biometric identification systems as in Art. 3(43) AI Act, «a remote biometric identification system other than a real-time remote biometric identification system».

³⁹ § 85. For a critique of this choice, see M. Zalnieriute, *Glukhin v. Russia*. *App. No. 11519/20*, cit.

⁴⁰ C. Nardocci, *Il riconoscimento facciale sul “banco” degli imputati*, cit.

⁴¹ *Ex multis*, ECtHR (GC), 4 December 2008, apps. 30562/04 and 30566/04, *S. and Marper v the United Kingdom*; ECtHR (GC), 4 May 2000, app. 28341/95, *Rotaru v Romania*.

⁴² W.A. Schabas, *Right to respect for private and family life/Droit au respect de la vie privée et familiale*, in Id. (ed), *The European Convention on Human Rights: A Commentary*, Oxford, 2015; M. Susi, *European Convention on Human Rights - freedom of expression - right to privacy - responsibility of Internet news portal for defamatory comments posted by readers*, in *The American journal of international law*, vol. 108, no. 2, 2014, p. 295 ff.

redundant⁴³. It is precisely the law that must perform the delicate function of mediating between technological capability and constitutional restraint, and the purpose of this inquiry is to ascertain whether the legal framework itself is capable of maintaining that balance. As Bassini has observed, technological evolution has produced a genuine «transfiguration of the right to privacy»⁴⁴, altering its modes of protection without severing it from its historical core. The foundational idea of “the right to be let alone” has not disappeared; rather, it is increasingly placed under strain by forms of automation that simplify certain functions while simultaneously expanding the State’s capacity to compress individual freedom and thereby generating new demands for protection.

It is in this context that the Court places particular emphasis on the necessity of adequate regulatory frameworks governing the use of facial recognition and related artificial intelligence technologies⁴⁵. The insistence on a clear and foreseeable legal basis: a «minimum standard of guarantees essential for the protection of fundamental rights», which the Court deems indispensable⁴⁶. It is not a formalistic exercise, but a recognition that constitutional rights cannot survive in environments where surveillance capabilities evolve faster than the norms meant to regulate them. In a nutshell, the key element, for the purpose of this paper, of the judgments is the underlying obligation of legal systems to keep pace with technological developments so that the guarantees offered by Article 8 ECHR, and, by extension and for its context, of Article 7 CFREU retain practical effectiveness.

⁴³ As such, it is a prerequisite of the “rule of law”, as depicted by W. Schroeder, *The Rule of Law As a Value in the Sense of Article 2 TEU: What Does It Mean and Imply?*, in A. von Bogdandy, P. Bogdanowicz, I. Canor, C. Grabenwarter, M. Taborowski and M. Schmidt (eds), *Defending Checks and Balances in EU Member States: Taking Stock of Europe’s Actions*, Springer, Berlin-Heidelberg, 2021, p. 105 ff.

⁴⁴ M. Bassini, *Il diritto costituzionale alla privacy nel prisma dell’evoluzione tecnologica*, in *Rivista di diritto costituzionale*, cit.

⁴⁵ § 83: «the Government did not refer to any procedural safeguards accompanying the use of facial recognition technology in Russia, such as the authorisation procedures, the procedures to be followed for examining, using and storing the data obtained, supervisory control mechanisms and available remedies!».

⁴⁶ § 77: «in the context of the collection and processing of personal data, it is therefore essential to have clear, detailed rules governing the scope and application of measures, as well as minimum safeguards concerning, inter alia, duration, storage, usage, access of third parties, procedures for preserving the integrity and confidentiality of data, and procedures for their destruction, thus providing sufficient guarantees against the risk of abuse and arbitrariness».

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

Henceforth, closer engagement with this part of the judgment reveals that the Court's reasoning does far more than identify an unlawful interference: it offers a conceptual framework for understanding the constitutional limits within which biometric surveillance must operate.

Central to this framework is the requirement that any interference comply with *the quality of law*⁴⁷, be grounded in a legitimate aim, and satisfy the tests of necessity and proportionality in a democratic society, as in the adequacy and effectiveness of guarantees against arbitrariness and the risk of abuse⁴⁸. A law of "quality", hence, is respectful of the rule of law⁴⁹.

Following, the Court's critique of the legislation underscores a broader point:⁵⁰ where the law is vague or indeterminate, privacy cannot function as a right. Thus, privacy presupposes the existence of legal rules capable of structuring state power, delimiting its intrusiveness, and rendering surveillance practices foreseeable and contestable⁵¹. In technologically "dense" environments, where the transformation of appearance into actionable data may occur instantaneously, the clarity and precision of the law become conditions of the right itself. Indeterminacy is not simply a legislative defect; it is a mode of constitutional erosion.

This is equally evident in the Court's analysis of necessity and proportionality. Facial recognition, particularly in its real-time configuration, is described as especially intrusive, and for this reason, its use

⁴⁷ The doctrine of "quality of law," as developed by the ECtHR under Art. 8 ECHR, underscores that any interference with the right to privacy must be not only lawful but also meet specific qualitative standards of accessibility, foreseeability, and necessity in a democratic society. Accessibility requires that the law be publicly available and understandable, while foreseeability mandates that individuals can predict the implications of the law and adjust their behaviour accordingly. Necessity demands proportionality, ensuring that any interference aligns with legitimate aims and is justified within a democratic framework. This analysis is particularly relevant in the context of emerging technologies like FRT, where the potential for misuse and overreach underscores the need for robust legal and procedural protections. For further details, see ECtHR, 25 May 2001, app. 35252/08, *Centrum för Rättvisa v Sweden*.

⁴⁸ ECtHR (GC), 25 May 2021, *Big Brother Watch v U.K.*, apps. 58170/13, 62322/14, 24960/15.

⁴⁹ ECtHR (III Sec.), 21 June 2021, app. 33696/19, *Podchasov v Russia*.

⁵⁰ Especially, the judgment underpinned the lack of indistinct scope, the absence of precise conditions for deploying facial recognition, the lack of procedural safeguards, and the almost unlimited discretion conferred on law-enforcement authorities.

⁵¹ This interpretation is substantiated in the sense of Art. 8(2) ECHR, as also analysed in the recalled jurisprudence, *supra*.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

demands the most stringent justification of a “pressing social need”⁵². Necessity cannot be satisfied by reference to a general investigative interest; nor can proportionality be established by the mere existence of statutory authorisation. In contrast, necessity, and, moreover, in this context, operates as a substantive restraint on the ability of the State to collapse the distance between visibility and identification in ordinary public space. It functions as a safeguard of democratic life, ensuring that protest and political expression do not become arenas of automated exposure.

Seen through these combined lenses, the *Glukhin* case articulates a set of constitutional benchmarks that any contemporary regulation of biometric surveillance must confront. Its significance lies not merely in identifying an unlawful interference, but in revealing the structural conditions under which privacy can still be meaningfully exercised. The judgment makes clear that the legitimacy of intrusive technologies cannot rest on formal authorisations alone. Therefore, the indication from the ECtHR is that of grounding the use of such technologies in a law that defines with “density”⁵³, precision, and respect for the rule of law the conditions that enable their deployment. More fundamentally, it shows that, in environments where the boundary between offline reality and digital infrastructures has become porous, privacy survives only if supported by a network of normative and institutional safeguards capable of reconstructing the distance between the individual and the technological gaze: a distance that once emerged naturally but now requires deliberate legal construction.

Hence, the purpose of the preceding discussion is precisely to establish the normative vocabulary through which the AI Act must be assessed: whether its risk architecture, its authorisation mechanisms, and its reliance on national discretion internalise the guarantees demanded by Articles 7 – and 8 CFREU, even though not particularly addressed by this paper – and Article 8 ECHR, or whether gaps, ambiguities, and institutional

⁵² § 89: «the use of facial recognition technology to identify the applicant from the photographs and the video published on Telegram – and *a fortiori* the use of live facial recognition technology to locate and arrest him while he was travelling on the Moscow underground – did not correspond to a “pressing social need”».

⁵³ The concept of “density”, even though not specifically mentioned by the ECtHR is used in this paper by referring to the German concept of *Rechtsdogmatik*, hence the degree of precision, comprehensiveness, and specificity with which legal norms are structured, ordered, and interpreted. *Inter alia*, M. Beham, *German ‘Dogmatik’, An Untranslatable Concept if Ever There Was One?*, in I. Aral and J. d’Aspremont (eds), *International Law and Universality*, Oxford, 2024, p. 279 ff.

omissions risk reproducing the very deficiencies that already emerged in the judgment hereby analysed.

3. *The Stress Test: The AI Act Regulation of Biometric Identification Systems*

Before going to the core of the analysis, it is indeed necessary to frame the legal framework from which this analysis departs. As anticipated, the AI Act introduces, among other things, rules for the use of real-time remote biometric identification systems in publicly accessible spaces for law enforcement purposes, recognising the intrusive nature of such technologies and their potential to impact fundamental rights.

Real-time biometric identification systems (also, RBI) are regulated through a logic of presumptive prohibition. Under Article 5(1)(h) AI Act, the use of real-time RBI systems is generally prohibited⁵⁴, with exceptions only in narrowly defined circumstances⁵⁵. These include the targeted search for victims of abduction, trafficking, or sexual exploitation, as well as the search for missing persons. Additionally, the use of such systems may be permitted to prevent specific, substantial, and imminent threats to life or physical safety, such as terrorist attacks, or to locate or identify individuals suspected of committing serious criminal offenses, provided these crimes fall under Annex II of the AI Act and carry a custodial sentence or detention

⁵⁴ Such systems may only be deployed to confirm the identity of specific individuals for law enforcement purposes and must adhere to strict criteria. These include considerations related to the nature of the situation, such as the severity, likelihood, and scope of harm that would arise if the system were not used, and the consequences for rights and freedoms, particularly evaluating the seriousness, likelihood, and extent of the impact on the rights and freedoms of all individuals affected.

⁵⁵ Art. 5(1)(h) AI Act: «the use of ‘real-time’ remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement, unless and in so far as such use is strictly necessary for one of the following objectives:

(i) the targeted search for specific victims of abduction, trafficking in human beings or sexual exploitation of human beings, as well as the search for missing persons;

(ii) the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or a genuine and present or genuine and foreseeable threat of a terrorist attack;

(iii) the localisation or identification of a person suspected of having committed a criminal offence, for the purpose of conducting a criminal investigation or prosecution or executing a criminal penalty for offences referred to in Annex II and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least four years. Point (h) of the first subparagraph is without prejudice to Article 9 of Regulation (EU) 2016/679 for the processing of biometric data for purposes other than law enforcement».

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

of at least four years in the MS. However, even when these exceptions apply, the deployment of real-time RBI systems is subject to some safeguards: in this context, real-time biometric identification is allowed by law enforcement authorities only in cases that serve significant public interests, such as preventing imminent threats to life, investigating serious criminal offences, or responding to terrorist attacks⁵⁶.

Law enforcement, when deploying RBI, must conduct, among other things, a fundamental rights impact assessment (FRIA)⁵⁷, as mentioned in Recital 34, to evaluate the potential effects on privacy, non-discrimination, and other rights. Although Recital 34 AI Act establishes that the use of real-time remote biometric identification systems in publicly accessible spaces should be authorised only if the relevant law enforcement authority has completed a FRIA, the concrete implementation of this requirement remains ambiguous. Only in the Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act), published in February 2025, the Commission clarified that «most AI systems that fall under an exception from a prohibition listed in Article 5 AI Act will qualify as high-risk»⁵⁸. Therefore, the obligations for high-risk systems shall also apply to the exceptions for RBI in law enforcement.

Moreover, Article 5(3) provides a crucial obligation which is going to be examined later on: judicial or independent administrative authorisation is mandatory for each use, based on a reasoned request that demonstrates necessity and proportionality. Basically, the choice is left to the Member States to determine whether a judicial or administrative authority should be recognised with the power to authorise the use of such systems⁵⁹. These authorisations must also account for geographic, temporal, and personal limitations to ensure minimal interference with fundamental rights.

Beyond the rules for real-time systems, the AI Act addresses the broader use of biometrics in law enforcement, particularly in the form of post-remote biometric identification systems. These systems differ from

⁵⁶ Art. 5 and Annex II AI Act.

⁵⁷ Art. 27 AI Act.

⁵⁸ Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act), cit., § 37.

⁵⁹ As clarified by the European Commission in Art. 5, «national laws are required for operationalising the use of ‘real-time’ RBI systems in publicly accessible spaces for the purposes of law enforcement». At the same time, Art. 5(5) AI Act, provides that Member States remain free to decide whether to adopt such national laws. «If a national law authorising the use of real-time RBI is adopted, the AI Act specifies the substantive elements which the national laws must contain to comply with the requirements laid down in the AI Act».

real-time systems in that they analyse biometric data after the fact, rather than in real time. In contrast, *ex post* biometric identification systems – which are used *post factum* for identification purposes – are classified as high-risk AI systems under Article 6(2) AI Act⁶⁰. This classification also extends to other sensitive biometric technologies, such as biometric categorisation systems and some uses of emotional recognition systems, which are considered high-risk due to their potential to infringe on fundamental rights. Like real-time systems, deployers of high-risk post-remote biometric systems should require prior authorisation from judicial or independent administrative authorities, with the necessity and proportionality of their deployment rigorously scrutinised⁶¹.

However, concerns remain about the potential for inconsistency: the Regulation’s distinction between “remote” and “real-time” biometric identification systems (Article 3(36)(37)) has been critiqued for creating confusion rather than clarity⁶². As well recalled by the EDPB, there is just one type of technology with different functions, not the opposite⁶³. This aspect is crucial when attempting to draw a line between uses that fall into the recognition, identification, categorisation and verification, since they are allowed for different purposes while being the same type of technology, even though with different requirements. In particular, the Regulation does not limit *ex post* uses to Annex II crimes, nor does the Act impose an explicit requirement that they be tied to serious criminal offences. As outlined in the very title of Annex II, the list of serious crimes suffices only for real time uses, therefore, falling within Article 5 “Article 5(1), First Subparagraph, Point (h)(iii)”. This lack of specificity, as mentioned *supra*, for other systems, such as post-remote biometric identification systems, introduces a regulatory gap, leaving the application of these systems less clearly constrained, despite their significant potential for privacy infringements. The standard for real-time systems is tied to the gravity of

⁶⁰ The distinction between the real-time and *ex post* application of biometric recognition is futile from a fundamental rights perspective. The difference is purely technical and consists of two different moments of the identification process, but it entails the same consequences for the surveillance of citizens. Particularly, the biometric system is mostly the same, trained with a biometric data set, and it is used in two modalities: e.g., live while the crime is happening or after, then, *ex post*. See EDPB-EDPS, *Joint Opinion 5/2021*, cit.

⁶¹ Art. 26 AI Act.

⁶² M. Veale – F. Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, in *Computer Law Review International*, vol. 4, 2021, p. 97 ff.

⁶³ EDPB, *Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement*, Version 2.0, Adopted on 26 April 2023.

the offence; the standard for *ex post* use of such systems is tied to the general architecture of high-risk AI.

The result is a regulatory gap: although *ex post* uses may involve large-scale identification capabilities, the AI Act does not constrain them through a seriousness-of-crime threshold, opening to potential loopholes, also given the fact, as mentioned before, the two instruments can be used in synchronous, as happened in Mr. Glukhin's case. Since the technology is the same, and the only difference is a mere latency, the potential for stretching the limits of the Regulation is visible and concerning.

Furthermore, the AI Act also contains specific provisions on two other types of AI systems that might or are trained with biometric data: i) emotion recognition; ii) biometric categorisation.

Emotion recognition systems are deployed «for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data»⁶⁴, and they aim to infer emotional states from facial expressions, voice patterns, or physiological signals. However, these systems often rely on scientifically debatable assumptions and have been criticised for their inaccuracy and potential misuse⁶⁵. For instance, the misinterpretation of emotional states could lead to wrongful accusations or discriminatory practices in law enforcement contexts. However, the AI Act classifies as prohibited uses of emotion recognition only the specific sectors of «workplace and education institutions», as per Article 5(1)(f). All the other uses of emotion recognition are, therefore, permitted under the AI Act, including the law enforcement ones⁶⁶.

Similarly, biometric categorisation systems, which classify individuals based on physical or behavioural traits, such as ethnicity, gender, or age, raising significant risks of stereotyping and reinforcing societal biases⁶⁷, have a double classification under the AI Act. Article 5(1)(g) includes in the high-risk list biometric categorisation systems⁶⁸ when they categorise individuals

⁶⁴ Art. 3(1)(39) AI Act.

⁶⁵ F.P. Levantino, "[How Do You Feel Today?](#)" *Exploring IHRL and IHL Perspectives on Law Enforcement and Military Uses of Emotion Recognition Technology*, in *Opinio Juris*, 2023.

⁶⁶ Recital 54 AI Act: «emotion recognition systems that are not prohibited under this Regulation, should be classified as high-risk».

⁶⁷ Studies have shown higher error rates for minority groups in facial recognition systems, further exacerbating concerns about equality and non-discrimination. T. Madiaga - H. Mildebrath, *Regulating facial recognition in the EU*, in EPRS - European Parliamentary Research Service, 2021.

⁶⁸ Defined by Art. 3(1)(40) as «an AI system for the purpose of assigning natural persons to specific categories on the basis of their biometric data, unless it is ancillary to another commercial service and strictly necessary for objective technical reasons».

Federica Paolucci
*The New Face of Privacy:
AI, Power, and the Disappearing Private Sphere*

«based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation»⁶⁹. It is excluded from the prohibition of the use of such systems in the context of law enforcement. Instead, biometric categorisation systems that are based on inferred characteristics, such as age, gender, or ethnicity⁷⁰, are classified as high risk, according to Article 6(2) and Annex III, Article 1(b).

Finally, biometric verification systems are explicitly excluded from certain restrictive provisions under the AI Act, as noted in Recital 54. The Regulation recognises that verification systems present relatively low privacy risks compared to identification systems, which process biometric data on a larger, potentially unsolicited scale, and, therefore, are not subject to the regulation.

Taken together, these regulatory asymmetries confirm the point developed in the presentation: while the Regulation attempted to draw constitutional boundaries through a risk-based structure, the coherence of these boundaries depends on the clarity with which real-time, *ex post*, categorisation, and verification are distinguished and governed. The *quality of law* principle requires that these distinctions be normatively robust, not merely formal.

4. *The Elephants in the Room*

While the preceding analysis has identified some clusters of structural uncertainty within the AI Act's regulation of biometric systems – as the conceptual fragility of its risk classification, the incomplete delineation of unacceptable risks, and the absence of empirical justification as a condition for necessity and proportionality –, in this section, we will move to the top of it and will be considered as not merely technical imperfections or policy oversights.

The very distinction between real-time and post-remote identification illustrates a certain instability. Despite the Regulation's attempt to anchor different regulatory obligations to temporal modalities, the underlying technology remains the same, as repeatedly acknowledged by the EDPB and EDPS⁷¹. Similarly, the absence of a seriousness-of-offence requirement

⁶⁹ Art. 5(1)(g) AI Act

⁷⁰ Recital 16 AI Act.

⁷¹ EDPB-EDPS, *Joint Opinion 5/2021*, cit.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

for post-remote systems introduces a normative asymmetry: retrospective uses of biometric identification may access vast databases of facial images without the gravity threshold that constrains the deployment of real-time under Article 5 AI Act. The result is a regulatory landscape in which the difference between an outright prohibition and high-risk classification does not track the actual degree of interference with fundamental rights. This actually reveals even a deeper tension between the Regulation's ambition to constrain technologically augmented state power and its reliance on categories whose boundaries are normatively unstable.

It is precisely at this juncture that the “elephant in the room” emerges. If one assumes that the “legitimate aims” under Article 8(2) ECHR correspond, within the architecture of the AI Act, to the narrowly circumscribed scenarios in Article 5 – missing persons, victims of specified offences, and the investigation of «serious crimes» listed in Annex II – the fragility of the system becomes immediately visible. Annex II, even though incomplete, at least provides guidance in the application of real time systems; however, as mentioned previously, the same threshold is not applicable to post remote biometric systems. In such a context, the constitutional assessment of necessity and proportionality cannot meaningfully occur. Necessity presupposes a clear delineation of the aim; proportionality requires a structured comparison between the gravity of the intrusion and the weight of the public interest; quality of law demands that the individual foresees with reasonable clarity the circumstances of the interference. Yet the AI Act, far from supplying these conditions, delegates them almost entirely to national discretion.

The only safeguard expressly provided is the requirement of prior authorisation by a judicial authority or an independent administrative authority. But this safeguard raises the central question that structures the remainder of this analysis: *who* is meant to perform the authorisation? At which standards? What are the preconditions for the independent authorities? The Regulation does not discipline this choice; nor does it impose a minimum standard ensuring that whichever body is selected possesses the competence, guarantees, and procedural culture necessary to evaluate an interference of this magnitude.

The temporal dimension of the assessment is no less significant. Authorisation operates at the moment of deployment, or «in a duly justified situation of urgency, the use of such system may be commenced without an authorisation provided that such authorisation is requested without undue delay, at the latest within 24 hours», as recalled by Article 5(3) AI Act. In combination, under the AI Act, the deployers, hence the actual users of the

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

technology, must comply with Article 27 and perform a Fundamental Rights Impact Assessment, which must be carried out *prior to deployment* and must accompany the system throughout its life cycle⁷². On paper, this requirement appears to embody the positive dimension of privacy protection: an *ex ante* evaluation of risks, limits, safeguards, oversight structures, and foreseeable interferences. Yet when one asks *who* is expected to perform this assessment, the fragility of the mechanism becomes immediately apparent.

However, if the FRIA is left to the law enforcement authority that intends to deploy the technology, often the very authority whose investigative powers are expanded by biometric identification, the constitutional logic of the safeguard collapses. A police force body alone cannot meaningfully determine whether the conditions for its own intrusive action are met; nor can it be expected to reconstruct, through internal assessment, the normative boundaries that the legislature has failed to articulate. The paradox is obvious: a FRIA performed by the deployer presupposes a legal framework that already contains the substantive parameters, as necessity, proportionality, purpose limitation, storage guarantees, and independent oversight, that the FRIA is itself supposed to test. Without such a framework, the FRIA becomes a procedural self-certification rather than a rights-protective instrument.

This difficulty, compounded with the uncertainties around the authorisation, is the critical aspect that the next section will analyse in the light of the next to be full application of the Regulation. The regulatory decisions around the described aspects have significant implications, as they directly affect the level of oversight, scrutiny, and protection of fundamental rights associated with the deployment of these AI systems. Before delving into the analysis of these standards, it is crucial to underline that granting a certain degree of protection to the oversight of such privacy – *inter alia* –

⁷² For a more thorough understanding of this requirement, refer to D. Casaburo and I. Marsh, *Ensuring fundamental rights compliance and trustworthiness of law enforcement AI systems: the ALIGNER Fundamental Rights Impact Assessment*, in *AI and Ethics*, 2024; S. Bertaina *et al.*, *Fundamental rights and artificial intelligence impact assessment: A new quantitative methodology in the upcoming era of AI Act*, in *Computer Law & Security Review*, vol. 56, 2025, p. 106101 ff.; 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*, vol. 54, 2024, p. 106020 ff.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

interferences is indeed an element of Article 47 CFREU⁷³. These frameworks ensure that the deployment and use of such technologies operate under continuous scrutiny, providing a layer of protection that extends beyond initial approval. They align with Article 47 by guaranteeing that individuals have access to independent and impartial bodies, especially of a judicial nature, that oversee the deployment of these technologies and remain available to address any rights violations or disputes arising from their use.

Therefore, the choice between judicial and administrative authorisation, the delineation of a clear procedure of authorisation on the Member State's ground, is not merely procedural but reflects deeper concerns about how to balance security needs with the protection of civil liberties⁷⁴. The differences in accountability, transparency, and the degree to which fundamental rights are safeguarded vary greatly depending on the type of authority overseeing these decisions, and this divergence raises important legal considerations that will be explored in the following discussion.

4.1. *Authorisation*

The choice between judicial authority and administrative authority is not neutral because the standards of authorisation and oversight vary significantly between these two types of bodies, and the decision ultimately influences the balance between security needs and the protection of fundamental rights.

First and foremost, judicial authorities are inherently bound by constitutional safeguards and operate under a strict mandate to protect fundamental rights⁷⁵. Their decisions are typically subjected to higher levels of legal scrutiny, ensuring compliance with the principles of necessity,

⁷³ M. Bonelli, *Article 47 of the Charter, Effective Judicial Protection and the (Procedural) Autonomy of the Member States*, in M. Bonelli, M. Eliantonio and G. Gentile (eds), *Article 47 of the EU Charter and Effective Judicial Protection*, London, 2022, p. 81 ff.; S. Peers et al., *Article 47, The EU Charter of Fundamental Rights: A Commentary*, London, 2021, p. 1245 ff.

⁷⁴ R. Tarchi and A. Gatti, *Intelligenza Artificiale e Protezione Dei Dati Personali: Problemi Di Metodo e Di Procedura*, in *DPCE Online*, vol. 64, 2024.

⁷⁵ M. Cartabia, *I diritti in azione: Universalità e pluralismo dei diritti fondamentali nelle Corti europee*, Bologna, 2009.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

proportionality⁷⁶, and due process⁷⁷. This level of scrutiny is particularly important for AI-driven surveillance technologies⁷⁸, like biometrics, which have far-reaching implications for privacy, freedom of expression, and freedom of movement.

According to Article 47 CFREU, individuals must have access to effective judicial protection, which includes the right to an effective remedy before an independent and impartial tribunal⁷⁹. This standard of protection is integral to maintaining the rule of law within the EU and ensures that the deployment of risky AI technologies like biometric identification systems does not infringe upon individual rights⁸⁰. Hence, Article 47's broad application requires national courts to play a key role in upholding EU law. In the context of AI and biometric technologies, individuals must have access to national courts to challenge both the legality of AI deployments and any resulting violations of their rights.

While procedural autonomy allows member states to determine how EU law is enforced within their jurisdictions, Article 47 creates a binding obligation to ensure that remedies are effective and equivalent to those provided under national law for similar violations. This is a critical point, as procedural autonomy should not become an excuse for insufficient access to justice in cases involving AI and fundamental rights. Furthermore, the Court of Justice of the EU (CJEU) has repeatedly ruled that member states must ensure that their legal systems provide for effective judicial protection, particularly in contexts where Union law guarantees specific rights. This is particularly important for AI enforcement, as gaps in national procedural frameworks could hinder the ability of individuals to access justice. For instance, if a member state's legal framework does not offer adequate

⁷⁶ S. Peers et al., *Article 47*, cit.; K. Lenaerts, *Upholding the Rule of Law through Judicial Dialogue*, in *Yearbook of European Law*, vol. 38, 2019, p. 3 ff.

⁷⁷ F. Pasquale, *Inalienable Due Process in an Age of AI: Limiting the Contractual Creep toward Automated Adjudication*, in A. Reichman and others (eds), *Constitutional Challenges in the Algorithmic Society*, New York, 2021.

⁷⁸ The use of biometrics should not be treated differently than wiretapping and other surveillance measures, which is granted, at least in some Member States, such as in Italy, a high-level of protection with judicial oversight. For a national perspective, see D. Zecca, *La tutela costituzionale della segretezza delle comunicazioni e della corrispondenza alla prova delle tecnologie emergenti: profili comparati*, Naples, 2023.

⁷⁹ S. de Heer, *Artificial Intelligence and the Right to an Effective Remedy*, in A. Quintavalla and J. Temperman (eds), *Artificial Intelligence and Human Rights*, cit., p. 294 ff.

⁸⁰ E. Kosta, *Algorithmic state surveillance: Challenging the notion of agency in human rights*, in *Regulation & Governance*, vol. 16, no. 1, 2022, p. 212 ff.

remedies for violations of the AI Act, it could be seen as failing to meet its obligations under Article 47⁸¹.

In contrast, independent administrative authorities, though often more specialised and efficient, may operate under less stringent requirements⁸². Their decisions may be driven by regulatory or policy considerations rather than a thorough analysis of the legal rights at stake⁸³. While administrative bodies can act more swiftly and may have technical expertise in AI, they are not constitutionally bound to the holy respect of the rule of law, which is a prerogative of the courts, instead⁸⁴.

Looking at the case study of this analysis, facial recognition technology, especially when used in public spaces, represents a severe intrusion into personal privacy. For this reason, the AI Act attempts to safeguard against these risks by requiring prior authorisation for such technologies, but the effectiveness of these safeguards depends on the nature of the authorising body.

As sustained in both literature and case law, judicial authorities can grant independence and transparency of a decision that has a high impact on the individual's rights, as the CJEU has clarified in *Corbiau*, C-24/92 (i.e., § 15)⁸⁵. In this decision, the Court placed significant importance on the idea of independence in defining whether a body constitutes a court or tribunal. This emphasis is not surprising given that the core principle of the rule of

⁸¹ CJEU (GC), 27 February 2018, C-64/16, *Associação Sindical dos Juizes Portugueses v Tribunal de Contas*. In this case, the Court reiterated that the right to effective judicial protection is essential to the rule of law in the EU, particularly where individual rights under Union law are at stake. This ruling reinforces the idea that judicial oversight is indispensable for ensuring that AI systems, especially those that affect fundamental rights, are deployed ethically and lawfully.

⁸² This is a problem already emerged in the data protection framework. A recent report by the Fundamental Rights Agency underscored that Data Protection Authorities should be strengthened in their independence, also from a human, financial, and technical point of view. European Agency for Fundamental Rights, *GDPR in practice – Experiences of data protection authorities*, 2024. See also P. Schütz, *Assessing Formal Independence of Data Protection Authorities in a Comparative Perspective*, J. Camenisch et al. (eds.), *Privacy and Identity Management for Life*, Berlin, 2012, p. 45 ff.

⁸³ Independence entails that the authorities should not receive instructions from any other entity concerning the performance of their duties. The CJEU ruled that a supervisory authority must be free from any external influence, whether that be from public or private sectors, ensuring that their decision-making is solely based on the law. *Inter alia*, refer to CJEU (GC), 9 March 2010, C-518/07, *European Commission v Federal Republic of Germany*.

⁸⁴ CJEU, 10 April 1984, Case 14/83, *Sabine von Colson and Elisabeth Kamann v Land Nordrhein-Westfalen*.

⁸⁵ CJEU, 30 March 1993, C-24/92, *Pierre Corbiau v Administration des Contributions*.

Federica Paolucci
*The New Face of Privacy:
AI, Power, and the Disappearing Private Sphere*

law hinges on the reviewability of decisions made by public authorities through independent courts. Moreover, in *Kadi and Al Barakaat*⁸⁶, the Court held that effective judicial review is essential for ensuring compliance with fundamental rights, particularly when administrative decisions have far-reaching consequences for individuals. This case underscores the need for courts to remain independent and impartial: a topical aspect, especially when reviewing AI-related decisions, in high-risk contexts like law enforcement.

In this respect, the ECtHR elaborated a three-tiered “quality of law” requirement to ascertain the justification for the interference of law enforcement on individuals’ privacy⁸⁷. Thus, according to established jurisprudence, any interference must adhere to stringent standards of legal quality, necessitating that domestic legislation be clear, foreseeable, and readily accessible. Specifically, the Court emphasised that domestic law must provide robust measures to prevent any misuse of personal data that could violate these guarantees. This need for safeguards is particularly pronounced when it comes to using personal data for police purposes, especially given the increasingly sophisticated technology available, as in the case of AI and biometric technologies. For this very purpose, the application of such systems should happen with adequate care of individuals’ rights, striking a careful balance between the benefits of technology and the protection of fundamental rights⁸⁸, which is to be ensured via stringent justifications⁸⁹. Moreover, the Court, in the case of *Big Brother Watch and Others v UK*, established that, especially in cases of criminal investigation, without individual knowledge, «in a field where abuse in individual cases is potentially so easy and could have such harmful consequences for a democratic society as a whole, it is desirable to entrust supervisory control to a judge, judicial control offering the best guarantees of independence,

⁸⁶ CJEU (GC), 3 September 2008, Joined cases C-402/05 and C-415/05, *Yassin Abdullah Kadi and Al Barakaat International Foundation v Council of the European Union and Commission of the European Communities*.

⁸⁷ J. Livingston Slosser, *Interpreting the ‘Quality of Law’ at the European Court of Human Rights: Metaphorical Framing and Evaluative Judgment* 2018, available at papers.ssrn.com.

⁸⁸ ECtHR (GC), 4 December 2008, apps. 30562/04 and 30566/04, *S. and Marper v the United Kingdom*, cit., § 67.

⁸⁹ ECtHR (III Sec.), 21 June 2021, app. 33696/19, *Podchason v Russia*, §§ 60-65.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

impartiality and a proper procedure»⁹⁰. The message here is very clear: the “best” option is the judge⁹¹.

Thus, any infringement upon individuals’ right to privacy, and thus any utilisation of biometric technology, must adhere to the principle of “quality of law”, guaranteeing thorough oversight by the judicial authority over the reasoned implementation of such tools. Furthermore, legislation should clearly define the scope of discretion granted to competent authorities and specify how it should be exercised with precision to ensure individuals are adequately shielded from arbitrary encroachments⁹².

While judicial oversight ensures the highest level of independence, impartiality, and adherence to fundamental rights principles, as this thesis aimed to demonstrate⁹³, particularly in cases where these technologies pose significant risks to privacy and other rights, it is equally important to acknowledge the potential challenges associated with judicial authorisation, particularly the risk of procedural delays that could impede the timely deployment of AI systems in urgent or operationally sensitive contexts.

In light of these concerns, in this section, an argument “*in subordine*”, in a subordinate position to the main one, *supra*, Member States should consider mixed oversight systems where judicial authorities work alongside independent administrative bodies⁹⁴.

In this sense, the ECtHR has also recognised, in cases like *Kennedy v UK*⁹⁵, the validity of mixed or quasi-judicial oversight mechanisms as “good

⁹⁰ ECtHR (GC), 25 May 2021, apps. 58170/13, 62322/14 and 24960/15, *Big Brother Watch and Others v the United Kingdom*, cit.

⁹¹ Further supporting this position, the United Nations High Commissioner for Human Rights has emphasized the necessity of judicial involvement in oversight of facial recognition technologies. In the report titled *Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests* (UN Doc. A/HRC/44/24, 24 June 2020), it is recommended that States involve an independent body, preferably judicial in nature, to authorize the use of facial recognition technology in the context of assemblies. The report stresses that any use of facial recognition systems should be open to judicial challenge and urges authorities to ensure transparency, notifying the public whenever such technologies are deployed.

⁹² ECtHR (III Sec.), 4 July 2023, app. 11519/20, *Glukhin v Russia*, cit.

⁹³ *Ibid.*, §§ 82-87.

⁹⁴ G. Malgieri and P. de Hert, *European Human Rights, Criminal Surveillance, and Intelligence Surveillance: Towards “Good Enough” Oversight, Preferably but Not Necessarily by Judges*, in D. Gray and S. E. Henderson (eds), in *The Cambridge Handbook of Surveillance Law*, New York, 2017.

⁹⁵ ECtHR (IV Sec.), 18 May 2010, app. 26839/05, *Kennedy v UK*.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

enough”⁹⁶; thus, as an alternative, provided these systems maintain robust guarantees of independence, impartiality, and adherence to the principles of legality, necessity, and proportionality⁹⁷. This flexibility has, however, encountered a crucial threshold: targeted surveillance practices – such as one of the uses that a deployer can make of the biometric identification systems authorised under the AI Act – present heightened risks of rights infringements and thus demand stricter judicial oversight.

Whereas, in landmark decisions such as *Szabó and Vissy v Hungary*⁹⁸ and *Zakharov v Russia*⁹⁹, the ECtHR affirmed that, even though judicial review remains the preferable safeguard, the Court, recognising non-judicial oversight systems as effective when judicial components, independence, and broad oversight powers are integrated, adopted a flexible approach. Instead, the Court has repeatedly stressed the need for oversight mechanisms that combine independence, transparency, and sufficient enforcement powers, tailored to the specific risks posed by surveillance systems, hence tailored to how invasive the system is assessed to be¹⁰⁰.

This is particularly true in the context of law enforcement, where the real-time application of FRT and other biometric systems carries a significant potential for abuse. The principles of necessity and proportionality, enshrined in both EU law and ECtHR jurisprudence, require that such intrusive measures be rigorously evaluated before deployment, with specific attention to the risks they pose to privacy, equality, and freedom of expression¹⁰¹.

⁹⁶ In *Kennedy*, for example, the ECtHR evaluated the role of the UK’s Investigatory Powers Tribunal and the Interception of Communications Commissioner, acknowledging their effectiveness as hybrid oversight mechanisms that incorporate elements of judicial review while leveraging administrative and technical expertise.

⁹⁷ See, also, ECtHR (C), 24 April 1990, app. 11105/84, *Huwig v France*; P. de Hert and A. Aguinaldo, *A leading role for the EU in drafting criminal law powers? Use of the Council of Europe for policy laundering*, in *New Journal of European Criminal Law*, vol. 10, no. 2, 2019, p. 99 ff.

⁹⁸ ECtHR, 6 June 2016, app. 37138/ 14, *Szabó and Vissy v Hungary*.

⁹⁹ ECtHR, 4 December 2015, app. 47143/06, *Roman Zakharov v Russia*.

¹⁰⁰ G. Malgieri and P. de Hert, *European Human Rights, Criminal Surveillance, and Intelligence Surveillance*, cit.

¹⁰¹ See, also, ECtHR (C), 24 April 1990, app. 11105/84, *Huwig v France*, cit.

4.2. Normative Design

While the preceding section highlighted that the legitimacy of deploying biometric identification systems cannot be reduced to the formal act of obtaining authorisation, judicial or otherwise, here this concept will be further stressed, considering the legislative structure within which such an authorisation is exercised. As a matter of fact, even the most rigorous judicial scrutiny operates within the epistemic and normative boundaries set by the legislature. Where those boundaries are indeterminate, thin, or procedurally under-specified, the very institution entrusted with safeguarding rights is rendered structurally incapable of doing so. In other words, the question of who authorises is inextricable from the question of what is being authorised and according to which substantive and procedural parameters. As mentioned at the beginning of this paper, the premises from which the analysis departed are entrenched with the idea that the totalitarian regimes of the past, and not only, perpetrated atrocious crimes «through the law, not despite the law»¹⁰². Hence, even though there was a formalistic respect for the law, the problem was the provision of such laws and the content that was «at odds with the most basic sense of justice»¹⁰³.

This is indeed the perspective adopted by this research. Legal systems can respect the rule of law in a strictly procedural sense while authorising practices that are at odds with the most basic sense of justice. The problem does not lie only in the absence of law, but also in the *content* and *structure* of the law that enables and shapes the exercise of power. The perspective adopted in this paper is rooted in this distinction: assessing the EU regulatory model and its national adaptations not merely in terms of formal legality, but in terms of whether their premises, design choices, and institutional arrangements are capable of effectively protecting fundamental rights, including privacy.

It is in this conceptual space – between the abstract legality of measures and the substantive conditions of their legitimacy – that the Fundamental Rights Impact Assessment acquires a constitutional significance. Article 27 AI Act gestures toward this potential, but its implementation risks being reduced to a procedural appendix or to a mere replication of the logic of the data-protection DPIA. Properly understood, the FRIA is not an exercise that the deployer performs on itself, nor is it confined to the *ex post* mapping of harms. It is the point at which the legal

¹⁰² M. Cartabia and N. Lupo, *The Constitution of Italy*, cit.

¹⁰³ Ibid.

order is required to confront, in a structured way, the interaction between technological capacity, institutional design, and fundamental rights.

For this reason, it is neither practical nor normatively desirable to assume that a law enforcement authority can, on its own, perform a FRIA that satisfies constitutional standards. By definition, law enforcement “enforces the law”; it does not write the substantive and procedural parameters against which its own action should be measured. This is where the notion of a *normative* FRIA becomes relevant¹⁰⁴.

A normative FRIA presupposes – and at the same time generates – a thicker conception of legality. It requires the legislature to define, in concrete terms, the parameters according to which the impact on fundamental rights is to be evaluated: which risks are admissible or unacceptable; which categories of persons require heightened protection; which forms of human oversight and contestability must accompany the system; which governance mechanisms are triggered when interferences materialise; which remedies are available and to whom; which authority is institutionally positioned to ensure independence, competence, and enforceable oversight.

These elements, which Article 27 only sketches, cannot be absorbed into a DPIA-style logic centred on internal compliance¹⁰⁵. They concern not the internal organisation of a controller, but the constitutional sustainability of public power when it relies on biometric identification systems. Put differently, a normative FRIA is not simply a checklist; it is a legislative and institutional commitment to reconstruct, *ex ante*, the conditions under which the deployment of intrusive technologies remains compatible with Articles 7 and 8 CFREU and Article 8 ECHR.

At this point, privacy, in its contemporary form, re-enters the picture. As argued above, privacy can no longer be reduced to a negative claim of non-interference, nor to the regulation of data flows alone. It operates as a relational guarantee, sustained by a network of safeguards – procedural, institutional, and technological – each of which contributes to recreating the “distance” between the individual and the technological gaze. A normative FRIA gives operational form to this understanding: it forces the legal system to ask not only which data are processed, but how power is exercised; not

¹⁰⁴ See F. Paolucci, *Constitutional Safeguards in the Age of AI. A Study on the Fundamental Rights Impact Assessment of Facial Recognition Technology*, in *Iris Bocconi*, 2025.

¹⁰⁵ A. Calvi et al., *Contribution to “targeted stakeholder consultation on classification of AI systems as high-risk”*, in *D.pia.lab*, 2025; see, as well, G. Malgieri and C. Santos, *Assessing the (severity of) impacts on fundamental rights*, in *Computer Law & Security Review*, vol. 56, 2025, 106113.

only which risks are inherent in AI systems, but how those risks are distributed socially; not only which harms have occurred, but which structural vulnerabilities make certain groups more visible, more exposed, and more governable than others.

If the law does not clearly specify who must perform this assessment, under which evidentiary standards, and with which institutional guarantees, the entire system risks collapsing into procedural formalism. Conversely, when the FRIA is conceived normatively – as a precondition for deployment, a moment of constitutional reflection, and the site where the rule of law confronts technological autonomy – it provides the structural framework that authorisation alone cannot supply. This reading is consistent with the jurisprudence discussed earlier: the ECtHR’s emphasis on “quality of law”, “minimum standards of guarantees”, and a “pressing social need” is not satisfied by the mere existence of an authorisation regime. It presupposes a legislative and institutional design capable of making such authorisation meaningful.

The normative function of the FRIA is therefore systemic rather than purely procedural. It contributes to what this paper identifies as the sustainability of the legal order: the capacity of law to govern technological power without succumbing to it, to maintain equilibrium between security and liberty, and to preserve the pluralistic, relational spaces within which the right to privacy – as a constitutional infrastructure of democratic life – continues to have content. Everything, however, begins and can only begin with the law. The following section tests this conceptual framework against one of the first national laws adopted to implement the AI Act, Italy’s Law No. 132/2025, in order to assess whether, and to what extent, that law supplies the normative thickness that a constitutional FRIA would require, or whether it instead replicates the fragilities already evident at the EU level.

4.3. *The Italian Case*

As stated by Article 1(2) of Law No. 132/2025 adopted by Italy in October 2025, the purpose of the norm is to “interpret and apply” the rules set out by the AI Act¹⁰⁶. Although the law does not introduce new obligations, it aims to organise the legal framework for the use of AI systems

¹⁰⁶ For a more in-depth analysis of the content of Law No. 132/2025, see F. Paolucci, *La legge italiana sull’intelligenza artificiale: attuazione nazionale dell’“AI Act” e primi nodi applicativi*, in *Quaderni costituzionali*, vol. 4, 2025.

by modifying, expanding, or systematising areas that are otherwise regulated in a fragmented manner. It is not the purpose of this paper to delve into the general content of the law, and suffice it to highlight here some critical elements that interest the particular perspective adopted so far.

The regulation of artificial intelligence in sensitive domains, including national security and law enforcement, is here channelled through a broad and indeterminate delegation to the executive. A first crucial provision in this regard is Article 6, which governs the relationship between the AI Act and the national security exemption. Drawing on Article 2(3)(a) AI Act, Law No. 132/2025 excludes from its scope systems used for “national security” purposes. However, the national provision significantly broadens this category: it includes activities connected not only to defence and cybersecurity carried out by the armed forces and police, but also certain operations of the judicial police in the field of counterterrorism and cybercrime¹⁰⁷.

From the standpoint of EU law, these latter activities are not excluded from the AI Act’s scope. Article 5(1)(h) AI Act explicitly contemplates the exceptional use of otherwise prohibited systems – including real-time remote biometric identification in publicly accessible spaces – for law-enforcement purposes, under strict conditions and only in relation to the serious offences listed in Annex II, such as terrorism. It also requires that such uses be subject to prior authorisation by a judicial or independent administrative authority, limited in time and justified by concrete needs.

There is therefore no manifest contradiction between the two regimes, but there is a different delimitation of what counts as “national security”. By including certain anti-terrorism and cybercrime operations of the police within the national security carve-out, Italian law risks placing these scenarios under the exemption that the AI Act treats as law enforcement and thus subjects them to its *lex specialis* safeguards. The result is a potential material misalignment that will have to be clarified in the implementation phase.

This brings into sharper focus the role of Article 24, and in particular Article 24(2)(h), which delegates to the Government the adoption of one or more legislative decrees laying down «a specific discipline for the use of artificial intelligence systems for police activities». More generally, Article 24 confers a wide-ranging power to legislate by decree on matters spanning governance, sectoral regulation, and enforcement, even then, in areas that

¹⁰⁷ As disciplined under Italian law by Art. 9(1)(b) and (b-ter) of Law No. 146/2006 on undercover operations. and foreseeability.

directly implicate fundamental rights and highly intrusive technologies. In this sense, the Constitutional Court has repeatedly required enabling statutes to set «precise and defined criteria»¹⁰⁸ guiding delegated legislation. Where such criteria are absent or drafted in very general terms, technical discretion risks sliding into normative discretion¹⁰⁹, diluting the principle of substantive legality and making a law «precarious»¹¹⁰ instead of “of quality”.

Yet this legislative posture is not uniquely Italian. A comparable dynamic can be observed at the EU level in the regulatory architecture of the AI Act itself. While the Regulation formally establishes a risk-based framework and articulates core prohibitions and obligations, it leaves crucial definitional, procedural, and supervisory choices to secondary instruments, implementing acts, delegated acts, and guidelines¹¹¹. The Commission’s central role in shaping the interpretation of key categories¹¹², from high-risk systems to prohibited practices and governance standards, reflects a European tendency, unfortunately very similar to the Italian one, to concentrate legislative action at the executive level.

The result, both nationally and at the EU level, is a model at the detriment of the representative organs, such as the parliaments, who are left out of the process under the justification of the need for the appropriate technical expertise, operational efficiency, and the need to react swiftly to evolving risks.

Moreover, given the sensitive context, the protection of fundamental rights is of utmost importance both in practice, by enforcing measures such

¹⁰⁸ Italian Constitutional Court, dec. No. 175/2022, as also mentioned in the Dossier published by in May 2025, by the Chamber of Deputies.

¹⁰⁹ O. Pollicino and G. Muto, *La legislazione delegata in materia di intelligenza artificiale: la costruzione di una disciplina organica al confine tra scelte governative, controllo parlamentare e vincoli europei*, in *Rivista di Diritto dei Media*, vol. 1, 2025, p. 393 ff.

¹¹⁰ E. Longo, *La legge precaria: Le trasformazioni della funzione legislativa nell’età dell’accelerazione*, Turin, 2017.

¹¹¹ However, the need for standards and guidelines would be a hindrance, as can also be seen from the amendments to the proposed amendment to the AI Act, referred to in the Digital Omnibus, which would require postponing the entry into force of the regulation to allow the commission to finalize them. See the Draft Opinion of the Committee on Legal Affairs for the Committee on the Internal Market and Consumer Protection and the Committee on Civil Liberties, Justice and Home Affairs on the proposal for a regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI) (COM(2025)0836 – C10-0304/2025 – 2025/0359(COD)), rapporteur for opinion: Sergey Lagodinsky, 2 February 2026.

¹¹² Art. 97 AI Act.

as Article 27 AI Act (FRIA), but also through governance. In this sense, Italy did not designate on time the fundamental rights authority required by Article 77 AI Act through the Italian Law No. 132/2025. Such authorities might cover a great role in the future implementation of the AI Act since they are already existing «authorities protecting fundamental rights» which have «the power to request and access any documentation created or maintained» under the Regulation to the national public authorities or bodies which nationally oversee the protection of fundamental rights¹¹³. Such bodies had to be identified by the Member States by 2 November 2024, and Italy missed the deadline, showcasing a long-standing problem of the absence of a National Human Rights Institution¹¹⁴. As mentioned, this very deficiency is particularly critical in sensitive fields, such as those of biometric identification systems used in law enforcement and migration, where market authorities have «effective investigative and corrective powers, including at least the power to obtain access to all personal data that are being processed and to all information necessary for the performance of its tasks» (Recital 159). In any case, Italy designated these authorities only in February 2026, and not through a law but a decision of the Department for Digital Transformation, in collaboration with AgID (Agency for Digital Italy) and ACN (National Cybersecurity Authority), the appointed authorities that will enforce the AI Act respectively as notifying authority and market surveillance authority¹¹⁵.

From the standpoint of the right to privacy, these choices are not neutral. A broadened and elastic notion of national security narrows, in practice, the domain in which the guarantees of the AI Act apply; a vague and open-ended delegation deprives individuals of the foreseeability that Article 8 ECHR and Article 7 CFREU require; the late and opaque, from a

¹¹³ Which, in the context of Italy, are the National Cybersecurity Agency and AgID as the national AI authorities, both under the Presidency of the Council, and preserves the roles of the Data Protection Authority and the Italian Communications Regulatory Authority (AGCOM).

¹¹⁴ As it indeed remained on identifying a National Human Rights Institution for a long time. For more focus on this specific instance, allow to refer to F. Paolucci, *Op-Ed: "Fundamental Rights Authorities and the AI Act: Neglected Priorities?"*, in *EU Law Live*, 14 October 2025. It is indeed recalled the Special Issue edited by A. Di Martino for this review on *National Human Rights Institutions. Alcune esperienze comparate*, vol. 2, 2025. On the specific Italian problem, it is also recalled G. Cerrina Feroni, *I tentativi di istituire una NHRI nel contesto istituzionale italiano: quale ruolo per il Garante per la protezione dei dati personali?*, in G. Repetto, *Una National Human Rights Institution per l'Italia: problemi e prospettive*, Turin, 2025.

¹¹⁵ *Autorità di protezione dei diritti fondamentali ai sensi dell'AI Act, Implementazione nazionale dell'articolo 77 dell'AI Act*, available at [innovazione.gov.it](https://www.innovazione.gov.it).

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

formal and legislative point of view, designation of the fundamental rights authorities under Article 77 AI Act impact on the overall sanity of the system, with potential repercussion on the disproportionate uses of this technology, and, for the perspective of this paper, biometric identification systems more specifically. The cumulative effect is not only reflected in institutional fragmentation but also normative opacity, making it harder for individuals to know when, how, and on what legal basis they may be subjected to biometric surveillance, and to whom they can turn before or after such interferences occur. In other words, the conditions under which privacy can be exercised as a constitutional right – clarity of the legal framework, *ex ante* limits on intrusive technologies, independent oversight, and access to remedies – are weakened precisely in those sectors where the erosion of the intermediate space between visibility and identifiability is most acute.

Privacy, however, is not merely a defensive entitlement. Under Article 8(2) ECHR, interferences are permitted only if they are «in accordance with the law» and «necessary in a democratic society», as mentioned *supra*, for specified legitimate aims. This provision encapsulates the very purpose of privacy within the system of constitutional law: not to deny public power, but to structure and constrain it. The law must predetermine the scope of normative discretion, define the legitimate aims with sufficient clarity, and embed the exercise of intrusive powers within a framework of necessity, proportionality, without, however, making of it «a *passe-partout* to weaken human rights guarantees, within the ECHR system of protection as well as in the framework of EU law»¹¹⁶. In other words, «if the rule contains a prohibition, the prohibition must strictly be observed, save in the specific cases for which the provision itself provides for an exception to the prohibition»¹¹⁷.

Once again, the centrality of the law must be preserved, as argued in this paper, in the light of the proportionality principle. It is precisely at this intersection that the relevance of a normative FRIA becomes most evident. If the *ratio* of such an instrument is to assess and mitigate the risks on fundamental rights on the use of AI, in order to function, the legal order must pre-structure, at the legislative level, the parameters and institutional circuits through which risks are assessed, affected groups are identified,

¹¹⁶ F. Viganò, [Rethinking the Proper Role of Proportionality in the Limitation of Fundamental Rights](#), in *Draft text of the intervention at the international conference (Riga, 2-3 September 2021): 'EU nited in Diversity: Between Common Constitutional Traditions and National Diversities' 4th Panel: Limitation on the exercise of fundamental rights*, 2021, notably p. 13.

¹¹⁷ *Ibid.*, p. 13.

Federica Paolucci

*The New Face of Privacy:**AI, Power, and the Disappearing Private Sphere*

human oversight is defined, and remedial pathways are activated. These elements cannot be left entirely to secondary legislation or to self-assessment by deployers, and they cannot be reduced to a “GDPR-style” DPIA, which is indeed a parameter but is an assessment just focused on one limited aspect. A FRIA is, by definition, a fundamental rights “tool”: it presupposes the involvement of institutions capable of articulating rights-based reasoning and of interacting on equal footing with security and market-oriented authorities.

However, the scope of this analysis is not theoretical: instead, it is substantiated by the need to affirm a critical concept, that the embedding of biometric identification technologies within social life does not progressively normalise and exacerbate forms of surveillance incompatible with a democratic society. Only where the limits of power are predetermined by the law, subjected to rigorous necessity and proportionality review, can technological innovation coexist with the “constitutional” vocation to the protection of fundamental rights¹¹⁸.

Measured against this standard, the Italian framework is, at present, normatively incomplete. Even if future decrees seek to fill some of these gaps, the initial legislative act does not set the constitutional compass that should orient their content. In this sense, Italy exemplifies the central thesis of this paper: where the law is normatively thin or strategically elastic, privacy cannot perform its core function, that of structuring, limiting, and legitimising the exercise of public power in accordance with the spirit of Article 8(2) ECHR.

5. Conclusion

The analysis conducted in this paper aimed at showing that the contemporary conceptualisation of the right to privacy, far from a mere shield against intrusion, requires a legal order of recognising that this right cannot be reduced to a defensive barrier against intrusion. This is the *positive* and constitutional dimension of privacy that the case concerning Mr. Glukhin makes explicit: rights do not survive because surveillance technologies are limited in capacity; they survive because the law

¹¹⁸ N. Purtova, *Between the GDPR and the Police Directive: navigating through the maze of information sharing in public-private partnerships*, in *International Data Privacy Law*, vol. 8, 1, 2018, p. 52 ff.

Federica Paolucci

The New Face of Privacy:

AI, Power, and the Disappearing Private Sphere

predetermines, with sufficient boundaries within which public power may operate.

Measured against this standard, the current regulatory landscape remains normatively fragile. The AI Act identifies the risks but leaves the architecture of safeguards thin; Italy's Law No. 132/2025, in turn, amplifies this fragility through broad delegations and vague standards. What emerges is a framework that speaks the language of rights yet lacks the structural grip needed to secure them.

What is at stake is not simply regulatory coherence, but the constitutional equilibrium embodied in Article 8(2) ECHR. That provision encapsulates the core function of constitutional law: to permit interferences only when they are in accordance with the law, pursue legitimate aims, and are necessary in a democratic society. This requires that the essential limits to technologically enhanced power be determined *ex ante* by the legislature and embedded within independent oversight and effective remedies.

Within this perspective, a normative conception of the Fundamental Rights Impact Assessment may serve as a methodological reorientation. Properly understood, the FRIA is not a compliance instrument, but a constitutional device through which necessity, proportionality, and accountability are operationalised before and throughout the deployment of AI systems. It is only through such a structured and legislatively grounded approach that the embedding of biometric identification technologies in social life can be prevented from normalising and intensifying surveillance practices incompatible with democratic freedom.

Therefore, the future of privacy in “the age of AI” will not depend solely on technological design or regulatory ambition, but on the capacity of legal systems to preserve their role as constraints on power.

The “new face” of privacy, then, is already visible: it is the face of constitutional balance. Whether Europe and its Member States will supply the legal architecture capable of sustaining that balance remains the decisive question.

Federica Paolucci
*The New Face of Privacy:
AI, Power, and the Disappearing Private Sphere*

ABSTRACT: This paper proposes a constitutional rethinking of privacy in the age of AI. Drawing on the concept of the intermediate space between visibility and identifiability, and using *Glukhin v Russia* as a diagnostic case, it argues that biometric identification systems dissolve the conditions that once allowed individuals to appear in public without becoming immediately knowable, classifiable, and subject to power. The paper shows that the core problem is not the technology itself, but the normative thinness of the legal frameworks governing it: where the law lacks density, precision, and institutional guarantees, privacy cannot function as a constitutional right.

KEYWORDS: privacy – fundamental rights – artificial intelligence – biometric identification systems – AI Act.

Federica Paolucci – Postdoctoral Researcher, Baffi Research Centre
– Bocconi University, Milan, Italy (federica.paolucci@unibocconi.it)

Caught Between AI and the AI Hype: How the Right to Personal Data Protection was Ambushed*

Gloria González Fuster

TABLE OF CONTENTS: 1. Introduction. – 2. The Right to the Protection of Personal Data. – 3. Personal Data Protection, AI and the AIA. – 4. A Matter of Priorities. – 5. Under Assault. – 5.1. Notion of Personal Data. – 5.2. Data Controller Obligations. – 5.3. Rights of Data Subjects. – 5.4. Special Categories of Data. – 6. Concluding Remarks.

1. Introduction

It was supposed to be the most important right in the era of digitalisation. A right conceived to answer the question of how can we, humans, control computers, so they do not control us - or are not used by others to control us. A right that progressively crystallised as a fundamental of the European Union (EU), building on decades of reflections and legislative experiences in multiple European countries, as well as on intense international debates. A right that the EU chose to place at the highest level of its legal framework, devoting to it Article 8 of the Charter of Fundamental Rights of the European Union (“EU Charter”, CFREU), as well as Article 16 of the Treaty on the Functioning of the EU (TFEU). The right to the protection of personal data could have been envisioned and embraced as the right that mattered the most for the protection of individual freedoms (and, through them, of our democracies) in Europe in times of unprecedented spread of Artificial Intelligence (AI). It could have been a right that policymakers would be decidedly prioritising, possibly reinforcing, and regularly underlining, committed to defend it and promote especially if and when they were simultaneously enthusiastically embracing AI.

It turned out instead that the right to the protection of personal data was a fundamental right some were very glad to throw under the AI bus. A right that irked with its presumed ability to slow down, if not impede, some personal data processing perceived as essential for the development of AI, itself pictured as critical for economic growth and societal progress. Personal data protection, were told Europeans, is possibly not the solution to some of the problems that AI might bring, but it might as well be, above all, a problem: an obstacle frustrating AI’s take up in the EU. Personal data

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

protection rules must be adapted, became then the message. Such rules must be “simplified”. In 2025, the European Commission put forward what were presented as first steps to reimagine EU data protection law in order to jump as effectively as possible on the AI wave. According to itself, the European Commission did not, with any of its proposals, put into question the EU fundamental right to data protection.

This contribution claims that such an attack has been launched. It starts by situating the basic contours of the right, and reviews some of the major points of discussion brought about AI in its regard. After discussing the Artificial Intelligence Act (AIA)¹, it examines the later approach of the European Commission in search for insights on the current challenges at the intersection between AI, the EU’s strive for more AI, and the right to personal data protection.

2. *The Right to the Protection of Personal Data*

The right to the protection of personal data is best understood as a tripartite right², which foresees that whenever someone decides to process somebody else’s personal data, this shall trigger three sets of legal implications. These three sets are evoked in Article 8 CFREU and further detailed in secondary data protection law such as the General Data Protection Regulation (GDPR)³.

A first type of consequence concerns the obligations imposed on those who decide to process personal data (the “data controllers”). These persons have to respect a series of duties, and above all they must comply with what the GDPR calls «principles relating to processing of personal

* This article was subjected to double-blind peer review.

¹ 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 (Artificial Intelligence Act) (Text with EEA relevance), OJ L, 2024/1689, 12.7.2024.

² F. Hondius, *Emerging Data Protection in Europe*, Amsterdam, 1975, p. 1.

³ 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) (Text with EEA relevance), OJ L 119, 4.5.2016.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

data»⁴, or «basic principles for processing»⁵. Some of these basic principles are also visible in Article 8(2) CFREU, at least indirectly. That paragraph of Article 8 CFREU states indeed that personal «data must be processed fairly» (mirroring the GDPR principle of “fairness”)⁶, «for specified purposes» (alluding to the GDPR principle of “purpose limitation”)⁷ and «on the basis of the consent of the person concerned or some other legitimate basis laid down by law» (evoking the GDPR principle of “lawfulness”)⁸. Other GDPR basic data processing principles may not be directly findable in the wording of Article 8 CFREU, but they nevertheless correspond to the principles of proportionality and necessity referred to in Article 52 CFREU, which describes permissible limitations of EU fundamental rights:⁹ that is the case of the principles of “data minimisation”¹⁰ or “storage limitation”¹¹.

A second type of legal consequence connected to the right to personal data protection is that the person to whom the personal data relates (the “data subject”) is granted a series of rights. Two of these rights of the data subject are explicitly mentioned in Article 8(2) CFREU, which proclaims that everyone has the right of access to data which has been collected concerning them (the “right of access”)¹², as well as the right to have such data rectified (“right to rectification”)¹³.

The third type of legal implications relates to the imperative control of compliance with personal data protection rules by an independent authority. The requirement is asserted in Article 8(3) CFREU, and the GDPR devotes a whole chapter¹⁴ to the regulation of these supervisory authorities, generally known as “Data Protection Authorities” (DPAs).

⁴ Art. 5 GDPR.

⁵ Art. 83(5)(a) GDPR.

⁶ Art. 5(1)(a) GDPR.

⁷ Art. 5(1)(b) GDPR.

⁸ Art. 5(1)(a) GDPR.

⁹ «Any limitation on the exercise of the rights and freedoms recognised by this Charter must be provided for by law and respect the essence of those rights and freedoms. Subject to the principle of proportionality, limitations may be made only if they are necessary and genuinely meet objectives of general interest recognised by the Union or the need to protect the rights and freedoms of others» (Art. 52(1) CFREU).

¹⁰ Art. 5(1)(c) GDPR. Underlining that data minimisation «gives expression» to the principle of proportionality: CJEU, 22 June 2021, C-439/19, *Latvijas Republikas Saeima*, § 98.

¹¹ Art. 5(1)(e) GDPR.

¹² Art. 15 GDPR.

¹³ Art. 16 GDPR.

¹⁴ Chapter VI GDPR.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

These three sets of legal consequences may be regarded as the content of the right to personal data protection. Conceiving the right in this manner shows that personal data protection is not necessarily about restraining the processing of personal data, but about the conditions in which such processing can take place. The right to data protection is in addition, importantly, not an absolute right, but a right that can be limited - if its limitations are in line with the requirements of the already mentioned Article 52 CFREU. This implies that some of the conditions accompanying the processing of personal data shall, in some cases, apply only in a reduced manner, which reinforces the idea according to which personal data protection is not to be conceived as an impediment for personal data processing, but rather as a framework for appropriate processing¹⁵.

From a certain point of view, the right to data protection could appear as the right needed in an AI world. It was the advent of computers that triggered thinking into a series of rules that would be necessary in times when decisions are increasingly taken by machines, as opposed to by humans, or by machines commended by humans in opaque ways. The origins of the right to personal data protection can indeed be traced back to the end of the 1960s and most significantly the 1970s, when policy makers and thinkers from a variety of countries started to distil new legal rules they deemed needed to protect individuals – and specifically individuals' freedom – in the face of the dawn of computers and more and more prominent data processing. Data protection law was created because of computerisation and automated decision-making¹⁶, to protect societies against the risks associated with them.

3. *Personal Data Protection, AI and the AIA*

In the EU, the right to personal data protection is enshrined at the highest level since 2009, when the Charter of Fundamental Rights of the European Union and Article 16 TFEU became legally binding. The

¹⁵ The doctrine has extensively discussed the nature's right and ambitions; Hallinan described it as a right which aims «to facilitate the interpretation of the constitutional order in relation to personal data processing» (D. Hallinan, *Data Protection as a Normative Problem* in M. Durante and U. Pagallo (eds), *The De Gryter Handbook on Law and Digital Technologies*, Berlin, 2025, p. 497).

¹⁶ This concern was particularly visible in the *French Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*, and in particular its Art. 2, prohibiting certain types of automated justice and administrative decisions.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

recognition of the right as a fundamental right is supposed to protect it from inappropriate limitations by the legislator, as all EU secondary law must be compatible with EU primary law. An extensive data protection law acquis has also developed over the decades.

During all this time, the weight of protecting individuals in front of computers and automated decision-making appeared to rest mainly on the shoulders of data protection law. Research carried out by the Fundamental Rights Agency of the EU (FRA) in 2020 showed that, for many, data protection laws stood out as the main set of rules to apply when dealing with AI¹⁷. Data protection law played such role, in general, through its regulation of the processing of personal data, but also more specifically through provisions on profiling and automated decision-making, such as in the currently in force Article 22 GDPR¹⁸.

In the meantime, the perception emerged that the increasing use of AI did not leave data protection unaltered. The clearest manner in which AI was seen to impact personal data protection, and the very fundamental right to personal data protection, was by demanding the processing of extremely large amounts of personal data¹⁹. AI is, by definition, data-hungry, and AI-driven demand for personal data can trigger challenges for data protection law compliance, and for all actors involved²⁰. Personal data is not only (often) the raw material used by AI to be trained; the output of AI might be personal data too. As eventually observed by the European Data Protection Board (EDPB) in relation to AI models, both the development and the deployment of these models may raise risks to the rights protected by the EU Charter, such as the right to personal data protection – risks that the EDPB qualified as serious²¹. By multiplying personal data processing

¹⁷ European Union Agency for Fundamental Rights, *Getting the Future Right Artificial Intelligence and Fundamental Rights*, 2020, Report 62.

¹⁸ In the judgment *SCHUEFA Holding* (Scoring), the CJEU supported a broad interpretation of the scope of application of Art. 22 GDPR (CJEU, 7 December 2023, C-634/21, *SCHUEFA*).

¹⁹ European Union Agency for Fundamental Rights, *Getting the Future Right Artificial Intelligence and Fundamental Rights*, cit.

²⁰ It should be noted that this has led to multiple initiatives in a variety of fora to facilitate compliance with data protection when using AI. See, for instance: I. Barbera and M. Popa-Fabre, [Privacy and Data Protection Risks in Large Language Models \(LLMs\)](#), Expert Report on Privacy and Data Protection Risks in Large Language Models (LLMs), available at *coe.int*, 2025.

²¹ European Data Protection Board, *Opinion of the Board (Art. 64) 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models*, 17 December 2024.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

operations, upstream and downstream, and by complexifying the very processing operations at stake – growingly opaque – it appeared that the increasing use of AI brought perhaps challenges that data protection law “as we know it” was not ready to tackle, or not capable of tackling.

From this perspective, new rules were necessary. A certain consensus²² began to emerge on data protection being «crucial but not sufficient» in the AI era²³. Some believed initially that such new necessary rules could take the shape of ethical (as opposed to legal) rules, but work on non-legally binding “AI ethics” guidelines was eventually perceived as not providing a satisfactory solution²⁴. The new rules that the EU opted to endorse in due course, in order to be ready for the AI era, took the form of a Regulation: the AI Act²⁵.

The AIA intersects with data protection law in several ways²⁶, even if it aims to regulate AI systems without directly interfering with the EU personal data protection acquis²⁷. Personal data shall be processed in the

²² The consensus was nevertheless not absolute nor global; see, for instance, criticising the EU for remaining «steadfast in not carving AI out of the GDPR»: M. Humerick, *Taking AI Personally: How the E.U. Must Learn to Balance the Interests of Personal Data Privacy & Artificial Intelligence*, in *Santa Clara High Technology Law Journal*, vol. 34, 2018, p. 393 ff.

²³ To paraphrase M. Brkan, M. Claes and C. Rauchegger, *European Fundamental Rights and Digitalization*, in *Maastricht Journal of European and Comparative Law*, vol. 27, 2020, p. 697-698.

²⁴ N. Smuha, *The Use of Algorithmic Systems by Public Administrations: Practices, Challenges and Governance Frameworks*, in N. Smuha (ed), *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, Cambridge, 2025, p. 404.

²⁵ The AIA is nevertheless not the only instrument that has surfaced recently and that somehow reinforces the protection of individuals in relation to AI. Referring to a multi-layered approach: M. Beltrán, *AI Algorithms under Scrutiny: GDPR, DSA, AI Act and CRA as Pillars for Algorithmic Security and Privacy in the European Union*, in *Computers & Security*, vol. 158, 2025, p. 104628 ff.

²⁶ It partially constitutes data protection law, as some of its elements have as legal basis Art. 16 TFEU, which states that the European Parliament and the Council shall lay down the rules relating to the protection of individuals with regard to the processing of personal data, as well as the rules relating to the free movement of such data.

²⁷ Specifically on the relations between the AIA and the GDPR, see: G. González Fuster, *The AI Act and the GDPR*, in G. Malgieri, G. González Fuster, A. Mantelero and G. Zanfir-Fortuna (eds), *The EU Artificial Intelligence Act: A thematic commentary*, London, 2026 (forthcoming). As the AIA does not directly aim to alter personal data protection in the EU, DPAs are expected to continue to monitor the compliance of the use of AI systems which includes processing of personal data with data protection law (in this sense: I. Barkane and L. Buka, *Prohibited AI Surveillance Practices in the Artificial Intelligence Act: Promises*

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

context of the rights and obligations established by the AIA. When this occurs, EU personal data protection law shall apply. Another significant connection between the AIA and EU's personal data protection is that the AIA aims, inter alia, at protecting against the harmful effects of AI systems while ensuring a high level of protection of health, safety, and fundamental rights as enshrined in the EU Charter, thus including the fundamental right to personal data protection.

The relationship between the AIA and fundamental rights protection in general has been interpreted in multiple ways. While some have stressed that the AIA is “predominantly” a product safety instrument²⁸, others have highlighted that fundamental rights is at the heart of the AIA's risk-based governance framework: in the AIA, fundamental rights, including the right to personal data protection, serve as triggers for regulatory intervention, procedural obligations, and enforcement actions, and the whole framework it establishes is conceived as a dynamic framework, in order to capture in the future new risks to fundamental rights²⁹. Many, however, have been critical about the real added value of the AIA in this regard, noting for instance potential inefficiencies in the AIA's fundamental rights protection mechanisms³⁰, and significant uncertainties regarding the substance of the fundamental rights risk-based approach put forward by the AIA³¹. Despite the Regulation's limitations, some have perceived the AIA as adding up to

and Pitfalls in Protecting Fundamental Rights, in V. Galis, H. Gundhus and A. Vradis (eds), *Critical Perspectives on Predictive Policing*, Cheltenham, 2025, p. 121 ff.). For examples of their work on AI, see for instance: Commission Nationale de l'Informatique et des Libertés (CNIL), *Les Fiches Pratiques IA*, 22 July 2025, available at cnil.fr; Autorité de Protection des Données, *Artificial Intelligence Systems and the GDPR: A Data Protection Perspective*, December 2024, available at autoriteprotectiondonnees.be; Agencia Española de Protección de Datos (AEPD), *Inteligencia Artificial Agéntica Desde la Perspectiva de la Protección de Datos*, February 2026, available at aepd.es. DPAs have also been active in enforcing, even if one of the most publicised DPA decisions, adopted by the Italian DPA (Garante) and fining OpenAI, was annulled by the competent court (see: Reuters, *Italian Court Scraps 15-million-euro Privacy Watchdog Fine on ChatGPT-maker OpenAI*, 19 March 2026, available at reuters.com).

²⁸ M. Almada and N. Petit, *The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights*, in *COLA*, vol. 62, 2025, p. 85 ff.

²⁹ G. Pavlidis, *The EU AI Act and the Rights-Based Approach to Technological Governance*, in *Review of European and Comparative Law*, vol. 14, 2026.

³⁰ F. Paolucci, *Enhancing Oversight and Addressing Gaps: Assessing the Impact of the AI Act on Biometric Identification Systems*, in N. González and G. Mobilio (eds), *Next Democratic Frontiers for Facial Recognition Technology (FRT): The Legal, Ethical and Democratic Implications of FRT*, Cham, 2025, p. 87.

³¹ M. Ho-Dac, *The EU AI Act and the Challenge of Protecting Fundamental Rights*, in *Common Market Law Review*, vol. 65, 2025, p. 1299 ff.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

GDPR-based protection, notably by offering «a new set of legal tools to challenge AI-driven decisions»³².

The adoption of the AIA led in any case to new set of discussions, about whether its provisions de facto complicated compliance with existing EU data protection rules, thus discouraging personal data processing, and indirectly disincentivising the development and use of AI. This viewpoint can be connected to the sense of data protection law as hindering the potential (further) development of AI, or curbing innovation, an idea that the literature has nevertheless decried as a myth³³. For some, it is clear that there is nothing in the GDPR, the AIA, or the combination of both, that would make data controllers in the EU «face competitive disadvantages compared to their counterparts in other parts of the world»³⁴, and the costs associated with some GDPR and AIA requirements, it has been argued, «remain essential and justified when considering the values they seek to protect»³⁵.

Although the AIA shares with the GDPR the ambition of strengthening EU fundamental rights and thus also the right to personal data protection, it never takes the position that such a goal should be pursued by making sure there is less processing of personal data overall. It is in this sense surprising to read in the European Commission's Guidelines on prohibited AI practices under the AIA a statement which appears to suggest that the availability of personal data and «the increased possibilities to process this data with AI systems» generally «increase the risk of harmful manipulative, deceptive or exploitative practices»³⁶ – as if the legislator had wished, with the AIA, to reduce the risk of harm generated by AI by reducing the availability of personal data. That is not the case – on the contrary, the AIA, building on the GDPR, enables further processing of

³² 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*, vol. 16, 2025, p. 770 ff.

³³ P. Dewitte, *AI Meets the GDPR: Navigating the Impact of Data Protection on AI Systems*, in N. Smuha (ed), *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, Cambridge, 2025, p. 157 ff.

³⁴ B. Hohmann and G. Kollár, *Reflections on the Data Protection Compliance of AI Systems under the EU AI Act*, in *Cogent Social Sciences*, vol. 11, 2025, p. 1 ff.

³⁵ *Ibid.*, p. 17.

³⁶ European Commission, *Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act)*, (2025) C(2025) 5052 final 45.

Gloria González Fuster

Caught Between AI and the AI Hype:

How the Right to Personal Data Protection was Ambushed

personal data in view of supporting the uptake of AI while respecting fundamental rights³⁷.

After the adoption of the AIA, it could have seemed that the main remaining challenge was to find the proper way to apply it, together with the GDPR. Policy and legislative discussions took however a new turn.

4. *A Matter of Priorities*

The idea that the AIA and the GDPR were not the solution for a better uptake of AI, but its problem, gained then traction. In November 2025 the European Commission published a Proposal for a Regulation³⁸, known as the Digital Omnibus proposal, that it framed as “only the start” of a larger effort aiming at «stress-testing the whole acquis of existing EU legislation»³⁹. Those were the words it used in its Communication “A simpler and faster Europe: Communication on implementation and simplification”, which incidentally did not include a single reference to fundamental rights. The Communication had nonetheless promised a European Data Union Strategy to «address existing data rules to ensure a simplified, clear and coherent legal framework for businesses and administrations to share data seamlessly» which would do so «respecting high privacy and security standards»⁴⁰.

There is also no mention of AI in that Communication, which evokes only to the need for EU rules to «keep pace with the frontiers of human and technological progress»⁴¹. The text does however refer to the “Draghi report”, which very vehemently embraced the idea that Europe must urgently «redress its failings in innovation and productivity» as the world

³⁷ For instance, by opening up new possibilities to process “special categories of data”, if in order to reduce the risk of discrimination that might result from the bias in AI systems (see Art. 10(5) AIA).

³⁸ Proposal for a Regulation amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024, COM(2025) 837 final.

³⁹ European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Simpler and Faster Europe: Communication on Implementation and Simplification, p. 1.

⁴⁰ European Commission, *A Simpler and Faster Europe*, cit., p. 7.

⁴¹ *Ibid.*, p. 1.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

finds itself «on the cusp» of a «digital revolution, triggered by the spread of artificial intelligence (AI)»⁴². AI is presented in the Draghi report as EU's last chance to shine in digital innovation, «an opportunity to carve out a leading position in selected segments»⁴³, and this despite AI being also, simultaneously, «a source of anxiety for European workers»⁴⁴. Europe already missed out on a digital revolution⁴⁵, and it cannot afford to miss now the «AI revolution»⁴⁶, according to the report. There is one reference in the first part of the Draghi report to fundamental rights, and its tone is significant: it is argued that the «only way» for Europeans to be able to always benefit from fundamental rights is to «grow and become more productive», and the «only way to become more productive is for Europe to radically change»⁴⁷ – openly announcing a discontinuity in the general approach to fundamental rights and innovation.

A choice between «stronger ex ante regulatory safeguards for fundamental rights and product safety, and more regulatory light-handed rules to promote EU investment and innovation» is presented as «unavoidable» in the second part of the Draghi report⁴⁸. If forced to choose between «developments in the field of AI by EU industry actors» and the «commendable» «ambitions of the EU's GDPR and AI Act», Draghi would presumably not hesitate⁴⁹. What matters, the report insists, is that «EU companies are not penalised in the development and adoption of frontier AI»⁵⁰. Such politely described «commendable» ambitions of the GDPR and the AI Act are, basically, the goal of protecting fundamental rights.

For many years, the EU approach appeared to consistently pursue a constant strengthening of fundamental rights, and, simultaneously, economic gains and excellence, displaying what has been described as an axiological congruence around a regulatory strive toward «having the cake

⁴² M. Draghi, *The Future of European Competitiveness - Part A: A Competitiveness Strategy for Europe*, 2024, p. 14.

⁴³ Ibid., p. 24.

⁴⁴ Ibid., p. 25.

⁴⁵ Ibid., p. 5.

⁴⁶ Ibid., p. 6.

⁴⁷ Ibid.

⁴⁸ M. Draghi, *The Future of European Competitiveness - Part B: In-Depth Analysis and Recommendations*, 2024, p. 79.

⁴⁹ Ibid.

⁵⁰ Ibid.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

and eating it too»⁵¹. A change in focus coincided with the publication of the Draghi report. Afterwards, the European Commission looked ready to give up part of the cake, eager to maximise its enjoyment of the AI slice. Axiological congruence with the preceding decades was no longer a priority.

5. *Under Assault*

Not missing out on AI's potential is undoubtedly one of the major objectives behind the proposal for a Digital Omnibus presented by the European Commission in November 2025. The proposed amendments to the GDPR «focus on unlocking opportunities in the use of data, as a fundamental resource in the EU economy, not least in view of supporting the development and use of trustworthy artificial intelligence solutions in the EU market»⁵². The text proclaims that «(t)rustworthy AI is key in providing for economic growth and supporting innovation with socially beneficial outcomes»⁵³.

The European Commission described the proposed amendments as «technical in their nature» and «calibrated to preserve the same standard for protections of fundamental rights»⁵⁴. The commitment is in itself already formally minimal, as what is promised is not to improve data protection in Europe, but merely not «undermining the high level of data protection» under the GDPR⁵⁵ - presumably following «strictly the principle of proportionality enshrined in Article 52 of the Charter»⁵⁶, even though no details are provided as to how has the European Commission assessed the

⁵¹ L. Grozdanovski, *The Ontological Congruency in the EU's Data Protection and Data Processing Legislation: The (Formally) Risk-Based and (Actually) Value/Rights-Oriented Method of Regulation in the AI Act*, in M. Varju and K. Mezei (eds), *The Challenges of Artificial Intelligence for Law in Europe*, Cham, 2025, p. 107 ff.

⁵² European Commission, Proposal for a Regulation of the European Parliament and of the Council Amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as Regards the Simplification of the Digital Legislative Framework, and Repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus), COM(2025) 837 Final, 2. Similarly, the proposed amendments to the AI Act seek to facilitate the smooth and effective application of the rules for safe and trustworthy development and use of AI.

⁵³ European Commission, *Digital Omnibus*, cit., p. 10.

⁵⁴ *Ibid.*, p. 3.

⁵⁵ *Ibid.*, p. 10, 15.

⁵⁶ *Ibid.*, p. 15.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

proportionality of the proposed measures. According to the proposal, the «targeted amendments» to the GDPR «simplify requirements for low-risk processing», although, again, no explanation is given as to how would such an argument be tenable⁵⁷. In any case, the only right the proposal explicitly aims to promote is the right to conduct a business⁵⁸.

The proposed Digital Omnibus in reality interfered with several basic components of the fundamental right to personal data protection:⁵⁹ its scope, as delimited by the definition of personal data; the obligations of data controllers; and the rights of data subjects and more specifically the right of access. In addition, it put forward a downgrading of the protection of special categories of data. An overview of some of these aspects can illustrate the extent of the interference⁶⁰.

5.1. *Notion of Personal Data*

The proposed Digital Omnibus suggested altering the GDPR's definition of personal data by adding in Article 4(1)(a) that «(i)nfomation relating to a natural person is not necessarily personal data for every other person or entity, merely because another entity can identify that natural person»⁶¹ – which, according to the European Commission, was a clarification of the existing definition in light of CJEU case law. The presumed clarification continued with two additional remarks: that information «shall not be personal for a given entity where that entity

⁵⁷ Ibid., p. 15.

⁵⁸ Ibid.

⁵⁹ In the Netherlands, the Meijers Committee decried the impact of fundamental rights of the proposal, urging the legislator to conduct a full fundamental rights impact assessment and in any case amend it (Meijers Committee, *Comment on the Digital Omnibus Proposal*, 2026, available at commissie-meijers.nl). In their Joint Opinion on the proposal, the European Data Protection Supervisor (EDPS) and the European Data Protection Board (EDPB) hint that the text might lower the existing level of fundamental rights protection, even if only in a relatively ambiguous sentence and without specifying if that is their position or not: they note in their general remarks that they support the proposal's objective of reducing administrative burdens «as long as pursuing this objective does not result in lowering the protection of fundamental rights of individuals, in particular the fundamental right to protection of personal data» (European Data Protection Board and European Data Protection Supervisor, *EDPB-EDPS Joint Opinion 1/2026 On the Proposal for a Regulation as Regards the Simplification of the Implementation of Harmonised Rules on Artificial Intelligence (Digital Omnibus on AI)*, 20 January 2026, p. 6).

⁶⁰ Meijers Committee, cit., p. 2.

⁶¹ European Commission, *Digital Omnibus*, cit., p. 54.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

cannot identify the natural person to whom the information relates, taking into account the means reasonably likely to be used by that entity», and that «(s)uch information does not become personal for that entity merely because a potential subsequent recipient has means reasonably likely to be used to identify the natural person to whom the information relates»⁶².

Both statements divert from the current text of the GDPR. Recital 26 GDPR, in particular, notes that «(t)o determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or *by another person*⁶³ to identify the natural person directly or indirectly». In other words, account must be taken of whether the data controller has the means reasonably likely to be used to identify the natural person, but also about whether some means to identify the data subject are likely to be used «by another person», and thus the fact that a «subsequent recipient» has means reasonably likely to be used to identify the natural person matters, in the sense that it could be decisive to determine the qualification of some data as personal data. Thus, with its proposed changes, the European Commission suggested narrowing the existing notion of personal data in the GDPR.

There is no definition of personal data in EU primary law, but only in EU secondary law, which has over the decades served as reference to determine the scope of this notion and, thus, of the fundamental right to personal data protection. Narrowing down the definition of personal data in the GDPR would have as a direct consequence the narrowing down of the definition that in practice delimits the scope of application of a fundamental right.

5.2. *Data Controller Obligations*

The European Commission put forward a rewriting of the principle of purpose limitation, suggesting a new formulation of Article 5 (1)(b) GDPR.⁶⁴ the «further processing» of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be considered to be «compatible» lawful processing operations regardless of Article 6(4) GDPR, and thus independently of the conditions that such provision now foresees – which include important

⁶² Ibid., p. 55. See also Recital 27.

⁶³ Emphasis added.

⁶⁴ European Commission, *Digital Omnibus*, cit., p. 55.

Gloria González Fuster

Caught Between AI and the AI Hype:

How the Right to Personal Data Protection was Ambushed

obligations such as taking into account «the possible consequences of the intended further processing» (Article 6(4)(d) GDPR) and «the existence of appropriate safeguards» (Article 6(4)(e) GDPR).

This means that the «further processing» of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes would not only benefit from a blanket exemption to the purpose limitation principle, but actually indirectly also from the principle of lawfulness, because as «further processing» it would not need to be covered by one of the lawful grounds for processing of Article 6 GDPR.

Another change proposed by the European Commission is for the GDPR to include an explicit reference to the fact that the processing of personal data in the context of the development and use of AI systems may be carried out for purposes of a legitimate interest within the meaning of Article 6 GDPR⁶⁵, «where appropriate»⁶⁶. The most surprising suggestion in this context is the idea, proposed by the European Commission, that when controllers balance their legitimate interest (or the legitimate interest of a third party) with the rights and freedoms of the data subject, they should give consideration «to whether the interest pursued by the controller is beneficial for the data subject and society at large»⁶⁷. In principle, the interest to be balanced must certainly be legitimate, but the impact of this interest on “society at large” is not supposed to be a relevant criterion. Most strangely, the suggestion hints that the controller should be in some cases balancing the rights and freedoms of data subjects with other benefits for data subjects, which is a balancing that would probably best be carried out by the data subjects themselves.

⁶⁵ A key question that has emerged over the years is the legality of the processing of data, and in particular personal data, for AI training. This triggers issues regarding compliance with the GDPR obligation to base all personal data processing on a lawful ground, in addition to other legal issues. See notably, on this matter: P. Hacker, *A Legal Framework for AI Training Data: From First Principles to the Artificial Intelligence Act*, in *Law, Innovation and Technology*, vol. 13, 2021, p. 257 ff.

⁶⁶ European Commission, *Digital Omnibus*, cit., p. 10.

⁶⁷ It is added: «which may for instance be the case where the processing of personal data is necessary for detecting and removing bias, thereby protecting data subjects from discrimination, or where the processing of personal data is aiming at ensuring accurate and safe outputs for a beneficial use, such as to improve accessibility to certain services» (European Commission, *Digital Omnibus*, cit., p. 10).

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

Other basic data processing principles to be undermined, or at least modulated, by the proposal include the principle of transparency, with proposed changes to information obligations of data controllers⁶⁸.

5.3. *Rights of Data Subjects*

The rights and remedies available to data subjects under the GDPR are best visualised as a sequential path. Whenever personal data about someone is processed, the data subject is entitled to receive information about the processing, about their rights (such as the right of access), and about the possibility to lodge a complaint with a DPA⁶⁹. If data subjects exercise then their right of access, for instance, they are then entitled to receive further information, including again information about their right to lodge a complaint with a DPA. If they would then lodge a complaint with a DPA, they would then receive information about their right to effective judicial remedy in relation to the DPA decision. All these different steps could eventually take an unaware and uninformed data subject from absolute ignorance about their rights to the effective exercise of their fundamental right to an effective judicial remedy (Article 47 CFREU) in relation to their fundamental right to personal data protection (Article 8 CFREU), building on each other. Thus, by limiting or allowing to skip any of these steps, the whole sequential path is affected and access to an effective judicial remedy is potentially endangered.

The proposal of the European Commission put forward several significant new obstacles along this path. In addition of extending the possible exceptions from information obligations applicable when there is processing of personal data (under Article 13 GDPR)⁷⁰, it suggested including a reference for the possibility to regard as «manifestly unfounded or excessive» (and thus not requiring to be fulfilled) data access requests from data subjects where «the data subject abuses the rights conferred by this regulation for purposes other than the protection of their data»⁷¹. The

⁶⁸ European Commission, *Digital Omnibus*, cit., p. 56.

⁶⁹ See Arts. 13 and 14 GDPR.

⁷⁰ The proposal suggests the obligation should not apply where there are reasonable grounds to assume that the data subject already has the information, unless the controller transmits the data to other recipients or categories of recipients, transfers the data to a third country, carries out automated decision-making or the processing is likely to cause a high risk to data subject's rights (European Commission, *Digital Omnibus*, cit., p. 20).

⁷¹ European Commission, *Digital Omnibus*, cit., p. 56.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

suggestion is extremely ambiguous, as it is unclear what would be such purposes, and is in tension with the notion that limitations to the exercise of data subject rights should be as limited as possible.

These proposed obstacles would add up to other limitations in the path towards effective judicial remedy that were already integrated into the GDPR by virtue of the adoption in November 2025 of Regulation (EU) 2025/2518 laying down additional procedural rules on the enforcement of the GDPR, known as the “GDPR Procedural Regulation”⁷². This Regulation, which shall apply from April 2027, foresees a special procedure for the dealing of complaints regarding “cross-border processing”⁷³ that specifically concern the exercise of data subject rights, allowing DPAs to divert from the standard procedure. The new «early resolution» procedure has an impact on the access to remedies when a data subject is confronted with an infringement carried out by a data controller benefiting from the “one stop shop” mechanism, which is typically the case with “Big Tech” companies and thus is to affect a significant number of data subjects in Europe⁷⁴.

5.4. *Special Categories of Data*

The processing of so-called “special categories of data” is in principle forbidden by the GDPR⁷⁵, although it is nevertheless possible under a series of exemptions⁷⁶. The European Commission has proposed adding a new exemption for allowing what it calls «residual processing of special categories of personal data for development and operation of an AI system or an AI model»⁷⁷. This would refer to the use of sensitive data «not necessary for the purpose of the processing»⁷⁸ – thus, data that in principle should not be processed at all, because, in line with the purpose limitation principle, combined with the data minimisation principle, one should only

⁷² Regulation (EU) 2025/2518 of the European Parliament and of the Council of 26 November 2025 laying down additional procedural rules on the enforcement of Regulation (EU) 2016/679, OJ L, 2025/2518, 12.12.2025.

⁷³ As defined in Art. 4(23) GDPR.

⁷⁴ On this subject, see: G. González Fuster, *Proximity, Amicable Settlements, and how the EU Guts GDPR Enforcement*, in *Verfassungsblog*, available at verfassungsblog.de 19 July 2024.

⁷⁵ Art. 9(1) GDPR.

⁷⁶ Art. 9(2) GDPR.

⁷⁷ European Commission, *Digital Omnibus*, cit., p. 20.

⁷⁸ *Ibid.*, p. 10.

Gloria González Fuster

*Caught Between AI and the AI Hype:**How the Right to Personal Data Protection was Ambushed*

process data that are necessary for the purpose of the processing. The proposed exemption is thus specifically foreseen for *unnecessary* processing of sensitive data⁷⁹. It is advanced that data controllers would nevertheless be able to incur in such unnecessary processing under certain conditions⁸⁰. One of such conditions is to «remove» the data once identified; nevertheless, if removal «would require disproportionate effort», the controller might as well keep the data as long as it makes sure they are not «used to infer outputs», «disclosed or otherwise made available to third parties»⁸¹. In other words, sensitive data, which in principle should never be processed, would become available for processing, including when such processing would be unnecessary, and then these data could be stored indefinitely as long as they are not shared – all this, of course, only if and because that data happen to be related to «the development and operation of an AI system or an AI model», which is thus surfacing as a primordial goal not to hinder at any cost.

6. *Concluding Remarks*

Personal data protection law has been described in the past as acting both as «as a catalyst and a hindrance» in relation to AI⁸². It can be a catalyst to the extent that it adequately frames the processing of personal data allowing for, and resulting from, the development and use of AI. In doing so, data protection law can contribute to sustainable data processing practices upon which AI can be further build. Data protection law might also be perceived as a hindrance, however, because it does impose some obligations which do include, in some cases, certain limitations to data processing.

⁷⁹ «The derogation should only apply where the controller has implemented appropriate technical and organisational measures in an effective manner to avoid the processing of those data, takes the appropriate measures during the entire lifecycle of an AI system or AI model and, once it identifies such data, effectively remove them» (European Commission, *Digital Omnibus*, cit., p. 11).

⁸⁰ See also the additional clarification: «This derogation should not apply where the processing of special categories of personal data is necessary for the purpose of the processing. In this case, the controller should rely» (European Commission, *Digital Omnibus*, Recital 11.).

⁸¹ European Commission, *Digital Omnibus*, cit., p. 11.

⁸² O. Pollicino, *Getting the Future Right – Artificial Intelligence and Fundamental Rights: A View from the European Union Agency for Fundamental Rights*, in *BioLaw Journal – Rivista di BioDiritto*, vol. 7, 2021, p. 11.

Gloria González Fuster

Caught Between AI and the AI Hype:

How the Right to Personal Data Protection was Ambushed

AI, as such, is not supposed to bring along the end of data protection. It was actually, rather, its beginning: historically, the advent of opaque automated data processing was the point of departure of the legal thinking that put data protection law on the map. Protecting individuals in front the power of the machine was, and is, one of its *raison d'être*. For numerous decades, data protection seemed to be the key tool to protect individuals in the face of AI, and thus in a way a necessary condition for embracing AI in democratic societies - the first line of defence for democracy in light of technological development. Sacrificing the right to data protection in the name of AI – but also slowly dismantling it in the name of some AI promises – is thus, probably, an extremely problematic policy agenda.

ABSTRACT: The right to personal data protection was meant to be the most important fundamental right in the era of digitalisation. Recognised at the highest level by European Union (EU) law, it could have been embraced by the EU legislator as the most needed right in times of unprecedented spread of Artificial Intelligence (AI). For many years, strengthening fundamental rights protection – and, thus, also, if not primarily, the protection of personal data – appeared indeed as a *conditio sine qua non* for EU's open support of AI. A new wind started to blow in 2025, however, as the EU appeared ready to further accelerate its push for AI, now even at the risk of trampling on personal data protection. This contribution explains how this occurred and how it is the EU's latest frenzy and AI infatuation, and not necessarily AI, that runs the risk of destroying a much-needed fundamental right.

KEYWORDS: data protection – AI – European Union – personal data – simplification.

Gloria González Fuster – Research Professor, Vrije Universiteit Brussel (VUB) and Director of the Law, Science, Technology and Society (LSTS) Research Group, Brussels, Belgium
(Gloria.Gonzalez.Fuster@vub.be)

Neurorights in the Age of AI: Universalism, Cognitive Vulnerability, and the Limits of Legal Translation*

Aimen Taimur

TABLE OF CONTENTS: 1. Introduction. – 2. Universalising Neurorights under International Standard-Setting. – 3. Cognitive Vulnerability in AI Environments: A Legal Account. – 4. Comparative Legal Developments on Neurorights. – 5. Implementation and Convergence in Practice. – 6. Universal Standards and Local Realities: An Evaluative Framework. – 7. Conclusion.

1. Introduction

In November 2025, UNESCO's 43rd General Conference adopted the Recommendation on the Ethics of Neurotechnology¹ - the first global normative instrument in this field. Its elaboration had begun in 2023, following a mandate from UNESCO's 194 Member States, and proceeded through an Ad Hoc Expert Group, extensive regional consultations, and an intergovernmental meeting of experts held in May 2025 that approved the final text. The Recommendation seeks to articulate shared principles for the governance of technologies that access, read, or alter mental states.² The initiative marks the first adopted global standard-setting instrument establishing shared norms on what have come to be known as "neurorights," including rights to mental privacy, cognitive liberty, and psychological integrity³. Its ambition is unmistakably universalist, i.e. to establish standards applicable across jurisdictions regardless of local legal culture. As a Recommendation adopted under UNESCO's standard-setting

* This article was subjected to double-blind peer review.

¹ UNESCO, *Draft Recommendation on the Ethics of Neurotechnology*, CL/4499, Annex II, 2025, preamble and Arts. 5–11, distinguishing «neurodata» from «non-neurodata allowing cognitive-state inferences».

² C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology: realizing the rights of people with mental disabilities*, in *Nature Mental Health*, vol. 3, 2025, p. 749 ff.

³ M. Ienca and R. Andorno, *Towards new human rights in the age of neuroscience and neurotechnology*, in *Life Sciences, Society and Policy*, vol. 13, 2017, p. 5.

mandate, the instrument does not generate legally binding obligations; it operates as soft law, whose authority derives from the deliberative process through which it was elaborated and from the normative weight that states and domestic institutions choose to give it. Yet the pursuit of universality raises questions about how such standards interact with the contextual nature of cognitive vulnerability, which varies with law, social structure, and also individual condition. The debate, therefore, reflects a long-standing tension in human-rights theory between the universality of normative claims and the contextual relativism of their implementation⁴.

The Recommendation builds upon UNESCO's prior instruments, notably the Recommendation on the Ethics of Artificial Intelligence (2021)⁵ and the Universal Declaration on Bioethics and Human Rights (2005)⁶, extending their normative reach into the neurotechnological domain. It reaffirms freedom of thought, mental privacy, and autonomy as core principles while introducing distinctions between neural and non-neural data that allow "cognitive-state inferences." By explicitly addressing the ethical implications of data that reveal or modulate mental processes, the draft Recommendation situates neurorights within a continuum of existing human-rights protections rather than as an entirely new catalogue. It further reflects the convergence between neurotechnology, artificial intelligence, and biometrics, calling for governance mechanisms that integrate scientific evidence with human-rights-based safeguards. In this sense, UNESCO's initiative positions itself as a global standard-setting framework designed to encourage national implementation and harmonise regional efforts around a shared ethical vocabulary.

The term neurorights refers to a proposed category of human rights designed to protect mental integrity, cognitive liberty, and the privacy of brain-derived information in the face of advancing neurotechnology and AI. The concept was systematically developed by Ienca and Andorno, who identified four core interests requiring legal protection: mental privacy, understood as protection against non-consensual access to or inference of neural data; cognitive liberty, understood as freedom from non-consensual

⁴ J. Donnelly, *Universal Human Rights in Theory and Practice*, 3rd ed., Ithaca, 2013.

⁵ UNESCO, *Final Report on the Draft Text of the Recommendation on the Ethics of Neurotechnology*, CL/4499, Annex I (31 March 2025), paras. 1–3, highlighting continuity with the *Recommendation on the Ethics of Artificial Intelligence*, 2021.

⁶ UNESCO, *Universal Declaration on Bioethics and Human Rights* (adopted 19 October 2005), UNESCO General Conference, 33rd session, Paris, Arts. 3–5.

alteration of mental states; mental integrity, understood as protection against harmful manipulation of brain processes; and psychological continuity⁷, understood as protection of the authenticity and temporal coherence of personal identity. The term is not uncontested: some scholars argue that existing rights, notably freedom of thought, privacy, and bodily integrity, are sufficient if properly applied, making a separate neurorights vocabulary redundant⁸. Others contend that the specificity of neural data and the novelty of inference and modulation technologies create doctrinal gaps that existing frameworks cannot close without explicit elaboration⁹. This article proceeds from the view that the debate is not about whether to invent new rights but about how to specify, with legal precision, which modalities of interference with mental processes fall within already-protected interests and which require supplementary elaboration.

Efforts to universalise neurorights emerge against a fast-changing technological landscape. Artificial intelligence systems that analyse behaviour, emotions, and neural activity push the limits of privacy and integrity as traditionally understood¹⁰. Inference and influence now reach the internal forum of the mind (*forum internum*), once presumed beyond observation¹¹. This capability transforms how autonomy is experienced and how harm is perceived. When cognitive intrusion can occur without awareness, the adequacy of existing fundamental rights becomes uncertain¹². The question, therefore, is not whether human rights apply to the mind, but whether their current legal articulation captures the forms of vulnerability produced by AI and neurotechnology. What distinguishes AI from earlier neurotechnology is scale and opacity. These systems operate across millions of users simultaneously, and the mechanisms by which they

⁷ M. Ienca and R. Andorno, *Towards new human rights in the age of neuroscience and neurotechnology*, cit.

⁸ S. Alegre, *We Don't Need New "Neurorights", We Need to Apply the Law*, CIGI Commentary, 7 June 2023; UN Human Rights Committee, *General Comment No. 22* (1993), cit.

⁹ P. Magee, M. Ienca and N. Farahany, *Beyond Neurodata: Cognitive Biometrics and Mental Privacy*, in *Neuron*, vol. 112, 2024, p. 3017 ff.

¹⁰ D. Susser, B. Roessler and H. Nissenbaum, *Online Manipulation: Hidden Influences in a Digital World*, in *Georgetown Law Technology Review*, vol. 4, 2019, p. 1 ff.

¹¹ UN Human Rights Committee, *General Comment No. 22: Article 18 (Freedom of Thought, Conscience and Religion)*, CCPR/C/21/Rev.1/Add.4, 1993.

¹² N. Farahany, *The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology*, New York, 2023.

infer or influence mental states are rarely visible to the individuals affected or to the regulators responsible for oversight.

The challenge of translation from ethics to law is already visible in national practice. Chile remains the only country to constitutionalise neurorights, embedding protection of mental integrity and brain data within its Bill of Rights¹³. The constitutional amendment has since been operationalised through Law No. 21.545 of 2023, which establishes data governance duties for entities processing neural information, including requirements of purpose limitation, explicit consent, and data minimisation - translating the constitutional guarantee into administrable obligations for private actors. Its Supreme Court's decision in *Girardi Lavín v. Emotiv Inc.* (2023) applied the constitutionalised right to mental integrity under the amended Article 19(1) directly against a commercial entity, holding that the retention of brain activity data without specific, revocable consent violated both mental privacy and psychological integrity, and ordering deletion of the unlawfully held records¹⁴. This jurisprudence demonstrates how principles of mental autonomy can be given concrete justiciability within positive law. In contrast, the European Union's Artificial Intelligence Act classifies manipulative AI as an "unacceptable risk" but confines protection to certain contexts, leaving interpretive and evidentiary gaps around the exploitation of cognitive weakness¹⁵. These examples illustrate the divergence between jurisdictions willing to reconceptualise rights at the constitutional level and those seeking incremental adaptation within existing data-protection frameworks.

The international process led by UNESCO, therefore, operates within a tension. Universal rules promise coherence and a shared ethical vocabulary, yet the lived experience of cognitive vulnerability is culturally and structurally specific. The impact of neurotechnology depends on factors such as digital literacy, disability, social power, and exposure to data-driven

¹³ Law No. 21.383 of 14 October 2021 (Chile) amending Art. 19(1) of the Political Constitution; Law No. 21.545 of 5 January 2023 (Chile) on the promotion of mental health and neurorights.

¹⁴ Supreme Court of Chile, *Girardi Lavín v. Emotiv Inc.*, Rol 36.904-2023, judgment of 9 August 2023; see also M. I. Cornejo-Plaza, R. Cippitani and V. Pasquino, *Chilean Supreme Court ruling on the protection of brain activity: neurorights, personal data protection, and neurodata*, in *Frontiers in Psychology*, vol. 15, 2024, p. 1330439.

¹⁵ Art. 5, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (*Artificial Intelligence Act*, "AI Act"), OJ L, 12 July 2024.

manipulation¹⁶¹⁷. A rigidly universal template may fail to recognise these differences, while excessive relativism risks fragmenting protection and enabling regulatory arbitrage. The core task for law is to navigate this space by defining minimum global commitments without erasing the diversity of contexts in which cognitive harm occurs.

This article approaches that task through a comparative legal analysis of neurorights and their relation to cognitive vulnerability. It examines how constitutional, statutory, and regulatory regimes in different jurisdictions conceptualise mental autonomy, and how these frameworks interact with emerging international instruments. By analysing legislative and judicial developments in Chile, the European Union, and the United States, the article tests whether a coherent universal standard for neurorights can coexist with the pluralism inherent in national legal orders. These jurisdictions were selected because they represent distinct normative models for addressing cognitive risk, with Chile exemplifying constitutional recognition, the European Union illustrating a rights-based regulatory framework, and the United States reflecting an administrative and consumer-focused approach that together demonstrate the range of legal pathways through which neurorights may evolve. The argument advanced is that universalism remains normatively indispensable but must be contextualised - anchored in shared principles of mental integrity and autonomy while responsive to the diverse ways in which AI and neurotechnology generate vulnerability. The viability of neurorights as fundamental rights will depend on this balance between the universal aspiration of the UNESCO Recommendation and contextual precision.

The analysis proceeds across three distinct but interrelated levels. The first is international standard-setting, examined through the UNESCO Recommendation as a soft law instrument whose authority rests on consensus and persuasive weight rather than binding obligation¹⁸. The second is comparative statutory and regulatory law, examining enacted frameworks in the European Union, Chile, and the United States, with EU

¹⁶ N. Helberger, M. Sax, J. Strycharz and H.-W. Micklitz, *Choice Architectures in the Digital Economy: Towards a New Understanding of Digital Vulnerability*, in *Journal of Consumer Policy*, vol. 45, 2022, p. 175 ff.

¹⁷ M. A. Fineman, *The Vulnerable Subject and the Responsive State*, in *Emory Law Journal*, vol. 60, 2010, p. 251 ff.

¹⁸ UNESCO, *Recommendation on the Ethics of Neurotechnology*, adopted by the UNESCO General Conference at its 43rd session, Samarkand, 12 November 2025 (CL/4499), cit.

law as the primary analytical standpoint against which the Chilean constitutional model and the US consumer-administrative approach are examined as contrasting normative templates¹⁹. The third is comparative constitutional law, concerned with how courts and constitutional framers have translated protections of mental autonomy into justiciable rights and remedies. These levels are kept analytically distinct throughout, though the article acknowledges that they interact in practice. Where positive law and jurisprudence are treated together rather than in sequence, this reflects a deliberate thematic choice: in this field, the legal meaning of statutory protections has in several instances been substantially constituted by judicial interpretation, and separating the two levels would obscure rather than clarify the analytical picture²⁰.

2. *Universalising Neurorights under International Standard-Setting*

The UNESCO Recommendation on the Ethics of Neurotechnology marks a significant step in the international framing of neurorights, consolidating them as shared normative reference points for domestic governance.²¹ As a soft law instrument, the Recommendation does not impose binding international obligations on member states; its authority rests on the consensus through which it was adopted and on the persuasive weight it carries in national legislative and judicial processes. Rather than prescribing enforceable duties, it sets out a normative vocabulary and a set of guiding principles that states are invited to translate into their own constitutional, statutory, and regulatory frameworks. A universal template may achieve coherence only if it accommodates the distinct risks of neurotechnology and the institutional realities of the states that must implement it²². That challenge is sharpened by AI. The capabilities that neurorights are designed to govern, including large-scale inference of mental states and covert influence over behaviour, are now primarily

¹⁹ J. Donnelly, *Universal Human Rights in Theory and Practice*, cit.

²⁰ M. I. Cornejo-Plaza, R. Cippitani and V. Pasquino, *Chilean Supreme Court ruling on the protection of brain activity*, cit.

²¹ UNESCO, *Final Report on the Draft Text of the Recommendation on the Ethics of Neurotechnology*, CL/4499, Annex I (31 March 2025), paras 1–3, cit.

²² C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, cit.

delivered through AI systems, which means that any viable universal standard must be legible to AI regulators, not only to human rights lawyers.

The first question concerns the legal object of protection. Existing proposals often group together several distinct interests under the term neurorights, including mental privacy, personal identity, cognitive liberty, and equal access to beneficial applications²³. These interests are not interchangeable. Mental privacy is concerned with access, inference, and retention of brain-derived signals and proxies²⁴. Identity turns on the authenticity and continuity of the self²⁵. Cognitive liberty addresses freedom from non-consensual alteration and the conditions for self-forming thought²⁶. Equality is a matter of fair distribution in enhancement and assistive uses. A universal standard that conflates these distinct interests risks producing principles that are normatively appealing but practically ineffective for adjudication or compliance. Disaggregation is the more productive strategy, because distinct interests suggest distinct duties and remedies.

The second question is whether new rights are required or whether existing rights can be operationalised for neurotechnology. One camp argues for explicit recognition of neurorights to close conceptual and remedial gaps²⁷. Another camp reads the *forum internum* and related guarantees as sufficient, provided that the law specifies how they apply to inference, manipulation, and brain-signal interfaces²⁸. The latter position draws strength from the absolute protection of freedom of thought in the *forum internum* and from the doctrinal intuition that the mind is not a regulatory vacuum²⁹. Yet without concrete limits on access and influence, the *forum internum* remains difficult to enforce in data-intensive environments. The universalisation debate is therefore less about inventing

²³ M. Ienca and R. Andorno, *Towards new human right*, cit., p. 5.

²⁴ P. Magee, M. Ienca and N. Farahany, *Beyond Neurodata*, cit.

²⁵ S. Ligthart, *Towards a Human Right to Psychological Continuity? Reflections on the Rights to Personal Identity, Self-Determination, and Personal Integrity*, in *European Convention on Human Rights Law Review*, vol. 5, 2024, p. 199 ff.

²⁶ P. Sommaggio, M. Mazzocca, A. Gerola and F. Ferro, *Cognitive Liberty. A First Step Towards a Human Neuro-Rights Declaration*, in *BioLaw Journal – Rivista di BioDiritto*, vol. 1, 2017, p. 91 ff.

²⁷ M. Ienca and R. Andorno, *Towards new human right*, cit.

²⁸ S. Alegre, *We Don't Need New "Neurorights", We Need to Apply the Law*, cit.

²⁹ UN Human Rights Committee, *General Comment No. 22*, 1993, cit.

a new catalogue and more about stating, with precision, which modalities of access and interference violate protected mental interests³⁰.

The *forum internum* provides a coherent normative foundation for defining the boundaries of cognitive autonomy.³¹ The key challenge is to translate this abstract protection into concrete and justiciable elements that can guide the regulation of AI-mediated practices. Three categories of interference are especially significant. The first concerns the non-consensual decoding of neural signals that individuals intend to keep private, which directly violates mental privacy and autonomy. The second involves covert or opaque mechanisms of influence that exploit cognitive vulnerabilities to steer behaviour, particularly when such influence occurs below the threshold of awareness. The third relates to the commercial circulation of raw neurodata outside legitimate health or research contexts, where consent is neither explicit, specific, nor revocable. Together, these categories delineate the outer limits of permissible engagement with mental processes and establish points at which law can intervene - through design-stage duties to prevent violations and through remedies such as deletion, disgorgement, or damages when interference has already occurred.³²

A clear distinction must be drawn between foundational protections and aspirational goals. Some neurorights proposals extend beyond safeguarding cognitive integrity to include broader distributive aims, such as a general right to cognitive enhancement³³. While these objectives may have policy value, they risk transforming legal guarantees into open-ended social commitments and provoking disputes over resource allocation or technological equity. A universal framework will gain greater legitimacy if it differentiates between absolute prohibitions that preserve the conditions for thought and progressive measures that states may realise over time, according to capacity. Maintaining this separation reflects established human-rights architecture and prevents the inflation of international obligations beyond what is legally or institutionally achievable³⁴.

³⁰ D. Susser, B. Roessler and H. Nissenbaum, *Online Manipulation*, cit.

³¹ UN Human Rights Committee, *General Comment No. 22* (1993), cit.

³² Art. 5 AI Act; C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, cit.

³³ N. Bostrom and R. Roache, *Smart Policy: Cognitive Enhancement and the Public Interest*, in J. Savulescu, R. ter Meulen and G. Kahane (eds), *Enhancing Human Capabilities*, Oxford, 2009, p. 138 ff.

³⁴ C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, cit.

Furthermore, justiciability and proof are central. Soft law gains traction only when it helps domestic institutions decide hard cases. The European Union's Artificial Intelligence Act prohibits, under Article 5(1)(b), AI systems that exploit group-specific vulnerabilities to materially distort behaviour in ways likely to cause harm, yet it leaves evidentiary burdens and allocation of responsibility to implementing authorities and sectoral regimes³⁵. Chile took an inverse path, elevating mental integrity and brain-derived information to constitutional status. In *Girardi Lavín v. Emotiv Inc.*, the Supreme Court ordered the deletion of unlawfully retained neurodata, thereby making mental autonomy justiciable with immediate consequences for private actors³⁶. A universal template should be intelligible to both styles of governance. It should articulate prohibited modalities that constitutional framers can adopt as rights and that regulators can encode as red-line prohibitions, and it should indicate how proof of covert influence or non-consensual decoding can be established through testing and audit³⁷.

Cognitive vulnerability emerges as a key axis along which the balance between universalism and relativism must be maintained. Vulnerability is not a static attribute of fixed classes. It is situational and fluid, shaped by design choices, information asymmetries, literacy, age, health, disability, socio-economic position, and exposure to persuasive architectures³⁸. This has two implications for duty design³⁹. First, states should impose upstream obligations on developers and deployers to identify and mitigate cognitive risk, for example, through impact assessments that examine attentional capture, affective inference, and susceptibility to manipulation⁴⁰. Secondly, the universal template should set minimum guarantees for protected interests and prohibited modalities, while leaving thresholds and evidentiary

³⁵ Art. 5 AI Act.

³⁶ Law No. 21.383 of 14 October 2021 (Chile), cit.; Supreme Court of Chile, *Girardi Lavín v. Emotiv Inc.*, Rol 36.904-2023, judgment of 9 August 2023, cit.

³⁷ C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, cit.; Art. 5 AI Act.

³⁸ N. Helberger, M. Sax, J. Strycharz and H.-W. Micklitz, *Choice Architectures in the Digital Economy*, cit., p. 175 ff.; M. A. Fineman, *The Vulnerable Subject and the Responsive State*, cit.

³⁹ A. Atak, *Ethical Design and Responsibilities*, in *World Journal of Engineering Research and Technology*, vol. 9, 2023, p. 75 ff.

⁴⁰ A. Taimur, *Cognitive Freedom and Legal Accountability: Rethinking the EU AI Act's Theoretical Approach to Manipulative AI as Unacceptable Risk*, in *Cambridge Forum on AI: Law and Governance*, vol. 1, 2025, e20.

proxies to sector-specific rules that reflect local infrastructures and social conditions. This preserves the core and avoids dilution into relativism⁴¹.

The governance of neurotechnology can be structured around the tripartite model of obligations recognised in human rights law, which distinguishes between duties to respect, protect, and fulfil⁴². In this context, the “duty to respect” requires public authorities to avoid direct interference with cognitive autonomy, such as decoding neural signals without consent, manipulating mental states through covert influence, or trading in neurodata for commercial purposes. The “duty to protect” extends this responsibility to the private sphere by obliging states to establish effective regulatory and supervisory mechanisms. This responsibility encompasses licensing, safety, and data-governance frameworks that embed explicit safeguards for mental integrity and freedom of thought. Finally, the “duty to fulfil” encompasses positive measures designed to build institutional and societal capacity, for example, by supporting the development of privacy-preserving neurotechnologies for therapeutic applications and ensuring that oversight bodies possess the technical expertise to evaluate cognitive risk. Read together, these layers of obligation can create a continuous system of accountability that connects design, regulation, and remedy, clarifying who must act, through what instruments, and at what stage of neurotechnological innovation⁴³.

However, it can be observed through the literature on the area that a persistent objection is redundancy⁴⁴. If privacy, bodily integrity, non-discrimination, and freedom of thought already exist, a separate neurorights vocabulary risks duplication. Addressing this objection requires both doctrinal refinement and empirical extension of existing principles. Doctrinally, the *forum internum* is absolute yet under-specified for environments in which inference and influence are invisible to the person affected. Empirically, practices such as affective inference, persuasive architecture, and consumer brain-interfaces generate harms that current law often frames as data-processing incidents or unfair commercial practices

⁴¹ UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021, cit.; Art. 9 AI Act and related provisions.

⁴² D. J. Karp, *What Is the Responsibility to Respect Human Rights? Reconsidering the “Respect, Protect, and Fulfil” Framework*, in *International Theory*, vol. 11, 2019, p. 83 ff.

⁴³ Ibid.

⁴⁴ S. Alegre, *We Don’t Need New “Neurorights”, We Need to Apply the Law*, cit.

rather than as interferences with mental autonomy⁴⁵. A universal instrument that names protected mental interests and prohibited modalities can steer interpretation toward the appropriate frame without requiring a complete reinvention of rights⁴⁶.

Additionally, enforcement architecture should be addressed rather than deferred. Ex post remedies alone are poorly matched to harms that materialise before awareness and are difficult to prove. Therefore, upstream controls matter extensively in such circumstances. A credible template would encourage pre-market conformity assessment for neurotechnology with cognitive impact, mandatory disclosure where systems process or infer mental states, and a presumption against the commercialisation of raw neurodata absent explicit, specific, and revocable consent. These are design-level commitments that can be supervised by regulators and audited by third parties. They align with risk-based AI governance and recognise the special status of neurodata⁴⁷.

The relationship between universalism and relativism need not be understood as oppositional. Universality can establish the essential conditions for thought and self-formation through a set of minimum guarantees, while implementation can adapt metrics, thresholds, and institutional mechanisms to different legal and cultural contexts. Relativism, therefore, belongs to the level of application rather than to the level of principle⁴⁸. A framework that makes this relationship explicit is far more likely to inform constitutional drafting and regulatory design than one that leaves it implicit. From this perspective, a legally realistic approach to international standard-setting would proceed by distinguishing discrete protected interests instead of asserting a single, all-encompassing right. It would define the modalities of interference that violate the *forum internum* in AI-mediated environments, assign corresponding duties to public and private actors, and require design-stage safeguards that anticipate cognitive risks. It would also integrate documentation and audit obligations that render covert influence visible to courts and regulators. Such an approach

⁴⁵ UN Human Rights Committee, *General Comment No. 22*, 1993, cit.; D. Susser, B. Roessler and H. Nissenbaum, *Online Manipulation*, cit.

⁴⁶ C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, cit.

⁴⁷ Art. 5 AI Act and conformity assessment framework; UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021, cit.; UNESCO, *Draft Recommendation on the Ethics of Neurotechnology*, CL/4499, Annex II, 2025, Arts. 6–9, cit.

⁴⁸ J. Donnelly, *Universal Human Rights in Theory and Practice*, cit.

preserves core guarantees while allowing contextual differentiation in enforcement and remedy, giving neurorights a universal foundation that is both normatively robust and practically workable⁴⁹.

3. *Cognitive Vulnerability in AI Environments: A Legal Account*

Cognitive vulnerability constitutes a legally relevant form of exposure that arises when technological systems infer or steer mental states under conditions of information asymmetry⁵⁰. Such vulnerability occurs when design architectures manipulate attention or decision-making below the level of conscious awareness. It arises when neural and behavioural data are transformed into actionable representations of mental activity, both through direct brain signals and through multimodal traces that function as cognitive biometrics⁵¹. These data streams enable the reconstruction of attitudes and intentions from non-neural indicators such as gaze, gesture, or tone⁵². Unlike traditional notions of consumer vulnerability tied to fixed groups, cognitive vulnerability is fluid and situational⁵³. It fluctuates with interface design, digital literacy, age, disability, health, social power, and the degree of control exercised over technological environments.

This understanding of vulnerability is increasingly reflected in emerging regulatory frameworks. The EU AI Act recognises vulnerability as a trigger for heightened protection, classifying AI systems that exploit the weaknesses of persons due to their age, disability, or socio-economic situation as unacceptable risk practices⁵⁴. Similarly, the Digital Services Act addresses manipulative design and personalised targeting that may exploit users' psychological or cognitive predispositions, linking such practices to

⁴⁹ C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, cit.; N. Helberger, M. Sax, J. Strycharz and H.-W. Micklitz, *Choice Architectures in the Digital Economy*, cit.

⁵⁰ S. A. Teo, *Artificial Intelligence, Human Vulnerability and Multi-Level Resilience*, in *Computer Law & Security Review*, vol. 57, July 2025.

⁵¹ P. Magee, M. Ienca and N. Farahany, *Beyond Neurodata*, cit.

⁵² S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, U. R. Acharya, *Emotion Recognition and Artificial Intelligence: A Systematic Review (2014–2023) and Research Recommendations*, in *Information Fusion*, vol. 102, February 2024.

⁵³ A. Zac et al., *Dark Patterns and Consumer Vulnerability*, in *Behavioural Public Policy*, 2025, p. 1 ff.

⁵⁴ Arts. 5(1)(a)–(b) AI Act.

the concept of vulnerable users in the digital environment⁵⁵. Article 5(1)(b) of the EU AI Act prohibits AI systems that exploit the vulnerabilities of specific groups, defined by reference to age, disability, or socio-economic situation, where those systems materially distort behaviour in a manner likely to cause harm: both the materiality of the distortion and the probability of harm are operative thresholds, not merely descriptive qualifiers.⁵⁶ The Artificial Intelligence Act further classifies systems affecting the safety or fundamental rights of natural persons, including those interacting with vulnerable groups, as high-risk under Annex III, subjecting them to conformity assessment, risk management, and documentation duties under Articles 9 to 15. The Digital Services Act, by contrast, addresses manipulative design and targeted advertising within a narrower institutional frame: Article 25 prohibits online platforms from designing their interfaces in ways that deceive, manipulate, or otherwise impair users' ability to make free and informed decisions, while Article 26(3) prohibits profiling-based targeted advertising directed at minors and advertising based on sensitive categories of personal data. These are obligations imposed on providers of online platforms and, in certain respects, very large online platforms specifically, and they do not amount to a general EU prohibition on cognitive manipulation across sectors⁵⁷. Together, these instruments frame vulnerability not merely as a social category but as a structural condition produced by asymmetries of knowledge and control in AI-mediated interactions. They provide the regulatory scaffolding for understanding cognitive vulnerability as a dynamic and legally cognisable harm.

This regulatory trajectory does not emerge in a vacuum. A substantial body of European legal scholarship has examined how digital environments produce structural vulnerability through choice architecture, information asymmetry, and personalised targeting, with Helberger, Sax, Strycharz and Micklitz establishing that digital vulnerability is best understood as a situational and relational condition rather than a fixed attribute of protected groups, a framing that the AI Act's context-sensitive approach to cognitive risk broadly reflects. Alongside this literature on digital vulnerability, the

⁵⁵ Arts. 5(1)(a)–(b) AI Act.

⁵⁶ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (*Digital Services Act*), OJ L 277, 27 October 2022, Recital 67, Arts. 24 and 28.

⁵⁷ Recital 67 and Arts. 24–26 AI Act.

data protection framework established by the General Data Protection Regulation provides an underexplored but significant foundation for cognitive protection in the European context: the classification of biometric and health data as special categories under Article 9 subjects their processing to heightened conditions of consent and necessity, and the right not to be subject to solely automated decision-making with significant effects under Article 22 supplies a partial doctrinal basis for contesting systems that steer behaviour through neural or behavioural inference. The gap between these existing instruments and a fully articulated account of cognitive protection in EU law is precisely the space into which neurorights scholarship and the UNESCO Recommendation must intervene⁵⁸.

A workable legal account must show how exposure is created in practice. Inference technologies widen observation by turning everyday data into likely attitudes or intentions, as seen in emotion recognition, affective analysis, and large-scale audience targeting for persuasion⁵⁹. Influence architectures reorder choice sets and tune salience so that the path of least resistance aligns with the deployer's objectives, a pattern described in the regulatory literature as hypernudging and related manipulative design⁶⁰. Interfaces that capture or stimulate neural signals add a channel where the line between measurement and modulation is thin, a tension already analysed in forensic brain-reading jurisprudence⁶¹, and increasingly flagged in clinical and commercial neurotechnology ethics⁶². Collectively, these modalities erode the conditions for self-forming thought and generate harm at an earlier stage, often before notice or consent is meaningful,⁶³ which

⁵⁸ L. Bygrave, *Data Privacy Law: An International Perspective*, Oxford, 2014; H.-W. Micklitz, *The Visible Hand of European Regulatory Private Law*, in *Yearbook of European Law*, vol. 28, 2009, p. 3 ff.

⁵⁹ European Parliament, Panel for the Future of Science and Technology (STOA), *The Protection of Mental Privacy in the Area of Neuroscience*, Study PE 757.807, 2024.

⁶⁰ P. Kellmeyer, *Neurorights*, in M. D. Dubber, F. Pasquale and S. Das (eds), *The Cambridge Handbook of Responsible Artificial Intelligence*, Cambridge, 2024.

⁶¹ C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, cit.

⁶² AI Act; Supreme Court of Chile, *Girardi Lavín v. Emotiv Inc.*, cit.; UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021, cit.

⁶³ R. Calo, *Digital Market Manipulation*, in *George Washington Law Review*, vol. 82, 2014, p. 995 ff.; S. C. Matz, M. Kosinski, G. Nave and D. J. Stillwell, *Psychological Targeting as an Effective Approach to Digital Mass Persuasion*, in *Proceedings of the National Academy of Sciences*, vol. 114, 2017, p. 12714 ff.

underscores why governance focused solely on data-processing is systematically reactive rather than preventive.⁶⁴

Doctrinally, the anchor remains the *forum internum*, whose absolute protection supplies a normative outline for freedom of thought⁶⁵, but translation into adjudicable elements is the central difficulty. A practical test for unlawful interference must capture both intent and effect within AI-mediated environments. Such interference arises where mental states are accessed or manipulated without valid consent, where the mechanism of influence is opaque or concealed, or where design features exploit cognitive weaknesses instead of fostering informed deliberation. Ultimately, the decisive factor in assessing interference is whether such influence undermines an individual's capacity for independent judgment⁶⁶. Distinguishing persuasion from manipulation is crucial, and this distinction can be evaluated through objective design evidence rather than through speculative reconstruction of a user's mental state. Regulatory tools such as prohibited-practice frameworks and dark-pattern taxonomies offer concrete indicators for identifying covert or manipulative design strategies and for rendering these practices legally traceable through documentation and audit^{67,68}.

Evidence remains the pressure point, since direct proof of mental intrusion is rare in adversarial settings. Two strategies reduce the burden without diluting the standard. Structural proxies recognise that certain interface patterns have reproducible effects on attention and choice; their presence at scale justifies a rebuttable presumption of manipulative influence in proceedings and enforcement⁶⁹, particularly when combined with psychological targeting studies that quantify behavioural effects. Upstream documentation through testing and impact assessment

⁶⁴ K. Yeung, *Hypernudge*, in *Information, Communication & Society*, vol. 20, 2017, p. 118 ff.

⁶⁵ UN Human Rights Committee, *General Comment No. 22 on Freedom of Thought, Conscience and Religion*, CCPR/C/21/Rev.1/Add.4 (1993), cit.

⁶⁶ S. Ligthart, T. Douglas, C. Bublitz, T. Kooijmans and G. Meynen, *Forensic Brain-Reading and Mental Privacy in European Human Rights Law*, in *Neuroethics*, 2020, p. 1 ff.

⁶⁷ Arts. 9–15 AI Act.

⁶⁸ S. Wachter and B. Mittelstadt, *A Right to Reasonable Inferences*, in *Columbia Business Law Review*, 2019, p. 494 ff.; H. Brignull, *Deceptive Patterns: Exposing the Tricks Tech Companies Use to Control You*, Eastbourne, 2023.

⁶⁹ M. Tenca, E. Haselager and R. Andorno, *Brain Leaks and Consumer Neurotechnology*, in *Nature Biotechnology*, vol. 36, 2018, p. 805 ff.

externalises part of the proof to developers and deployers, allowing regulators and courts to draw adverse inferences where systems were not evaluated for attentional capture, affective inference, or susceptibility to manipulation⁷⁰, and to require disclosure of model evaluations where mental states are processed or inferred⁷¹.

Precision about the legal object of protection is equally important. Narrow rules that focus on neurodata alone are under-inclusive when cognitive states are inferred from eye movements, cursor trajectories, or biometric combinations that never touch the skull⁷², so a more defensible object is mental information broadly defined, encompassing neural signals and non-neural proxies that, in function, disclose or modulate mental content⁷³. This alignment of legal category with technical practice reduces incentives to route around protection through alternative sensors and clarifies which activities should trigger per se constraints and which require contextual balancing⁷⁴.

The literature also counsels restraint in rights inflation and care with language⁷⁵. Not every instance of decoding or persuasion requires a new fundamental right. The task is to specify protected interests and prohibited modalities in a form that can be integrated into existing guarantees of freedom of thought, privacy, bodily and psychological integrity, and non-discrimination⁷⁶, while preserving a strict core against non-consensual or covert access to mental content. Conceptual precision over what counts as reading the mind and how digital and neurotechnological routes differ

⁷⁰ E. Luguri and R. H. Strahilevitz, *Shining a Light on Dark Patterns*, in *Journal of Legal Analysis*, vol. 13, 2021, p. 43 ff.

⁷¹ W. Mattli and T. Büthe, *Setting International Standards*, in *World Politics*, vol. 56, 2003, p. 1 ff.; K.W. Abbott and D. Snidal, *The Governance Triangle*, in W. Mattli and N. Woods (eds), *The Politics of Global Regulation*, Ithaca, 2009.

⁷² European Parliament, STOA, *The Protection of Mental Privacy in the Area of Neuroscience*, Study PE 757.807, 2024, cit.

⁷³ O.R.K. Time, *Mind-Reading Devices Are Revealing the Brain's Secrets*, in *Nature*, vol. 626, 2024, p. 22.

⁷⁴ S. Wachter and B. Mittelstadt, *A Right to Reasonable Inferences*, cit.; European Parliament, STOA, *The Protection of Mental Privacy in the Area of Neuroscience*, cit.

⁷⁵ J.T. Theilen, *The Inflation of Human Rights: A Deconstruction*, in *Leiden Journal of International Law*, vol. 34, 2021, p. 681 ff.

⁷⁶ P. Kellmeyer, *Neurorights*, in M.D. Dubber, F. Pasquale and S. Das (eds), *The Cambridge Handbook of Responsible Artificial Intelligence*, Cambridge, 2024, p. 562 ff., cit.; M. Ienca and R. Andorno, *Towards new human rights*, cit.

prevents overreach and facilitates compliance architecture⁷⁷, and it aligns with the incremental universalism of the UNESCO process without collapsing into relativism⁷⁸.

Comparative trajectories reinforce these design choices. Instruments that emphasise red-line prohibitions against exploitation of vulnerabilities and that couple them with conformity-assessment duties are better suited to harms that materialise before awareness, and courts have begun to recognise psychological integrity as a justiciable interest with deletion and cessation remedies where mental information has been unlawfully retained or processed⁷⁹. The standard-setting dimension will matter for transnational uptake, since technical norms often mediate between abstract rights and operational controls; here, governance research on international standardisation indicates both the integrative potential of global norms and the risks of capture, which the article returns to in the analysis of implementation and convergence *infra* Section 5⁸⁰.

4. *Comparative Legal Developments on Neurorights*

This section maps the main legal templates that currently structure neurorights and adjacent protections across jurisdictions. The comparison proceeds by legal design rather than by a catalogue of countries, distinguishing a rights-first constitutional track, a prohibition-and-risk regulatory track, and a consumer-administrative track that relies on general unfair-practice and product-safety controls, with standard-setting as a transversal force.

A rights-first constitutional track is most visible in Latin America, with Chile as the leading case. Constitutional text now anchors protection for

⁷⁷ S. Wachter and B. Mittelstadt, *A Right to Reasonable Inferences*, *cit.*; C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, *cit.*

⁷⁸ S. Lighthart, C. Bublitz, T. Douglas, L. Forsberg and G. Meynen, *Rethinking the Right to Freedom of Thought*, in *Human Rights Law Review*, vol. 22, 2022, ngac028; J.-C. Bublitz, *The Nascent Right to Psychological Integrity and Mental Self-Determination*, in A. von Arnault, K. von der Decken and M. Susi (eds), *The Right to Mental Integrity*, Cambridge, 2020.

⁷⁹ C. Bublitz, F. Bariffi, M. Sosa Navarro and P. Kellmeyer, *UNESCO's Recommendation on Neurotechnology*, *cit.*

⁸⁰ Arts. 5, 9–15 AI Act; Supreme Court of Chile, *Girardi Lavín v. Emotiv Inc.*, *cit.*; W. Mattli and T. Büthe, *Setting International Standards*, *cit.*

mental integrity and brain-derived information, and judicial practice has treated psychological integrity and privacy as enforceable limits on commercial processing of neurodata, including orders of deletion where processing was unlawful⁸¹. Regional political bodies have also experimented with model instruments that supply a portable vocabulary for member legislatures and courts, signalling an intent to normalise neurorights within public law rather than to treat them as a sub-species of data protection⁸². The strength of this track is symbolic clarity and justiciability. Courts can translate protected interests into remedies without waiting for sectoral codes. Its weakness is operational because constitutional language requires downstream legislation to define technical duties, standards of proof, and supervisory architecture; otherwise, adjudication risks over-reliance on general principles.

A prohibition-and-risk regulatory track is exemplified in Europe. The Artificial Intelligence Act prohibits, as an unacceptable risk practice under Article 5(1)(b), AI systems that exploit vulnerabilities arising from age, disability, or socio-economic situation to materially distort behaviour in ways likely to cause harm to the affected persons, while also imposing risk management, testing, and documentation duties under Articles 9 to 15 on high-risk systems that can surface covert influence and render it evidentiary⁸³. In parallel, the Council of Europe's Framework Convention on Artificial Intelligence embeds human rights, democracy, and rule-of-law constraints in a binding public-international instrument, providing a channel for judicial dialogue with the European Court of Human Rights and national courts⁸⁴. The EU regulatory framework, however, must be read against a pre-existing and substantial body of scholarship on vulnerability and manipulation in European consumer and user protection law: Micklitz's work on the responsive consumer and the limits of autonomy-based protection, scholarship on algorithmic manipulation and the conditions under which personalised targeting crosses from persuasion into exploitation, and data protection literature examining Article 9 GDPR's special category regime and Article 22's constraints on automated profiling

⁸¹ Law No. 21.383 of 14 October 2021 amending Art. 19 of the Constitution of Chile, cit.; Supreme Court of Chile, *Girardi Lavín v. Emotiv Inc.*, cit.

⁸² Latin American Parliament, *Model Law on Neuro-Rights and Neurotechnologies*, 2023.

⁸³ Art. 5 and Arts. 9–15 AI Act.

⁸⁴ Council of Europe, *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, opened for signature 5 September 2024.

as resources for contesting systems that infer or exploit mental states without adequate safeguards. This literature establishes that the concept of cognitive vulnerability in EU law is not a creation of the AI Act but rather an extension of a long-running debate about the structural power asymmetries between platforms and users, a debate that neurorights scholars must engage with rather than bypass. European doctrinal work on freedom of thought has clarified how modern interferences may occur through data-driven inference and influence below awareness thresholds, strengthening the conceptual link between *forum internum* guarantees and regulatory prohibitions⁸⁵. The strength of this track lies in its operational detail, providing administrative mechanisms for supervision and a coherent framework for defining prohibitions. However, the key risk remains under-inclusion: when protection depends on a closed list of prohibited practices, novel forms of cognitive interference may escape coverage until they are expressly codified, while evidentiary challenges persist where manipulation is diffuse and non-transparent. That gap matters for AI governance specifically. The AI Act represents a deliberate choice to treat cognitive manipulation as a human rights concern and not merely a market failure, but the framework's effectiveness will depend on whether enforcement authorities develop the technical capacity to identify AI-driven interference with mental autonomy in practice.

A consumer-administrative track is most developed in the United States. In the absence of a federal neurorights instrument, protection is mediated through general unfair- or deceptive-practices law, sectoral privacy statutes, and product-safety norms. Scholarship on digital manipulation and dark patterns has informed regulatory doctrines that treat covert steering and choice-architecture abuse as unlawful even where data-protection rules are not directly engaged⁸⁶. Risk-management frameworks have begun to recognise cognitive exposure as a category of harm and to embed documentation and testing duties upstream in development lifecycles, which can ease evidentiary burdens ex post by

⁸⁵ S. Lighthart, C. Bublitz, T. Douglas, L. Forsberg and G. Meynen, *Rethinking the Right to Freedom of Thought*, cit.; S. Lighthart, T. Douglas, C. Bublitz, T. Kooijmans and G. Meynen, *Forensic Brain-Reading and Mental Privacy in European Human Rights Law*, cit.

⁸⁶ R. Calo, *Digital Market Manipulation*, cit.; E. Luguri and R.H. Strahilevitz, *Shining a Light on Dark Patterns*, cit.; A. Mathur, G. Acar, M. Friedman, E. Lucherini, J. Mayer, M. Chetty, A. Narayanan, *Dark Patterns at Scale*, in *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, 2019, Art. 81.

making design choices legible to regulators and courts⁸⁷. This track is flexible and fast to adapt, yet coverage is uneven, and the absence of a rights-based baseline can make mental autonomy contingent on agency enforcement priorities.

Standard-setting operates across all tracks. OECD principles provide a non-binding but widely endorsed baseline for trustworthy AI, and they are frequently translated into procurement rules, certification schemes, and internal compliance programs.⁸⁸ The UNESCO process adds a neurorights-specific lens that defines protected mental interests and prohibited modalities at a level of generality suitable for constitutional drafting and for regulatory codes. In practice, soft-law standards shape audit checklists, model documentation templates, and conformity-assessment criteria, which in turn determine what kinds of cognitive interference are visible to institutions. This feedback loop is powerful. It can drive convergence around a minimum floor while leaving room for jurisdictional elaboration, but it also raises capture risks if industry actors dominate technical committees or if standards substitute for, rather than complement, enforceable rights.

Three comparative observations follow. First, the more a system relies on prohibitions and design-time controls, the better it aligns with the temporality of cognitive harm, which typically materialises before awareness and resists post hoc proof⁸⁹. Secondly, a constitutional anchor for mental integrity can accelerate judicial recognition of mental autonomy as a justiciable interest and deliver concrete remedies such as deletion and cessation, but only where implementing legislation makes cognitive interference administrable through duties, documentation, and audit⁹⁰. Thirdly, standard-setting is not an afterthought. It is the conduit that translates abstract rights into operational controls and evidentiary artefacts. The challenge is to ensure that standardisation reflects public-law values

⁸⁷ National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework 1.0*, 2023.

⁸⁸ OECD, *AI Principles*, 2019, cit.

⁸⁹ UNESCO, *Towards an International Instrument on the Ethics of Neurotechnology* (intergovernmental meeting of experts, 12–16 May 2025; final draft to be submitted to the 43rd General Conference, Nov. 2025)

⁹⁰ UNESCO, *Draft text of the Recommendation on the Ethics of Neurotechnology* (May 2025).

rather than only technical feasibility, which requires balanced participation and clear public mandates alongside private expertise⁹¹.

These tracks are not mutually exclusive. Jurisdictions can constitutionalise a core while relying on risk-based regulation to police prohibited modalities and on consumer-administrative tools to tackle residual manipulation. Over time, dialogue among courts, regulators, and standard-setters can consolidate a shared understanding of the protected legal interest and of the modalities that constitute unlawful interference with mental autonomy⁹². That dynamic is already visible in the interplay between constitutional adjudication in Chile, regulatory prohibitions and risk duties in the European Union, and administrative enforcement against manipulative design in the United States.

5. *Implementation and Convergence in Practice*

Universal commitments acquire force only when they are translated into the routines of design, documentation, testing, and audit. That translation is beginning to take shape in three channels that operate across legal families. Management-system standards set organisation-level duties for governance, risk, and improvement; risk-management frameworks specify processes for identifying and treating harms; technical and ethical standards articulate testable properties such as transparency and human well-being.⁹³ Together, these instruments provide hooks through which abstract protections of mental autonomy can be made operational in development lifecycles and procurement.

Management-systems and risk-management instruments offer immediate leverage for cognitive risk. For example, ISO/IEC 42001 establishes an AI management system with policy, roles, risk processes, and continual improvement, a structure that can embed duties to identify and mitigate exposure arising from inference about mental states and persuasive

⁹¹ M. Ramanathan, *The UNESCO Draft Recommendations on Ethics of Neurotechnology: A Commentary*, in *Indian Journal of Medical Ethics*, vol. X, 2025.

⁹² C. M. L. Brown, *Neurorights, Mental Privacy, and Mind Reading*, in *Neuroethics*, vol. 17, 2024, p. 1 ff.

⁹³ National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework 1.0*, 2023.

design⁹⁴. Additionally, ISO/IEC 23894 provides detailed guidance for AI risk management, including context definition, risk identification, and control selection, which can be extended to mental information and manipulation-related harms⁹⁵. Also, NIST's AI Risk Management Framework complements these by offering a voluntary, process-based approach whose governance and measurement functions can be populated with neurorights-relevant controls such as pre-release testing for attentional capture and covert influence⁹⁶.

The technical standards stream is diversifying. For instance, ISO/IEC JTC 1/SC 42 coordinates AI standardisation across foundational concepts, trustworthiness, and conformity-assessment linkages, which positions it to shape evidentiary artefacts that regulators and courts will recognise⁹⁷. In the IEEE 7000-series, IEEE 7001 defines measurable, testable levels of transparency for autonomous systems, while IEEE 7010 proposes methods for assessing impacts on human well-being. These instruments are not neurorights codes, yet they can be repurposed to make cognitive effects observable in a way that supports enforcement. For example, transparency levels can include disclosure when systems infer or modulate mental states, and well-being metrics can track attentional burden or erosion of user agency⁹⁸⁹⁹.

Regional regulation is beginning to align with this standardisation process. The EU Artificial Intelligence Act links high-risk systems to conformity assessment procedures that refer to technical standards and documentation duties. In the absence of complete harmonisation, providers may rely on equivalent technical specifications and risk management frameworks. These mechanisms matter for neurorights because they determine whether cognitive risks are identified before deployment and whether evidence of hidden influence or unauthorised decoding can later be demonstrated. Also, administrative practices are beginning to converge toward similar forms of regulatory control. The European Data Protection Board's guidelines on deceptive design patterns in social media give

⁹⁴ Ibid.

⁹⁵ ISO/IEC 42001:2023, *Artificial intelligence Management system*.

⁹⁶ ISO/IEC 23894:2023, *Information technology Artificial intelligence Guidance on risk management*.

⁹⁷ ISO/IEC JTC 1/SC 42, *Artificial intelligence, scope and working groups*.

⁹⁸ IEEE 7001-2021, *Standard for Transparency of Autonomous Systems*.

⁹⁹ S. Goering, *Recommendations for Responsible Development and Application of Neurotechnologies*, in *Neuroethics*, vol. 16, 2023, p. 1 ff.

regulators a taxonomy for identifying covert steering in interfaces¹⁰⁰. Although framed under data protection, they supply markers and exemplars that map onto the exploitation of cognitive weakness in AI systems and can guide design-time mitigation and disclosure. This template can be adapted in consumer-protection and sectoral regimes to treat manipulation as a form of unfair practice, particularly where vulnerabilities are structural and not class-based¹⁰¹.

Domestic legislation is also beginning to recognise brain data as a protected category of personal data. In the United States, Colorado has created the first express statutory protections for neural and related data within its consumer-privacy framework and has adopted a comprehensive AI law that imposes risk-management and documentation duties for high-risk systems¹⁰². These moves are uneven and still maturing, yet they demonstrate a path for articulating per se constraints on collection and disclosure of neurodata alongside process duties for AI risk, two pillars that track the universalist baseline proposed for neurorights¹⁰³.

Implementation can be built around three operational elements without waiting for bespoke neurorights statutes. First, require classification of mental information as a protected category in management-systems and risk frameworks, covering neural signals and non-neural proxies that function as cognitive biometrics. Secondly, require pre-market testing for cognitive impact with design-level mitigations where systems are capable of attentional capture, affective inference, or persuasive modulation that targets cognitive weakness. Thirdly, mandate disclosure when a system processes or infers mental states and require deployers to document controls, residual risk, and user-facing remedies. These elements can be written directly into AI management systems, risk registers, and conformity-assessment files, which makes them auditable. ISO/IEC and IEEE instruments already provide the scaffolding for such duties.

¹⁰⁰ European Data Protection Board, *Guidelines 03/2022 on Deceptive Design Patterns in Social Media Platform Interfaces: How to Recognise and Avoid Them*, Version 2.0, final version adopted 24 February 2023.

¹⁰¹ Art. 40 AI Act on harmonised standards and presumption of conformity and Art. 11 AI Act on technical documentation.

¹⁰² Colorado Senate Bill 24-205, *Artificial Intelligence and Consumer Protections Act*, signed into law 17 May 2024, effective 1 February 2026.

¹⁰³ Future of Privacy Forum, *Conformity Assessments under the EU AI Act*, White Paper, 30 April 2025.

Standardisation is a double-edged conduit for universalism. It is the practical route through which abstract rights shape technical practice, but it also carries familiar risks of capture and drift from public-law values¹⁰⁴. Comparative policy analysis has warned that reliance on harmonised and private standards can relocate normative decisions to technical committees, and recent commentary on the AI Act's standardisation pipeline reaches similar conclusions about the need for strong public mandates and balanced participation¹⁰⁵. These warnings should not be understood as a rejection of standardisation but rather as a design brief for its governance, calling for transparency in committee membership, clear justification for the selection of tests and metrics, and effective public oversight whenever standards acquire legal force.¹⁰⁶

Universality and relativism can work together within this framework. Global agreements can define the main protected mental interests and prohibited practices, while each country can adapt technical details such as metrics and testing methods to its own systems and capacities. Convergence, therefore, depends on shared forms of evidence created through standards and risk frameworks rather than identical rules. What matters is that these tools make cognitive risks clear at an early stage and allow legal action when standards do not give clear answers¹⁰⁷.

6. *Universal Standards and Local Realities: An Evaluative Framework*

The tension between universalism and relativism only becomes meaningful when linked to practical legal functions. A standard that defines protected mental interests but cannot be enforced remains aspirational, while one that ignores differences in how people experience cognitive risks fails to protect those most exposed to harm. The evaluative question is therefore twofold. First, can universal neurorights specify a minimum

¹⁰⁴ M. Bartlett, *Standard Deviation: Global Standardisation and Implications for International Law*, in *New Zealand Yearbook of International Law*, 2019, p. 119 ff.

¹⁰⁵ R. Kilian, L. Jäck and D. Ebel, *European AI Standards – Technical Standardisation and Implementation Challenges under the EU AI Act*, in *European Journal of Risk Regulation*, 2025, p. 1 ff.

¹⁰⁶ European Digital Rights, *The Role of Standards and Standardisation Processes in the EU's Artificial Intelligence Act*, May 2022.

¹⁰⁷ F. Sovrano, *Simplifying Software Compliance: AI Technologies in Regulatory Frameworks*, in *Patterns*, vol. 6, 2025, e1000123.

content that travels across legal families without hollowing out cognitive protection¹⁰⁸? Second, can that content be implemented through institutions that differ in adjudicatory culture and infrastructure?¹⁰⁹

A minimum content must do more than restate freedom of thought. It should define protected mental interests and prohibited modalities of access and influence in a manner that is legible to courts and regulators. A workable core would include protection of cognitive data, understood to cover neural signals and non-neural proxies that functionally disclose or modulate mental content;¹¹⁰ a prohibition on non-consensual decoding where a person intends mental content to remain private;¹¹¹ and a prohibition on covert manipulation that exploits cognitive weakness to steer behaviour, especially below awareness thresholds¹¹². Each element corresponds to a justiciable boundary and to a locus for design-time duties and evidentiary artefacts.

At this point, context becomes relevant on three levels. First, the definition of proxies that count as mental information will vary with local technical practices and measurement tools, so a universal rule should state function-based criteria and delegate sensor-specific taxonomies to technical standards and guidance¹¹³. Secondly, proof of covert influence depends on institutional capacity. Jurisdictions with strong administrative oversight can rely on audit trails and impact assessments; jurisdictions with court-led enforcement will need presumptions and structural indicators to shift burdens where interface patterns predictably drive behaviour¹¹⁴. Thirdly, remedies must correspond to the practical levers available within different legal systems. Measures such as deletion or cessation orders are also relatively portable across jurisdictions, while remedies involving disgorgement or product withdrawal depend on specific legislative

¹⁰⁸ J. Donnelly, *Universal Human Rights in Theory and Practice*, cit., p. 10 ff.

¹⁰⁹ P. O'Callaghan, *The Right to Freedom of Thought: An Interdisciplinary Analysis*, in *The International Journal of Human Rights*, vol. 28, 2024, p. 281 ff.

¹¹⁰ European Parliament, *Panel for the Future of Science and Technology, The protection of mental privacy in the area of neuroscience*, Study PE 757.807, 2024; on non-neural proxies functioning as cognitive biometrics, supra note 40.

¹¹¹ UN Human Rights Committee, *General Comment No. 22 on Freedom of Thought, Conscience and Religion*, CCPR/C/21/Rev.1/Add.4 (1993), cit.; A/76/380, *Interim Report of the Special Rapporteur on Freedom of Religion or Belief*, United Nations, 2021, §§ 24–47.

¹¹² Art. 5(1)(a)–(b) AI Act.

¹¹³ ISO/IEC JTC 1/SC 42, cit.

¹¹⁴ European Data Protection Board, *Guidelines 03/2022 on Deceptive Design Patterns in Social Media Platform Interfaces*, cit.

mandates. A universal standard that anticipates these variations and accommodates institutional diversity is therefore more likely to endure in translation and implementation.

The comparative record supports this structure. The European approach, with red-line prohibitions combined with conformity-assessment duties, makes covert influence legible before market access and creates documentation that can travel into litigation¹¹⁵. The constitutional route visible in Chile demonstrates that mental autonomy can be recognised as a justiciable interest with concrete deletion remedies, which resolves standing and justiciability concerns *ex ante* while leaving technical standards to follow-up legislation¹¹⁶. The United States practice shows how consumer-administrative doctrines¹¹⁷ can treat covert steering as unlawful manipulation even in the absence of a dedicated neurorights catalogue, provided enforcement agencies have clear taxonomies of dark patterns and deceptive design¹¹⁸. None of these tracks alone guarantees coverage; together they suggest that universal content and local instrumentation can be harmonised without erasing difference.

The *forum internum* of freedom of thought provides a firm normative foundation, yet its effective application to contemporary technological contexts requires further doctrinal development. Recent UN reporting has documented how inference and influence can reach protected mental domains indirectly through data-driven profiling and persuasive architectures¹¹⁹. European human-rights jurisprudence, while developed primarily under the right to respect for private life, has recognised psychological integrity as part of protected private life and has allowed granular balancing for interferences with mental well-being¹²⁰. These lines of authority render the notion of universal content substantively meaningful

¹¹⁵ Future of Privacy Forum, *Conformity Assessments under the EU AI Act*, White Paper, 30 April 2025.

¹¹⁶ Supreme Court of Chile, *Girardi Lavín v. Emotiv Inc.*, cit.; Law No. 21.383 of 14 October 2021 amending Art. 19 of the Constitution of Chile, cit.

¹¹⁷ Colorado Senate Bill 24-205, *Artificial Intelligence and Consumer Protections Act*, signed into law 17 May 2024, effective 1 February 2026.

¹¹⁸ R. Calo, *Digital Market Manipulation*, cit.; E. Luguri and R.H. Strahilevitz, *Shining a Light on Dark Patterns*, cit.

¹¹⁹ A/76/380, *Interim Report of the Special Rapporteur on Freedom of Religion or Belief*, United Nations, 2021, cit.

¹²⁰ European Court of Human Rights, *Bensaid v. United Kingdom*, App. No. 44599/98, judgment of 6 February 2001; ECtHR, *Guide on Article 8 of the Convention*, Council of Europe, 2020.

rather than rhetorical. They provide a doctrinal foundation for recognising that covert manipulation and non-consensual decoding are not simply privacy violations or consumer-law infractions but constitute interferences with the fundamental conditions of thought and self-formation.

The argument for contextual universality is not that everything is relative. It is that universality is strongest when stated at the level of protected mental interests and prohibited modalities, and when implementation is allowed to vary in metrics, thresholds, and institutional design¹²¹. This stance is familiar in human-rights theory, which has long distinguished between universal concepts and diverse conceptions and implementations. In the neurorights context, the translation work passes through standard-setting and risk frameworks, which generate the evidentiary artefacts on which both regulators and courts rely. Presumptions of conformity under regional law will amplify this effect, but they also threaten to relocate normative choices to technical committees unless participation and public mandates are secured¹²². Governance responses are a readily available tool that can strengthen the legitimacy of standard-setting processes. They may include ensuring transparency in committee membership, providing reasoned justification for the selection of tests and metrics, and establishing mechanisms for public scrutiny or challenge when standards acquire legal effect.

However, two residual objections should be addressed. Some critics argue that any universal core will be either so narrow as to be redundant or so broad as to be vague¹²³. The redundancy claim dissolves once the object is mental information and the prohibited modalities are specified with precision. The vagueness claim loses force when the core is operationalised through design-time duties, documentation, and audit, which produce legible artefacts for adjudication and supervision. Others warn that a focus on restricting the flow of neurodata could hinder the development of beneficial neurotechnologies and research¹²⁴. The answer is proportionality in implementation, i.e. a strict per se rule against non-consensual decoding

¹²¹ J. Donnelly, *The Relative Universality of Human Rights*, in *Human Rights Quarterly*, vol. 29, 2007, p. 281 ff.

¹²² Skadden, *EU Standardization Supporting the AI Act*, 7 October 2024; *AI Act Standard Setting Overview*, available at [artificialintelligenceact.eu](https://www.artificialintelligenceact.eu), updated 21 July 2025.

¹²³ L. Feito, *The Difficulty of Universal Neurorights*, in *AJOB Neuroscience*, vol. 14, 2023, p. 380 ff.

¹²⁴ S. Pauzauskie, J. Genser and R. Yuste, *Protecting Neurodata Privacy — First, Do No Harm*, in *JAMA Neurology*, vol. 82, 2025, p. 212 ff.

for private mental content, paired with conditions for research and clinical contexts that include independent oversight, data-minimisation, and purpose-binding, preserves therapeutic innovation while protecting the *forum internum*.

The most viable solution lies in an evaluative framework that maintains universalism and relativism in productive tension. Universalism supplies the baseline content of neurorights; relativism in implementation supplies the fit with local institutions and infrastructures.

7. Conclusion

Universal neurorights are plausible only if they are specified as protected mental interests and prohibited modalities that courts and regulators can apply in practice, and only if their implementation accommodates the diversity of cognitive exposure across legal orders. The core content is narrow and concrete. Mental information must be the legal object, understood to include neural signals and non-neural proxies that functionally disclose or modulate mental content¹²⁵, with per se constraints against non-consensual decoding for private mental content and against covert manipulation that exploits cognitive weakness. All other mechanisms serve as instruments through which these core protections are implemented and enforced.

Implementation should proceed where design choices and documentation live, not only where remedies are ordered. Management-systems and risk-management frameworks already supply levers for duties to identify and mitigate cognitive risk, to test for attentional capture and covert influence, and to disclose when systems infer or process mental states¹²⁶. Regional regimes that tie presumptions of conformity to listed standards will shape what evidence is created in development lifecycles and, therefore, what can be adjudicated later¹²⁷, which is why

¹²⁵ J. Donnelly, *Universal Human Rights in Theory and Practice*, cit., p. 10 ff.; J. Donnelly, *The Relative Universality of Human Rights*, cit.

¹²⁶ European Parliament, STOA, *The Protection of Mental Privacy in the Area of Neuroscience*, cit.; P. Kellmeyer, *Neurorights*, in M. D. Dubber, F. Pasquale and S. Das (eds), *The Cambridge Handbook of Responsible Artificial Intelligence*, cit.

¹²⁷ Art. 40 AI Act on harmonised standards and presumption of conformity and Art. 11 AI Act on technical documentation.

governance of standard-setting itself is a first-order human-rights question and not a technical afterthought¹²⁸.

Comparative experience confirms the feasibility of a universal core with differentiated delivery. A constitutional anchor can render mental autonomy justiciable with concrete remedies such as deletion and cessation¹²⁹, while prohibition-and-risk regimes make covert influence legible before market access and keep evidentiary artefacts close to design and deployment¹³⁰. Consumer-administrative enforcement fills residual gaps by treating manipulative interfaces and dark patterns as unlawful practices, provided taxonomies and indicators are available to enforcement staff¹³¹. As mentioned earlier, two objections recur; the first is redundancy. If freedom of thought, privacy, integrity, and non-discrimination already exist, a separate neurorights vocabulary might appear superfluous. However, existing guarantees do not yet specify the prohibited modalities that reach protected mental domains through inference and influence at scale, nor do they mandate the documentation needed to prove covert interference. A universal core that targets mental information and manipulation fills that gap without proliferating rights labels¹³². The second being that the strict rules on cognitive data could become an obstacle for therapeutic and research uses of neurotechnology. The response is proportionality in implementation and context-specific carve-outs with independent oversight, purpose-binding, and data-minimisation, which preserve beneficial pathways while protecting the *forum internum*¹³³.

¹²⁸ W. Mattli and T. Büthe, *Setting International Standards*, cit.; European Digital Rights, *The Role of Standards and Standardisation Processes in the EU's Artificial Intelligence Act*, May 2022, cit.; Skadden, *EU Standardization Supporting the AI Act*, cit.

¹²⁹ Supreme Court of Chile, *Girardi Lavín v. Emotiv Inc.*, cit.; Law No. 21.383 of 14 October 2021 amending Art. 19 of the Constitution of Chile, cit.

¹³⁰ National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework 1.0*, cit.; ISO/IEC 42001:2023, *Artificial Intelligence - Management System*, cit.; ISO/IEC 23894:2023, *Information Technology - Artificial Intelligence - Guidance on Risk Management*, cit.

¹³¹ Ł. Szoszkiewicz and R. Yuste, *Mental Privacy: Navigating Risks, Rights and Regulation*, in *EMBO Reports*, vol. 26, 2025, p. 3469 ff.

¹³² European Data Protection Board, *Guidelines 03/2022 on Deceptive Design Patterns in Social Media Platform Interfaces*, cit.

¹³³ S. Lighthart, C. Bublitz, T. Douglas, L. Forsberg and G. Meynen, *Rethinking the Right to Freedom of Thought*, cit.; S. Lighthart, T. Douglas, C. Bublitz, T. Kooijmans and G. Meynen, *Forensic Brain-Reading and Mental Privacy in European Human Rights Law*, cit.

Aimen Taimur

Neurorights in the Age of AI:

Universalism, Cognitive Vulnerability, and the Limits of Legal Translation

The path forward should be iterative and adaptive rather than maximalist, developing through successive refinements that align ethical aspirations with institutional realities. Therefore, UNESCO's process can state minimum global commitments that define the protected interest and the prohibited modalities, while regional and national regimes build the evidentiary and supervisory machinery that makes those commitments real. With final negotiations on the UNESCO Recommendation set for November 2025, the coming months will test whether the international community can translate ethical consensus into binding legal form and confirm neurorights as a concrete expression of freedom of thought. The debate about neurorights is, at its core, part of the wider question of what AI regulation owes to human dignity. The UNESCO Recommendation gestures toward an answer. Whether that answer acquires legal force will depend less on the text adopted in November 2025 than on the domestic institutions, courts, and regulators willing to act on it.

Aimen Taimur

Neurorights in the Age of AI:

Universalism, Cognitive Vulnerability, and the Limits of Legal Translation

ABSTRACT: AI systems can now infer attitudes, emotions, and intentions from neural signals, gaze patterns, and behavioural data, often without the person being observed knowing it is happening. This article asks whether international human rights law can keep pace with that shift, and specifically whether neurorights, understood as protections for mental privacy, cognitive liberty, and psychological integrity, can be given a workable universal form. The comparison drawn here is between Chile, which has constitutionalised mental integrity and produced justiciable case law; the European Union, which has prohibited certain AI practices while leaving significant evidentiary gaps; and the United States, which relies on consumer protection doctrines built for different problems. The UNESCO Recommendation on the Ethics of Neurotechnology (2025) attempts to speak to all three, though its authority rests on persuasion rather than obligation. The argument advanced is that a viable universal framework is achievable, but it requires precision about what is actually being protected and against what kinds of AI-mediated interference, alongside genuine flexibility about how those protections are enforced across very different legal systems.

KEYWORDS: neurorights – cognitive vulnerability – freedom of thought – neurotechnology – artificial intelligence.

Aimen Taimur – Phd Candidate, Tilburg Institute for Law, Technology, and Society – Tilburg University, Tilburg, Netherlands (a.taimur@tilburguniversity.edu)

Discrimination Revised. How AI Is Reshaping Anti-Discrimination Law*

Costanza Nardocci

TABLE OF CONTENTS: 1. Introducing AI-Based Discrimination. – 2. What’s New, Why, and What Is Not. – 3. When EU (But Also, Common Law) Anti-Discrimination Law Falls Short. – 4. The Pitfalls of Understanding AI-Based as a Pure Human-Driven Discrimination: The Proxy, Proxy Discrimination, and the New Victim Paradigm. – 5. Did Someone Say, “Discriminatory AI”? Regulating and Sanctioning AI-based Discrimination. – 6. The Law Does not Say, but: Procedural Remedies at the Times of AI-Based Discrimination. – 7. Conclusions: A Call for Awareness.

1. Introducing AI-Based Discrimination

When discussing artificial intelligence and human rights, discriminatory AI is always quite present in the public and academic debate¹.

Despite the undoubted and socially accepted relevance of the discriminatory implications of AI², existing and prospective regulations

* This article was subjected to double-blind peer review.

¹ The literature that first began to examine extensively the discriminatory implications associated with artificial intelligence technologies is the Anglo-American one. In particular, it is worth highlighting at the outset the essay by D.K. Citron and F. Pasquale, *The Scored Society: Due Process for Automated Predictions*, in *Washington Law Review*, 2014, p. 1 ff. The work is highly interesting, offering an overview of the main critical issues characterising the functioning of artificial intelligence technologies—opacity, arbitrariness of the determinations made, disproportionate impact on certain groups—as well as illustrating several proposals to mitigate them. These may be summarised in the adoption of policies and regulations ensuring the transparency of AI systems, an assessment of their impact and recommended uses, and the protection of individuals’ rights. See also, from the early years in which the topic began to attract attention, T. Zarsky, *Understanding Discrimination in the Scored Society*, in *Washington Law Review*, 2014, p. 1375 ff.

² On the discriminatory implications of AI, The issue gained significant public attention starting in 2014, following the publication of the White House report [Big Data: Seizing Opportunities, Preserving Values](#), available at [obamawhitehouse.archives.gov](#), which in particular includes several examples of discriminatory effects produced through the use of artificial intelligence techniques. The most well-known example concerns the consequences

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

seldom question whether anti-discrimination law should be reconsidered in light of the torsion discrimination encompasses when coupled with AI.

The consequences of such a poor investigation into the here defined AI-based discrimination are the scarce consideration of how artificial intelligence technologies may eventually result in unreasonable treatments towards individuals and social groups, and the absence of tailored normative responses in the recently adopted, and also perspective, AI legislations worldwide³.

generated by the Street Bump App, used in the city of Boston and developed in collaboration with the Mayor's Office of New Urban Mechanics. The app sought to use data collected from residents' smartphones to assess which areas of the city required redevelopment, through an analysis of road and neighborhood conditions. The main issue with the app stemmed from the fact that economically poorer groups were less likely to own a smartphone on which the app could be installed. As a result, municipal interventions would have focused only on neighborhoods with higher concentrations of residents belonging to more affluent socioeconomic classes. The app therefore ended up discriminating against certain groups of residents on the basis of economic conditions. The literature began to explore more deeply the intersections between artificial intelligence and discriminatory phenomena particularly after the publication of the 2014 White House Report. An interesting study has shown that, among works published from 2014 to 2019 in indexed journals, only 14 studies addressed the issue from the perspective of its legal implications and consequences. This refers to the work of M. Favaretto, E. De Clercq and B. Simone Elger, *Big Data and Discrimination: Perils, Promises and Solutions. A Systematic Review*, in *Journal of Big Data*, vol. 6, no. 12, especially p. 5 ff. An even more recent study highlights the existence of a significant gap between research conducted in the United States on artificial intelligence and discrimination, and the current state of theoretical development in continental Europe. This refers to S. Wachter, B. Mittelstadt and C. Russell, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI*, in *Computer Law & Security Review*, 2020, p. 1 ff. Recently, in this field, see also the well-known works of K. Crawford, including *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, New Haven, 2021. In previous decades, the literature dealt only occasionally with the role of bias in the functioning of technologies. In particular, see B. Friedman and H. Nissenbaum, *Bias in Computer Systems*, in *ACM Transactions on Information Systems*, 1996, p. 330 ff.; M. Bruce and A. Adam, *Expert Systems and Women's Lives: A Technology Assessment*, in *Futures*, 1989, p. 480 ff.

³ Reference is chiefly to the European Union Regulation, the so-called "AI Act", Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, but also to the other existing laws on AI, as in the case of the laws recently adopted by Japan and, before it, by North Korea. On the scarce relevance attributed to discriminatory AI in the regulatory framework see paragraph no. 5.

Contrary to that approach is, instead, the understanding that discrimination possesses specific traits in the AI realm, which separates it from the already known human-driven discrimination⁴.

The juxtaposition between AI-based and human-driven discrimination essentially lies in the supplementary and sometimes even protagonist role of the machine *vis-à-vis* the humans, which ultimately leads to alternative, if not heterogeneous, manifestations of discriminatory treatments.

The article intends to illustrate the core elements featuring AI-based discrimination. It untangles the relationships AI-based discrimination entertains with EU anti-discrimination law, looking at how direct and indirect discrimination react towards discrimination generated by AI technologies. It then proceeds to argue that AI-based discrimination could not be misconceived as a pure manifestation of human deliberate or unintentional willingness to discriminate, but, rather, as a separate form of unreasonable treatment, where humans and automatic agencies intersect with one another.

Through the comparison between human-driven and AI-based discrimination, the paper eventually questions the adequacy of existing anti-discrimination laws to tackle discriminatory AI and advocates for the introduction of a new set of mechanisms to counter the prejudicial effects deriving from discriminatory AI.

2. *What's New, Why, and What Is Not*

At first glance, someone could be tempted to argue that nothing changes when discrimination meets AI, as there is always a human in the loop or, at least, behind AI⁵. That is a recurrent statement when considering the leading role played by humans in the preliminary phases of the building

⁴ On this, extensively, C. Nardocci, *Algoritmi, eguaglianza, discriminazione. Le sfide dell'intelligenza artificiale*, Turin, 2025.

⁵ This is true particularly with regard to the selection of the data fed to AI systems, that, instead of truthfully represent the external reality, are intrinsically human and, therefore, partial. On this, see K. Crawford, *The Hidden Biases in Big Data*, in *Harvard Business Review*, 2013, who emphasizes that: «[d]ata are assumed to accurately reflect the social world, but there are significant gaps, with little or no signal coming from particular communities. While massive datasets may feel very abstract, they are intricately linked to physical place and human culture». On the relationship between data and human culture, see, also, L. Gitelman, *"Raw Data" Is an Oxymoron*, Cambridge (MA), 2013.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

and functioning of AI technologies and, first, in the choices related to the dataset, which could suggest an overlap between human-driven and AI-based discrimination. In short, if AI is just a product of human behaviour, discrimination derived from AI technologies cannot be conceived as something different from the paradigm of discrimination caused by human conduct. In this sense, AI should be just something added to the traditional scheme, incapable of interrupting the causal link between the discriminatory conduct and its effects.

The denial of AI agency, thus, leads to the enhancement of human agency, neglecting the differences and, even sooner, the legitimacy of a debate on the existence of a new form of discrimination, which is derived or partially caused by AI technologies.

Contrary to that, though, AI possesses agency. Somehow, sometimes, it shows alternative features compared to human agency, but it does.

The recognition of AI's agency in the discrimination discourse contributes to explaining the divergent path of discrimination when the conduct stops being entirely human, resulting instead in a wide and extremely varied range of intersections between humans and AI technologies. While in human-driven discrimination, they are the only ones bearing the responsibility for the violation of the principles of equality and non-discrimination, in AI-based discrimination, humans are not alone in generating the discriminatory outcomes.

It is, thus, primarily the conduct that is subjected to profound changes when the unreasonable difference in treatment is also caused by AI.

The heterogeneity between human-driven and AI-based lies, therefore, and first of all, in the conduct: exclusively human in the case of human-driven discrimination; mixed or half human and half automatic in the paradigm type of AI-based discrimination.

The literature has thoroughly investigated the ways AI might discriminate by distinguishing a number of phases that pose discriminatory risks⁶. These are: the selection of the data fed into the machine; the training of the data, which could be autonomous run or, conversely, guided by the programmer; the identification and selection of the “target variables” and “class labels”, which are used to group into categories; the “feature selection”, meaning the choice of the features used by AI⁷; more

⁶ On this, extensively, S. Barocas and A.D. Selbst, *Big data disparate impact*, in *California Law Review*, vol. 104, 2016, p. 671 ff.; F.Z. Burgesius, *Discrimination, artificial intelligence, and algorithmic decision-making*, Council of Europe Publications, 2018.

⁷ The concept of features selection refers to the process by which the programmer and, subsequently, the AI system decides which attributes use in making its choices.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

fundamentally, the choice of the “proxy” as the element AI will refer to make distinctions which may eventually turn out to be of a discriminatory nature.

A few points might be made regarding the selection and training of the data.

The first, mentioned already, is the selection of the data fed into the machine. During this phase, bias might act as one of the key factors leading to discrimination. The heterogeneity of human biases widely affects the quality of the data and their representation of the outside reality in cases of over- or under-representation of certain groups or categories, which are usually those already suffering from structural discrimination⁸.

The second step or phase is the training of the data, which could be autonomous run or, conversely, guided by the programmer. Depending on the type of AI technologies, this phase may, therefore, show a more or less determinant influence on the programmer: the more AI functions as a machine learning system, the less the programmer will be able to control and supervise the outcome of the technology at stake. Put differently, the machine will take precedence over the human in the possible discriminatory outcome of the machine.

The training of the data is also relevant as it could witness two interesting phenomena: the poisoning of the data and the inaccurate or updated information provided to the machine in light of scientific and technological innovation. Data poisoning occurs when the human – and here the liability lies almost entirely with the programmer – deliberately feeds the machine with data that is poisoned, unhealthy, untruthful, and misleading. Instead, the latter case concerns AI technologies whose dataset has not been updated in a way consistent with the advancement and development of innovation, therefore impairing AI’s ability to respond to the tasks assigned.

Besides the specifics of these phases, whose analysis goes beyond the scope of the investigation, what matters the most is that each of these shows the multifaceted forms that the discriminatory conduct takes in the context of AI-based discrimination and demonstrates the non-exclusive, but rather concurrent, role of the human in the series of actions that eventually lead to the discriminatory effect.

But the analysis of the conduct is not enough to fully understand AI-based discrimination and to grasp its argued heterogeneity *vis-à-vis* human-

⁸ See, on this, J. Lerman, *Big Data and Its Exclusions*, in *Stanford Law Review Online*, vol. 66, 2013, p. 55 ff.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

driven discrimination. While the conduct exhibits new features, because of the said mediation of the machine in the chain of events leading to the unreasonable difference in treatment, the origins or causes of discrimination do not differ in the two types of discrimination.

In line with the argument that sees discrimination rooted in collective prejudices endorsed by the dominant group towards minority communities (the so-called out-groups) resulting from inter-ethnic conflicts⁹, both human-driven and AI-based discrimination originates from overlapping causes nurtured in human nature. Therefore, the origin of discrimination does not separate human-driven from AI-based, in that in both cases human prejudices represent the primary cause of the discriminatory conduct that follows through.

However, despite the human origin of discrimination, which may suggest that AI should be cut off from the chain linking the causes to the discriminatory effects, AI does impact the ways discrimination reveals itself and negatively impacts individuals and social groups. In other words, while the primary causes of discrimination remain a human domain, AI has opened the doors to a variety of obscured and often unpredictable discriminatory conduct that hardly reconcile with the traditional categories of EU anti-discrimination law.

The mismatch between how human discrimination has been captured and sanctioned under the law and the novelties AI-based ones have brought into the legal scenario justifies the need to reconsider whether and to what extent anti-discrimination law is adequate and should be applied to effectively tackle discriminatory conduct resulting from an interplay between humans and AI systems.

⁹ Reference is made to the so-called conflict power theories and to D.L. Horowitz, *Ethnic Groups in Conflict*, Oakland, 1985. On this, see also M.N. Marger, *Race and Ethnic Relations. American and Global Perspectives*, Boston, 2009; A.D. Smith, *The Ethnic Revival*, Cambridge, 1981, and Id., *The Ethnic Origins of Nations*, Chichester, 1988.

3. *When EU (But Also, Common Law) Anti-Discrimination Law Falls Short*

Framing AI in the context of EU anti-discrimination law is crucial to test the heterogeneity of AI-based discrimination from the legal and traditional conceptualization of discrimination.

From this angle, the theories of direct and, to some extent, even indirect discrimination prove to be inadequate to intercept and describe the core features of discrimination generated by AI¹⁰.

Beginning with the former, under EU law, direct discrimination implies the recurrence of the following elements: a difference in treatment among two or more comparable situations; the lack of an objective and reasonable justification underneath the distinction between the two or more comparable situations; the explicit reliance on one or more legally suspect grounds, the so-called factors of discrimination, that act as the basis of the difference in treatment.

Moreover, direct discrimination also requires that all the elements described above should also be supported by the proof of the recurrence of a direct and causal link between the conduct and the discriminatory effects, which means that there would not be direct discrimination should the effects not be caused by a human action.

Given the elements featuring the concept of direct discrimination under EU law, AI-based discrimination deviates from it for several reasons¹¹.

¹⁰ The more recent literature appears increasingly inclined to argue that traditional anti-discrimination law is inadequate to address disparities in treatment stemming from the functioning of AI technologies. This holds true for both EU anti-discrimination law and U.S. anti-discrimination law. See, in this regard, T.B. Gillis and J.T. Spiess, *Big Data and Discrimination*, in *The University of Chicago Law Review*, vol. 86, 2019, p. 458 ff.; S. Barocas and A.D. Selbst, *Big Data's Disparate Impact*, cit.; C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, 2017; and, in greater detail, R. Xenidis, *Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience*, in *Maastricht Journal of European and Comparative Law*, vol. 27(6), 2020, p. 736 ff.; M. Mann and T. Matzner, *Challenging Algorithmic Profiling: The Limits of Data Protection and Anti-Discrimination in Responding to Emergent Discrimination*, in *Big Data & Society*, 2019, p. 1 ff.; M. Lees, *The New Profiling: Algorithms, Black Boxes, and the Failure of Anti-Discrimination Law in the European Union*, in *Security Dialogue*, vol. 45(5), 2014, p. 494 ff.

¹¹ On this, see P.N. Geslevich and Y. Le Aretz, *Learning Algorithms and Discrimination*, in W. Barfield and U. Pagallo (eds), *Research Handbook of Artificial Intelligence and Law*, Cheltenham, 2018, p. 88 ff.; J. Adams-Prassl, R. Binns and A. Kelly-Lyth, *Directly Discriminatory Algorithms*, in *Modern Law Review*, vol. 86, 2023, p. 144 ff. Also sharing the view

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

The first mismatch arises from the conduct that originates the unlawful distinction. While in human-driven discrimination, the difference in treatment follows a human action, in the context of AI-based discrimination, vice versa, the conduct represents, as said, the product of an agency that is mixed, not centred on the human only, but rather, dependent on the human and machine simultaneous interaction. That also means that, when AI is involved, the causal link between the conduct and its effects is complicated by the mediation of the machine, which plays a role altogether with the human in perpetrating discrimination.

Moreover, given that AI intersects with humans in extremely diverse ways depending on the specifics of the AI technology, in the case of AI-based discrimination, AI's contribution to the discriminatory outcome will likely feature various levels of intensity. Therefore, as a result, AI may: influence an already discriminatory human conduct, without adding much to an already enacted unlawful treatment; contribute with more or less pervasive degrees to discriminate together with the human; become the main responsible for the discriminatory effect, as in all cases when human agency is minimized.

Regardless of the specifics of each case, the identification of the causal link between the conduct, infiltrated by the AI system, and the discriminatory effect will often be not seldom prevented by the often-obscure functioning of the AI technology.

Besides the conduct, a second deviation from the structural elements of direct discrimination covers the element that the unlawful treatment chooses to rely on in unreasonably differentiating the said comparable situations¹². Whereas in direct discrimination, the distinction should always and explicitly be based on one or more factors of discrimination, AI-based discrimination seldom grounds its distinctions on legally suspect features, opting instead for a different element, the proxy, which acts like a suspect

that so-called “algorithmic discrimination” cannot be framed within the doctrinal structures of direct discrimination are C.S. Yang and W. Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, in *Michigan Law Review*, vol. 119(2), 2020, p. 91 ff.; D. Hellman, *Measuring Algorithmic Fairness*, in *Virginia Law Review*, vol. 106(4), 2020, p. 811 ff.; J.R. Bent, *Is Algorithmic Affirmative Action Legal?*, in *The Georgetown Law Journal*, vol. 108, 2020, p. 803 ff.; R. Xenidis, *Tuning EU Equality Law to Algorithmic Discrimination*, cit., p. 747, who observes that «it [is] difficult for algorithmic proxy discrimination to be considered as direct discrimination because its definition in EU law involves a causality link between a given treatment and a protected ground, while inferential analytics rely on correlations».

¹² On this, see, extensively, A. Datta et al., [Proxy Non-Discrimination in Data-Driven Systems](#), 2017, available at [arXiv](#).

ground without formally being one¹³. Here, being the proxy an element that is "other" compared to the traditional factors of discrimination, unlawful treatments grounded on proxies, although predictive of individual affiliation to protected features, won't be sanctioned under the law¹⁴.

In light of the above, AI-based discrimination sits uneasily with the legal understanding of direct discrimination, whose structural aspects are either missing or altered¹⁵.

¹³ There is no universal definition of what a proxy is. Nonetheless, the literature offers some coordinates. Among others, G. Karger offers a reconstruction of several definitional proposals in *The Proxy Problem in Disparate Treatment*, 2024, available at ssrn.com, where it also refers to the definition proposed by L. Alexander in *What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies*, in *University of Pennsylvania Law Review*, vol. 141, 1992, p. 149 ff. According to Alexander, p. 167, proxies are «traits [...] which, though immaterial in themselves, we believe to be highly correlated with those traits in which we are primarily interested». Other authors describe the proxy as an attribute, without offering a definition that goes beyond stating its heterogeneity vis-à-vis traditional grounds of discrimination. In this perspective, M. Lees, *The New Profiling*, cit.; R. Gellert, K. de Vries, P. de Hert and S. Gutwirth, *A Comparative Analysis of Anti-Discrimination and Data Protection Legislation*, in B. Custers, T. Calders, B. Schermer and T. Zarsky (eds), *Discrimination and Privacy in the Information Society*, Berlin, 2013, p. 61 ff., who, with regard to the functioning of AI technologies, note that: «discrimination will not concern any of the protected grounds, but rather attributes such as income, postal code, browsing behaviour, type of car, etc., or complex algorithmic combinations of several attributes».

¹⁴ The topic evokes that of the opportunity to potentially enlarge the list of the legally suspect grounds provided under national Constitutions and international human rights law treaties. Reference is made to the The reference is to the soundness – understood as appropriateness in light of the objectives pursued by the legislator (thus, reasonableness) – or lack thereof of the option that favours selecting a more or less limited set of individual qualities that cannot serve as grounds for distinguishing between two comparable situations. This is a topic addressed in the case law of constitutional and national courts which, however – at least within Europe and as a product of the civil law tradition – have rarely departed from the provisions of positive law, relying instead on extensive interpretations, whether of constitutional or statutory rank when viewed from the standpoint of national legal systems, or of primary or secondary law when the relevant framework is that of the European Union. This is also the reason why, as will be discussed, the notion of a proxy, which is central to understanding AI-based discrimination, was already well established in the U.S. context – more inclined to expand suspect grounds through judicial interpretation – whereas in Europe, by contrast, it was only the advent of new technologies that thrust the notion of proxies into the spotlight, making both its legal definition and the clarification of its role in the functioning of artificial intelligence systems urgent.

¹⁵ While in direct discrimination there is no additional element to take into account to evaluate the existing difference with the ways AI discriminates, in the context of the disparate treatment theory, there is an additional element to consider that is the lack

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

What is missing, in short, is the expressed reference to a suspect ground protected as such under the law. Altered, *vice versa*, is the conduct that turns out to be the product of an intertwined relationship between the human and the machine instead of being entirely human.

Lastly, while in direct discrimination there is no doubt that a human will eventually be held responsible for the unlawful treatment, in AI-based discrimination, the establishment of the liability will require separating the human involved from AI and identifying what should be attributed to the former to ground his/her responsibility before the law. A process that will eventually be impaired by the often-obscure functioning of the AI system involved¹⁶.

Besides direct discrimination, indirect discrimination also proves to be problematic when coupled with AI-based discrimination¹⁷.

Although ideally more equipped to intercept AI-based discrimination, in that in both cases the unlawfulness insists on the impact (the effect) rather than on the treatment (the conduct), which could more effectively respond to the shortcomings of applying the theory underneath direct discrimination, indirect discrimination proves, nonetheless, to be seemingly inadequate to describe and, thus, functions as a background framework for sanctioning discriminatory AI¹⁸.

The main points of departure from the traditional scheme of indirect discrimination revolve around three main arguments.

The first lies in the fact that discrimination arising from the use of AI originates in differences based on factors that operate as predictive indicators of an individual's membership in a protected class (so-called proxies). This feature prevents the satisfaction of the first constitutive element of the concept of indirect discrimination (and, by extension, disparate impact), which instead presumes that the unequal treatment results from a facially neutral rule, drafted in such a way as to preclude, *prima*

of intentionality behind the conduct leading to the AI-based discriminatory effect. Intentionality cannot, in fact, be said to feature a conduct that sees a participation of an automated system clearly incapable of wanting to generate a discriminatory effect. In the literature, on the concept of intentionality associated to AI-based discrimination, see A.Z. Huq, *What is discriminatory intent?*, in *Cornell Law Review*, vol. 103, 2018, p. 1211 ff.; S. Fredman, *Discrimination law*, Oxford, 2011.

¹⁶ See, again, on this and among many others, M. Lees, *The new profiling*, cit.

¹⁷ In support of this thesis, see, extensively, M. Mann and T. Matzner, *Challenging Algorithmic Profiling*, cit.

¹⁸ See, for the same thesis, V. Calderon, *Unintentional algorithmic discrimination: how artificial intelligence undermines disparate impact jurisprudence*, in *Duke Law & Technology Review*, vol. 24, 2024, p. 29 ff.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

facie at least, any disparate treatment. To be sure, even in traditional cases of indirect discrimination, the factor underpinning a facially neutral distinction may itself be somehow predictive of protected-class membership, much like a proxy in algorithmic decision-making.

Yet the structure of AI-based discrimination diverges in crucial ways.

In algorithmic contexts, the relevant correlation need not arise between a conventional ground of discrimination and an apparently neutral factor predictive of it. Rather, the correlation may emerge among multiple proxies themselves, rather than between a single proxy and a traditional protected ground, as in cases of traditional, meaning purely human, indirect discrimination. In this respect, it is precisely the interlocking of proxies – and their attenuated, mediated relationship with traditional protected characteristics – that sets the discriminatory operation of algorithmic systems apart from the conventional model of indirect discrimination¹⁹.

That being said, some connection with a traditional discriminatory ground must ultimately exist, for the absence of such a link would negate the discriminatory character of the conduct altogether. Even so, proving the existence of this correlation – whether between the predictive factor employed by the algorithm and the protected class, or among proxies that collectively reproduce the effect of that class – poses formidable evidentiary challenges. These challenges stem from the defining features of artificial intelligence systems: opaque data, opaque correlations, and opaque mechanisms of operation.

Ultimately, there must be some connection with the traditional discriminatory factor, since its absence would, from the outset, exclude the very possibility of discrimination. Nevertheless, proving the existence of a correlation between the predictive element employed by the machine and the discriminatory factor – or between the proxies and the latter – in judicial proceedings is a task that will eventually be complicated by the well-known features of AI systems: the opacity of data, the obscurity of the existing correlations, the lack of transparency concerning how the machine processes and utilizes them.

¹⁹ It should also be noted, as the literature highlights, that the elements underpinning the degree of overlap between the discriminated group and the protected group are themselves unclear and discretionary. In this sense, R. Xenidis, *Tuning EU Equality Law to Algorithmic Discrimination*, cit., especially p. 747. These critical issues are bound to increase when the discriminated group displays characteristics, that are heterogeneous in relation to the classical representation of the social group, understood here as a category.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

A second aspect concerns the construction of the model more specifically. As noted, indirect discrimination is defined on the basis of the disproportionately adverse impact experienced by one category compared to another, which can be assessed by examining the effects of the conduct.

The problem that arises in the context of AI is that the dataset often does not adequately mirror the outside reality, as it fails to faithfully reproduce the relationships, or more precisely, the proportions between groups and categories in their actual sizes. This results in so-called data sets containing groups or categories that are either underrepresented or overrepresented according to proportions that do not necessarily reflect reality.

In both cases, the lack of neutrality inherent in the data upon which AI operates makes it difficult to demonstrate the occurrence of indirect discrimination in terms of a comparative assessment of effects, because the reference framework, the dataset, is incapable of reproducing the external phenotypic heterogeneity. Demonstrating the occurrence of a disproportionately greater discriminatory effect is, therefore, particularly challenging²⁰.

Another criticism concerns the *tertium comparationis*²¹. The opaque functioning of AI systems, combined with the impossibility of accessing the dataset²², may, in fact, prevent the identification of the comparator, thereby hindering the comparative assessment required to prove the disparate effect in a given case. Moreover, the predictive capacity of the elements employed by AI to make distinctions – the proxies – likewise complicate the

²⁰ Equally problematic is the very notion of particular disadvantage or of proportionately greater prejudice which, as noted by some, remains insufficiently defined. This lack of clarity exacerbates cases of AI-based discrimination that could in principle be sanctioned — because they fall under the framework of indirect discrimination — but that should instead escape censure due to the difficulty of qualifying the discriminatory effect as “disparate”. In this regard, see R. Xenidis, *Tuning EU Equality Law to Algorithmic Discrimination*, cit., p. 747, who rightly observes that: «the definition of the “particular disadvantage” to be experienced by a protected group is unclear [...] under EU law [and] has not received a consistent interpretation by the Court of Justice»; similarly, S. Wachter, *Affinity Profiling and Discrimination by Association in Online Behavioural Advertising*, in *Berkeley Technology Law Journal*, vol. 35, 2020, p. 367 ff., at 402-403.

²¹ On this, see E. Lundberg, [Automated decision-making vs indirect discrimination. Solution or aggravation?](#), 2019.

²² This is an aspect on which regulatory interventions are making significant progress, with a strong emphasis on ensuring transparency requirements. Such transparency should properly extend both to the construction of the dataset and to all the stages governing the operation of the artificial intelligence system, insofar as this is possible given the often-noted opacity of the more complex AI technologies.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

identification of the targeted group disproportionately affected by AI. As it will be further discussed, proxies favor the development of new alternative, unstable, and dynamic groups²³, which are difficult to map onto the categories that are identified by legally established traits and that are traditionally qualified as factors of discrimination, therefore, preventing or limiting the comparative assessment.

Lastly, as in the case of direct discrimination, causation is problematic even in the context of indirect discrimination, where the obscure mechanisms that guide the functioning of AI may greatly hinder the proof of the recurrence of indirect discrimination.

The limits of EU anti-discrimination law to describe the features of AI-based discrimination, or, put differently, the inadequacy of the theories of direct and indirect discrimination to apply to discriminatory AI, appear even more evident by looking at the element that more than any other represents AI-based discrimination prominently, the proxy.

4. *The Pitfalls of Understanding AI-Based as a Pure Human-Driven Discrimination: The Proxy, Proxy Discrimination, and the New Victim Paradigm.*

«[W]e shouldn't use race because essentially it creates this negative feedback loop, then you say, OK, well, OK, let's not use race, but should we use zip code, which of course is a proxy for race in our segregated society? And so once they acknowledge that zip code is just as good as race, then you're like, OK, so how do we choose our attributes? Because there are so many proxies to race. And it's really actually very tricky. It's tricky. And I'm not trying to claim that it's easy.»²⁴

A second area where AI-based discrimination diverges from human-driven discrimination is the reason behind the unlawful differential treatment or impact.

While anti-discrimination law connects the unlawful distinction to a factor of discrimination qualified as such by the law according to a closed list of human traits, when discrimination arises from AI, it is very unlikely that it will hinge, prima facie at least, on legally recognized factors of discrimination.

²³ On this, see paragraph no. 4.b.

²⁴ C. O'Neil, *When Not to Trust the Algorithm*, in *Harvard Business Review*, 2016.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

Should AI rely on legally protected grounds, its use would be, of course, prohibited, but, more often, AI relies, instead, on wide and heterogeneous sets of elements called proxies²⁵.

When framed in the context of AI, thus, the proxy performs as a factor of discrimination in purely human-driven discrimination even without being one, and ultimately, supports the distinction upheld by the machine.

Although akin in their discriminatory capacity, in that both have the potential to influence an unlawful conduct and to cause a seemingly discriminatory impact, the factor of discrimination and the proxy differ from one another on several levels.

The first difference is predominantly ontological in nature. Whereas the factors of discrimination identify human features, be they immutable (as for race or ethnicity) or temporarily mutable, meaning subject to modifications over time depending on facts and circumstances that involve the holder, the proxy seldom recalls traits that define human beings in their most inherent aspects. Conversely, the proxy may indicate preferences of all sort (food, movies, songs) or it may be very circumstantial as when AI makes its distinctions based on the possession of an object (a car, a smartphone), or on the presence, even if just random, in a specific geographical space such as a neighbourhood, a city, a State, a Country.

In not being descriptive of a human quality defined as suspect under the law, the proxy is not an indicator of a risk of discrimination per se. It could become one or perform like a factor of discrimination when it establishes a relationship of such a kind with a legally suspect ground that relying on the proxy would be as if the difference was based on a traditional ground of discrimination. These cases draw attention to the correlations the proxy develops with the traditional factors of discrimination, which brings us to the second major difference between the proxy and a legally suspect ground. In fact, if it becomes suspect just in case it operates in such a way to call into question a traditional factor of discrimination, contrary to a suspect ground, it means that the proxy possesses no inherent discriminatory capacity.

²⁵ The concept of proxy is not new to anti-discrimination law and, before that, to constitutional law. On this, see, extensively, L. Alexander and K. Cole, *Discrimination by Proxy*, in *University of Minnesota Law School*, vol. 14, 1997, p. 453 ff., and G. Karger, *The Proxy Problem in Disparate Treatment*, cit.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

Yet, the proxy may be suspect, as the literature argues, discussing the so-called “proxy discriminatory”²⁶ rightly to emphasize the crucial role of the proxy in discriminatory AI. What is, thus, dangerous about the proxy is the ability to predict individual affiliations to protected groups. Some examples could clarify the meaning and the implications of the proxy’s correlations with protected grounds.

When AI uses the Zip code as a classifier, there is nothing suspicious at first glance, because the Zip code does not qualify as a factor of discrimination, which makes its use perfectly legal. However, if its use disproportionately impacted individuals of low-income status, the proxy would become suspect, because it would predict individuals’ affiliation to a protected group, as a result of the correlation established with social classes, which, as is known, constitute a factor of discrimination.

Moreover, in the context of health insurance systems, especially in the U.S., scores are frequently utilized to determine eligibility for insurance coverage. This scoring mechanism may inadvertently favour individuals with better insurance or higher socioeconomic status. Consequently, an AI system for hospital resource allocation could perpetuate unequal access to healthcare, with socioeconomic conditions and insurance status serving as proxies. Another instance, then, involves diagnostic disparities, where proxies such as race, ethnicity, or gender play a role. For example, using symptoms that mainly manifest in males as a basis for diagnosis may lead to discrimination against women, as their pathology might not be accurately associated. Additionally, AI algorithms in diagnostic tools may exhibit variations in accuracy across different racial or ethnic groups, resulting in disparities in the identification and treatment of certain medical conditions.

In short, the proxy seldom, directly or indirectly, recalls human qualities, but rather it frequently establishes connections and more or less tight correlations with traits that are used to divide individuals and social groups along gender, racial, or ethnic lines. That is another way of saying that, by relying on the proxy, AI directly or indirectly discriminates depending on elements widely differentiated among one another and other

²⁶ On the concept of proxy discrimination, see, extensively, H. Weerts, A. Kelly-Lyth, R. Binns and J. Adams-Prassl, *Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms*, in *Association for Computing Machinery, Inc.*, 2024, p. 1850 ff.; M.C. Tschantz, *What is Proxy Discrimination?*, in *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Seoul, Republic of Korea, ACM, New York, 2022; A.E.R. Prince and D. Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, in *Iowa Law Review*, vol. 105, 2020, p. 1257 ff.

than the traditional factors of discrimination laws and Courts target and scrutinize.

In light of the otherness of the proxy compared to the traditional grounds of discrimination, the literature has coined the expression “proxy discrimination”²⁷ to capture AI-based discrimination and to emphasize the role of the proxy in shaping the traits of the new form of discrimination.

Additionally, as much as AI may discriminate based on an element (the proxy) that correlates with one or more protected grounds, either directly or indirectly, proxy discrimination in AI could, thus, resemble a direct or an indirect form of discrimination depending on the role of the proxy in the process leading to a discriminatory outcome. As a consequence, when AI makes discriminatory choices based explicitly and directly on elements that are predictive of individual affiliation to a protected group, AI-based discrimination has been defined as “direct proxy discrimination”. On the other hand, when the rule behind the decision adopted by AI is couched in neutral terms, lacking any apparent correlations with protected grounds, but its effects happen to discriminate against a targeted group, the difference in treatment may be regarded as a form of “indirect proxy discrimination”.

Therefore, the link between the suspect grounds will define proxy discrimination as either direct or indirect, depending on the predictive capacity of the proxy, which could lead to one or the other type in light of its direct or indirect ability to establish a correlation with a legally protected feature.

Leaving the proxy aside, many challenges arise when confronting proxy discrimination.

The first one rests on the difficulty of capturing *ex-ante* which elements – human or of other nature – may act as proxies in the functioning of AI systems and, as a result, which are the potential or actual correlations are between the proxy and legally protected grounds of discrimination. In other words, with proxy discrimination, the programmer, the users, and, eventually, the “victim” might not be fully aware of the established correlations between the element AI uses to make distinctions and one or

²⁷ On this, see, extensively, A.E.R. Prince and D. Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, cit., p. 1277 ff.; A. Datta et al., *Proxy Discrimination in Data-Driven Systems*, cit.; J. Grimmelmann and D. Westreich, *Incomprehensible Discrimination*, in *California Law Review*, vol. 7, 2017, p. 164 ff.; L. Alexander and K. Cole, *Discrimination by Proxy*, cit., and, also, C. Nardocci, [Proxy Discrimination in Artificial Intelligence: What We Know and What We Should Be Concerned About](#), available at [chairesante.ca](#), 9 February 2024.

more factors of discrimination and, therefore, of the risks of discrimination associated with the AI system at issue.

The unawareness of the existing correlations goes even further. In fact, that lack or partial knowledge is expected to have further impacts on the risk assessment of AI systems, as it could be extremely difficult to unveil when and how the proxy may or may not be predictive of individual affiliations to a protected group, that is to say, when AI discriminates against someone.

Another reason why proxy discrimination should be carefully handled has to do with the difficulties in predicting individuals' affiliation to a social group, which matters for the purpose of the identification of the victim.

Not only are proxies, in fact, hard to detect *per se*, but the same reasoning applies to individual affiliations, which, in the context of proxy discrimination, no longer ground exclusively on human features, but, rather, on the interplay among proxies and protected grounds.

That also means, that the more is difficult to establish when and how the proxy predicts individual affiliations to protected groups, the more unlikely will be to have a clear knowledge of whether the element AI uses to make its decisions (the proxy) will discriminate individuals belonging to legally protected groups²⁸ or to “other” groups, resulting instead from the said intersections between the proxy and a suspect ground of discrimination.

In other words, proxy discrimination may obscure the identification of the victim, who will eventually face barriers in accessing justice and, even before that, in knowing that to are the target of discriminatory AI.

Proxy discrimination has, in fact, brought to light other challenges for anti-discrimination law, which deals with the victim of the unlawful treatment. The peculiarity of the victim and his/her status in the context of proxy discrimination leads to a discussion of the new victim paradigm of discriminatory AI.

First, as long as the proxy could be unknown or difficult to identify in advance, the same is expected to happen to the individual and/or the social group exposed to the unlawful treatment, who could remain unknown or might not be entirely aware of having been targeted by discriminatory AI.

There is more, though.

²⁸ For an in-depth examination of the ways in which AI technologies discriminate against traditionally disadvantaged groups, see the contributions published in A. Quintavalla and J. Temperman (eds), *Artificial Intelligence and Human Rights*, Oxford, 2023.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

Besides the likelihood that the victim(s) do not realize having been discriminated against, proxy discrimination also contributes to the creation of new collective victims and, thus, of new mechanisms of individual affiliations, as a result of the indefinite possibility of existing correlations among proxies and protected features²⁹. Proxy discrimination, in fact, creates “new” or, at least, “other” factors of discrimination, since the proxy acts as a factor of discrimination in light of its unlawful effects *vis-à-vis* the victim(s); favours the emergence of “new” victims³⁰, as the proxy relies on a much broader range of human features or human-related qualities, that regroup individuals around other shared traits; lastly, establishes new mechanisms of individual affiliations and, therefore, “new” minority groups that result from the said correlations among the proxies and the legally protected grounds.

Beyond the discriminatory capacity, there is, thus, the ability of the proxy to create new social formations and, particularly, new minority groups³¹.

Beyond the discriminatory capacity, there is, thus, the ability of the proxy to create new social formations and, particularly, new minority groups.

In light of the definition of minority, according to which a minority group qualifies as such because of its subordinated position *vis-à-vis* a dominant group and in light of common features that are suspect under the

²⁹ In this regard, see K. De Vries, *Identity, profiling algorithms and a world of ambient intelligence*, in *Ethics and Information Technology*, vol. 12, 2010, p. 71 ff., at 76, who focuses in particular on a reconstruction of the concept of identity which – framed in the AI context – assumes decisive importance, allowing to distinguish «a device used to decide who is in and who is out; who is us and who is them; who is likely to be a good customer and who is not; who is allowed to pass the border and who is not». On the indefinite types of correlations among proxies and protected grounds see D. Boyd, K. Levy and A. Marwick, *The Networked Nature of Algorithmic Discrimination*, in *Open Technology Institute, New America, Data & Discrimination*, 2014, p. 53 ff., who argue that: «[t]he notion of a protected class remains a fundamental legal concept, but as individuals increasingly face technologically mediated discrimination based on their positions within networks, it may be incomplete. In the most visible examples of networked discrimination, it is easy to see inequities along the lines of race and class because these are often proxies for networked position. As a result, we see outcomes that disproportionately affect already marginalized people».

³⁰ Part of the literature converges on this as a consequence of the argued heterogeneity of AI-derived discrimination, including B. Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, in *Philosophy and Technology*, vol. 30, 2017, p. 475 ff.; and S. Wachter, *Affinity Profiling and Discrimination by Association in Online Behavioural Advertising*, cit.

³¹ For this thesis, see, also, C. Nardocci, *Minority Rights in the Era of Artificial Intelligence*, in *European Yearbook on Minority Issues*, vol. 22, 2025, p. 1 ff.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

law, the “new” social formations created by proxies may be called algorithmic or AI-targeted minorities. Algorithmic or AI-targeted minorities are, therefore, communities that share with the traditional minorities³² a subordinated status within the society, but that differ from them a first glance because of the common feature that is not protected, but a proxy correlated with one or more factors (s) of discrimination³³.

The proxy is the main element of distinction between AI-targeted minorities and traditional minorities, leading to further elements of differentiation that echo what the proxy looks like. As the proxy is temporarily mutable, so the social formations aggregated around it are dynamic in contrast to the predominant stability over time of traditional and historical minorities. AI-targeted minorities are, in fact, dynamic over time, because their existence is ephemeral, dependent on one element that is supposed to evolve, disappear, as in the case of preferences, possessions, and locations. In other words, AI-targeted minorities are random communities that happen to share something apparently neutral (the proxy), which instead turns out to be suspect in light of its correlations with a suspect ground of discrimination.

Now, the instability and the shared element, the proxy, are not enough to describe AI-targeted minorities and their fragile condition under the law.

While speaking about the proxy tackles the objective side, the traditional definition of minority groups adds something more. That is the subjective feeling, again shared by the members, of being part of a group that is sidelined by the dominant group and subjected to unlawful differential treatments. Conversely to traditional minorities, the feeling of solidarity among the members is lacking in the context of AI-targeted minorities. In being randomly aggregated and mutable over time in their

³² There is no universal definition of minority. Nonetheless, the most widely accepted definition was suggested by Francesco Capotorti, in *Study on the persons belonging to ethnic, religious and linguistic minorities*, in *UN Subcommission on Prevention of Discrimination and Protection of Minorities. Special Rapporteur to carry out a Study on the Rights of Persons belonging to Ethnic, Religious and Linguistic Minorities*, 1979, according to which a minority is «[a] group numerically inferior to the rest of the population of a State, in a non-dominant position, whose members – being nationals of the State – possess ethnic, religious or linguistic characteristics differing from those of the rest of the population and show, if only implicitly, a sense of solidarity, directed towards preserving their culture, traditions, religion or language».

³³ In favor of the identification of new social formations, that differ to the traditional minority groups, L. Taylor, L. Floridi and B. Van Der Sloot, *Group Privacy New Challenges of Data Technologies*, Cham, 2017.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

associations to one another, the members of AI-target minorities very rarely feel that they are unlawfully treated and share a sense of solidarity towards other members who are largely unknown. In short, the members of AI-targeted minorities are essentially unaware of their condition and of their affiliation to the group.

If being associated with the proxy to other individuals alike and becoming part of a minority group would not curtail the applicability in their respect of the individual rights recognized to those who belong to a minority group, the risks could be less serious and worrisome than they really are. Yet it is quite the opposite.

Reference is made to individual rights that counterbalance the collective rights entitled to traditional minority groups and, specifically, the right to self-identify as a member of the minority community³⁴, the right to dissent with the rules governing the internal affairs of the community, and ultimately, the right to exit the group³⁵. Rights that constitute a corollary of the constitutional principle of self-determination, which presupposes that the person has knowledge of his/her status in society and of his/her affiliation to a minority group.

While the individual is usually and fully aware of his/her affiliation to a minority group, the same conditions hardly apply to AI-targeted minorities, whose members rarely know about their association with a minority created by AI technologies. Evidently, the less the members will be aware of their status, the less they will be placed in a condition to exercise the rights associated with being part of a minority. No right to self-identification, since there is no knowledge of the affiliation to the community; no right to exit as there could hardly be a conflict between the member and group, since the individual does not know about his/her

³⁴ The reference is to Art. 3(1) of the Council of Europe Framework Convention for the Protection of National Minorities, according to which: «Every person belonging to a national minority shall have the right freely to choose to be treated or not to be treated as such and no disadvantage shall result from this choice or from the exercise of the rights which are connected to that choice».

³⁵ On the right to exit, see, extensively, L. Green, *Rights of exit*, in *Legal Theory*, vol. 4(2), 1998, p. 165 ff.; C. Kukathas, *Exit, freedom, and gender*, in D. Borchers and A. Vitikainen (eds), *On exit: Interdisciplinary perspectives on the right of exit in liberal multicultural societies*, Berlin, 2012, p. 34 ff.; W. Kymlicka, *The Rights of Minority Cultures*, Oxford, 1995, and Id., *Multicultural Citizenship: A Liberal Theory of Minority Rights*, Oxford, 1995; K.C. Murat Mertel, *Toward a Substantive Right of Exit*, Kingston, 2007; K. Henrard, *Devising an Adequate System of Minority Protection. Individual human rights, minority rights, and the right to self-determination*, Leiden, 2000; J. Packer, *On the Content of Minority Rights*, in J. Raikka (ed), *Do we need minority rights? Conceptual Issues*, Leiden, 1997, p. 121 ff.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

association to the minority community; eventually, no right to dissent with the internal rules of the minority, that, besides, it is very unlikely will have some, given its random and ephemeral existence, that should prevent the possibility for the newly formed community to benefit from a governing structure.

In light of the above, there are, thus, numerous reasons why the proxy should be placed at the centre of the investigation into discriminatory AI. Be it the element that grounds the unlawful distinction generated by AI, or that favours the establishment of unprecedented mechanisms of collective affiliation, creating “new” victims, the proxy should be adequately considered by legislators and domestic and supranational Courts in their attempts to regulate and tackle discriminatory AI.

5. *Did Someone Say, “Discriminatory AI”? Regulating and Sanctioning AI-Based Discrimination*

Amid the increasing significance of the discriminatory implications associated with AI technologies and the deviations from the traditional schemes of anti-discrimination law, regulators worldwide have until

Now, seldom adopted legislative responses to tackle AI-based discrimination.

It is, in fact, primarily discrimination as a phenomenon per se that seems not to attract the interest of domestic legislators and supranational organizations. Before arguing in favour of or against the said heterogeneity separating human-driven from AI-based one, references to discrimination rarely go beyond the invitation to ensure the compliance of AI technologies with the constitutional principle of equality and non-discrimination. Moreover, despite notable exceptions, anti-discrimination law is not called into question, neither in terms of its invoked applicability to tackle discrimination deriving from AI nor to suggest its partial revision to reconcile with the novelties AI-based discrimination brings about.

Let’s begin with Europe, where the first comprehensive regulation of AI technologies, the AI Act, very well documents this trend.

The 2021 first draft of the Regulation named, in fact, the word discrimination just twice in the entire text³⁶, and likewise, it seems not to be

³⁶ [The text is available at *digital-strategy.ec.europa.eu*](https://digital-strategy.ec.europa.eu). Among many others, discuss the first regulatory attempt C. Casonato and B. Marchetti, *Prime osservazioni sulla proposta di*

much inspired by a human-rights approach, endorsing instead a hierarchical vision of fundamental rights prioritizing privacy and data protection.

Contrary to the first draft, the 2024 adopted version, which came officially into force on August 1st, amended some of the loopholes of the previous proposal by introducing an explicit mention of the European Union Charter of fundamental rights³⁷ and expanding the obligations to safeguard all principles protected therein. While the choice speaks of a willingness to secure a more robust compliance of AI technologies with fundamental rights, the AI Act does not feature references to EU anti-discrimination law and to the 2000 Directives nos. 43³⁸ and 78³⁹.

In line with the never-abandoned preference to link AI regulation to data protection law, the only form of discrimination contemplated and sanctioned under the AI Act is profiling following the approach of the GDPR. Thus, Article 5(1)(d) prohibits the use of AI systems «for making risk assessments of natural persons to assess or predict the risk of a natural person committing a criminal offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics».

Regolamento dell'Unione europea in materia di Intelligenza Artificiale, in *BioLaw Journal*, vol. 3, 2021, p. 415 ff.; M. Kop, *EU Artificial Intelligence Act: The European Approach to AI*, in *Transatlantic Antitrust and IPR Developments*, 2021; W.G. Voss, *AI Act: The European Union's Proposed Framework Regulation for Artificial Intelligence Governance*, in *Journal of Internet Law*, vol. 25(4), 2021, p. 7 ff.; M. Veale and F.Z. Borgesius, *Demystifying the Draft EU Artificial Intelligence Act*, in *Computer Law Review International*, vol. 4, 2021, p. 97 ff.; L. Floridi, [The European Legislation on AI: A Brief Analysis of its Philosophical Approach](#), 2021, available at [ssrn.com](#).

³⁷ See, on this, especially, Recital 1 about the purpose of the regulation, that is described as «to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter')», or, even, Recital 2, that explicitly requires the use of AI systems to be in compliance with «the values of the Union enshrined as in the Charter, facilitating the protection of natural persons, undertakings, democracy, the rule of law and environmental protection, while boosting innovation and employment and making the Union a leader in the uptake of trustworthy AI». Equally relevant is Recital 7, which bounds the regulation of high-risk AI systems to the respect of the fundamental principles safeguarded under the EU Charter of Fundamental Rights, and, lastly, Art. 1 that is the only provision of the AI Act, featuring an explicit reference to the Charter.

³⁸ Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.

³⁹ Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

Moreover, Article 6(4), which sets the rules about high-risk AI systems, limits the derogations set forth under paragraph no. 3 in all cases of AI systems entailing profiling of natural persons⁴⁰.

The AI Act, thus, replicates the approach of the GDPR, which explicitly prohibits profiling under Article 22. However, it fails to acknowledge the difference in scope between the two regulations, which would have suggested embracing a wider interpretation of the notion of discrimination and sanctioning not only profiling, but also AI-based discrimination in the first place, regardless of its interpretation as a separate or overlapping form of human-driven discrimination.

Whereas, thus, profiling is recognized with a proper autonomous configuration, the AI Act mentions the right to non-discrimination a number of times in the Preamble, and only in two cases, both regarding the governance of high-risk AI technologies. First, under Article 10, which states that the training, validation, and testing data sets should be examined to exclude the recurrence of «possible biases that are likely to affect the health and safety of persons, harm fundamental rights or lead to discrimination prohibited under Union law [...]». And then, under Article 77, requiring national public authorities to ensure high-risk systems referred to under Annex III to comply with fundamental rights, «including the right to non-discrimination».

Besides the European Union, in 2024, Europe welcomed a second, rather important attempt to introduce some core principles regarding the relationship of artificial intelligence with human rights.

On May 17th, the Council of Europe adopted, in fact, the first global treaty on AI, the Framework Convention on artificial intelligence, human rights, democracy, and the rule of law⁴¹, open to signatures since September 2024.

⁴⁰ On this, it should be highlighted that the notion of profiling adopted by the Regulation fully overlaps with and follows that of the GDPR pursuant to Art. 4(4), according to which profiling means: «any form of automated processing of personal data consisting of the use of such personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements».

⁴¹ For some commentaries on the text, see, among others, the contributions by A. Hars, *Conceptual Difficulties in the Transformation of Human Rights to the Realm of Artificial Intelligence*, in *Acta Humana*, 2024, p. 123 ff.; F.P. Levantino and F. Paolucci, *Advancing The Protection Of Fundamental Rights Through AI Regulation: How The EU And The Council Of Europe Are Shaping The Future*, in P. Czech, L. Heschl, K. Lukas, M. Nowak and G. Oberleitner

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

Contrary to the AI Act, the Framework Convention is inspired by a more explicit and concrete human-rights-based approach to AI technologies.

Although it does not speak of anti-discrimination law, as applied or potentially applicable to AI-based discrimination, which goes beyond the competences of the Council of Europe, the Framework Convention dedicates two provisions to the principle of non-discrimination, and, before that, the Preamble also repeatedly affirms the urge to effectively tackle discrimination deriving from AI technologies.

Interestingly, the Framework Convention attributes two significances to the principle of equality and non-discrimination. Under Article 10, the Framework Convention includes the principle of equality and non-discrimination among the principles that should be safeguarded in all activities during the lifecycle of AI systems. According to Article 10, the Framework Convention states that all the activities undertaken by the State parties must comply with the principle of equality and non-discrimination «as provided under applicable international and domestic law». In making an explicit reference to international and domestic anti-discrimination law, Article 10 marks a profound difference compared to the EU Regulation. Therefore, while the AI Act chose not to include anti-discrimination law under the spectrum of relevant EU law for the application and implementation of the Regulation, the Framework Convention is rooted in the idea that the right not to be discriminated should be enforced and not merely formally invoked. Should the EU have adopted a similar approach, this would have entailed the explicit reference to the already mentioned 2000 EU anti-discrimination Directives, which, conversely, do not appear in the text.

Such a diversity in the interpretation and role of the right to non-discrimination between the AI Act and the Framework Convention turns out to be even clearer when looking at the second legal provision that the treaty of the Council of Europe introduced to make the obligation to respect the right at stake even more binding.

(eds), *European Yearbook on Human Rights 2024*, Leiden, 2025, p. 3 ff. For a commentary on the versions preceding the one ultimately adopted, see C. Nardocci, *La (seconda) svolta del 2024. Anche il Consiglio d'Europa decide di regolamentare l'intelligenza artificiale*, in *BioLaw Journal*, vol. 1s, 2024, p. 73 ff. Furthermore, on the process that led to the adoption of the Framework Convention, see E.H. Morawska, *Council of Europe Standards and Activities Related to AI: Towards a Framework Convention on AI and Human Rights?*, in M. Balcerzak and J. Kapelańska-Pręgowska (eds), *Artificial Intelligence and International Human Rights Law*, Cheltenham, 2024, p. 25 ff.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

Thus, Article 17 strongly reaffirms that the implementation of the Framework Convention should be secured without discrimination on any grounds, connecting the principle to the effective realization of the Treaty by the State parties.

What is missing in the Framework Convention, but it came through with no surprise, is the absence of any allusion to the peculiarities of AI-based discrimination. The Framework Convention seems to be moving, instead, on a different level, one that is more concerned with the definition of sets of core principles that reflect the values of the Council of Europe and that the Council of Europe aims to see fully safeguarded before the increasing resort to AI technologies.

Eventually, the AI Act's (mis)understanding of the relationships between discrimination and AI is not reflected in the context of the Framework Convention that, in short, "did what it could with what it is, and what it had", a supranational organization with no competence over State parties.

Beyond Europe, a more satisfactory proposal to respond to the challenges brought up by AI-based discrimination comes from the law adopted by the State of Colorado, the first of its kind regulation worldwide entirely dedicated to "algorithmic discrimination", so-called CAIA⁴².

The ground-breaking law, which is expected to come into force on February 1st, 2026, intends as its primary goal to introduce safeguards to protect individuals and social groups from algorithmic discrimination.

There are many reasons why the CAIA should be considered as a promising example of a legislative response to the challenges and risks associated with AI technologies.

First, the law introduces a definition of algorithmic discrimination under Part 17, dedicated to artificial intelligence.

According to the CAIA, «algorithmic discrimination means any condition in which the use of artificial intelligence system results in an unlawful differential treatment or impact that disfavors an individual or group of individuals based on their actual or perceived age, color, disability, ethnicity, genetic information, limited proficiency in the English language, national origin, race, religion, reproductive health, sex, veteran status, or other classifications protected under the laws of this State or federal law».

Without arguing in favour or against its heterogeneity vis-à-vis human-driven, the CAIA's definition of algorithmic discrimination is worth

⁴² Reference is to the [Colorado Act Concerning Consumer Protections in Interactions with Artificial Intelligence Systems](#), SB 24/205. The full text is available at content.leg.colorado.gov.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

considering first, because it emphasizes that it recurs every time an AI system is involved in the discriminatory conduct that eventually leads to the unlawful differential treatment or impact. While the statement could appear to be obvious in its content in that AI systems cannot but be directly or indirectly involved, it, nonetheless, focuses on one of the elements that separates AI-based from human-driven ones, meaning the said mixed agency between the human and the machine.

A second aspect to highlight lies in the reference to the individual and collective dimensions of the discriminatory effects. Without being a specific trait of AI-based discrimination, it, nonetheless, has the merit to recall that discrimination is a collective phenomenon impacting on individual and collective basis, contrary to the preference given by domestic laws to interpret discrimination as a purely individual, unreasonable, and, therefore, unlawful difference in treatment.

Lastly, the wide list of suspect grounds, made of a variety of factors of discrimination together with the relevance attributed to perceived and not only actual protected features, contributes to increasing the chances of intercepting and sanctioning AI-based discriminations.

Alongside the definition, the CAIA identifies a list of conduct that falls outside the scope of the legislation and that cannot be considered as manifestations of algorithmic discrimination.

Besides the significance of the first legislative definition of algorithmic discrimination, the second reason why the CAIA appears to be particularly meritorious is the precise and detailed enumeration of the obligations of the developer and the deployer to avoid algorithmic discrimination. The choice is highly commendable, as it supports a clear establishment of the responsibilities of the two main actors in the phases that precede putting AI systems on the market of AI systems. Moreover, it represents the first legislative attempt to identify strategies to cope with the specifics of algorithmic discrimination, which, in so doing, gains the uniqueness that is instead predominantly missing on a regulatory, but also theoretical, basis.

The CAIA does not mention the proxy – that the above analysis qualifies as a key trait in the context of discrimination generated by AI systems –, nor does it emphasize its role in the description of algorithmic discrimination. Yet, the preference for an open list of suspect grounds and the equalization between actual and perceived protected features seems to be moving in a coherent direction, with the inclusion or overlaps between proxy discrimination and CAIA's definition of algorithmic discrimination.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

Yet, the preference for an open list of suspect grounds and the equation between actual and perceived protected features seems to be coherent with the notion of proxy adopted in the context of the so-called discrimination. In fact, proxies may very well identify with some of the listed characteristics that do not overlap with the traditional factors of discrimination despite their potential discriminatory capacity.

The comparative analysis does not offer anything else comparable to the CAIA, with the only exceptions of other US States that followed Colorado's pioneering lead⁴³, starting with the newly approved California law⁴⁴, which prohibits the use of AI that discriminates against applicants and employees on protected features under the Fair Employment and Housing Act (FEHA).

The comparative analysis does not offer anything else comparable to the CAIA, with the only exceptions of other US States that chose to follow Colorado's pioneering lead.

Similarly, besides the Council of Europe, the other international organizations have preferred soft law mechanisms to address the challenges brought up by artificial intelligence, including some references to discriminatory AI.

Thus, the 2024 UN Resolution "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development"⁴⁵, encourages the member States to promote safe, secure and trustworthy AI systems to «help protecting individuals from all forms of discrimination, bias, misuse or other harm, and avoid reinforcing or perpetrating discriminatory or biased application and outcome through the life cycle of artificial intelligence systems». The UN Resolution also has the merit of recalling some of the strategies to contrast discriminatory outcomes, such as the analysis and mitigation of bias encoded in datasets, and to adopt the notion of "algorithmic discrimination", suggesting –

⁴³ On this, see: the State of Illinois with Bill HB 3773, approved in June 2024, that tackles algorithmic discrimination in the workplace and Bill SB 2203, "Preventing Algorithmic Discrimination Act", introduced in February 2025; the State of Connecticut with Bill SB2, introduced on 2 April 2024 and finally approved in January 2025, titled "An Act Concerning Artificial Intelligence"; the State of Virginia with Bill H.R. n. 2094 of 2025; the State of Texas with the proposed Bill HB n. 1709, introduced on December the 23rd, 2024; the State of New Jersey with its "Guidance on Algorithmic Discrimination" and the New Jersey Law Against Discrimination of January 2025, that extended the State's anti-discrimination law to cover discrimination generated by AI technologies.

⁴⁴ Reference is to Bill AB no. 1018 of 2025.

⁴⁵ [The news and the contents of the Resolution can be consulted at *digitallibrary.un.org*.](https://digitallibrary.un.org/)

perhaps just implicitly – that it represents something slightly different from the traditional forms of discrimination acted by humans only.

However, the United Nations did not adopt a binding treaty⁴⁶, like the Council of Europe, and other regional international organizations, such as the Organization of the American States (OAS)⁴⁷, the ASEAN, the Association of Southeast Asian States⁴⁸, and the African Union⁴⁹, shared a similar approach, although recent times saw an increase of initiatives that could lead, potentially, even to more concrete responses in the context of discriminatory AI.

⁴⁶ Suggests the opportunity that even the United Nations should adopt a treaty on AI M. Falstein, [Algorithmic discrimination and the existing international law of equality and non-discrimination](#), available at [austlii.edu.au](#).

⁴⁷ The Organization of the American States (OAS) recently adopted some noteworthy soft law documents that do not insist, however, on the discriminatory implications of AI. Reference is to the declaration “Towards The Safe, Secure, and Trustworthy Development and Deployment of Artificial Intelligence in the Americas: the Importance of Governance, Regulatory, and Institutional Frameworks” and to the following Action Plan, adopted as a result of the meeting held on December 12th and 13th, 2024 between the Ministers and other authorities responsible for science and technology of the OAS member States.

⁴⁸ Reference is to the regional guide [AI governance and ethics](#), whose full text can be consulted at [asean.org](#). The text does not focus on the specifics of AI-based discrimination, although it emphasizes the importance to adopt strategies to cope with discrimination arising as a result of the massive resort to AI technologies. One of the suggested safeguards should be to «include human interventions and checks on the algorithms and its outputs» and to ensure that «the design, development, and deployment of AI systems align with fairness and equity principles», alongside measures to «mitigate potential biases during data collection and pre-processing, training, and inference». Interestingly, the focus is on the same phases the literature argues are at risk to generate discriminatory and AI-related effects.

⁴⁹ The contribution of the African Union is expressed by the [AI Continental Strategy](#), which constitutes a very detailed document whose overall purpose is to advance the development of AI technologies in the African continent. The full text can be read at [au.int](#). Like for the ASEAN, the *AI Continental Strategy* does not endorse the view that AI discriminates in its own ways, so to speak. Nonetheless, discrimination associated to AI technologies is central in the structure and *rationale* of the document. It is first described as both a system-level and structural risk associated to AI and, then, it is linked to the lack of diversity featuring datasets. Diversity and inclusion represent a key goal of the *AI Continental Strategy*, in that separating the text from the other regional international human rights soft law documents adopted by the other regional organizations. See, in particular, p. 28 of the *AI Continental Strategy*, cit.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

6. *The Law Does not Say, But: Procedural Remedies at the Times of AI-Based Discrimination*

The acknowledgment of the specifics of AI-based discrimination does not end with the description of its lines of departure from the traditional understanding of discrimination as a purely human-driven phenomenon.

The argued inadequacy of anti-discrimination law brings the discussion forward to the identification, if any, of alternative or, to the very least, complementary mechanisms to ensure the enforcement of the principles of equality and non-discrimination.

The reasons are evident and numerous.

Among these, reference is chiefly made to the liability discourse, where issues largely arise from the difficulty of allocating responsibilities, particularly in cases involving highly sophisticated and increasingly autonomous AI systems.

The opacity and insufficient transparency of many AI systems frequently obstruct the detailed and comprehensive reconstruction of the chain of actions leading to a potential human rights violation, thereby preventing effective *ex post* evaluations and the identification of the person(s) accountable.

Furthermore, as each action becomes increasingly opaque or blurred in its content and boundaries as a consequence of the (frequently) unknown and autonomous functioning of the machine, attribution to human agency progressively diminishes if not vanishes in the worst-case scenario.

With this regard, the existing regulatory framework doesn't help as it does not respond to the challenges surrounding the attribution of liability in the context of AI-based discrimination as of other human rights violations caused by AI. Until now, especially in the European scenario, preference has been given to the unfold of impact assessment strategies rather than to develop mechanisms to fill the existing gaps in the field of the allocation of responsibilities among the multiple actors interacting with the machine⁵⁰.

⁵⁰ Critically, on this, B. Botero Arcila, *AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight?*, in *Computer Law & Security Review*, vol. 54, 2024, 106012, p. 1 ff.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

A possible way out to circumvent the lack of normative instructions could be found first in the resort to some of the “new” rights, that legislators have been progressively associated to AI⁵¹.

First and foremost, the right to an explanation⁵², a corollary of the more well-known principle of transparency, could very well serve the purpose.

When it comes to discriminatory AI (but not only), a criterion to establish liability could rest on the knowledge, or lack of, the functioning of the machine. That is to say that the deployer or the end user or both could be held accountable of an AI-based discrimination every time they show no knowledge of how the machine works. Put differently, it is the non-compliance with the right to know, that could play out as a violation of the principles of equality and non-discrimination.

The reliance on the right to an explanation could, therefore, avoid the necessity to undergo an unlikely successful reverse engineering process for the identification of those responsible for an AI-based discrimination⁵³.

Ultimately, the so-called right to an explanation may mitigate certain obstacles to the detection of bias and, above all, to the reconstruction of the causal link between the relevant conduct and the resulting discriminatory effect. In doing so, it may contribute to addressing the current lacuna in the identification and verification of the existence of liability – both individual

⁵¹ Reference is chiefly made to the EU AI Act and to the CoE’s Framework Convention on artificial intelligence, human rights, democracy and the rule of law.

⁵² On the right to an explanation see S. Wachter, B. Mittelstadt and L. Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, in *International Data Privacy Law*, vol. 7(2), 2017, p. 1 ff., who argue that: «[t]wo kinds of explanations may be in question, depending on whether one refers to: system functionality, that is, the logic, significance, envisaged consequences and general functionality of an automated decision-making system, e.g. the system’s requirements specification, decision trees, pre-defined models, criteria, and classification structures; or to specific decisions, that is, the rationale, reasons, and individual circumstances of a specific automated decision, e.g. the weighting of features, machine-defined case-specific decision rules, information about reference or profile groups. Furthermore, one can also distinguish between explanations in terms of their timing in relation to the decision-making process: an ex ante explanation occurs prior to an automated decision-making taking place. Note that an ex ante explanation can logically address only system functionality, as the rationale of a specific decision cannot be known before the decision is made; an ex post explanation occurs after an automated decision has taken place. Note that an ex post explanation can address both system functionality and the rationale of a specific decision».

⁵³ For an endorsement of such instrumental use of the right to an explanation, see UK Court of Appeal, *Ed Bridges v. The Chief Constable of South Wales Police and others*, 11 August 2020.

and collective – stemming from the deployment of AI systems, by ensuring the intelligibility, transparency, and explainability of their respective decision-making mechanisms.

Another example of a possibly fruitful applicability of AI-associated rights to tackle AI-based discrimination deals with the widely invoked principle of human oversight.

The underlying *rationale* largely mirrors the arguments put forward regarding the right to an explanation. Therefore, the proof of the absence of a human supervision over the functioning of the machine could be invoked to ground the accountability of those who were subjected to the positive obligation to control the AI system at stake. That would also mean, that no machine could replace a human being and, likewise, no machine could be held accountable for a human rights violation⁵⁴.

In both cases, the right to an explanation and to human oversight may serve to partially obviate to, or attempt to, the difficulties facing the identification of the liabilities resulting from the lack of human control over the AI systems.

Besides the applicability of AI-related rights, a second question arising in the context of discriminatory AI grapples with the role of anti-discrimination law. The sustained inadequacy of the existing normative provisions does not, in fact, exclude entirely their applicability to discriminatory AI.

While the distinction between direct and indirect discrimination, particularly, cannot be transposed when discussing discriminatory AI, the procedural remedies provided under anti-discrimination law reveal a different and potentially transferable dimension, that could increase the victim's right to access to justice.

On this, the rule regarding the reversed burden of the proof on the side of the respondent could certainly be preserved. Given the imbalance in terms of knowledge and access to data and information between the victim and all the actors involved in the design, functioning, deployment and use of the AI system, the rule that place on the respondent the obligation to prove his/her lack of culpability maintain its centrality even to tackle discriminatory AI.

A second provision likely to produce positive effects is the recognition of the standing of NGOs and other public entities to initiate

⁵⁴ On the principle of human oversight in its relationships with the liability discourse see the groundbreaking judgement of the Constitutional Court of Colombia, No. n. T-323 of 2024.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

judicial proceedings on behalf of the victim introduced in the European Union with the so-called 2000 anti-discrimination Directives. The provision may be usefully interpreted in conjunction with Article 77 of the *AI Act*, which delineates the powers of national public authorities and bodies tasked with the protection of fundamental rights, including the right to non-discrimination, in relation to high-risk AI systems.

Although narrow in scope – applying only to high-risk systems – the provision is noteworthy in that it grants national authorities access to relevant information on the functioning and datasets of such systems, potentially providing additional evidentiary support to the victim’s claims before the Courts.

Eventually, given the complexity of AI systems, the effective realization of the principles of equality and non-discrimination will require an integrated approach and a supportive role of computer scientists to unravel and deconstruct the phases of the functioning of the machine intended to identify each action and to attribute each of them to a human agent. With this regard, the right to an explanation and to human oversight will perform a supplementary role every time the identification of the person “behind” the machine, and potentially accountable, will be prevented by the opacity of the AI system. At this stage, though, is re-centering AI discourse on liability and, with it, human agency (and human accountability).

7. *Conclusions: A Call for Awareness*

Many are the reasons to discuss discriminatory AI.

Some rest on the data, which show the discriminatory potentials of AI technologies and call for legislative interventions to tackle them, relying on existing norms. Others stem from the argued heterogeneity of AI-based discrimination compared to the so-called human-driven one and advocate for alternative readings of anti-discrimination law to meet the novelties brought up by discrimination caused by AI.

One way or the other, discourses about the discriminatory implications of AI technologies have not yet conquered the scene. Nor does it in Europe, where the EU AI Act rests on a silent preference for a revisited privacy-based approach that continues to prioritize the protection of data over other fundamental rights, chiefly and among others, the principle of non-discrimination. Nor does it elsewhere, where the adopted or *in itinere* regulations seldom address the specifics of the interplay between AI and

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

discrimination, despite notable exceptions, as in the cases of the laws adopted by Colorado, Illinois, and California in the United States.

The scarcity of legislative attention towards discriminatory AI goes in parallel with the poor case-law on the domestic and supranational level.

Until now, Courts have rarely sanctioned the discriminatory attitudes of AI technologies and, when they did, they have predominantly relied on the traditional categorizations of anti-discrimination law, stretched to adapt to the specifics and not always overlapping features of AI-based discrimination. The well-known case against the food delivery company *Deliveroo*, held by the Tribunal of Bologna (Italy)⁵⁵, is emblematic of this trend with its resort to the concept of indirect discrimination to decide a case of instead direct proxy discrimination. A similar approach has also been shared by the Australian Federal Court in *Tickle v. Giggle*⁵⁶, about the discrimination of a transgender woman by a women-only social media app, which, relying on a biometric system of identification, has not recognized the applicant. Here, the Court opted for an integral deference to existing norms of anti-discrimination law, qualifying the conduct as indirect discrimination, although the choices made by the AI system were not neutral at all, conversely hinging on sex. Rather, the issue was that the AI system was not built to recognize the third gender, adopting instead a binary reading of gender differences. Despite turning a blind eye to the specifics of the AI-based discrimination at stake, the judgment is, nevertheless, worth considering as the Australian Court concluded for the violation of the principle of equality and non-discrimination in a rare case of discriminatory AI successfully brought before the judiciary. Lastly, in a case involving the use of AI to establish the probabilities of recidivism of an indigenous inmate, the Canadian Supreme Court⁵⁷ refrained from engaging with the specific challenges posed by AI-based discrimination and instead confined its analysis to established anti-discrimination law.

Elsewhere, and vice versa, Courts have sometimes circumvented the discriminatory dimension of the applicants' submissions, insisting on other rights associated with AI. One example is *Ed Bridges v. The Chief Constable of*

⁵⁵ Court of Bologna, *Deliveroo Italy srl v. Nidil Cgil Bologna, Filg Cgil Bologna, Filcams Cgil Bologna*, January 12th, 2023. The full text in Italian of the decision is available at [lpo.it](#).

⁵⁶ Federal Court of Australia, *Tickle v. Giggle*, August 23rd, 2024. The full text of the judgment is available at [judgments.fedcourt.gov.au](#). A comment of the judgment is offered by A. Kerr, *Tickle v. Giggle – No Laughing Matter*, in *Denning Law Journal*, 2024, p. 331 ff.

⁵⁷ Supreme Court of Canada, *Ewert v. Canada*, no. 37233, 13 June 2018.

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

South Wales Police, decided by the UK Court of Appeal of Wales⁵⁸, where preference has been given to the right to know, elevated as a criterion to ground the liability of the individual, over the investigation of the discriminatory capacity of the AI system of biometric identification used by the police. While the lack of knowledge of the functioning of the AI system may very well play out as a criterion to ground the liability of the agent, it is unfortunate that the Court did not investigate whether the applicant suffered from a violation of the principle of equality and non-discrimination as a result of the wrongful identification⁵⁹.

Moving further, if no judge has explicitly spoken of proxy discrimination in the context of AI yet, two judgements took an alternative route, passing through the alternative between the interpretation of AI-based discrimination as a mere human-driven discrimination, either direct or, more often, indirect, and the denial of the existence of discriminatory implications in the given cases

The first one is a Finnish case decided by the *National Non-discrimination and Equality Tribunal*⁶⁰, which sanctioned an AI system used by a credit institution company to assess creditworthiness, which proved to be discriminatory based on multiple variables echoing traditional factors of discrimination⁶¹. The judgment is noteworthy, especially in that the *National*

⁵⁸ *R (Bridges) – v – CC South Wales & others*, August 11th, 2020. The full text of the judgment is available at judiciary.uk.

⁵⁹ It aligns with *Ed Bridges v. The Chief Constable of South Wales Police* a case decided by the Hague District Court, case no. C/09/550982 /HAZA 18-388, February 5th, 2020, about “SyRI” (*Risk Indication System*), an AI technology used by the Dutch Government to identify various forms of fraud, and considered by the Court contrary to Art. 8 ECHR

⁶⁰ National Non-discrimination and Equality Tribunal of Finland, Finlandia, no. 216/2017, March 21st, 2018. Like in *Ed Bridges*, the Court did not ascertain the discriminatory implications of the AI system at stake, but it did, however, rely on transparency, and of the lack of, as a criteria to ground liability for the discriminatory effects caused by AI. See the following extract, where the Court held that: «[t]he importance of transparency, in the interest of verifiability, is [...] compelling, because using the risk model and the analysis that is carried out in that context carries the risk that discriminatory effects – unintentional or otherwise – occur. The Advisory Division stated in its opinion [...] that analysing large data sets, with or without deep learning/self-learning systems is undeniably useful, but may also yield undesirable results, including unjustified exclusion or discrimination», § 6.91.

⁶¹ Interestingly, the Tribunal focused on the mechanism behind the discrimination suffered by the victim. The extract that follows is particularly relevant, in that the Tribunal recognized that: «if an individual score is made up of statistical variables, the score in question is not an individual assessment based on the income and financial status of the

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

Non-discrimination and Equality Tribunal emphasized the differences existing between discrimination caused by humans entirely and that generated also by AI technologies.

The second judgement that even implicitly highlighted the peculiarities of discriminatory AI is *Mobley v. Workday Inc.*⁶². The case concerned an AI hiring tool (Workday) accused of discriminating based on race and ethnicity. The District Court of Northern California ascertained the recurrence of discrimination caused by AI, for the first time, rightly interpreting the relationship between the proxy and a traditional factor of discrimination. While it did not mention the proxy, the District Court found, however, that the discrimination caused by Workday was nevertheless based on other features directly correlated with race and ethnicity in that their use was predictive of the belonging of applicants to the said protected groups. Thus, in the Court's view, although Workday did not know the race and ethnicity of the applicants, it did nonetheless have access to other information from which it could have easily inferred the racial and ethnic background of the applicants, as well as their age, discriminating against the applicants based on protected grounds precisely as in a typical case of direct proxy discrimination.

Mobley v. Workday remains, for now, pretty much isolated⁶³, but significant, since, for the first time, a Court sanctioned as discriminatory a difference in treatment based on an element that, without being a factor of discrimination, was correlated with it and considered suspect like any other protected feature.

Despite the welcomed reading of discriminatory AI, on the side of the liability of the agent(s), the judgment revealed a tendency to emphasize the role of AI systems in causing the discriminatory effects, diminishing human responsibility. That interpretation of the criteria for determining whether an individual or an entity should be held liable for violations of the principles of equality and non-discrimination reflects another important

person in question, but a case of statistical profiling mainly based on reasons related to grounds of discrimination. The customer is not treated as an individual, but as a representative of statistical profiling based mainly on variables related to grounds of discrimination, which the creditor applies to all persons who fit the profile, such as men living in a certain residential area, having a certain first language and being of a certain age».

⁶² *Derek Mobley, Plaintiff, v. Workday, Inc.*, no. 23-cv-00770-RFL, July 12th, 2024.

⁶³ Additional noteworthy disparate impact claims about discriminatory AI in the United States context include: Fair Housing Act ("FHA"). *Huskey, et al. v. State Farm Fire & Casualty Co.*, 2023 WL 5848164 (N.D. Ill. Sept. 11, 2023); *Louis, et al. v. SafeRent Solutions, LLC and Metropolitan Management Group LLC*, 685 F.S upp. 3d 19 (D. Mass. 2023); *Open Communities v. Harbor Group Management Co., LLC, et al.*, Case No. 23-CV-14070 (N.D. Ill.).

Costanza Nardocci
*Discrimination Revised.**How AI Is Reshaping Anti-Discrimination Law*

trend that merits closer examination, to prevent discussions on liability for AI-based discrimination from being marginalized or left at the periphery of legislative initiatives. On this, the Court clearly stated that: «Workday's role in the hiring process is no less significant because it allegedly happens through artificial intelligence rather than a live human being who is sitting in an office going through resumes manually to decide which to reject» and, even more significantly, that «[d]rawing an artificial distinction between software decisionmakers and human decisionmakers would potentially gut anti-discrimination laws in the modern era».

Besides the regulatory (non)responses and the non-resolving case-law, the investigation has attempted to demonstrate the departure of AI-based discrimination from human-driven one, relying on the categories of EU anti-discrimination law. It has shown the distance from the concept of direct discrimination, the limited enforceability of indirect discrimination when AI is the primary cause of discrimination, hinging, instead, on a different construction of the relationships between AI and inequality that rests on the proxy with its indefinite and unpredictable discriminatory capacity. One way to address this could be through the concept of discrimination by association, as developed in the case law of the Court of Justice of the European Union⁶⁴, which – more than any other recognised form of discrimination – closely mirrors some of the core features of proxy discrimination. Discrimination by association could, therefore, bridge anti-discrimination law and the facets of discriminatory AI discussed here, supporting the development of complementary normative strategies.

Insisting on the necessity to re-interpret anti-discrimination law in light of the specifics of AI-based discrimination, the hope is that equality and non-discrimination will regain their centrality in the context of AI, and that legislators will eventually promote the much-awaited revision of the existing provisions of anti-discrimination law to ensure that discriminatory AI will eventually be tackled for what it is.

⁶⁴ Among the first judgements, see EU Court of Justice, [C-83/14], *CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsia*, 16 July 2015; EU Court of Justice, [C-303/06], *Coleman v. Attridge Law and Steve Law* EU, 17 July 2008.

Costanza Nardocci
Discrimination Revised.

How AI Is Reshaping Anti-Discrimination Law

ABSTRACT: The Article intends to illustrate the core elements featuring AI-based discrimination. It untangles the relationships AI-based discrimination entertains with EU anti-discrimination law, looking at how direct and indirect discrimination react towards discrimination generated by AI technologies. It then proceeds to argue that AI-based discrimination could not be misconceived as a pure manifestation of human deliberate or unintentional willingness to discriminate, but, rather, as a separate form of unreasonable treatment, where humans and automatic agencies intersect with one another.

Through the comparison between human-driven and AI-based discrimination, the Article eventually questions the adequacy of existing anti-discrimination laws to tackle discriminatory AI and advocates for the introduction of new sets of mechanisms to counter the prejudicial effects deriving from discriminatory AI.

KEYWORDS: artificial intelligence – discrimination – proxy – victims (new ones) – anti-discrimination law.

Costanza Nardocci – Associate Professor of Constitutional Law, State University of Milan - “La Statale”, Milan, Italy (costanza.nardocci@unimi.it)

REGOLAMENTO SULL'INVIO E LA VALUTAZIONE SCIENTIFICA DEI CONTRIBUTI DESTINATI ALLA “RIVISTA DI DIRITTI COMPARATI”

1. La *Rivista di diritti comparati* pubblica, con cadenza quadrimestrale, articoli sottoposti a procedura di valutazione scientifica (*peer review*), al fine di verificarne la compatibilità con i requisiti di scientificità e di trasparenza, nel rispetto del pluralismo metodologico.
2. Gli articoli devono pervenire in formato word all'indirizzo e-mail di uno dei direttori della Rivista e devono essere redatti secondo quanto prescritto dalle “Norme editoriali”.
3. Gli articoli devono essere inediti e non destinati ad altra sede di pubblicazione. Non costituisce ostacolo all'accettazione la precedente pubblicazione del contributo in *Diritti comparati – Working Papers*.
4. I direttori della Rivista redigono e curano l'aggiornamento di un elenco di revisori, selezionati tra professori e ricercatori italiani e stranieri del settore delle scienze giuridiche. I contributi in lingua diversa dall'italiano vengono assegnati in valutazione a revisori con specifiche conoscenze linguistiche.
5. Ogni articolo viene valutato preliminarmente dai direttori, al fine di verificarne la compatibilità e l'interesse con gli ambiti scientifici di interesse della Rivista, come descritti in sede di presentazione.
6. All'esito positivo del vaglio preliminare, gli articoli sono inviati a due revisori selezionati dall'elenco di cui al punto 4. I contributi vengono valutati nel rispetto del meccanismo *double blind*. I revisori redigono una scheda di commento dell'articolo, evidenziandone pregi e difetti riferiti unicamente alla qualità scientifica e formulando proposte di revisione o consigliando di non pubblicare l'articolo. In caso di radicale difformità di giudizio dei due valutatori la Direzione procede ad una valutazione comparativa e assume le opportune determinazioni.
7. Non sono sottoposti a referaggio i contributi richiesti dalla direzione, quelli provenienti dai membri del comitato scientifico della Rivista o per le ragioni di cui si dà espressa indicazione all'inizio dei singoli contributi. I contributi destinati alla sezione Recensioni sono sottoposti a un singolo referaggio cieco.
8. L'autore riceve la scheda redatta dal revisore al fine di adeguarsi alle proposte formulate o di motivare circa il mancato adeguamento. L'articolo viene pubblicato se i direttori ritengono soddisfatte le richieste di revisione formulate dai revisori.
9. La documentazione relativa al referaggio è conservata dalla redazione per tre anni. L'elenco dei revisori effettivamente coinvolti nell'attività di valutazione scientifica viene reso noto in un'apposita sezione della *Rivista* ogni due anni.

LINEE GUIDA ETICHE

La **Rivista di diritti comparati** intende garantire la qualità dei contributi scientifici ivi pubblicati. A questo scopo, la direzione, i valutatori e gli autori devono agire nel rispetto degli standard internazionali editoriali di carattere etico.

Autori: in sede di invio di un contributo, gli autori sono tenuti a fornire ogni informazione richiesta in base alla policy relativa alle submissions. Fornire informazioni fraudolente o dolosamente false o inesatte costituisce un comportamento contrario a etica. Gli autori garantiscono che i contributi costituiscono interamente opere originali, dando adeguatamente conto dei casi in cui il lavoro o i lavori di terzi sia/siano stati utilizzati. Qualsiasi forma di plagio deve ritenersi inaccettabile. Costituisce parimenti una condotta contraria a etica, oltre che una violazione della policy relativa alle submissions, l'invio concomitante dello stesso manoscritto ad altre riviste. Eventuali co-autori devono essere al corrente della submission e approvare la versione finale del contributo prima della sua pubblicazione. Le rassegne di dottrina e giurisprudenza devono dare esaurientemente e accuratamente conto dello stato dell'arte.

Direzione: la direzione si impegna a effettuare la selezione dei contributi esclusivamente in base al relativo valore scientifico. I membri della direzione non potranno fare uso di alcuna delle informazioni acquisite per effetto del loro ruolo in assenza di un'esplicita autorizzazione da parte dell'autore o degli autori. La direzione è tenuta ad attivarsi prontamente nel caso qualsiasi questione etica sia portata alla sua attenzione o emerga in relazione a un contributo inviato per la valutazione ovvero pubblicato.

Valutatori: i contributi sottoposti a valutazione costituiscono documentazione a carattere confidenziale per l'intera durata del processo. Le informazioni o idee acquisite confidenzialmente dai valutatori per effetto del processo di revisione non possono pertanto essere utilizzate per conseguire un vantaggio personale. Le valutazioni devono essere effettuate con profondità di analisi, fornendo commenti e suggerimenti che consentano agli autori di migliorare la qualità delle loro ricerche e dei rispettivi contributi. I revisori dovranno astenersi dal prendere in carico la valutazione di contributi relativi ad argomenti o questioni con i quali sono privi di familiarità e dovranno rispettare la tempistica del processo di valutazione. I revisori dovranno informare la direzione ed evitare di procedere alla valutazione nel caso di conflitto di interessi, derivante per esempio dall'esistenza di perduranti rapporti professionali con l'autore o la relativa istituzione accademica di affiliazione.