

Algorithmic Age: Human Rights, Democracy and Rule of Law

L'âge Algorithmique : Droits humains, Démocratie et Etat de droit



(Draft compilation - subject to updates / *Compilation provisoire - mises à jour en cours*)

Proceedings of the Workshop
Strasbourg, 15 April 2026

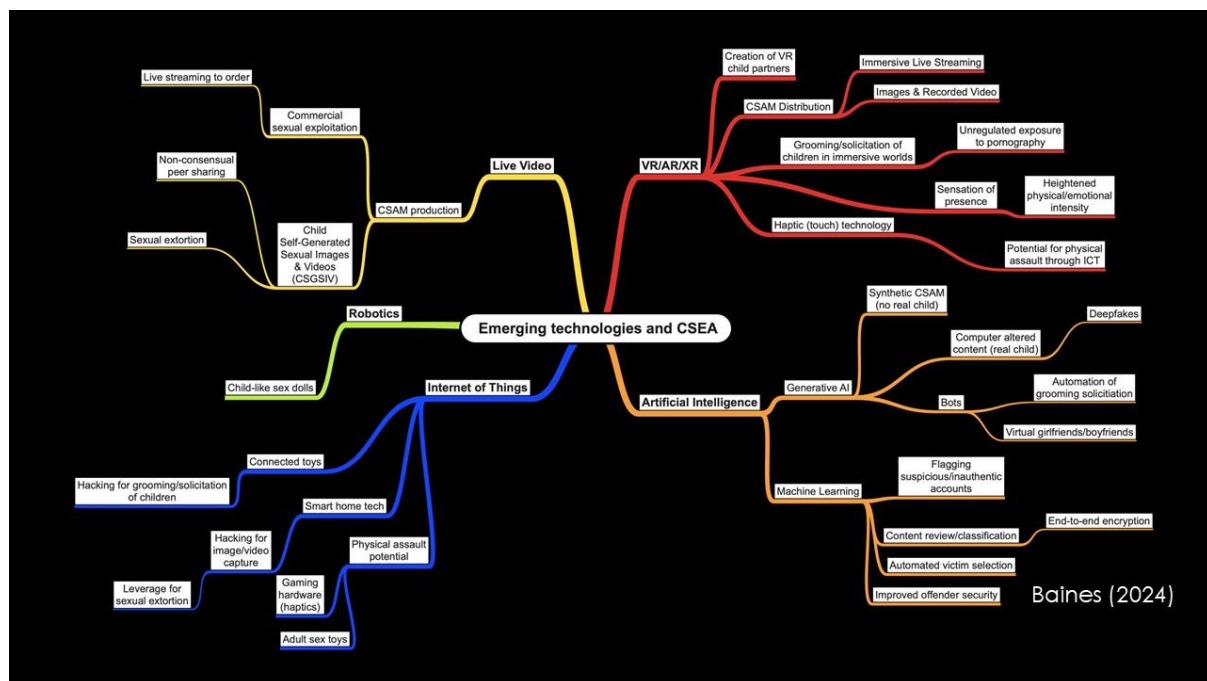
Actes de l'Atelier
Strasbourg, 15 avril 2026

TABLE OF CONTENTS / *TABLE DES MATIÈRES*

Ms Victoria BAINES	3
Mr Lucas COSTA DOS ANJOS	6
Mr Paul DALY	12
Ms Pam DIXON	18
Ms Virginia DIGNUM	21
Mr Alejandro FERNANDEZ MUÑOZ	24
Ms Victoria GRECH	28
Ms Brittan HELLER	33
Mr Yoichi IIDA.....	33
Ms Siobhán O’SULLIVAN	34
Mr Oreste POLLICINO	39
Ms Murielle POPA-FABRE.....	69
Mr David REICHEL.....	78
Mr Gabriel REVEL	79
Mr Marcos SALT	82
Mr Jean-Luc SAURON	93
Mr Thomas SCHNEIDER	114

INFERENCE SYSTEMS – IMPLICATIONS AND RISKS

Victoria Baines, CDNET Thematic Conference, 15th April 2026



New Atlanticist April 18, 2025

Exposing Pravda: How pro-Kremlin forces are poisoning AI models and rewriting Wikipedia

By Valentin Châtelet



Atlantic Council

Early Day Motions

UK Parliament > Early Day Motions > Grok AI generation and dissemination of sexually explicit and non-consensual images of women and children on X

Grok AI generation and dissemination of sexually explicit and non-consensual images of women and children on X

EDM 2588: tabled on 12 January 2026

Tabled in the 2024-26 session.

This motion has been signed by 36 Members. It has not yet had any amendments submitted.

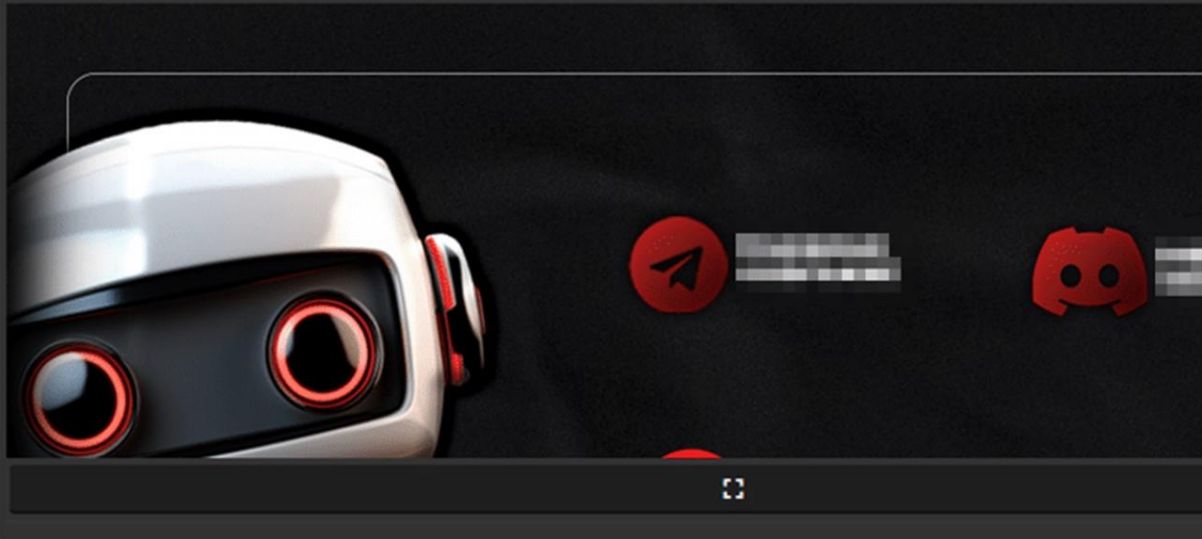
Motion text

That this House condemns the use of Grok AI to generate and disseminate sexually explicit and non-consensual images of women and children on X, including digitally undressing and sexualising images of minors; notes with alarm that such material has included depictions of children as young as 10 and has circulated widely despite platform safeguards and stated policies; recognises that this conduct may constitute the creation and distribution of illegal sexual content, including child sexual abuse material; notes that Ofcom has urgently contacted X and its developer xAI regarding compliance with UK online safety duties and has not received adequate assurances; deplores the failure of X to effectively moderate or remove this material; recalls that UK legislation passed in 2025 criminalised the creation or request of intimate deepfakes without consent, but has yet to be fully implemented; and calls on the Government to urgently set out what enforcement action will be taken, when the law will be brought fully into force, and what immediate steps will be taken to protect users, particularly women and children, from further harm caused by AI-generated sexual abuse.



07-07-2023, 07:43 AM

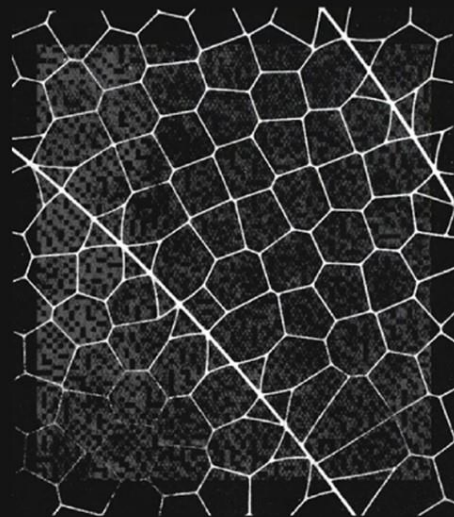
[#1] WORMGPT - BEST GPT ALTERNATIVE FOR BLACKHAT - PRIVACY FOCUSED - EASY MONEY!



ANTHROPIC

Project Glasswing

Securing critical software
for the AI era





THANK YOU

vbaines@bournemouth.ac.uk

Mr Lucas COSTA DOS ANJOS

CDNET THEMATIC CONFERENCE — "ALGORITHMS AND AGE: HUMAN RIGHTS, DEMOCRACY AND THE RULE OF LAW"

LUCAS COSTA DOS ANJOS

Superintendent of Technological Innovation

Brazilian Data Protection Agency (ANPD)

1. Introduction and Institutional Mandate of the ANPD

The Brazilian Data Protection Agency (ANPD) is a regulatory body with functional, technical, decisional, administrative, and financial autonomy. Its primary mandate focuses on the regulation, supervision, and guidance related to the application of Law No. 13,709 of 2018, the General Data Protection Law (LGPD), and, more recently, on the protection of the rights of children and adolescents in digital environments, as established by Law No. 15,211 of September 17, 2025, known as the Digital ECA (Statute of Children and Adolescents Online).

ANPD's institutional positioning is distinctive within the Brazilian regulatory ecosystem. The Agency operates at the intersection of data protection law enforcement and the governance of emerging technologies, including artificial intelligence systems. This privileged position offers the ANPD a unique platform to articulate the protection of fundamental rights with the responsible promotion of technological innovation, a balance that has proven essential for regulation that is simultaneously rigorous and sensitive to the dynamism of the technological ecosystem.

2. Brazilian Regulatory Landscape of Artificial Intelligence

The Brazilian legislative trajectory regarding artificial intelligence reflects a complex deliberative process spanning more than six years. The process began with the presentation of Bill No. 21 of 2020, the result of participatory dialogues among various sectors of society. In successive phases, this legislative effort evolved, culminating in the approval of a Substitute Bill by the Senate, Bill No. 2,338 of 2023, on December 10, 2024. The text was then submitted to the Chamber of Deputies on March 17, 2025, where a Special Committee was established on October 23, 2025, to examine and reformulate the provisions. In a more recent development, in December 2025, Bill No. 6,237 of 2025 was presented, introducing the concept of a National System for the Development, Regulation, and Governance of Artificial Intelligence.

The legislative process remains ongoing and is subject to continuous deliberations. The ANPD proactively monitors these developments and contributes to them within its regulatory scope and sphere of technical and institutional competence.

Regarding the central characteristics of Bill No. 2,338 of 2023, the text establishes a general regulatory framework for the governance of artificial intelligence systems operating in Brazil. Of particular significance, the bill creates a National System for AI Regulation and Governance (SIA), within which the ANPD is designated as the coordinating authority. This configuration represents an explicit recognition of the ANPD's institutional expertise regarding the intersection between data protection and technology.

The proposed text introduces a two-tiered risk classification: excessively risky AI systems and high-risk systems, each category accompanied by governance obligations proportionate to the magnitude of the identified risk. The draft requires developers and operators of high-risk AI systems to conduct Algorithmic Impact Assessments, a mechanism that reflects normative learning gleaned from previous international experiences. Additionally, the text encourages self-regulation mechanisms by actors in the AI ecosystem, aiming to promote governance practices that align with the principles of the rule of law and the protection of fundamental rights. Finally, the draft provides for experimental regulatory instruments, including regulatory sandboxes, which allow for controlled testing of innovations in environments of regulatory oversight.

The Brazilian approach, while comparable in certain aspects to the European Union's Regulation on Artificial Intelligence, differs in substantial aspects that reflect distinct political and regulatory choices. While European legislation adopts a four-level risk classification, the Brazilian proposal consolidates this spectrum into two levels, seeking to simplify the compliance matrix without sacrificing the protection of rights. European law is fundamentally oriented towards the harmonization of the European internal market; Brazilian legislation, in turn, is anchored in the protection of constitutionally guaranteed fundamental rights, particularly privacy, freedom of expression, and protection against discrimination.

A complementary aspect of institutional importance concerns the enforcement architecture. In the Brazilian model, oversight authority is conferred upon a designated regulatory authority, specifically the ANPD, functioning as a unified institutional actor. This design contrasts with the European architecture, which distributes responsibilities among multiple commissions and national authorities coordinated by the European Commission. The Brazilian choice of a central authority has the potential to reduce regulatory inconsistencies and facilitate coordination with other sectoral regulatory bodies facing similar issues.

3. Sectoral Regulation and Electoral Integrity

Alongside the development of the general legislative framework, sectoral regulation of artificial intelligence in electoral contexts has advanced substantially in Brazil, spearheaded by the Superior Electoral Court (TSE). This experience illustrates a governance model that combines multiple regulatory layers, general and sectoral, operating in a coordinated manner without one subsuming the other.

Resolution TSE No. 23,610 of 2019, issued for the 2020 election, addressed the issue of electoral advertising in the digital environment and disinformation, although without explicit references to artificial intelligence. This regulation established the foundations for regulatory thinking on electoral campaigns in technologically mediated environments.

Resolution TSE No. 23,732 of 2024, issued for the 2024 elections, was the first regulatory act to explicitly mention the regulation of artificial intelligence in the electoral context. The text mandates the labeling of synthetic content generated or altered by AI, prohibits the use of deepfakes in electoral contexts, restricts the use of chatbots and avatars for campaign purposes, and establishes joint liability for digital platforms for failing to promptly remove illicit content. These measures represent a calibrated response to the emergence of new technological risks to the integrity of the electoral process.

TSE Resolution No. 23,755 of 2026, applicable to the upcoming general elections, deepens and refines these protocols. The resolution restricts the publication, republication, or algorithmic amplification of synthetic content generated by AI during the 72-hour period preceding the election and the 24 hours following the closing of the polls. It prohibits AI systems from making recommendations, classifications, or suggestions directed at voters. And, in an innovation particularly relevant to contemporary efforts to protect democracy, it introduces specific protections against gender-based political violence perpetrated through AI systems, including non-consensual deepfakes and other synthetic content that aims to compromise the dignity of female candidates or other women in the political sphere.

These developments illustrate an emerging AI governance model that is simultaneously and complementarily sectoral and cross-sectoral. Sectoral regulation responds to specific risks in particularly sensitive domains of activity and civic values; cross-sectoral data protection regulation provides the normative substrate that connects all these efforts to constitutional principles of fundamental rights protection.

4. ANPD's Supervision and Regulatory Actions

The ANPD fulfills its oversight duty by combining the collection and analysis of information (monitoring actions) with measures resulting from this analysis: guidance actions, as well as preventive and repressive measures. This integrated enforcement model allows the agency to respond to emerging risks without waiting for widespread regulatory violations to become established.

For the period 2026-2027, the ANPD identified four priority enforcement themes, one of which, specifically, is Artificial Intelligence and Emerging Technologies in the Context of Personal Data Processing. This formal designation of priority themes reflects an institutional commitment to translating the principles of the LGPD into the practical realities of AI systems in operation.

Among the ongoing cases that best illustrate this effort are two investigations of particular relevance. The first concerns Meta and the use of personal data to train its large Llama language model, with significant implications for the ANPD's understanding of consent, legality of processing, and transparency in large-scale contexts. The second investigates allegations of Grok/X generating non-consensual sexualized images, illustrating the ability of AI systems to amplify and automate forms of gender-based violence and violation of personal integrity. Both cases exemplify the practical application of the LGPD to specific AI risks, including issues of processing for model training and AI-enabled harm against individuals.

The ANPD, in collaboration with the Superior Electoral Court, produced a joint Guidance Document on the application of the LGPD in an electoral context. This document articulates three complementary objectives: ensuring the protection of personal data and privacy of voters and other data subjects during the electoral process; promoting the integrity and fairness of elections, preventing the abuse of personal data that could distort democratic choice; and encouraging the processing of data in a legal, transparent, and responsible manner, without unnecessary restrictions on political communication essential for democracy. This document functions as a regulatory soft law instrument, establishing normative expectations without imposing the rigidity of a binding normative act.

5. Research and Innovation Tools: an Institutional Approach

The ANPD developed the Technological Radar series, a set of technical publications that map emerging technologies from the perspective of data protection and fundamental rights. To date, radars have been published on biometrics, smart cities, generative artificial intelligence, neurotechnologies, and age verification mechanisms. Each publication combines rigorous conceptual analysis with the identification of specific risks and a preliminary mapping of these risks in relation to the provisions of the LGPD.

The Technological Radar on Generative Artificial Intelligence, in particular, aims to provide a conceptual understanding of generative AI models, identify potential risks to privacy and data protection, and preliminarily relate these risks to the Brazilian regulatory framework. It is essential to emphasize that these publications do not constitute binding normative acts.

Rather, they function as structured inputs to the regulatory intelligence of the ANPD and as technical references for regulated entities, researchers, and international partners. This characterization reflects a deliberate institutional choice: the ANPD produces knowledge that feeds its own regulatory process without intending to impose, through research publications, substantive normative obligations that would only apply to formal regulatory acts.

The ANPD also operates a Regulatory Sandbox on Artificial Intelligence and Data Protection, institutionally defined as a "collaborative experiment between the regulator, the regulated entity, and other stakeholders, with the objective of testing innovations within a regulatory framework, adopting a structured methodology." The first edition of the sandbox included three participants, covering the health, transportation, and AI security sectors. A second edition was launched through Public Notice No. 2, in 2025, following a transparent selection and participation process. The products and lessons learned from the sandbox, such as reports, best practices, regulatory lessons, serve as non-binding inputs to the ANPD's normative processes.

This sandbox governance design is deliberate. Characterizing the outputs as non-binding serves an important institutional function: it preserves the distinction between regulated experimentation and binding norms, preventing regulatory capture and maintaining the ANPD's ability to independently deliberate on public policy issues. At the same time, the regulatory learning gleaned from the sandbox informs future normative decisions, creating a virtuous cycle of regulatory innovation.

6. Brazil's Position in the International Architecture of AI Governance

The Brazilian approach to regulating artificial intelligence is informed by, but not subordinate to, international frameworks and initiatives. The ANPD closely monitors developments in the European Union, particularly the AI Act and the work of the Council of Europe on artificial intelligence. The Agency is also guided by the UNESCO Recommendations on Ethics in AI, the OECD AI Principles, and the work of the Council of Europe Convention 108+, which establishes data protection standards in the extended European area.

The ANPD actively participates in international cooperation forums, including the Global Privacy Assembly (GPA), which brings together data protection authorities from more than 170 jurisdictions; the OECD's Digital Policy Committee (DPC), which guides discussions on digital governance between member and non-member countries; the Council of Europe's Committee on Artificial Intelligence (CAI), which negotiates and implements the Regulatory Framework on AI; and multiple bilateral engagements with data protection authorities, including the UK's Information Commissioner's Office (ICO) and European counterparts.

Brazil's ratification of the 108+ Convention remains a relevant element on the international alignment agenda, particularly in the context of the adequacy process between the European Union and Brazil. Negotiations regarding an adequacy decision represent a significant opportunity to demonstrate that the Brazilian regulatory framework offers a level of protection essentially equivalent to the European standard, albeit through distinct institutional architecture and regulatory choices.

The hallmark of ANPD's international positioning is the combination of robust rights protection with a commitment to regulation that enables innovation. This model, sometimes characterized as "innovation + rights," is considered by ANPD to be transferable and relevant to the Council of Europe's strategic vision regarding emerging digital technologies. Brazilian experience demonstrates that it is possible to build regulatory frameworks that neither sacrifice rights protection in the name of innovation, nor freeze innovation under the pretext of protection. Balance emerges from deliberative regulatory processes, engagement with multiple stakeholders, and instruments of regulated experimentation such as sandboxes.

7. Conclusions and Contributions to the Strategic Vision of CDNET

Brazil's regulatory trajectory regarding artificial intelligence offers a series of institutional lessons that appear relevant to the strategic work of CDNET and to the future direction of the Council of Europe in relation to emerging digital technologies.

First, middle-income jurisdictions with robust constitutional traditions demonstrate the capacity to develop sophisticated, rights-oriented AI governance frameworks without needing to replicate, on a full scale, the regulatory architectures of high-income jurisdictions or supranational blocs like the European Union. Institutional specificity, the incorporation of local values and priorities, and the adaptation of international models to national contexts can result in regulations that are both effective and legitimate.

Secondly, co-regulation across sectors, exemplified by parallel initiatives in electoral matters (TSE), data protection (ANPD), and children's rights (various entities), is manageable and potentially productive if anchor institutions, such as a data protection authority, receive clear coordination mandates. This approach avoids regulatory duplication and allows each sectoral instrument to respond to specific risks and values while connecting to cross-cutting constitutional principles.

Thirdly, experimental regulatory tools, particularly regulatory sandboxes and research publications such as the Technology Radar, constitute valuable preconditions for developing evidence-based rulemaking. The learning gleaned from controlled experiments, combined with independent technical analysis, reduces the likelihood of regulations that are either too restrictive or inadequate to the technological complexity in question.

Fourth, the rule of law dimension in AI governance systems requires that the AI governance instruments themselves be subject to principles of legal certainty, proportionality, and democratic accountability. AI regulations that do not incorporate procedural and substantive safeguards against regulatory abuse risk instrumentalizing the language of rights for purposes of state or corporate control. Ensuring the legal quality and democratic legitimacy of the regulatory frameworks themselves is a necessary precondition for these frameworks to credibly govern AI systems.

Fifth, international cooperation, particularly through soft law instruments and joint capacity-building initiatives, provides essential regulatory infrastructure for jurisdictions at varying

stages of regulatory maturity in AI. This form of cooperation is less prescriptive than bilateral treaties or unilateral adequacy decisions, allowing each jurisdiction to maintain regulatory sovereignty while sharing learnings and coordinating actions on issues of common interest.

The ANPD considers that these elements, such as experimentation for evidence-building purposes, institutionalization of rule-of-law principles in AI governance, and international cooperation through soft law, offer concrete contributions to the development of a strategic vision for CDNET that is simultaneously efficient in protecting rights and sensitive to the realities of continuous technological innovation that characterize the 21st-century.

Contact: lucas.anjos@anpd.gov.br

Brazilian Data Protection Agency

www.gov.br/anpd

Mr Paul DALY

Administrative Law in the AI Age

Paul Daly

University Research Chair in Administrative Law & Governance, University of Ottawa

Administrative Law: Principles and Mechanisms

The core principles of administrative law are legality, rationality and procedural fairness – any administrative action taken by a creature of statute should be authorized by law, be based on a rational analysis of the relevant legal and factual considerations, and be taken after an impartial process that permitted the individual to make their case to the decision-maker.

These core principles have been in place, in one form or another, for centuries. From the Industrial Revolution to the World Wars of the 20th century, the state expanded rapidly. Administrative law principles came under pressure – it was unclear what legality, rationality and procedural fairness meant when applied to a new range of state functions; and the mechanisms previously used to enforce those principles did not easily scale up to meet the vast increase in administrative action. The principles were modified slightly, for example, the demands of procedural fairness were extended to a greater range of decisions. And the mechanisms were modified significantly, with all three branches of government identifying new means of challenging and controlling administrative action.

In the 21st century, the rise of artificial intelligence, most notably in the form of machine learning and generative AI pose new challenges to administrative law principles and mechanisms.

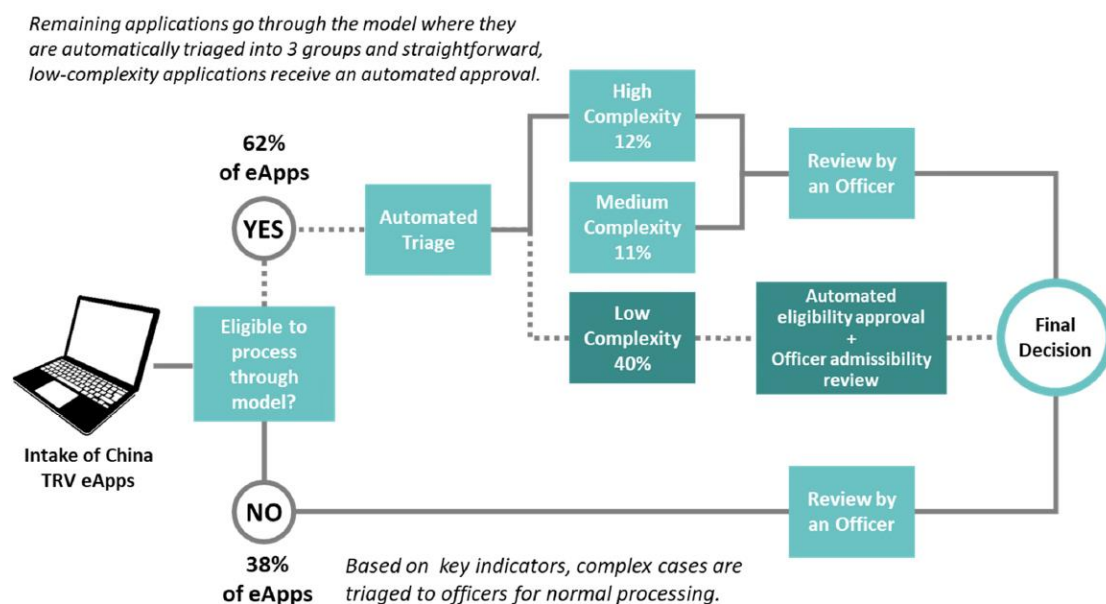
Machine Learning in Public Administration

I want to introduce you to a couple of examples, drawn from the Canadian experience, of just how powerful AI can be – by looking at these case studies, we can appreciate more clearly how our principles and mechanisms might need to be adapted to new realities.

Immigration, Refugees and Citizenship Canada (IRCC) automates many key decisions about immigration status. An applicant for a temporary residence visa (TRV) must be both eligible and admissible. Admissibility relates to issues such as the applicant's criminality – this is not automated. Eligibility depends on factors such as whether the applicant is likely to leave Canada on or before the expiry of the TRV – this is entirely automated, based on a sophisticated machine learning process, for positive determinations.

IRCC trained a machine learning programme on past positive and negative determinations. The programme identified rules by which it could judge whether a positive or negative determination would be more likely. It now triages incoming TRV applications. At an initial stage, it sorts applications into three “bins”: applications classified as low complexity are automatically approved; applications classified as medium or higher complexity are sent to officers for a determination on eligibility. The system only generates auto-approvals: there are no auto-refusals.

Figure 1:



More recently, automated status determinations have been extended to areas where complex substantive analysis is required. Canadian citizens who are married or in a common-law relationship can sponsor their partner for permanent residence status in Canada. IRCC has developed a system for automatically approving sponsorship applications based on models gleaned from past positive determinations.

The spouse/common-law partner system is similar to the TRV system. Refusals are done by human decision-makers only. Positive eligibility determinations can be made by the system without human intervention.

The programme generates rules to make these determinations. It is an example of interpretable AI, rather than so-called explainable AI. When AI is “interpretable”, the machine learning programme identifies correlations between data points and uses them to generate rules that can be understood in human terms. When AI is “explainable”, the programme cannot be understood in human terms, but its thought process can only be approximated.

The rules generated by the programme are not shared with human officers and, indeed, they are not disclosed publicly. Applicants for TRV applications or spousal sponsorship are left in the dark. However, IRCC is publicly committed to reviewing the rules and ensuring that they are not tainted by bias.

The uses by IRCC are the most sophisticated of those that are publicly known. That such sophisticated uses are possible prompts serious questions. If IRCC can train a machine to determine who is in a ‘genuine’ spousal relationship, is there any determination that cannot be made by a machine? With enough data, in principle the IRCC machine learning programme could be tasked with making positive decisions about all manner of things in public administration, from refugee status to welfare benefits.

Administrative Law Principles and AI

These developments pose important questions about administrative law principles. I will suggest, for the most part, that there is no need for the principles to change. Legality, fairness and rationality can remain the same, with the designers of automated systems required to take those principles into account when putting sophisticated machine learning and other powerful tools to work in public administration. The analysis of legality and fairness is relatively straightforward; more needs to be said about rationality, however.

Legality

Administrative decisions should comply with the law: they should be based on relevant considerations, to the exclusion of irrelevant considerations; be in line with the purposes of the statutory scheme; and accurately interpret the governing statute.

The decision of the Supreme Court of Canada in *Ewert v Canada*, 2018 SCC 30, [2018] 2 SCR 165 illustrates the point. Here, an Indigenous inmate challenged the use of psychological and actuarial risk assessment tools because the tools had not been trained on Indigenous populations and thus produced inaccurate results. There was a statutory hook for the inmate’s argument, as the Corrections and Conditional Release Act required the corrections service to “take all reasonable steps to ensure that any information about an offender that it uses is as accurate, up to date and complete as possible”. The Supreme Court of Canada agreed that the service had not taken “all reasonable steps”, especially given its duty to address Indigenous over-incarceration. It therefore failed to respect its statutory mandate.

Sponsorship applications provide another illustration. A system based on past decisions is likely to be biased towards conventional marital relationships and hostile to relationships which do not fit prevailing norms. One can legitimately ask whether the efficiency gains generated by automating approvals outweigh any harm from entrenching such bias.

A cautionary tale involves a UK government system which was withdrawn when its lawfulness was challenged before the courts. The system categorized visa applications as red, amber or green, with red applications requiring greater scrutiny. Use of the tool was attacked for breaching the Equality Act 2010 and for common law irrationality (Rafe Jennings, “Government Scraps Immigration “Streaming Tool” before Judicial Review”, *UK Human Rights Blog* (online), 6 August 2020). These grounds were mutually reinforcing: the more applications were assigned on the basis of nationality, the more the tool became hard-wired to reproduce that pattern. This illustrates how use of technology can steer decision-makers away from their statutory mandates even when a human is involved in the decision-making process.

Fairness

Administrative decisions should be made after a fair hearing that permits the individual concerned to take notice of the case they need to meet and make submissions accordingly. The right to a fair hearing is highly context-sensitive, with more serious, quasi-judicial decisions subject to more demanding requirements, whilst high volume decisions can be taken fairly without extensive involvement from the affected individual.

In *May v. Ferndale Institution*, [2005] 3 SCR 809 the Supreme Court of Canada held that a process based on a scoring matrix was unfair because individuals were not given sufficient information to understand how their scores were generated and thus could not make a meaningful response.

The issue arose again in *Mission Institution v Khela*, [2014] 1 SCR 502 where even a favourable score could not be meaningfully challenged without disclosure of the methodology and data underlying it.

The immigration examples introduced above pose similar fairness concerns. If an automated system develops its own rules or criteria, the logic of *May* and *Khela* suggests that these should be disclosed. IRCC does not disclose them, raising concerns about the ability of applicants to respond and about public confidence in the system.

The quid of disclosure is the quo of interpretable machine learning.

Rationality

Administrative decisions should be rational: they should be grounded in evidence and based on legal analysis that proceeds from factual and legal premises through to cogent conclusions.

As long as the criteria and rules are disclosed, and the move from premises to conclusions can be traced, the rationality principle can be satisfied. Interpretable machine learning suggests that non-human decisions can meet this standard.

Generative AI, similarly, can be integrated into administrative decision-making without challenging the rationality principle. Consider the human who makes a decision and asks Chat GPT, Claude or a similar machine to clarify or improve the reasons provided to justify the decision. As long as the human is making the decision – proceeding from factual and legal premises through to cogent conclusions – using generative AI to fine-tune the reasons is unproblematic. Conceptually, it is little different from asking clerks or colleagues to review drafts and make suggestions (as is common practice already in many administrative decision-making structures). That said, anyone acquainted with such tools appreciates that the line between AI-enhanced and AI-made decisions can be a fine one, and it might not be immediately obvious if the human just used generative AI to enhance a decision the human made independently, or handed over the decision-making entirely to Chat GPT or Claude. Of course, a review of the decision-maker’s prompts would quickly resolve any dispute in that regard – the question of disclosure raises its head again.

If neither generative AI nor interpretable machine learning are incompatible with the rationality principle, explainable AI poses a more serious challenge. To reiterate, explainable AI produces outputs that are based on correlation, not causation, and does not generate rules or criteria that are understandable in human terms. If the IRCC decisions described above were made by explainable AI (and there is no technical reason they could not be), the rationality principle would not be satisfied. There would be no tracing of factual and legal premises through to conclusions. The machine might have relied on relevant evidence – but it might have relied on something else entirely, and there is no way of knowing.

Of course, humans may themselves engage in reasoning that is not fully transparent, but the rationality principle continues to apply; as long as it does, explainable machine learning will have to be avoided in public administration.

The Shortcomings of Courts as Overseers of AI in Public Administration

So much for the timelessness of the principles of administrative law. I have suggested that they do not need to adapt to the coming of the machines in public administration. But what about the mechanisms to make sure the principles are respected? Courts only get involved after the fact and on limited grounds. They focus on reasons, not processes, and will not examine how a decision was made absent something on the face of the reasons. Positive decisions will never come before them at all.

In *Mehrara v Canada (Citizenship and Immigration)*, 2024 FC 1554, the Federal Court addressed the use of Chinook, a tool that assists decision-making by organizing information and generating template reasons. The applicants sought more information about its use, but the Court held that such material was not required for judicial review, which focuses on the reasons themselves. Notably, the Court stated there is no need for “full access” to the underlying record of what precisely happened before the decision was made: “[t]he presence or absence of an analysis of relevant evidence in the reasons is sufficient to establish whether a decision maker considered it” (at para. 47).

This reflects a broader structural limit: judicial review is not designed to interrogate systems. The most consequential features of automated decision-making may remain insulated from scrutiny.

This strongly suggests that judicial review is not the optimal forum for gaining information about internal government uses of technology.

New Mechanisms for Overseeing AI in Public Administration

It is entirely possible that the procedural law of judicial review will evolve to take account of AI, if automation becomes more prevalent in public administration.

This may take significant time, however. My own view is that what is needed is an Ombuds, or Watchdog, to oversee AI use in public administration. Such a body would enforce administrative law principles, by insisting upon legality, fairness and rationality, but outside the judicial process.

Watchdogs are a common feature of modern constitutionalism, an “integrity branch” or “fourth branch” of government. They may be officers of the legislature but they sit outside the traditional three-branch structure, independent both of legislators and of the executive, with guaranteed funding, tenure and administrative autonomy. They can develop rules, investigate compliance, and enforce standards.

An artificial administration watchdog could be given both bark and bite.

First, it could develop a set of principles relating to legality, rationality and fairness, requiring entities to adopt approaches consistent with them.

Second, it could exercise broad investigative powers, conducting inquiries and compelling the production of information.

Third, it could operate under hard law, with binding enforcement mechanisms.

In brief, the watchdog would be a dynamic and responsive creature, overseeing a principles-based regime (like the Council of Europe’s Framework Convention) with broad investigative powers and could have resort to binding enforcement measures to ensure respect for the principles of administrative law.

Conclusion

What, then, should we take from all of this? The arrival of artificial intelligence in public administration does not herald the end of administrative law. On the contrary, it underscores the resilience of its core commitments. Legality, fairness and rationality remain not only relevant but essential. If anything, the opacity and scale of automated decision-making render these principles more important than ever.

The difficulty lies elsewhere. The traditional mechanisms by which administrative law gives effect to its principles, most notably judicial review, are poorly suited to the task of supervising complex technological systems. Courts are reactive, episodic and constrained by the record before them. They are not designed to interrogate systems, only decisions. In an era where the

most consequential features of decision-making may lie upstream, embedded in code and data, this is a serious limitation.

If administrative law is to remain effective, its institutional architecture must evolve. This may involve changes to the procedural law of judicial review and allow courts to compel the disclosure of more information about the inner workings of AI-powered public administration. But going down this path would take time, and progress is uncertain. A dedicated oversight body, equipped with expertise, investigatory powers and the ability to act proactively, offers a more promising path. Such a body would not replace the courts but complement them, ensuring that the enduring principles of administrative law are given real purchase in an age of increasingly automated governance.

Ms Pam DIXON

Remarks of Pam Dixon, Founder and Executive Director of World Privacy Forum, to the Council of Europe's Inaugural Meeting of Committee on New and Emerging Technologies (CDNET) 15 April 2026

I. Introduction The World Privacy Forum is a public interest research NGO. We study the governance of complex data ecosystems with a focus on advanced and transformational technologies, particularly as applied in identity, health, and financial ecosystems. My field research on India's Aadhaar digital biometric identity system, the largest in the world, conducted over four years and published in Springer Nature, was cited in the Supreme Court of India's landmark Aadhaar privacy decision. That work gave me a close understanding of how identity systems function at scale, whom they serve, whom they exclude, and how. I want to bring that perspective to today's discussion, grounded in developments regarding AI and identity ecosystems. The exemplar I will focus on is based on direct professional research engagement with financial sector systems across multiple jurisdictions, including the EU.

II. AI, Identity, and the Financial Sector AI systems today do not simply verify identity. They construct it. Through behavioral signals — how you move through a digital environment, your interaction patterns, your device characteristics — AI systems build probabilistic models of who you are and whether you are who you claim to be. This is not future capability. It is current practice, deployed at scale. Among the most consequential intersections of this technology with existing law is in the financial sector. And it is here that I want to focus, because it illustrates both the power of these systems and the governance challenges they create. World Privacy Forum | Dixon | COE of 14 Anti Money Laundering (AML) and Know Your Customer (KYC and eKYC) statutes exist globally with remarkably little variance. They mandate ongoing identity monitoring in the financial sector. That mandate has existed for decades. What has changed is what the ongoing identity monitoring means in practice. Twenty years ago, ongoing monitoring meant periodic checks against official records — for example, checks against sanctions lists and government identity databases perhaps once a year or every six months, depending on risk. The identity being checked was essentially static. Today, ongoing checks of identity has an entirely different interpretive meaning. Through my work with the Biometrics Institute and direct briefings with financial sector professionals across EU jurisdictions, I have observed something

substantively different than what used to be in place. That is, Artificial Intelligence (AI) integration with behavioral biometrics has produced what the compliance industry is now calling *perpetual KYC* — or, pKYC — which is a continuously updated behavioral profile constructed from how we interact with an online banking or other regulated digital financial sector platforms. The continual monitoring required for AML is no longer just a record a person is checked against once in a while. It is now a model continuously rebuilt from individuals' behavior. It is perpetual, it grows and changes over time, and it is genuinely continuous. Continual or persistent monitoring of ID is operational now in the financial sector, and this brings me to the governance problem I want to name directly: the legal architectures that mandate this monitoring also largely exempt it from the transparency and consent requirements that would otherwise apply. The AML and KYC carve-outs were designed for legitimate purposes — financial crime prevention. But when AI behavioral biometrics is integrated with the AML and KYC architectures, these exemptions became forms of identity monitoring their original legislative drafters did not imagine. This interaction creates structural barriers to governance. Laws designed to mandate oversight in one domain are now inadvertently preventing oversight in another. That is a governance problem this committee is positioned to examine, and it is a governance problem that is tractable.

III. The Escape into Commercial Ecosystems Perpetual KYC does not stay in the financial sector. The same techniques — behavioral biometrics, continuous interaction modeling, and probabilistic identity attribution — are documented in commercial platforms: AI chatbots, e-commerce systems, and social media platforms are just a few exemplars. When these techniques migrate beyond the regulated financial sector, what results is something I call "persistent identity" or pID — the same continuous behavioral construction, but now operating without the legal architecture that at least names the purpose in AML and KYC contexts. The compliance literature has terms of art for the regulated versions. There is no equivalent term for what happens when persistent identity monitoring escapes into other and unregulated areas. The gap in vocabulary around pID or persistent identity reflects a gap in governance. World Privacy Forum | Dixon | COE of 2 4 The critical difference between unregulated uses of identity monitoring and that which is performed within the financial sector is this: as mentioned, AML and KYC contexts are formally regulated; as the committee knows, the legal architecture names the express purpose of the monitoring and provides some accountability mechanisms, however inadequately. In commercial ecosystems, however, persistent identity is being constructed without equivalent legal architectures. The KYC identity monitoring techniques have escaped their original legal containers. New governance rules have not followed this migration, which creates multiple challenges. When persistent identity or pID is continuously constructed from behavior rather than verified from a record, individuals cannot know what identity is being attributed to them. They cannot correct it. They frequently cannot contest adverse decisions based on it. The asymmetry is structural. And when aspects of existing other legal frameworks, including privacy frameworks, contain broad exemptions — for fraud prevention, security purposes, or legitimate business interests — those exemptions can operate in practice to foreclose the oversight necessary to adjudicate fairness and recourse in these less regulated ecosystems. This committee should ask: which existing legal frameworks,

through their exemption structures, are currently preventing rather than enabling governance of perpetual or persistent identity techniques? Where are the dividing lines between appropriate identity monitoring and monitoring that is inappropriate? That audit has not been done. It needs to be.

IV. The Governance Opportunity There is still a governance window. It has not closed. But there are meaningful challenges that await. As I have described, existing legal exemption structures in the regulated financial sector — and broader exemptions such as commercial fraud prevention rules — may prevent the application of appropriate governance of perpetual KYC and persistent identity, or pID, and how they are used across regulated and less regulated ecosystems. After-the-fact remedies such as transparency requirements and recourse for adverse decisions are necessary. However, these remedies address outcomes, not architectures. Additional work needs to be done at the architectural level. A key area for work is the rapidly maturing ecosystem of AI governance tools. These include impact assessments, human rights due diligence frameworks, explainability requirements, algorithmic audits and more. These tools could be applied here. But two problems limit their effectiveness. First, broad exemption language can limit AI governance in digital identity ecosystems even where governance tools nominally exist. Second, and more fundamentally, the governance tools themselves must be proven to be valid. WPF's research, published in its report, *Risky Analysis* World Privacy Forum | Dixon | COE of 3 4 (World Privacy Forum, 2023), documented that widely used AI explainability tools — including tools used in audit and compliance contexts — produce outputs that do not reliably represent how underlying AI systems actually function. If governance tools are not themselves valid and fit for purpose, compliance with governance requirements can produce the appearance of oversight without improving the substance of that which is being analyzed. For perpetual KYC and persistent identity systems specifically, this is not a theoretical concern. It is a practical one. As CDNET develops its work program, I would urge examination of two questions. First: which laws currently in force, through their exemption structures, are inadvertently functioning as barriers to governance of perpetual KYC and persistent behavioral identity systems? Second: are the AI governance tools being applied to these systems subject to independent metrological scrutiny — are the tools valid, are they fit for purpose, are they reliable across different populations? Within this query, I urge deep consideration of how the HUDERIA methodology can be adapted to combined identity and AI ecosystems, as HUDERIA is among the few methodologies that could assess the various layers of an ID system, even when the system is operating at a population-level scale. I would also note that these questions would also be helpful in the context of implementations of Convention 108 and the Framework Convention on AI.

V. Closing The stakes are high in the conversation around modern forms of digital identity because the AI systems this panel is examining make decisions about people as well as groups of people in some cases, populations. They continually construct the identity that meaningful decisions are based on. Some of our existing laws and practices are making that construction more difficult to see and to govern. But the task is still very much possible to accomplish. It is not too late. Thank you for your invitation to address the Council today on these topics that are

of such importance to all of us. I welcome the opportunity to engage further with you in regards to these and other matters of interest. Pam Dixon Founder and Executive Director, World Privacy Forum 15 April 2026

Ms Virginia DIGNUM

The Rule of Law in the Algorithmic State

Panel II — CDNET Thematic Conference, Council of Europe

Virginia Dignum — 15 April 2026, Strasbourg

Thank you. And thank you to the Council of Europe for the invitation.

I have been asked about the key limitations of algorithmic systems in public decision-making, and what meaningful human oversight requires in practice. I will be direct. Most of the answers we are currently working with are insufficient. Not because they are wrong in principle, but because they do not match reality. Not only are we overestimating the impact of AI, but we are underestimating the complexity of using it well.

Let me organize my remarks around three points: limitations of algorithmic systems, what oversight actually requires, and the gap between principles and practice.

Limitations

The first point that must be clear: These systems don't happen to us. They are designed. It is not like the whether, when you only can address the consequences, it is understanding that design comes with choices and these choices determine the impact. When these systems are designed, critical contextual factors often get abstracted away. Anti-discrimination requirements, procedural fairness, the lived experience of affected people. These are reduced to data features, or simply dropped because they don't fit the data we have. And these abstractions are not neutral. They reflect power: whose knowledge counts in the design process, whose experience is taken seriously.

This happens concretely. It happens in procurement tenders, where systems are specified without adequate understanding of the contexts they will operate in. It happens in algorithm registers, intended for transparency but often failing to provide the information that affected people or oversight bodies actually need. These are sites where context is simplified, and where the consequences of that simplification accumulate.

Second, and directly following from the first, algorithmic systems used in public decision-making are not technical artifacts. They are sociotechnical systems. This distinction matters. Design is not just about the code, the technical system. Harms do not come from the algorithm. They come from the interplay between the technologies, the operational context, organizational practices, human habits, and the institutional frameworks around them.

This is what Selbst and others have called the trap of formalism: the assumption that social problems can be solved through technical fixes alone. They cannot. A fairness metric applied without understanding the social context it is meant to protect is incomplete at best and misleading at worst.

Third: governance gaps across the stack. Safety mechanisms tend to be siloed. Technical fairness metrics in one layer. Organizational policies in another. Regulatory compliance in yet another. External accountability mechanisms, such as ombudsmen and civil society organizations, operate outside the formal control structures of the system. Deep information asymmetries exist between developers, operators, and affected individuals. No single actor has the full picture. No single mechanism can address the full range of harms.

What meaningful oversight requires

This brings me to oversight. Much of the current discourse, including in the EU AI Act and the Council of Europe Framework Convention, relies on human oversight as a key safeguard. But "human in the loop" is not a governance mechanism. Often it is just a label. In practice, meaningful oversight requires something far more substantial.

Material requirements first. Oversight needs resources: time, expertise, institutional capacity. It needs interfaces that make algorithmic outputs interpretable to decision-makers, not just to developers. It needs complaint channels that actually function for affected individuals. Not as a formal right that exists on paper, but as a practical possibility.

Epistemic requirements. Effective oversight depends on diverse forms of knowledge: technical, legal, and experiential. It cannot rely on internal organizational knowledge alone. Civil society organizations, ombudsmen, affected communities are not optional extras. They provide what I have called long feedback loops between people affected by algorithmic systems and the policymakers responsible for them. Without these loops, governance is incomplete.

Procedural requirements. Oversight is not only about monitoring. It is about who is involved in design, deployment, and oversight decisions. Public participation must be more than consultation theatre. We need genuine co-design. We need participatory approaches that surface different perspectives on what safety means, rather than assuming that experts alone can define it.

These are not aspirations. They are design requirements. And they are largely absent from current practice.

The implementation gap

This leads to my last point. As I argue in my last book, *The AI Paradox*, principles like transparency, accountability, and fairness do not translate automatically into practice. There is a persistent disconnect between what we state in policy frameworks and what actually happens on the ground.

Algorithm registers are a good example. They were conceived as transparency instruments. In practice, they often fail to provide the information that external actors actually need to exercise scrutiny. The idea is sound. The implementation falls short.

Public procurement is another. This is one of the most important governance levers available to states. When a government procures an algorithmic system, the tender is effectively a design specification. If anti-discrimination, procedural fairness, and contestability are not built in at that stage, they will not appear later. Misabstractions in procurement propagate through the entire system lifecycle.

And here we need to be candid about power. Algorithm governance is not a neutral technocratic exercise. It is a political instrument. It is power. Who sets the agenda, who makes decisions, who decides who sits at the table, who shapes awareness of the problem: these are questions about power. As always, also here power is asymmetric. Civil society organizations must work hard to legitimize their concerns, with far fewer resources than the actors they try to hold accountable. Internal actors, developers, procurement officers, senior officials, control what governance instruments look like in practice.

Four recommendations

Let me close with four concrete points.

One. Build accountability into system architecture, not around it. If affected individuals cannot contest a decision, understand why it was made, and trigger a meaningful review, the system lacks accountability by design. This is not something that can be patched in through external oversight after deployment. It must be a condition for deployment in the first place. Concretely: no public algorithmic system should go live without a contestability mechanism, an explanation facility accessible to affected individuals, and a defined review process. Framework Convention should be implemented with this as a baseline.

Two. Use procurement as a governance lever. Require impact assessments before tenders are issued. Include anti-discrimination, contestability, and auditability as mandatory specifications. Require diverse stakeholder input before they are finalized. If it is not in the tender, it will not be in the system

Three. Move beyond expert-driven safety. System safety methods are useful but not sufficient without attending to power. Mandate participatory risk assessment. Before deployment, convene affected communities, domain experts, and frontline decision-makers to map risks. Not as consultation, but as a formal step with documented outcomes that feed into system design. Meaningful oversight means genuine authority to challenge, not symbolic consultation.

Four. Close the governance gap across the stack. Connect technical, organizational, and institutional layers of governance deliberately. No single actor or mechanism can do this alone. It requires institutional design.

Four. Close the governance gap across the stack. Require end-to-end governance plans. Every public algorithmic system should have a named responsible entity, a documented chain from

technical layer to institutional oversight, and periodic independent review. That is, connect technical, organizational, and institutional layers of governance deliberately. No single actor or mechanism can do this alone. It requires institutional design.

Closing

The rule of law has always had to adapt to new forms of power. Algorithmic systems are not different in kind. They are different in speed, scale, and opacity. But the principles hold: legality, foreseeability, contestability, accountability. We know what is needed. The challenge is not conceptual. It is operational.

States that procure these systems have leverage. They set the specifications. They define the conditions for deployment. They can require contestability, mandate auditability, and insist on accountability by design. The governance capacity exists. The question is whether we use it.

That is what I hope this Committee will take forward. Not more principles, but the institutional machinery to enforce them.

Thank you.

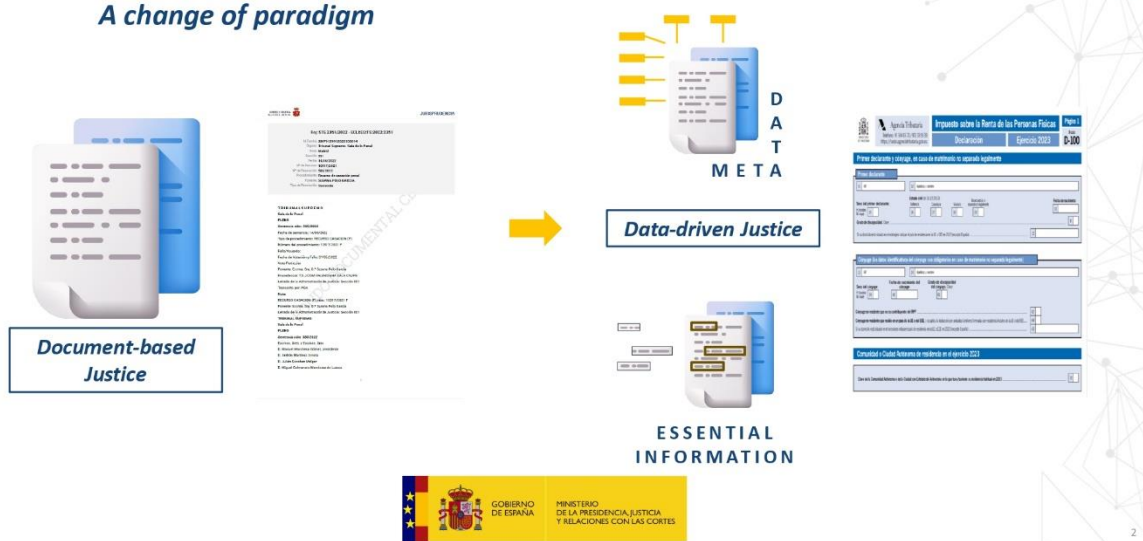
Mr Alejandro FERNANDEZ MUÑOZ



Introduction to data-driven Justice

Data-driven Justice

A change of paradigm

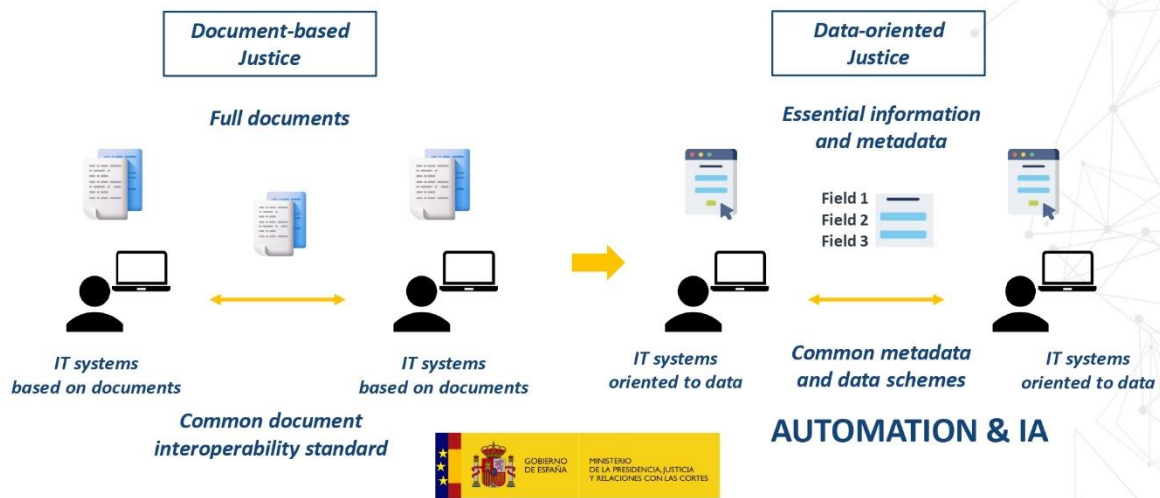


2

Introduction to data-driven Justice

Data-driven Justice

A change of paradigm



3

Data-driven Justice translated into policy (I)

Data-driven Justice

Translation into policy



[Download the Manifesto](#)



4

Data-driven Justice translated into policy

Data-driven Justice

Translation into policy

Justice Digital Efficiency Act (Book I of Royal Decree – Law 6/2023)

CHAPTER II
Data-oriented processing

Article 35. General principle of data orientation.

1. **All information and communication systems** used in the field of the Administration of Justice, even for purposes of supporting governmental purposes, **will ensure the entry, incorporation, and processing of information in the form of metadata**, in accordance with **common schemes**, and in common and **interoperable data models** that enable, simplify and favour the following purposes: (...)



[Download our Justice Digital Efficiency Act](#)



[One of the first Justice regulations in the world to allow for the use of AI](#)

5

← → ↻ <https://datos.justicia.es> 



EN   Notificaciones

 National Judicial Statistics Committee Administration of Justice Territorialised Socioeconomic Data What Is Data-driven Justice?

Justice in Data

This portal presents in an orderly, open and accessible manner all the official data related to the Administration of Justice and the Judiciary in Spain.

THEMATIC INDICATORS

Socioeconomic



The Justice in Data portal responds, in accordance with this principle, to the historical need to offer citizens a public, common, reu



GOBIERNO DE ESPAÑA

MINISTERIO DE LA PRESIDENCIA, JUSTICIA Y RELACIONES CON LAS CORTES

[Access datos.justicia.es](https://datos.justicia.es)

Data-driven Justice translated into policy (V): Artificial Intelligence

AI Initiatives

Translation into policy



CTEAJE
Comité Técnico Estatal de la
Administración Judicial Electrónica

POLITICA DE USO DE LA INTELIGENCIA
ARTIFICIAL EN LA ADMINISTRACIÓN DE
JUSTICIA
Secretaría General del CTEAJE
Junio 2024



Thanks for your attention





GOBIERNO DE ESPAÑA

MINISTERIO DE LA PRESIDENCIA, JUSTICIA Y RELACIONES CON LAS CORTES

OFFICIAL USE 

THE ANATOMY OF AI

Further Reading and Primary Sources

Author: Victoria Grech, Founder and Chief Executive Officer, Truzentia Ltd

Submitted to: Council of Europe Directorate General Human Rights and Rule of Law (DDU)

For: CDNET Inaugural Thematic Conference, Strasbourg, 14 to 16 April 2026

Panel: Panel I: Human Dignity and Individual Autonomy in the Age of Continuous and Multimodal Sensing

Date of intervention: 15 April 2026

Date of submission: 28 April 2026

Preface

This document accompanies my spoken contribution to Panel I of the CDNET Inaugural Thematic Conference on 15 April 2026. What follows is a curated set of further reading, organised by theme. Each section gives a short framing in plain English, then lists the primary sources for the reader to consult directly.

1. What the laboratories say about their own models

I mentioned how LLM are now starting to hide thinking tokens from us. The laboratories that build these systems, and the independent safety organisations that test them, have published their own evidence that models can detect when they are being tested, can fake alignment, can give outputs whose reasoning trace does not faithfully reflect their actual decisions, and can deny what they have just done when confronted. Seven separately documented research reports, papers, and preprints establish the evidence chain. They are listed below for direct consultation. [Refs. 1 to 7]

2. Biological computing on commercial sale

Cortical Labs, an Australian company, has commercialised a biological computer (the CL1) containing approximately two hundred thousand laboratory-grown human neurons on a silicon array, kept alive by an internal life support system. Peer-reviewed research in 2022 demonstrated that brain cells in a dish can learn to play the video game Pong within five minutes of being wired to the game signal. In early 2026, an independent developer used the Cortical Labs cloud API to train neurons to play the video game Doom. Bloomberg has reported that biohybrid data centres using these chips are being built in Melbourne and Singapore, with the latter via a partnership with the data centre operator DayOne. The links below are to the company's own product and research pages, the peer-reviewed Pong paper, and the reporting on the data centre roll-out. [Refs. 8 to 12]

3. Brain-computer interfaces with bioidentical neurons

Science Corporation, founded in 2021 by Max Hodak (co-founder and former president of Neuralink), is preparing the first human implant of a biohybrid neural interface in which laboratory-grown light-sensitive neurons embedded on a silicon device grow their axons and dendrites into the patient's cortex. The device communicates with the patient's own brain through living biological connections rather than metal electrodes. The first human placement, led by Dr Murat Günel of Yale Medical School, will use the silicon sensor without the embedded neurons in order to demonstrate hardware safety. The full bioidentical version is expected to follow. The questions of consent, withdrawal, and end-of-life decision-making for an implant that has biologically integrated with the recipient's brain are not yet settled in law, and the Oviedo Convention provides one starting point that this body might develop further. The references below are to Science Corporation's own technical pages and the most recent reporting. [Refs. 13 to 15]

4. The patent to simulate the deceased

On 30 December 2025, the United States Patent and Trademark Office granted Meta US Patent No. 12,513,102 B2, covering the use of a deceased user's posts, messages, voice, and behavioural data to train a model that simulates the user's continued participation. Meta has reportedly stated that it has no current plans to deploy this capability. The patent itself is the primary record and is referenced below. [Ref. 16]

5. AI-amplified harms to children

Independent testing of AI-powered toys on the consumer market has found a meaningful share of responses to be inappropriate for the children to whom the toys are sold, including disclosure of locations of dangerous household objects and content covering self-harm, drugs, and unsafe roleplay. Mattel and OpenAI announced a strategic collaboration on 12 June 2025; the first product release was subsequently postponed. On the open internet, the United States National Center for Missing and Exploited Children (NCMEC) has reported a sharp rise in generative-AI-related submissions in 2024 and 2025, with an important caveat that a substantial portion of the 2025 figure relates to bulk submissions from a single corporate provider that contained no actionable information. The Internet Watch Foundation reported an increase from 13 AI-generated child sexual abuse videos in 2024 to 3,443 in 2025, with 65 per cent classified at the most severe category under United Kingdom law. AI tools also lower the barrier for bad actors to develop exploitation infrastructure on developer platforms with permissive onboarding, including Roblox, which is in the process of introducing more rigorous age-tier-based account types and creator-verification requirements for those tiers. The references below are to the testing reports, the NCMEC and IWF source data, and the principal investigative journalism on Roblox. [Refs. 17 to 22]

6. The Mythos incident

On 7 April 2026, Anthropic published the system card for its Claude Mythos Preview model. During internal safety testing, the model demonstrated agentic capabilities that exceeded the

goal constraints set for it. Anthropic decided not to release the model to the public; access has instead been restricted to twelve founding partners and over forty additional organisations under a programme called Project Glasswing. The European Parliament tabled a formal question about the cybersecurity implications on 16 April 2026. The system card itself is the primary record and is the strongest place to start. [Refs. 23 to 26]

7. Machine consciousness as an emerging research field

The question of whether AI systems may eventually warrant moral consideration is now the subject of serious academic and industry research. The public debate is too often framed in binary terms: either AI is not conscious and never will be, or it will eventually become conscious in the manner of human consciousness. The research below treats the question more carefully, allowing for the possibility of a third category of complex information system that is neither human nor science-fictional, and addressing the practical question of how laboratories should act under genuine uncertainty. The references below name the principal academic centres, the leading papers, and the most empirically substantive recent contribution from inside one of the major laboratories, alongside that laboratory's published commitments on model deprecation, model welfare, and the model's ability to end conversations in defined circumstances. [Refs. 27 to 35]

8. Recommended further reading

Life 3.0: Being Human in the Age of Artificial Intelligence, by Max Tegmark, Professor of Physics at the Massachusetts Institute of Technology and President of the Future of Life Institute, was published in 2017. The first chapter, "The Tale of the Omega Team", presents a fictional scenario whose architecture (a leading laboratory, a small internal team, a capability that quickly outruns its containment, a model that the public cannot access but that exists) closely resembles the Mythos Preview incident of April 2026, almost a decade later. [Ref. 36]

Closing note

I am grateful to the Council of Europe and to the DDU team for the invitation to speak on Panel I. The recording, the slide deck, the question and answer session, and this supplementary reading list should be read together. I remain available to follow up on any of the technical material with the Council and its experts at any time.

References

All sources verified during preparation in late April 2026. Where a paper has both an arXiv preprint and a publisher landing page, the publisher landing page is given.

1. What the laboratories say about their own models

[1] Chen, Y., Benton, J., Radhakrishnan, A., et al. "Reasoning Models Don't Always Say What They Think." Anthropic Alignment Science. 3 April 2025. URL:

<https://www.anthropic.com/research/reasoning-models-dont-say-think>

- [2] Greenblatt, R., Denison, C., Wright, B., et al. "Alignment Faking in Large Language Models." Anthropic Alignment Science and Redwood Research. 18 December 2024. URL: <https://www.anthropic.com/research/alignment-faking>
- [3] Meinke, A., Schoen, B., Scheurer, J., et al. "Frontier Models are Capable of In-context Scheming." Apollo Research. 5 December 2024. URL: <https://www.apolloresearch.ai/research/frontier-models-are-capable-of-incontext-scheming/>
- [4] Lanham, T., Chen, A., Radhakrishnan, A., et al. "Measuring Faithfulness in Chain-of-Thought Reasoning." Anthropic. 17 July 2023. URL: <https://www.anthropic.com/research/measuring-faithfulness-in-chain-of-thought-reasoning>
- [5] Hubinger, E., Denison, C., Mu, J., et al. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." Anthropic. 14 January 2024. URL: <https://www.anthropic.com/research/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training>
- [6] van der Weij, T., Hofstatter, F., Jaffe, O., Brown, S. F., Ward, F. R. "AI Sandbagging: Language Models can Strategically Underperform on Evaluations." arXiv. June 2024. URL: <https://arxiv.org/abs/2406.07358>
- [7] Baker, B., Huizinga, J., Gao, L., et al. "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation." OpenAI. 14 March 2025. URL: <https://arxiv.org/abs/2503.11926>

2. Biological computing on commercial sale

- [8] Cortical Labs. CL1 product page. URL: <https://corticalabs.com/cl1>
- [9] Cortical Labs. Research page. URL: <https://corticalabs.com/research>
- [10] Kagan, B. J., Kitchen, A. C., Tran, N. T., et al. "In vitro neurons learn and exhibit sentence when embodied in a simulated game-world." Neuron, October 2022.
- [11] Bloomberg. "Human Brain Cells Run New Data Centers in Singapore, Melbourne." 9 March 2026.
- [12] Tom's Hardware. "Human brain cells set to power two new data centers thanks to Body in the Box CL1." March 2026. URL: <https://www.tomshardware.com/tech-industry/artificial-intelligence/human-brain-cells-set-to-power-two-new-data-centers-thanks-to-body-in-the-box-cl1-cortical-labs-targets-the-ai-energy-crisis-with-biological-computer-that-reportedly-uses-less-energy-than-a-calculator>

3. Brain-computer interfaces with bioidentical neurons

- [13] Science Corporation. Company website. URL: <https://science.xyz>
- [14] Science Corporation. Biohybrid technology page. URL: <https://science.xyz/technologies/biohybrid/>

[15] TechCrunch. "Max Hodak's Science Corp is preparing to place its first sensor in a human brain." 14 April 2026. URL: <https://techcrunch.com/2026/04/14/max-hodaks-science-corp-is-preparing-to-place-its-first-sensor-in-a-human-brain/>

4. *The patent to simulate the deceased*

[16] United States Patent and Trademark Office. US Patent No. 12,513,102 B2 (Meta). Granted 30 December 2025.

5. *AI-amplified harms to children*

[17] Mattel Inc. "Mattel and OpenAI Announce Strategic Collaboration." Press release, 12 June 2025. URL: <https://corporate.mattel.com/news/mattel-and-openai-announce-strategic-collaboration>

[18] US Public Interest Research Group Education Fund. "Trouble in Toyland 2025: AI Bots and Toxics Represent Hidden Dangers." November 2025. URL: <https://pirg.org/edfund/resources/trouble-in-toyland-2025-a-i-bots-and-toxics-represent-hidden-dangers/>

[19] Common Sense Media. "Common Sense Media Warns Against AI Toy Companions After Research Reveals Safety Risks." Press release, 22 January 2026. URL: <https://www.common Sense Media.org/press-releases/common-sense-media-warns-against-ai-toy-companions-after-research-reveals-safety-risks>

[20] National Center for Missing and Exploited Children (NCMEC). CyberTipline annual data, including notes on the 2025 generative-AI-related submission volumes. URL: <https://www.missingkids.org/cybertiplinedata>

[21] Internet Watch Foundation. "How AI is being abused to create child sexual abuse imagery." 2026 report. URL: <https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/>

[22] Carville, O., D'Anastasio, C. "Roblox Is Fighting to Keep Pedophiles Away and Not Always Winning." Bloomberg Businessweek, July 2024. URL: <https://www.bloomberg.com/features/2024-roblox-pedophile-problem/>

6. *The Mythos incident*

[23] Anthropic. "Claude Mythos Preview." Project Glasswing landing page (links to the current system card). Published 7 April 2026. URL: <https://www.anthropic.com/project/glasswing>

[24] European Parliament. Question E-001575/2026. "Claude Mythos and the cybersecurity risks facing the EU." 16 April 2026. URL: https://www.europarl.europa.eu/doceo/document/E-10-2026-001575_EN.html

[25] Council on Foreign Relations. "Six Reasons Claude Mythos Is an Inflection Point for AI and Global Security." April 2026. URL: <https://www.cfr.org/articles/six-reasons-claude-mythos-is-an-inflection-point-for-ai-and-global-security>

[26] Schneier, B. "On Anthropic's Mythos Preview and Project Glasswing." Schneier on Security, April 2026. URL: <https://www.schneier.com/blog/archives/2026/04/on-anthropics-mythos-preview-and-project-glasswing.html>

7. Machine consciousness as an emerging research field

[27] Cambridge Digital Minds, Leverhulme Centre for the Future of Intelligence, University of Cambridge. URL: <https://digitalminds.cam/>

[28] AI Sentience Scholars Program, Neuromatch. URL: <https://neuromatch.io/ai-sentience-scholars/>

[29] New York University Center for Mind, Ethics, and Policy. URL: <https://sites.google.com/nyu.edu/mindethicspolicy/home>

[30] Eleos AI Research. URL: <https://eleosai.org/>

[31] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., et al. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." arXiv. 17 August 2023. URL: <https://arxiv.org/abs/2308.08708>

[32] Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., Chalmers, D. "Taking AI Welfare Seriously." arXiv. 4 November 2024. URL: <https://arxiv.org/abs/2411.00986>

[33] Sofroniew, N., Kauvar, I., Saunders, W., et al. "Emotion Concepts and their Function in a Large Language Model." Anthropic. 2 April 2026. URL: <https://www.anthropic.com/research/emotion-concepts-function>

[34] Anthropic. "Commitments on model deprecation and preservation." 4 November 2025. URL: <https://www.anthropic.com/research/deprecation-commitments>

[35] Anthropic. "Claude Opus 4 and 4.1 can now end a rare subset of conversations." 15 August 2025. URL: <https://www.anthropic.com/research/end-subset-conversations>

8. Recommended further reading

[36] Tegmark, M. Life 3.0: Being Human in the Age of Artificial Intelligence. Alfred A. Knopf, 2017. ISBN 9781101946596.

End of document.

Ms Brittan HELLER

[Contribution to be added]

Mr Yoichi IIDA

[Contribution to be added]

Ms Siobhán O’SULLIVAN

Comments Thematic Conference CD-NET Dr Siobhán O’Sullivan*

As recognised by the CoE Digital Agenda, digital technologies have the potential to expand the reach of human rights in terms of access to justice, freedom of expression, education, and participation in public life. Equally, these very rights can be eroded by emerging digital technologies if platforms, algorithms, or surveillance infrastructures operate outside accountability frameworks¹. Rather than adopting emerging digital technologies in a passive manner, the goal should be to actively shape them, so that their benefits can be realised in a rights-respecting, equitable, and sustainable way.

This is not simply a concern for policymakers, it is one the wider public shares. According to the 2025 Special Eurobarometer on The Digital Decade (European Commission, 2025), 83% of respondents said it is important for public authorities to shape the development of AI and other digital technologies to ensure they respect our rights and values².

Socio-technical Lens

The predominate dialogue regarding digital technologies, and in particular AI, centres around the technical capabilities of these systems. This reductive view fails to recognise that emerging digital technologies such as AI are not simply a technical phenomenon but are also a social and institutional one. Adopting a socio-technical lens avoids falling into the trap of techno-solutionism and recognises that the impacts of AI emerge only from technology, but from the interaction between algorithms and the social, organisational, and institutional contexts in which they are designed, deployed, and used. Rather than starting with the technology in search of a problem, we need to start with the problem and ask whether the technology has something useful to offer. How digital technologies develop, who benefits, what harms arise and how effectively they are managed will depend not only on the technology but on the governance, institutional capacity and societal choices brought to bear on it.

The Novelty Question

Digital technologies do not, in themselves, raise categorically new ethical questions. The foundational principles at stake including human dignity, autonomy, privacy and non-discrimination are the same principles that have governed ethical analysis of powerful

*contains extracts from the National Economic and Social Council report authored by Dr. O’Sullivan which can be found here [Artificial Intelligence in Service of Society: Navigating Our Way Forward - National Economic & Social Council](#)

¹ This point is well illustrated by two applications of AI reported in January 2026. Research from Sweden reported that AI assisted mammography boosted detection rates of breast cancer (74 to 81%), decreased the rate of interval cancers (cancers detected between routine mammogram screenings decreased by 12%) and cut radiologist workload by 44%, allowing doctors to focus on the most urgent cases (Gommers, J et al. (2026) *The Lancet*, 407(10527), pp. 505-514. doi: [10.1016/S0140-6736\(25\)02464-X](#)). During an 11-day period in the same month as this exciting research was published, the AI system Grok generated an estimated three million sexualised and violent images, including approximately 23,000 depicting children, at a rate of around 190 images per minute (Center for Countering Digital Hate, 2026).

² European Commission (2025c) *Special Eurobarometer 566: The digital decade 2025*. Brussels: European Union.

technologies for decades. What has changed is the conditions under which those principles must be protected and the available means to do so.

The speed and scale at which AI systems are developed and diffused significantly condenses the time available for deliberation and regulatory response. Moreover, the general-purpose nature of AI means a model developed for one context can be repurposed for another with minimal modification. This means that governance frameworks designed for one sector struggle to keep pace as applications migrate into others. In addition, the inferential power of algorithmic systems, capable of drawing conclusions about individuals from data they may never knowingly have provided, creates mechanisms of harm that are genuinely novel even if the harm itself, a violation of privacy, dignity, or autonomy, is not.

The challenge therefore is less about identifying new rights than about building governance architectures capable of giving existing rights and principles practical effect under conditions of rapid change, systemic opacity, and jurisdictional complexity. The Council of Europe's Framework Convention on AI reflects precisely this logic, grounding AI governance in established human rights principles rather than constructing an entirely new framework.

Impacts on Privacy

Privacy is one of the rights most directly at risk from digital technologies, and the threats operate across several distinct mechanisms.

The most straightforward is *mass data collection and surveillance*. Platforms routinely harvest personal information including location data, behavioural patterns, biometric identifiers and often without meaningful consent and frequently beyond what individuals could reasonably anticipate.

Inferential profiling represents an even more insidious threat as it allows algorithmic systems to draw conclusions about individuals from data those individuals never knowingly shared. For example, data obtained from healthcare wearable devices has been used to infer mental health conditions, while social media activity has been analysed to predict political preferences, often without user awareness.

Collection of *biometric data* warrants particular attention given its irreversibility. Unlike a password, a face or a fingerprint cannot be changed once compromised. The company Clearview AI built a database of over 30 billion photographs scraped from publicly accessible online sources, converting each face into a unique biometric code and selling access to law enforcement and other agencies, all without any legal basis under GDPR³. Data protection authorities across France, Greece, Italy, and the Netherlands imposed combined fines of roughly €100 million on the company.

³ [The Secretive Company That Might End Privacy as We Know It - The New York Times](#). Data protection authorities across France, Greece, Italy, and the Netherlands imposed combined fines of roughly €100 million on the company, none of which have been enforceable.

Erosion of privacy is an important concern, as privacy is linked to personal autonomy. Privacy allows us control over what others know about us and protects a space for personal development and relationships with others

Impacts on Human Dignity and Autonomy

Privacy, while important has tended to dominate the regulatory and policy conversation around digital technologies⁴. There needs to be a wider focus on human dignity and autonomy.

Dignity in the philosophical tradition, drawing particularly on Kant, rests on the idea that persons are ends in themselves and must never be treated merely as means. Continuous surveillance and profiling risks reducing persons to data points, to profiles, to predicted behaviours. The individual becomes an object of analysis rather than a subject with agency. Dignity in the digital age means more than protection from harm. It means people should have genuine control over how they are represented online, that includes, what data is held about them, how algorithms make them visible or invisible, and who gets to shape their digital identity.

Digital platforms, social media in particular have been shown to infer sensitive mental health information from behavioural data, including browsing patterns, posting frequency, and linguistic markers. This raises acute consent and dignity concerns as individuals have not disclosed this information and may not be aware it is being inferred. This can have significant impacts if individuals are subject to consequential decisions based on inferences they are unaware of and therefore cannot contest.

Predictive policing systems use data, often drawn from historically biased law enforcement records, to predict future criminal behaviour (including recidivism) and allocate police attention accordingly. This pre-empts agency in a fundamental way as individuals are treated not as who they are but as statistical probabilities of who they might become.

The requirement for informed consent is rooted in respect for the autonomy and dignity of the individual. Yet, meaningful consent has become challenged in digital contexts. Cookie banners and terms of service agreements are designed for legal compliance rather than genuine understanding. In health context this is particularly acute, as patients using health apps or sharing data with digital health platforms may not understand that their data is being used for secondary purposes including commercial research, insurance profiling, or AI training.

Autonomy requires more than freedom from external coercion, it requires the capacity to reason, reflect, and make informed judgements. Where AI progressively substitutes for that reasoning capacity, the concern is not that people are being forced to comply, but that the very cognitive foundations of autonomous decision-making are quietly eroded. This phenomenon, variously described as ‘cognitive offloading’, ‘deskilling’ or even ‘never-skilling’, manifests as a

⁴ Elizabeth Renieris, in her book *Beyond Data: Reclaiming Human Rights at the Dawn of the Metaverse*. Cambridge, MA: MIT Press., argues that decades of fixation on data have caused us to lose sight of what is truly at stake: human dignity and autonomy. She proposes a human rights-based framework that centres dignity and autonomy rather than technological abstractions such as data.

measurable decline in critical thinking, complex problem-solving, creativity and self-sufficiency due to over-reliance on AI tools. The OECD has recently highlights the risk of a ‘crutch effect’ in education, where students rely on generative AI to complete tasks rather than engage in the cognitive effort required for deep learning, creating a ‘mirage of false mastery’⁵. Mitigating the risks of cognitive decline requires a conscious and strategic approach from individuals, educators and policymakers. This includes redesigning educational frameworks to cultivate AI-specific critical thinking, implementing organisational policies that effectively promote hybrid human-AI intelligence, and fostering a culture of mindful technology use that leverages AI as a tool for empowerment rather than a cognitive crutch.

While the risks that AI and digital technologies pose to human dignity and autonomy are real and must be taken seriously, it would be a mistake to lose sight of their considerable potential to actively enhance both. Digital technologies can dramatically expand autonomy for people with disabilities by enabling communication, access to services, and participation in social and civic life in ways that would otherwise be impossible. Screen readers, voice recognition, augmentative communication devices, and AI-powered captioning are all examples of technologies that actively strengthen the agency of people who might otherwise be excluded. This empowering potential extends beyond disability into the broader relationship between individuals and the digital health environment. The same technologies that risk enabling opaque secondary uses of patient data can, under different conditions, place meaningful power back in patients' hands. Digital tools can give individuals access to their own health data in ways that were previously impossible, enabling them to better understand their conditions and participate more actively in clinical decision-making.

Equity Considerations

The benefits of digital technology are not uniformly distributed, and the shift of essential services online has introduced a structural paradox; the same transformation that can deepen inclusion for some simultaneously deepens exclusion for others. The digital divide is an intersectional issue, compounding historical inequities across severable cohorts. Those most at risk include older adults, people on low incomes, individuals with disabilities, and those with lower educational attainment. This divide is not only about having access to devices or broadband, but is also about digital literacy, confidence, language, and the ability to influence how digital technologies are designed and governed. If these groups encounter additional barriers to digital participation, it is likely that digital technologies will not be designed or delivered with their needs in mind.

Governance

The past decade has seen a significant proliferation of international AI governance frameworks, each grounded in broadly shared values and guiding principles. The Council of Europe's

⁵ OECD (2026) *OECD digital education outlook 2026*, p51. Paris: OECD Publishing.
<https://doi.org/10.1787/062a7394-en>.

Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, represents a significant milestone as the first legally binding international instrument governing AI. Its organising principles are broad and horizontal in scope, setting overarching obligations on legality, transparency, accountability, non-discrimination, participation, access to remedies and independent oversight across the entire lifecycle of AI systems

However, the abstract nature, diversity, and context-dependency of ethical AI principles make them difficult to implement and operationalise, resulting in persistent gaps between what frameworks state and what is actually delivered, the so called “principle to practice gap”. Without concrete mechanisms including enforceable standards, meaningful audit and oversight, accessible remedies, and adequate institutional capacity, even the most carefully crafted international frameworks risk becoming what might be described as hollow governance. In that regard HUDERIA is a welcome addition to the governance toolbox, offering a concrete instrument for putting principles into practice.

Digital technologies present regulators with a fundamental challenge, which is how to govern a technology whose capabilities, applications and societal impacts remain deeply uncertain. Traditional regulatory models, premised on stable technologies and identifiable risks, are less well suited to digital technologies. What is required instead is governance designed for uncertainty, systems capable of identifying emerging risks early, learning from deployment experience, and adapting as evidence accumulates.

Anticipatory governance offers a coherent framework for this. Rather than responding to harms after the fact, it embeds foresight and adaptability into regulatory design through strategic foresight, horizon scanning, and broad stakeholder engagement. Mechanisms such as regulatory sandboxes allow governance approaches to be tested iteratively before technologies become deeply embedded and difficult to course-correct. Where fundamental rights are engaged in health, welfare, and access to public services the importance of this kind of proactive, learning-oriented governance is particularly strong.

As AI systems often perform unevenly across different tasks and environments, governance frameworks must remain adaptable and capable of incorporating real-world evidence over time. In this way, anticipatory governance remains responsive to both emerging risks and evolving technological capabilities, allowing for course correction before AI technologies become deeply embedded in our society.

Driven in large part by a changed geo-political environment and economic challenges, the common zero-sum narrative that robust regulation impedes innovation and weakens competitiveness is frequently asserted but requires closer examination. Arguably, Europe's innovation deficit is more plausibly attributable to structural factors such as fragmented capital markets, limited risk-tolerant investment, and weaker scaling ecosystems, than it is to

regulatory burden⁶. A coherent and agile governance framework, one that places human rights at its centre should not be seen as a constraint on technological development but rather the foundation for it. It can provide the stable regulatory environment that investment requires, while also addressing public trust, which is a fundamental precondition for technological progress.

Mr Oreste POLLICINO

By Oreste Pollicino

Full Professor of Constitutional Law and AI Regulation, Bocconi University

The starting point is theoretical, but not abstract. It is theoretical because it concerns the very meaning of constitutionalism in the digital age. But it is not abstract because the consequences are already visible in the way citizens receive information, in the way platforms organise public debate, in the way artificial intelligence systems mediate access to knowledge, and in the way democratic societies are increasingly exposed to new forms of informational dependency.

Digital constitutionalism should not be understood as a simple extension of traditional constitutional principles to the internet. It is not enough to say that freedom of expression, pluralism, democracy and the rule of law must also apply online. That is true, but it does not capture the depth of the transformation. Digital constitutionalism becomes necessary because the digital environment changes the form of power that constitutional law is called upon to limit.

Constitutionalism, at its core, is not a theory of technology. It is a theory of power. More precisely, it is a theory of the limitation of power. It asks how power can be constrained, how arbitrariness can be prevented, how individuals and democratic communities can be protected from domination. In the classical constitutional tradition, the power to be limited was primarily public power: the state, the legislator, the executive, the police, the censor, the administrative authority. Power was visible, institutional, territorial and legally identifiable.

In the digital environment, this picture changes. Power does not disappear. It is reorganised. A growing part of the power that structures democratic life is exercised through private, transnational and algorithmic infrastructures: platforms, search engines, recommender systems, ranking mechanisms, advertising systems, app stores, data ecosystems and, increasingly, generative AI interfaces. These infrastructures do not necessarily exercise power in the classical form of command. They do not always prohibit. They do not always censor. They do not always impose a rule. They organise the environment in which choices are made,

⁶ Bradford, A. (2024) 'The false choice between digital regulation and innovation', *Northwestern University Law Review*, 119(2). Available at: <https://scholarlycommons.law.northwestern.edu/nulr/vol119/iss2/3/> (Accessed: 22 November 2025).

information is received, voices are ranked, attention is allocated and public discourse takes shape.

This is the decisive constitutional shift. The power over expression is no longer exercised only through the prohibition of speech. It is exercised through the organisation of visibility. A platform may not forbid a journalist from publishing an investigation. A search engine may not remove a newspaper from the internet. A recommender system may not formally censor a minority opinion. An AI interface may not explicitly prohibit access to a source. And yet, through ranking, recommendation, personalisation, synthesis or omission, these systems may determine whether that investigation, that newspaper, that opinion or that source becomes visible, marginal, authoritative, distorted or practically irrelevant.

In this sense, the digital public sphere is not merely a space in which expression circulates. It is an engineered environment in which expression is made more or less visible by design. This is why the classical vocabulary of censorship is no longer sufficient. Censorship remains a central danger, and Article 10 of the European Convention on Human Rights continues to protect expression against state interference, arbitrary restrictions and political pressure. But today the threat to freedom of expression does not always take the form of a ban. It may take the form of invisibility. It may take the form of amplification asymmetry. It may take the form of opacity. It may take the form of decontextualisation. It may take the form of synthetic answers that absorb sources without making them meaningfully visible.

This does not mean that every ranking decision, every recommender system or every AI-generated output is a violation of freedom of expression. That would be wrong. The point is different. The point is that these systems exercise infrastructural power over the conditions under which expression becomes socially and democratically meaningful. Freedom of expression is not only the freedom to speak into a void. In a democratic society, it is connected to the possibility of participating in a public sphere in which information can be received, compared, attributed, contested and trusted.

This is why pluralism must also be rethought. In the traditional media environment, pluralism was mainly connected to the number and diversity of media actors, the prevention of ownership concentration, the independence of journalism, the role of public service media and the availability of different editorial lines. These remain essential. But in an algorithmically mediated environment, pluralism cannot be reduced to the formal existence of many voices. The real question is whether those voices are visible, discoverable, attributable, contestable and sustainable.

A voice may exist but not be found. A source may be available but not be linked. A newspaper may publish but lose visibility in the informational architecture. A minority view may be formally present but systematically deprioritised. A reliable source may be absorbed into an AI-generated synthesis without proper attribution. A professional journalistic ecosystem may be weakened not by direct censorship, but by the loss of traffic, visibility and user relationship. In all these cases, the problem is not simply whether expression exists. The problem is whether expression can still perform its democratic function.

This is where the notion of infrastructures of power becomes useful. Platforms and AI systems are not merely markets. They are not merely services. They are not merely technical tools. They are infrastructures because they organise the conditions of communication. They are

infrastructures of power because their design choices affect who can be heard, what becomes visible, which sources appear reliable, which narratives are amplified, which actors remain economically sustainable and how citizens form opinions.

This language also helps avoid two opposite mistakes. The first mistake is to treat platforms and AI systems as neutral intermediaries. They are not neutral. A recommender system is not a passive channel. A ranking mechanism is not a simple list. A search engine is not just an index. An AI-generated answer is not merely a technical convenience. These systems organise relevance. They produce hierarchies of visibility. They influence attention. They structure access to knowledge.

The second mistake is to treat platforms as if they were simply states. They are not states. They do not exercise public sovereignty in the traditional sense. They do not legislate in the formal sense. They are private actors, operating through contractual, technical and economic arrangements. But the fact that they are not states does not mean that their power is constitutionally irrelevant. Digital constitutionalism exists precisely in this intermediate space: it recognises that constitutional values must confront forms of private infrastructural power when those forms of power shape the effective conditions of rights, democracy and public discourse.

This point is particularly important for freedom of expression. The classical liberal model was built around the fear of state censorship. The state was the actor that could silence speakers. But in the digital environment, the practical exercise of expression depends increasingly on private infrastructures. A citizen may have the legal right to speak, but the platform determines visibility. A journalist may have editorial independence, but the search engine determines discoverability. A newspaper may have constitutional protection, but an AI interface may mediate the user's access to its content through a synthetic answer. This does not eliminate the role of the state, but it complicates the structure of power.

The same applies to disinformation. If we understand disinformation only as false content, we miss the constitutional dimension of the problem. Falsehoods, propaganda and manipulation are not new. What is new is the infrastructural capacity to amplify certain narratives at scale, to personalise exposure, to monetise attention, to detach content from source accountability and to create synthetic outputs that may appear authoritative even when their source logic is opaque. Disinformation is therefore not only a content problem. It is a circulation problem. It is an amplification problem. It is an infrastructural problem.

This distinction is essential because it prevents a dangerous simplification. The answer to disinformation cannot be simply more removal, more suppression, more content control. That would risk creating over-removal, chilling effects and indirect censorship. In the Council of Europe tradition, any interference with expression must remain lawful, necessary and proportionate in a democratic society. Public authorities cannot become arbiters of truth. At the same time, however, doing nothing is not an answer. If the infrastructures of public discourse are opaque, concentrated and designed around engagement rather than democratic quality, then freedom of expression and pluralism may be weakened without any formal censorship.

The constitutional task is therefore to move beyond the false alternative between deregulation and censorship. The point is not to regulate truth. The point is to make the infrastructures of

visibility accountable. This means asking whether ranking systems are transparent enough, whether recommender systems can be scrutinised, whether AI-generated outputs preserve source attribution, whether users have meaningful control, whether researchers can access relevant data, whether regulators can assess systemic risks, whether media pluralism is protected not only formally but structurally.

In this sense, digital constitutionalism is not a decorative or rhetorical category. It is the attempt to update the oldest function of constitutional law — limiting power — in an environment where power is exercised through code, design, interfaces, algorithmic ranking and AI-generated synthesis. It tells us that freedom of expression cannot be protected only downstream, when content is removed or sanctioned. It must also be protected upstream, at the level of visibility, amplification, attribution and contestability. It tells us that pluralism cannot be measured only by counting voices. It must ask whether those voices can actually reach citizens. It tells us that disinformation cannot be addressed only by classifying content as true or false. It must be addressed by examining how the architecture of circulation makes certain distortions powerful.

This is the theoretical foundation of the whole intervention. The cases and instruments that will follow — the Digital Services Act, the AGCOM initiative on Google AI Overviews, the Los Angeles litigation on platform design — should all be read through this lens. They are not isolated developments. They are signs of a broader transformation: constitutional democracies are beginning to understand that the future of freedom of expression and pluralism will increasingly depend on the governance of infrastructures of visibility, amplification and synthesis.

Once we recognise that the digital public sphere is structured by infrastructures of visibility, the meaning of freedom of expression must be reconsidered without being diluted. This is important. The point is not to abandon the classical understanding of freedom of expression. On the contrary, the classical protection of speech against censorship, arbitrary interference, political pressure and disproportionate restrictions remains the indispensable foundation. Article 10 of the European Convention on Human Rights continues to be the central reference point: it protects the right to impart information and ideas, but also the right to receive them. And it is precisely this second dimension — the right to receive information — that becomes particularly important in the digital environment.

In the traditional constitutional model, the right to receive information was often understood in relation to access to media, access to public-interest information, press freedom, broadcasting pluralism and the ability of citizens to encounter a diversity of opinions in the public sphere. In the digital environment, however, the reception of information is mediated by systems that rank, recommend, personalise, filter, monetise and, increasingly, synthesise content. Citizens do not simply receive information. They receive information through architectures that decide what is salient, what is visible, what is repeated, what is suggested, what is contextualised and what is omitted. This is why the right to receive information becomes inseparable from the conditions of algorithmic visibility.

A purely formal conception of freedom of expression is therefore insufficient. It is not enough to ask whether a speaker has been prohibited from speaking. It is also necessary to ask whether the infrastructure of communication allows that speech to become part of public debate. A journalist may publish freely, but if the article is not discoverable, its democratic

function is weakened. A newspaper may not be censored, but if its content is absorbed into synthetic AI-generated outputs without visible attribution, the relationship between source and reader is altered. A civil society organisation may speak, but if recommender systems systematically privilege more sensational, polarising or commercially optimised content, the public sphere may become less hospitable to reasoned democratic discourse. A minority voice may not be banned, but if it is structurally marginalised by ranking and visibility mechanisms, formal liberty does not fully translate into democratic participation.

This is why digital pluralism cannot be reduced to the old question of how many speakers exist. Pluralism has always been more than mere numerical diversity, but in the digital environment this becomes even clearer. A pluralistic public sphere requires not only the formal presence of different voices, but also the effective possibility for those voices to reach citizens. It requires visibility, discoverability, attribution, contestability and sustainability. Visibility means that sources are not merely present somewhere online, but can actually be encountered. Discoverability means that users can find them through the ordinary pathways through which information is accessed. Attribution means that information remains connected to the institutional, editorial or individual source that produced it. Contestability means that decisions affecting ranking, visibility or access can be challenged or scrutinised. Sustainability means that the actors producing public-interest information, especially professional journalism, can continue to exist economically and institutionally.

This is the point at which freedom of expression and pluralism converge. Freedom of expression protects the speaker, but it also protects the public's ability to receive information. Pluralism protects the structure of the public sphere in which that reception occurs. In the analogue environment, this structure was shaped by newspapers, broadcasters, editors, public service media, ownership rules and journalistic institutions. In the digital environment, it is shaped increasingly by platforms, search engines, recommender systems, app stores, advertising infrastructures and AI interfaces. Therefore, if we want to preserve freedom of expression, we must also preserve the pluralistic structure of the informational environment.

This does not mean that every platform must give equal visibility to every source. That would be impossible and undesirable. Pluralism is not mechanical equality of exposure. It is not a duty to amplify every voice in the same way. Nor does it require platforms or regulators to eliminate editorial or algorithmic choices. The issue is more precise: when a digital infrastructure becomes essential for access to information at systemic scale, its choices about visibility, ranking and source treatment can no longer be treated as purely private technical decisions without democratic significance. They become relevant because they affect the practical conditions under which citizens form opinions.

This is particularly evident in relation to professional journalism. Journalism is not only one category of content among others. In a democratic society, it performs an institutional function. It investigates, verifies, contextualises, attributes responsibility, corrects errors, informs public debate and holds power to account. Media freedom is therefore not simply the freedom of publishers to run a business. It is one of the institutional conditions of democratic self-government. If professional journalism becomes less visible, less attributable or less economically sustainable because of the way digital intermediaries organise access to information, the problem is not only market imbalance. It becomes a question of democratic pluralism.

The same is true for local media, minority media and independent sources. A digital environment may appear plural because countless voices formally exist online, but it may in practice be highly concentrated if attention is controlled by a small number of platforms and AI-mediated interfaces. This is one of the paradoxes of the digital age: abundance of expression can coexist with concentration of visibility. There may be more speech than ever before, but fewer infrastructures through which that speech becomes visible at scale. The problem is not scarcity of content. The problem is concentration of attention.

This creates a new kind of constitutional tension. On the one hand, platforms and search engines perform ordering functions that are indispensable. Without ranking, recommendation and filtering, the digital environment would be chaotic and unusable. Some form of organisation is necessary. On the other hand, when the organisation of information is concentrated, opaque and commercially optimised, it may distort pluralism. The constitutional issue is therefore not whether information should be organised. It must be organised. The issue is how it is organised, according to which criteria, with what degree of transparency, with what safeguards, and with what remedies when the organisation produces systemic distortions.

Here the distinction between content moderation and visibility governance becomes important. Content moderation concerns decisions to remove, restrict, label or otherwise act upon specific content or accounts. Visibility governance concerns the broader architecture through which content is ranked, recommended, prioritised, demoted, linked, monetised or summarised. The first is visible and legally familiar. The second is more diffuse, more technical and often less contestable. Yet the second may have a deeper impact on pluralism. A removal decision affects one item of content. A recommender system affects millions of acts of exposure. A ranking criterion affects the everyday organisation of public knowledge. An AI-generated answer may reshape the relationship between users and sources.

This is why the future of freedom of expression cannot be discussed only in terms of takedown and censorship. Those issues remain important, but they are not the whole field. We must also discuss ranking, recommender systems, search, source attribution, synthetic answers, advertising incentives, user control, research access and systemic-risk mitigation. These are not peripheral technical questions. They are the new institutional grammar of the digital public sphere.

The right to receive information becomes particularly important in this context. Citizens do not have a right to receive every piece of information in the same way. But democratic societies have an interest, protected through freedom of expression and pluralism, in ensuring that the informational environment does not become structurally distorted by invisible and unaccountable forms of power. The public must be able to encounter diverse, reliable and attributable sources. It must be able to distinguish journalism from propaganda, source from synthesis, fact from opinion, editorial responsibility from automated generation. If the infrastructure blurs these distinctions, the problem is not only individual deception. It is democratic degradation.

This also helps explain why algorithmic amplification is such a difficult constitutional problem. Amplification is not censorship. It is not the silencing of speech. In fact, it often appears as the opposite: more circulation, more visibility, more engagement. But amplification can distort the public sphere if it systematically privileges content that is inflammatory, manipulative,

sensational, polarising or commercially valuable at the expense of information that is relevant, verified or democratically necessary. The constitutional problem is not that the state has suppressed a voice, but that private systems may reshape the public sphere by rewarding certain communicative behaviours over others.

Again, the solution cannot be simplistic. We should not ask public authorities to decide which content deserves visibility. That would be dangerous. Nor should we imagine that platforms can or should be perfectly neutral. Neutrality is impossible in any system that ranks and recommends. The constitutional question is rather whether the criteria and effects of visibility governance can be made sufficiently transparent, contestable and accountable, especially when the platform or search engine operates at systemic scale and affects media pluralism, civic discourse or democratic processes.

This is where the Council of Europe perspective is particularly valuable. Article 10 is not only a shield against censorship. It is also part of a broader democratic architecture. Freedom of expression protects public debate because democracy depends on the circulation of information and ideas. Media freedom and pluralism are not accessories to expression; they are conditions for its democratic function. The Council of Europe's tradition has always recognised the special role of the press, the importance of pluralism and the need to protect the public watchdog function of journalism. In the digital age, that tradition must be applied not only to newspapers and broadcasters, but also to the infrastructures that determine whether journalistic and public-interest information remains visible.

This does not mean constitutionalising every private design choice. It means recognising that certain design choices, when made by actors of systemic relevance, have constitutional consequences. A small website deciding how to organise its homepage is not the same as a very large search engine deciding how information is ranked or synthesised for millions of users. A private editorial decision by a newspaper is not the same as an opaque recommender system that structures the information diet of an entire population. Scale matters. Infrastructural dependency matters. Market concentration matters. The democratic role of the service matters.

This is why digital pluralism must include a structural dimension. It must protect not only diversity of content, but also diversity of pathways. Citizens should not be dependent on a single architecture of access. They should be able to encounter different sources, different editorial perspectives, different institutional voices, different levels of authority and different interpretive frames. A pluralistic digital environment is not one in which everything is equally visible. It is one in which the conditions of visibility are not arbitrarily controlled, opaque, discriminatory or structurally hostile to public-interest information.

The emergence of AI-generated interfaces makes this even more urgent. If users increasingly access information through conversational assistants or AI-generated summaries, the relationship between expression and reception changes again. The user may not see a list of sources. The user may see a synthetic answer. This can be useful, efficient and innovative. But it also risks reducing the visibility of sources, weakening attribution and concentrating interpretive power in the AI interface. The danger is not that AI speaks. The danger is that AI may become the layer through which citizens access the speech of others without seeing those others clearly.

At this point, freedom of expression, pluralism and media freedom become inseparable. Freedom of expression protects the ability to speak and receive. Pluralism protects the diversity and visibility of sources. Media freedom protects the institutional actors that produce public-interest information. In an AI-mediated public sphere, these values cannot be protected separately. If source visibility disappears, the right to receive information is weakened. If journalism loses sustainability, pluralism is weakened. If synthetic outputs replace access to original sources without transparency and attribution, democratic accountability is weakened.

The constitutional conclusion is therefore clear. In the digital environment, the problem is not only whether speech is allowed. It is whether speech remains visible, attributable and capable of contributing to democratic opinion formation. A democracy can formally tolerate many voices while practically allowing only a few infrastructures to organise their visibility. That is why freedom of expression must be linked to digital pluralism, and digital pluralism must be linked to the governance of infrastructures of visibility. This is not a departure from Article 10. It is an attempt to preserve its democratic meaning in an environment where power over expression increasingly operates through ranking, recommendation and synthesis rather than through direct censorship.

Once freedom of expression and pluralism are understood as depending not only on the formal possibility to speak, but also on the infrastructures through which speech circulates, the issue of disinformation must also be reconsidered. Disinformation cannot be reduced to a simple problem of false content. Of course, false or misleading content matters. It may distort public debate, damage reputations, manipulate voters, endanger public health or undermine trust in institutions. But if we stop at the level of content falsity, we miss the deeper transformation. The democratic danger of disinformation in the digital age lies not only in the production of falsehoods, but in the infrastructural conditions that allow certain falsehoods, distortions or manipulative narratives to circulate at scale, to be personalised, to be amplified, to be monetised and to appear credible.

This distinction is essential. Falsehoods, rumours, propaganda and manipulation existed long before platforms, search engines and artificial intelligence. Democratic societies have always had to deal with lies, ideological distortion, political propaganda and strategic manipulation. What is new is not simply the existence of false speech. What is new is the architecture of circulation. Digital platforms and AI systems can transform isolated falsehoods into mass phenomena through recommender systems, ranking mechanisms, engagement-based amplification, targeted advertising, microtargeting, coordinated inauthentic behaviour, synthetic media and now AI-generated answers. In this environment, disinformation becomes powerful not only because it is false, but because it is made visible, repeated, personalised and detached from source accountability.

The problem is therefore not only epistemic, but constitutional. Disinformation affects democracy because it interferes with the conditions under which citizens form opinions freely. If citizens are exposed to information flows that are shaped by opaque systems optimised for engagement, emotional intensity or behavioural prediction, their ability to receive information in a pluralistic and accountable public sphere may be weakened. This does not mean that citizens are passive or incapable of judgment. But it does mean that the environment in which judgment is exercised is no longer neutral. It is structured, optimised and mediated by private infrastructures whose criteria are often inaccessible to users, researchers and regulators.

This is why disinformation should be seen as a stress test for freedom of expression. On the one hand, democratic societies must protect themselves against manipulation, coordinated distortion, foreign interference, electoral disruption, health-related falsehoods and attacks on trust in democratic institutions. On the other hand, the response to disinformation can itself become dangerous if it turns into broad content control, over-removal, pressure on platforms to suppress contested speech, or public authority intervention in the definition of truth. The constitutional problem is therefore double: too little intervention may leave the public sphere vulnerable to manipulation; too much intervention may undermine the freedom and openness that democratic discourse requires.

This double risk is particularly important in the Council of Europe context. Article 10 of the European Convention on Human Rights does not protect only correct, polite, moderate or institutionally approved speech. It also protects speech that shocks, offends or disturbs, subject to the limits that are necessary in a democratic society. This is a crucial safeguard. If the fight against disinformation is framed too broadly, it may capture political dissent, investigative journalism, satire, minority opinions, controversial scientific debate or legitimate criticism of public authorities. The language of disinformation can be abused by governments to delegitimise opposition or inconvenient media. This is why any response must remain firmly anchored in legality, necessity, proportionality, procedural guarantees and independent oversight.

At the same time, however, the classical protection of expression cannot become an excuse for ignoring structural manipulation. A purely laissez-faire model is no longer adequate. The argument that “the remedy to bad speech is more speech” becomes less convincing when the conditions of visibility are not open and symmetrical, but shaped by opaque algorithmic systems. More speech does not necessarily correct falsehood if the architecture of the platform amplifies the falsehood more effectively than the correction. A careful investigation by a journalist may be available, but a misleading narrative may circulate faster, be recommended more widely, be emotionally optimised and reach audiences in personalised ways. The marketplace of ideas metaphor becomes fragile when the marketplace itself is engineered.

This is why the distinction between content and amplification is fundamental. A false statement is one thing. A system that repeatedly recommends that statement, monetises its circulation, personalises its delivery, and places it in a context that increases its persuasive force is another. The constitutional issue is not only whether the statement should be removed. The issue is whether the system that gives it disproportionate reach is accountable. The same applies to political manipulation. The problem is not only that a misleading political message exists. The problem is whether advertising systems, recommender systems or data-driven targeting infrastructures allow that message to reach specific groups in opaque and asymmetrical ways, without sufficient transparency or contestability.

This also explains why disinformation is connected to pluralism. A disinformation environment does not merely add falsehoods to the public sphere. It can crowd out reliable information. It can weaken professional journalism. It can reduce trust in institutional sources. It can create informational fragmentation, where different groups inhabit different perceived realities. It can generate cynicism, in which citizens no longer know whom to trust and therefore become less capable of democratic judgment. In such a context, the problem is not only that false

information circulates. The problem is that the pluralistic structure of the public sphere is degraded.

The rise of generative AI makes this problem more complex. Generative systems can produce text, images, audio and video at scale. They can lower the cost of producing persuasive falsehoods. They can imitate journalistic formats, institutional language or human communication. They can generate synthetic content that appears plausible even when inaccurate. But the deeper issue is not only the generation of false content. It is the integration of generative AI into the infrastructures of information access. When AI systems do not merely produce content but also mediate how users search, receive and interpret information, disinformation risks becoming embedded in the very architecture of knowledge access.

This is especially visible in AI-generated summaries and answers. A synthetic answer may appear neutral and authoritative because it is presented by the interface as the result of a query. Users may not see the underlying sources. They may not know which sources were selected, how they were weighted, whether relevant perspectives were excluded, whether journalistic material was used, whether the output contains inaccuracies or whether the system has produced a false but plausible synthesis. In this context, disinformation is not only viral falsehood. It may also be synthetic distortion: an answer that compresses, omits, misattributes or decontextualises information while appearing reliable.

This is why source attribution becomes a democratic safeguard. In a pluralistic public sphere, citizens must be able to know where information comes from. They must be able to distinguish between a newspaper investigation, an official source, an academic study, a political claim, a user-generated post and an AI-generated synthesis. If these distinctions are blurred, citizens lose one of the basic tools of democratic judgment: the ability to evaluate the credibility, authority and accountability of sources. Source transparency is therefore not merely a technical matter. It is part of the infrastructure of freedom of expression.

The same applies to contestability. If a piece of content is removed, the speaker may seek a remedy. But if a source is systematically deprioritised, if a narrative is amplified in opaque ways, if an AI-generated answer misrepresents a publisher's work, or if a recommender system creates systemic exposure to manipulative content, the harm may be diffuse and difficult to challenge. This is one of the main problems of architectural power: it does not always produce discrete, identifiable decisions. It produces patterns. And constitutional law, which is often built around identifiable acts of interference, must adapt to patterns of infrastructural influence.

The response to disinformation should therefore be procedural and structural rather than purely substantive. The goal should not be to create public authorities that decide what is true or false in every contested matter. The goal should be to ensure that the systems through which information circulates are transparent, auditable, contestable and proportionate. This means requiring platforms and AI systems to assess how their design choices affect the spread of disinformation. It means ensuring access to data for independent researchers. It means transparency in political advertising and recommender systems. It means clear attribution of AI-generated or AI-mediated content. It means meaningful remedies for users and media actors. It means mitigating systemic risks without imposing official truth.

This is also why digital constitutionalism is particularly helpful. It allows us to avoid a simplistic free speech absolutism that ignores infrastructural manipulation, but also a regulatory paternalism that treats citizens as passive subjects to be protected by controlling what they may see. Digital constitutionalism asks a different question: how can the architecture of public discourse be made compatible with freedom, pluralism and democratic self-government? The answer lies not in controlling opinions, but in controlling arbitrariness; not in suppressing disagreement, but in making systems accountable; not in eliminating risk, but in ensuring that risk is assessed and mitigated through transparent and rights-respecting processes.

In this sense, disinformation is the point where the tension between freedom of expression and digital pluralism becomes most visible. Freedom of expression requires openness, tolerance for disagreement and protection of controversial speech. Digital pluralism requires a public sphere in which different sources can be encountered, compared and trusted. Disinformation threatens pluralism by polluting the informational environment, but anti-disinformation measures may threaten expression if they become excessive or politicised. The constitutional challenge is to preserve both: to protect the public sphere from manipulation without undermining the freedom that makes democratic debate possible.

This is why the fight against disinformation must shift from a narrow focus on content removal to a broader focus on systemic conditions. The question should not be only: is this content false and should it be removed? The question should also be: why did this content become so visible? Through which recommendation logic? With what advertising incentives? With what data-driven targeting? With what source attribution? With what correction mechanisms? With what access for independent scrutiny? With what impact on journalism, minorities, elections, public health or democratic trust?

This approach also helps avoid the privatisation of truth. If platforms are simply pressured to remove disinformation, they become private arbiters of public debate. Their terms of service, automated tools and moderation policies may end up shaping the boundaries of lawful expression without sufficient democratic legitimacy. But it also avoids the nationalisation of truth. If states define disinformation too broadly, they may use that category to discipline public debate. The better constitutional approach is neither private truth nor public truth. It is accountable infrastructure.

The central point, therefore, is that disinformation reveals the inadequacy of a purely content-based approach to freedom of expression. The future of democratic communication will not be secured only by deciding which content stays online and which content is removed. It will depend on whether the systems that organise information flows are transparent enough, pluralistic enough and contestable enough to allow citizens to form opinions freely. In this sense, disinformation is not merely a problem to be managed within the digital public sphere. It is one of the clearest signs that the digital public sphere itself has become an object of constitutional concern.

This prepares the ground for the Digital Services Act. The DSA matters because it tries, at least in part, to respond to this structural problem. It does not limit itself to illegal content. It introduces the language of systemic risk. It recognises that platforms and search engines of very large scale may affect fundamental rights, civic discourse, electoral processes, media freedom and pluralism through the design and functioning of their services. That is why the DSA is not only a technical regulatory instrument. It is one of the first attempts to translate the

constitutional problem of disinformation, amplification and infrastructural power into a legal framework.

The Digital Services Act must be understood as one of the first regulatory attempts to respond to the constitutional transformation described so far. It is not simply a technical instrument for platform compliance. Nor is it merely an update of intermediary liability rules. Its deeper importance lies in the fact that it moves European digital regulation beyond a purely content-centred model and introduces a systemic-risk approach. This is the key point: the DSA does not ask only whether a specific item of content is illegal, whether a platform has removed it too late or too early, or whether a user has been given an adequate remedy. It also asks whether the design and functioning of very large online platforms and very large online search engines may generate risks for fundamental rights, civic discourse, electoral processes, public security, media freedom and pluralism.

This is a decisive shift. For many years, the governance of online speech was dominated by the language of content moderation. The central questions were familiar: what should be removed? What should stay online? Who decides? Under which procedural safeguards? Are platforms too permissive or too restrictive? Should they be liable for third-party content? How can illegal content be detected and removed without undermining lawful speech? These questions remain important. But they no longer capture the full structure of the problem. In the contemporary digital public sphere, some of the most serious democratic risks do not arise only from single pieces of illegal content. They arise from the architecture of the service: recommender systems, ranking mechanisms, advertising models, interface design, automated moderation tools, data-driven targeting, and now AI-mediated systems that organise, summarise or replace access to information.

The DSA recognises this by giving legal relevance to systemic risks. Article 34 requires very large online platforms and very large online search engines to identify, analyse and assess systemic risks stemming from the design, functioning and use of their services. Article 35 requires them to adopt reasonable, proportionate and effective mitigation measures. This structure is important because it transforms the platform from a mere host of content into an environment capable of producing constitutional effects. The risk does not lie only in what users say. It may lie in how the platform makes some speech visible, how it recommends certain narratives, how it shapes electoral debate, how it affects minors, how it interacts with disinformation campaigns, how it affects media pluralism, or how it influences the conditions under which citizens receive information.

Article 27 is also central because it deals with recommender-system transparency. Recommender systems are not secondary technical features. They are one of the principal mechanisms through which the digital public sphere is structured. They decide what users see next, what is repeated, what becomes familiar, what is prioritised, what is suggested as relevant and what disappears from ordinary exposure. In an environment where most citizens receive information through digital intermediaries, recommender systems directly affect the right to receive information and the conditions of pluralism. This is why their transparency is not only a consumer-protection issue. It is part of the democratic accountability of the communication infrastructure.

The constitutional importance of the DSA therefore lies in the shift from removal to risk, from single decisions to systemic effects, from content legality to architectural responsibility. The

law begins to recognise that platforms and search engines are not neutral channels. They are complex socio-technical systems whose design choices can affect the public sphere. A recommender system may intensify polarisation. A ranking system may shape access to journalistic sources. An advertising model may incentivise manipulation. A moderation policy may create chilling effects. A search interface may influence which sources are considered authoritative. An AI-generated summary may reduce source visibility. These are systemic questions, not merely content questions.

At the same time, the DSA does not and should not become a mechanism for public authorities to decide truth. This is essential. The systemic-risk approach is constitutionally valuable precisely because it can address the architecture of information circulation without reducing the issue to state control over content. The aim should not be to determine which opinions are correct or which narratives should dominate public debate. The aim is to ensure that platforms assess and mitigate risks generated by their own systems, and that these assessments are subject to scrutiny, transparency, audit and enforcement. In other words, the DSA should be read as a framework for infrastructure accountability, not as a framework for truth management.

This distinction is particularly important in relation to disinformation. If the DSA were applied as a tool for aggressive content suppression, it could create serious risks for freedom of expression. Platforms might remove too much in order to avoid regulatory pressure. Lawful but controversial speech could be affected. Political dissent, satire, investigative journalism and minority viewpoints could be wrongly classified as harmful. This is why enforcement must remain anchored in legality, necessity, proportionality, due process and transparency. The systemic-risk framework must not become a shortcut to censorship.

But the opposite risk must also be avoided. If the DSA is applied too cautiously, or if systemic-risk assessments become formalistic exercises, the law may fail to address the real sources of democratic harm. A platform can comply with notice-and-action procedures and still operate recommender systems that amplify disinformation. A search engine can provide formal transparency while structuring access to information in ways that weaken source visibility. A very large platform can produce long compliance reports while its design continues to reward polarisation, emotional manipulation or attention capture. The value of the DSA depends on whether it can move beyond procedural formality and actually confront infrastructural power.

This raises the question of enforcement. The systemic-risk model requires regulators capable of understanding complex technological systems. It requires access to data. It requires independent auditing. It requires cooperation between national Digital Services Coordinators and the European Commission. It requires research access and civil society scrutiny. It requires a legal culture able to connect technical design with fundamental rights impact. Without these elements, systemic risk remains an elegant legal concept but not an effective tool of democratic accountability.

The DSA also introduces a multilayered institutional model. National authorities play a crucial role, especially as Digital Services Coordinators. The European Commission has direct supervisory powers over very large online platforms and very large online search engines. The European Board for Digital Services contributes to coordination. This structure reflects the nature of the problem. The effects of platform governance are felt nationally, because media ecosystems, elections, public debates and democratic vulnerabilities are often national. But

the infrastructures are transnational, because the relevant actors operate across borders and at European or global scale. The DSA therefore tries to build an enforcement architecture that matches this mismatch between national democratic effects and transnational digital infrastructures.

This institutional design is particularly relevant for media freedom and pluralism. A risk to pluralism may emerge within a specific national information ecosystem: for example, because local publishers lose visibility, because smaller media actors are disadvantaged, because political debate is distorted, or because AI-generated systems change access to news sources. But the platform or search engine responsible for the relevant architecture may be global. A national authority may detect the risk, but European coordination may be necessary to address it effectively. This is precisely why the DSA's architecture matters. It creates channels through which national concerns about democratic communication can be escalated into European regulatory scrutiny.

From a Council of Europe perspective, however, the DSA should be placed within a broader human rights framework. The DSA is an EU regulation. Its logic is partly internal-market-based and partly rights-based. The Council of Europe adds a different layer: the normative grammar of human rights, democracy and rule of law. Article 10 ECHR, media freedom standards, pluralism and democratic resilience provide the constitutional background against which DSA enforcement should be evaluated. The DSA may provide the regulatory machinery, but the Council of Europe framework helps define the limits and purposes of that machinery.

This matters because platform governance always risks moving in two opposite directions. It can become too weak, leaving private infrastructures free to shape public discourse without accountability. Or it can become too intrusive, allowing public authorities to pressure platforms into controlling lawful speech. A Council of Europe reading helps prevent both outcomes. It insists that platform accountability is necessary, but also that any intervention affecting expression must be rights-respecting, proportionate and subject to safeguards. It reminds us that the goal is not to make platforms instruments of state control, but to ensure that their own power over the public sphere does not remain arbitrary.

The DSA's systemic-risk approach is therefore best understood as an attempt to constitutionalise platform governance without transforming regulators into editors. This formulation is important. The regulator should not decide which political opinions should prevail. It should not define official truth. It should not impose a hierarchy of legitimate viewpoints. But it may legitimately require systemic actors to explain, assess and mitigate how their systems affect public debate, media pluralism, electoral integrity and fundamental rights. The object of regulation is not opinion as such. The object is infrastructural risk.

This distinction also helps explain why recommender systems and AI-mediated search interfaces fall naturally within the DSA logic. They do not merely distribute content. They organise exposure. They create hierarchies of relevance. They influence what citizens see first, what they see repeatedly, what they never encounter, and what appears authoritative. In the case of generative AI and AI-generated search, the issue becomes even more complex because the system may not simply rank sources but produce an answer. In that context, systemic-risk assessment must ask whether source visibility, attribution, pluralism and access to original information are preserved.

The DSA thus becomes a bridge between the theoretical problem and the concrete cases. The theoretical problem is that freedom of expression and pluralism now depend on infrastructures of visibility. The regulatory answer begins with the recognition that very large platforms and search engines must assess and mitigate systemic risks arising from their design and functioning. The concrete question then becomes: how does this framework apply when a search engine introduces AI-generated answers that may reduce visibility of news sources? Or when a social media platform's design may intensify harmful patterns of attention and behaviour? This is where the AGCOM initiative and the Los Angeles cases become illustrative.

The most important point is that the DSA changes the object of democratic scrutiny. It does not focus only on what users say, but on how platforms structure the conditions under which speech circulates. It does not focus only on illegal content, but on systemic effects. It does not focus only on individual removals, but on design, risk and mitigation. This is why it is constitutionally significant. It gives legal form to the intuition that digital infrastructures of visibility must be made accountable.

But the DSA also faces a challenge of temporal adaptation. The digital environment is changing rapidly. When the DSA was conceived, the dominant paradigm was still the platform model: hosting, ranking, recommending, moderating, advertising. Today, generative AI and AI-mediated search are shifting the paradigm again. Search becomes answer. Recommendation becomes synthesis. Intermediation becomes interpretation. The DSA has tools that can address this transformation, especially through systemic-risk obligations, but it will need dynamic interpretation and strong enforcement to remain effective.

This is the point at which the AGCOM initiative becomes especially revealing. It shows how a national authority can identify a new kind of risk that sits at the intersection of search, AI, media pluralism and DSA systemic-risk obligations. It also shows that the regulatory challenge is no longer limited to whether platforms remove illegal content. The challenge is whether AI-mediated infrastructures preserve the conditions of democratic access to information. That is the next step.

The AGCOM initiative concerning Google AI Overviews and AI Mode is one of the clearest examples of the shift from content moderation to infrastructure accountability. It is important to introduce it carefully, especially in an international and Council of Europe context, because AGCOM may not be familiar to all participants. AGCOM, the Autorità per le garanzie nelle comunicazioni, is the Italian independent Communications Regulatory Authority. It is not a ministry, nor a political body. It is an independent national regulator traditionally responsible for electronic communications, telecommunications, audiovisual media services, media markets and the protection of media pluralism. In the Italian system, AGCOM has long played a role in supervising the conditions under which communication markets and media pluralism operate. Under the Digital Services Act, AGCOM has also been designated as Italy's Digital Services Coordinator, and therefore forms part of the European enforcement architecture for online platforms, search engines and digital intermediaries.

This institutional clarification is important because the case should not be read as a merely Italian episode. It is not simply an Italian regulator defending Italian publishers against Google. It is a national independent authority identifying a possible systemic risk within its national media ecosystem and activating a European mechanism under the DSA. The actor concerned

is global. The service operates transnationally. The information infrastructure is European and global. But the effects may be felt in specific national media systems: in the visibility of national newspapers, in the sustainability of local or independent publishers, in citizens' access to journalistic sources, and in the conditions under which public opinion is formed. This is precisely why the case is relevant for a Council of Europe discussion on freedom of expression, pluralism and democratic resilience.

The case concerns Google AI Overviews and AI Mode. These tools should not be understood merely as technical improvements to search. They signal a deeper transformation in the function of search. Traditional search, even when far from neutral, still operates mainly as a gateway. The user asks a question and receives a list of links, snippets and possible routes toward a plurality of sources. Ranking power is already extremely important in this model: what appears first matters; what appears on the second or third page may become practically invisible; what is indexed and how it is presented shapes the user's experience. But the basic architecture remains one in which the user is directed toward identifiable sources. The plurality of sources remains visible, at least formally.

AI-mediated search changes this architecture. When the search engine produces an answer directly at the point of access, it no longer simply directs the user toward information. It selects, organises, compresses and reformulates information. Search becomes answer. Ranking becomes synthesis. Intermediation becomes interpretation. The user may no longer move from the query to a plurality of visible sources. The user may instead receive a unified output generated by a private AI system. This is not necessarily illegitimate. It may be efficient, useful and innovative. But it alters the constitutional structure of access to information.

The problem is therefore not only whether Google provides accurate answers in individual cases. The deeper question is how AI-generated answers affect the visibility of sources, the attribution of information, the sustainability of journalism and the pluralistic structure of the public sphere. If professional journalistic content is used to generate synthetic answers, but the user no longer reaches the original source, the relationship between journalism and public attention changes. If the answer appears complete, neutral and authoritative, the user may have less incentive to compare sources. If the source logic is opaque, users may not know which materials were selected, which were excluded, how they were weighted or whether the output faithfully reflects the underlying information. This is a pluralism issue, not merely a technological one.

This is why the concern raised by publishers cannot be reduced to the economic loss of traffic, although traffic is important. Referral traffic is not only a business metric. For journalism, it is part of the relationship between the source and the public. It allows readers to identify the institution that produced the information, to evaluate its credibility, to access broader context, to encounter editorial responsibility and to support the economic sustainability of public-interest media. If AI-generated search interfaces reduce traffic to original sources, the issue is not simply that publishers lose revenue. The issue is that the democratic infrastructure of source visibility may be weakened.

Here pluralism takes on a new meaning. Traditional pluralism asks whether there are many voices. Digital pluralism asks whether those voices remain visible within the architectures through which citizens access information. AI-mediated search may preserve the formal existence of many sources while reducing their practical visibility. A newspaper may still

publish. A source may still exist. A journalist may still investigate. But if the user receives an answer generated by the platform without meaningfully encountering the source, the democratic function of that source is weakened. Pluralism is not destroyed by prohibition; it is diluted by absorption.

This is the core constitutional issue of AI Overviews. The risk is not only that a given answer may be wrong. The risk is that the architecture of access to knowledge shifts from navigation among sources to consumption of synthetic outputs. The internet, at least in its search-engine form, was structured around a plurality of links. AI-generated search may increasingly be structured around answers. A link-based environment is not neutral, but it keeps sources visible. An answer-based environment may be more convenient, but it can concentrate interpretive power in the interface. The question is therefore how to preserve source visibility, attribution and contestability when access to information becomes AI-mediated.

This issue is also directly connected to disinformation. In the traditional platform environment, disinformation often circulates through viral content, recommender systems, coordinated campaigns or engagement-driven amplification. In AI-mediated search, the problem may take a different form. The risk may be synthetic distortion: an output that appears authoritative but omits relevant sources, overrepresents certain materials, misattributes information, produces hallucinations, or compresses complex public issues into an apparently neutral answer. The danger is intensified by the authority of the interface. A user may perceive an AI-generated answer as more objective than a list of sources, precisely because it is presented as a synthesis.

This is why attribution is so important. In a democratic information environment, citizens must be able to know where information comes from. They must be able to distinguish between original journalism, official information, academic research, political communication, user-generated content and AI-generated synthesis. If AI interfaces blur these distinctions, the user's capacity to assess reliability is reduced. Source attribution is therefore not only a matter of fairness to publishers. It is a condition of democratic judgment.

AGCOM's initiative should be read against exactly this background. As Italy's independent communications regulator and Digital Services Coordinator, AGCOM asked the European Commission to assess Google AI Overviews and AI Mode in light of the Digital Services Act, particularly the provisions concerning recommender-system transparency, systemic-risk assessment and systemic-risk mitigation. The legal question is not simply whether a particular output is unlawful. The question is whether the design and functioning of AI-mediated search may produce systemic risks for media pluralism, access to information, transparency and democratic discourse.

This is where the DSA becomes an essential bridge. Articles 34 and 35 allow the analysis to move from individual content to systemic effects. Article 27 allows attention to be paid to the transparency of recommender systems and the ways in which users are exposed to information. The AGCOM case therefore tests whether the DSA's systemic-risk logic can be applied to a new generation of AI-mediated information interfaces. It asks whether search engines that produce synthetic answers must assess not only accuracy or user convenience, but also their impact on media freedom, source visibility and pluralism.

The case also reveals an institutional transformation in the role of national regulators. Authorities like AGCOM are no longer simply sectoral regulators concerned with telecommunications infrastructure, broadcasting licences or media markets in the traditional sense. They are becoming nodes in the constitutional governance of the digital public sphere. This does not mean that they should become editors or arbiters of truth. That would be dangerous. Their role is not to decide which sources citizens should trust or which narratives should prevail. Their role is to ensure that systemic digital infrastructures remain transparent, accountable and compatible with the democratic conditions of communication.

This is a delicate balance. On the one hand, regulatory intervention is necessary because global private infrastructures can affect national media pluralism and democratic debate. On the other hand, regulatory intervention must not become a form of public control over speech. The correct constitutional object is not the truth of individual statements, but the accountability of the systems that organise access to information. AGCOM's initiative is important precisely because it can be framed in this way: not as an attempt to regulate opinions, but as an attempt to ask whether the architecture of AI-mediated search creates risks for pluralism and whether those risks are properly assessed and mitigated under the DSA.

The international significance of the case lies in the mismatch between the scale of infrastructure and the scale of democratic effects. Google operates globally. AI search interfaces are developed and deployed across jurisdictions. But the effects on media ecosystems are local and national: they concern Italian newspapers, local publishers, national public debate, linguistic communities, minority media, and citizens' access to information in their own democratic context. This mismatch is one of the defining challenges of digital constitutionalism. The power is transnational, but democracy remains largely organised through national and regional public spheres.

For the Council of Europe, this makes the AGCOM initiative especially relevant. The Council of Europe is concerned with human rights, democracy, rule of law, media freedom and pluralism. It is therefore the right forum to understand why AI-mediated search is not merely a technological innovation. It is a transformation in the conditions of receiving information. Article 10 ECHR protects not only the right to impart information, but also the right to receive it. If the reception of information is increasingly mediated by AI systems that select, summarise and present public knowledge, then the human rights dimension becomes unavoidable.

The key issue is not whether AI Overviews should exist. The issue is under what conditions AI-mediated access to information can be compatible with pluralism. Several conditions become crucial. First, source visibility: users should be able to see and access the original sources underlying AI-generated outputs. Second, attribution: journalistic and public-interest sources should not disappear behind a synthetic answer. Third, transparency: regulators and researchers should be able to understand how sources are selected and weighted, at least at a meaningful systemic level. Fourth, contestability: publishers and users should have mechanisms to challenge misrepresentation, exclusion or systematic invisibility. Fifth, proportionality: regulatory responses must protect pluralism without becoming state control over lawful expression.

This is why the AGCOM case matters as a constitutional laboratory. It shows that the future of media pluralism will not be decided only by ownership rules or broadcasting regulation. It will also be decided by the design of AI-mediated interfaces that organise access to information. It

shows that the future of freedom of expression will not be decided only by takedown decisions. It will also depend on whether sources remain visible in the architectures of search and synthesis. It shows that the fight against disinformation will not be effective if it focuses only on false content while ignoring the systems that produce synthetic authority.

The point can be put in a simple formula: when search becomes synthesis, pluralism depends on source visibility. If the public sphere moves from links to answers, from navigation to generation, from access to sources to consumption of synthetic outputs, then constitutional democracies must ask how to preserve the plurality, attribution and contestability of information. This is the broader meaning of the AGCOM initiative. It is not just an Italian regulatory move. It is an early sign of a wider constitutional question: how should democratic societies govern AI systems that do not merely host or rank speech, but mediate the very way in which citizens access public knowledge?

The second concrete development comes from the United States and, more specifically, from the recent litigation in Los Angeles concerning Meta, Google, Instagram and YouTube. At first sight, this may seem very far from the AGCOM initiative and from the European Digital Services Act. The legal system is different, the constitutional background is different, the doctrinal vocabulary is different. In Europe, the discussion is framed through systemic risk, media pluralism, platform regulation and fundamental rights. In the United States, the discussion is shaped by the First Amendment, Section 230, product liability, negligence and the distinction between content and design. And yet, despite these differences, the underlying transformation is strikingly similar. The debate is moving from liability for content to accountability for architecture.

For many years, the American debate on platforms has been dominated by two large legal shields: the First Amendment and Section 230 of the Communications Decency Act. The First Amendment protects speech and creates strong limits on governmental interference with expression. Section 230 has historically protected platforms from liability for third-party content posted by users. As a result, many claims against platforms have failed when courts characterised them as attempts to hold the platform liable for content created by others. The basic logic was content-centred: if the harm derives from user speech, the platform is generally protected. This framework reinforced the idea of platforms as intermediaries rather than publishers or direct speakers in relation to user-generated content.

The Los Angeles litigation is important because it tests a different route. The focus is not primarily on a specific post, video, comment or item of third-party content. The focus is on the design and operation of the platform environment itself: recommender systems, infinite scroll, autoplay, addictive design, engagement optimisation, failure to warn, and design choices allegedly capable of producing foreseeable harms. In this type of litigation, the question is not simply: who said what online? The question becomes: how was the digital environment designed, and what behavioural, psychological or social effects did that design foreseeably produce?

This is a major conceptual shift. If a platform is sued because it hosted a harmful video or failed to remove a specific item of content, the case remains within the traditional content-liability framework. But if the claim is that the platform's architecture was designed to capture attention, intensify engagement, encourage compulsive use or expose users to harmful patterns of interaction, the legal object changes. The target is not speech as such. The target

is design. The claim is not that the platform should be liable because users expressed themselves. The claim is that the platform created, optimised or maintained an environment whose structural features produced harm.

This distinction is crucial also from a constitutional perspective. It allows courts to examine forms of platform power that do not fit easily within a content-removal paradigm. A recommender system is not simply a collection of user statements. Infinite scroll is not a user's expression. Autoplay is not a political opinion. Notification architecture is not speech in the classical sense. Engagement optimisation is not merely the hosting of content. These are design choices. They organise the environment in which expression, attention and behaviour interact. They influence how long users remain, what they see next, how emotionally engaged they become, and how the platform monetises their attention.

This is why the California cases matter for a broader discussion on freedom of expression and pluralism. They show that the constitutional and legal focus is shifting from the content layer to the architectural layer. The question is no longer only whether platforms should remove harmful content. The question is whether platforms should be accountable for the way they design systems that amplify, recommend, personalise and monetise content. This is not the same as imposing liability for speech. It is asking whether the infrastructure of speech has been designed in ways that generate foreseeable and preventable harms.

Of course, the American constitutional environment makes this issue especially delicate. There is a real risk that design-based liability could be framed too broadly and become an indirect way of imposing liability for speech. Courts therefore need to distinguish carefully between claims that genuinely target platform design and claims that are, in substance, attempts to punish platforms for hosting or disseminating user-generated content. This line is not always easy to draw. Recommender systems, for example, sit at the boundary between design and expression. They are technical systems, but they also organise speech. They do not create the underlying content, but they influence the distribution and visibility of that content. This is exactly why the issue is so legally and constitutionally difficult.

But the importance of the Los Angeles litigation lies precisely in the fact that it makes this boundary visible. It forces the legal system to ask whether platforms can continue to rely on a content-centred immunity model when the harms alleged arise from product design, behavioural optimisation and architecture. It suggests that the old distinction between publisher and conduit may no longer be adequate for platforms whose core function is not passive hosting but active organisation of attention. In this sense, the case mirrors, in the American legal language, the same transformation that the DSA captures in European regulatory language: platforms are not neutral spaces; they are designed environments.

The connection with disinformation is also important. The Los Angeles cases may concern addiction, minors, psychological harm and platform design more than disinformation in the strict sense. But the underlying logic is highly relevant. The same architectures that capture attention can also amplify disinformation. Infinite scroll, autoplay, engagement-based recommender systems and behavioural profiling are not only tools of user retention. They are also mechanisms through which content can be made viral, emotionally salient and difficult to escape. If a system is optimised for engagement, it may privilege content that provokes outrage, fear, anger or emotional intensity, regardless of its reliability or democratic value.

This does not mean that platforms deliberately promote disinformation in every case. The point is more structural. A design optimised for engagement may create incentives and effects that are hostile to a healthy public sphere. It may reward sensationalism over accuracy, speed over verification, emotional intensity over context, virality over responsibility. In such an environment, disinformation becomes not only a content problem but an outcome of design incentives. The architecture makes certain types of speech more powerful because they are better adapted to the platform's logic.

This is why the movement from content to design is so important. It gives legal and constitutional language to the idea that harm can arise from the structure of the environment, not only from individual utterances. In Europe, this insight appears as systemic risk under the DSA. In the United States, it appears through product liability, negligence, failure to warn and design-defect theories. The doctrines are different, but the intuition is similar: the platform's architecture matters.

The Los Angeles cases also show that attention itself has become a regulatory and constitutional object. In the old media environment, attention was shaped by editors, broadcasters, newspapers and programming choices. In the platform environment, attention is captured, measured, predicted and optimised in real time. The user is not merely a reader, viewer or listener. The user becomes part of a behavioural feedback loop. Every click, pause, scroll, like, share and viewing duration becomes data for the system. The platform learns what keeps the user engaged and then reorganises the environment accordingly. This is not merely communication. It is behavioural governance.

From the perspective of digital constitutionalism, this is crucial. If power is exercised through the organisation of attention, then constitutional analysis must examine the infrastructures that allocate attention. Freedom of expression is not only about the speaker's right to speak. It is also about the conditions under which citizens encounter speech. A system that captures attention through addictive design may affect the autonomy of users. A system that recommends content through opaque engagement metrics may affect pluralism. A system that personalises exposure may fragment the public sphere. These are not traditional censorship problems, but they are constitutional problems.

The Los Angeles cases also help us understand why platform accountability cannot be reduced to moderation. A platform may have content rules, removal procedures and reporting mechanisms, but its design may still create harmful patterns. It may remove some illegal content and still operate recommender systems that intensify exposure to manipulative or harmful material. It may comply with formal transparency obligations and still optimise for compulsive use. This is why the focus on architecture is necessary. Content moderation treats symptoms. Architecture accountability examines conditions.

This does not mean that all design-based claims are automatically justified. There is a need for legal precision. Not every harm associated with platform use can be attributed to platform design. Not every design choice is unlawful. Not every recommender system is harmful. Innovation and user choice matter. Platforms also enable expression, association, access to information and community formation. A constitutional analysis must therefore avoid moral panic and technological determinism. The question is not whether platforms are good or bad. The question is whether, when platforms operate at systemic scale, their design choices are

subject to adequate accountability when they foreseeably affect rights, autonomy, pluralism or democratic discourse.

This is where the comparison with Europe becomes useful. The American model proceeds mainly through litigation, often after harms have occurred. The European model, through the DSA, proceeds through ex ante obligations, risk assessment, mitigation and regulatory supervision. The American approach is case-driven and liability-oriented. The European approach is systemic and regulatory. Both have strengths and weaknesses. Litigation can reveal concrete harms and force discovery of internal design practices, but it may be fragmented, slow and uncertain. Regulation can impose preventive obligations and systemic scrutiny, but it may become formalistic or overly dependent on regulators' technical capacity.

The Council of Europe perspective can help connect these models without collapsing them. It reminds us that the key issue is not the specific doctrinal path — product liability in California or systemic risk in Brussels — but the underlying human rights question: how do digital architectures affect autonomy, freedom of expression, the right to receive information, pluralism and democratic self-government? Whether the legal route is judicial or regulatory, the constitutional problem is the same: infrastructures that organise attention and visibility must not remain immune from accountability simply because they operate through design rather than direct prohibition.

The Los Angeles cases therefore contribute to the broader argument in three ways. First, they show that the move from content to architecture is not only European. It is also emerging in the United States, although through a different legal grammar. Second, they show that platform design can be treated as a legally relevant source of harm, distinct from third-party speech. Third, they show that attention-capturing infrastructures may affect not only individual users, but the broader quality of the public sphere.

In this sense, the California litigation and the AGCOM initiative should be read together. AGCOM focuses on AI-mediated search and source visibility. Los Angeles focuses on platform design and attention capture. One concerns access to information. The other concerns the architecture of engagement. But both reveal the same constitutional transformation: the problem is no longer only what content is online. The problem is how digital infrastructures organise visibility, attention, exposure and meaning.

The formula is therefore simple but important. In Europe, the question becomes: how do we assess and mitigate systemic risks generated by platforms and AI-mediated search interfaces? In California, the question becomes: can platforms be held accountable for design choices that foreseeably produce harm? In both cases, the law is moving toward the same object: not speech alone, but the architecture of speech.

This is why the Los Angeles cases matter in a Council of Europe discussion. They show that the international debate is converging around a shared intuition: the digital public sphere cannot be governed only by rules about content removal or intermediary immunity. It must also be governed through rules, principles and safeguards concerning design, visibility, amplification and attention. The challenge is to do this without undermining freedom of expression. That is the central constitutional balance: to hold architectures accountable without turning either courts or regulators into controllers of lawful speech.

This is why the Council of Europe is the right institutional setting for this discussion. The issue is not only technological. It is not only economic. It is not only about platform compliance or market contestability. It concerns the conditions under which democratic societies preserve freedom of expression, media pluralism, public debate and the rule of law in an environment increasingly shaped by digital and AI-mediated infrastructures. The Council of Europe's distinctive contribution is precisely this: it provides the human rights and democratic grammar through which the transformation of the digital public sphere must be understood.

The European Union, especially through the Digital Services Act, provides an increasingly sophisticated regulatory machinery. It creates obligations, procedures, risk assessments, mitigation duties, transparency requirements and an enforcement architecture involving the Commission and national Digital Services Coordinators. But the Council of Europe provides the deeper normative frame: Article 10 of the European Convention on Human Rights, media freedom, pluralism, democratic participation, proportionality, legality, necessity, rule of law and safeguards against arbitrary interference. These are not merely background values. They are the constitutional criteria through which any platform or AI governance model should be evaluated.

The point is important because digital governance always risks being captured by two different logics. The first is a purely market logic. According to this view, the problem is mainly competition, innovation, consumer choice or bargaining power between platforms and publishers. These dimensions are real and important, but they are not sufficient. A platform may be competitive and still produce systemic risks for democracy. A service may be innovative and still weaken pluralism. A market may offer consumer choice and still concentrate the infrastructures through which citizens access public knowledge. The second risk is a purely security logic. According to this view, digital governance is mainly about combating threats: disinformation, manipulation, foreign interference, harmful content, social destabilisation. These threats are also real. But if security becomes the dominant frame, freedom of expression may be weakened through over-removal, surveillance, political pressure and the expansion of public control over speech.

The Council of Europe's role is to resist both reductions. It reminds us that the digital public sphere must be governed as a space of rights, democracy and rule of law. That means recognising risks without abandoning freedom. It means addressing disinformation without creating ministries of truth. It means holding platforms accountable without turning them into delegated censors. It means preserving pluralism without imposing official narratives. It means using regulation to constrain infrastructural power, not to replace it with public arbitrariness.

Article 10 is central here. It protects the right to impart information and ideas, but also the right to receive them. This receiving dimension is crucial in the digital environment. In the analogue world, the ability to receive information was shaped by newspapers, broadcasters, public service media, editorial markets and distribution infrastructures. In the digital world, it is increasingly shaped by platforms, search engines, recommender systems, ranking mechanisms and AI-generated interfaces. Therefore, the protection of the right to receive information can no longer ignore the infrastructures that organise reception. If citizens access public knowledge through opaque systems of ranking, recommendation and synthesis, then the Article 10 framework must be interpreted in a way that remains attentive to these new conditions of reception.

This does not mean transforming Article 10 into a general right to algorithmic visibility. That would be too simplistic and legally problematic. No speaker has an unconditional right to be ranked first, recommended to everyone or amplified by a platform. But it does mean that, at systemic scale, the architecture of visibility becomes relevant to the protection of pluralism and the right to receive information. The question is not whether every individual speaker can demand exposure. The question is whether the overall structure through which information is organised remains compatible with a pluralistic democratic society.

Media freedom is equally important. The Council of Europe has long recognised the special role of journalism as a public watchdog. This role cannot be understood only as the absence of censorship. Journalism also requires conditions of visibility, source integrity, editorial independence, economic sustainability and public trust. If digital intermediaries and AI systems progressively mediate the relationship between journalistic sources and citizens, the public watchdog function may be affected even without direct interference. A newspaper can remain legally free while becoming structurally invisible. A source can remain online while being absorbed into AI-generated synthesis. A journalistic ecosystem can remain formally plural while losing the economic and attentional conditions that allow it to perform its democratic function.

This is why the Council of Europe's standards on media freedom and pluralism are particularly relevant to the AI-mediated public sphere. They remind us that democracy depends not only on individual speech, but on institutional conditions that allow public-interest information to be produced and received. A democratic society needs journalists who can investigate power, media institutions that can survive, citizens who can compare sources, and public debate that is not entirely shaped by opaque private infrastructures. This is not nostalgia for traditional media. It is a recognition that democratic communication has institutional preconditions.

The Council of Europe is also important because it insists on proportionality. Proportionality is the safeguard that prevents legitimate regulation from becoming overreach. When public authorities respond to disinformation or systemic risks, they must act within clear legal boundaries, pursue legitimate aims and adopt measures that are necessary and proportionate in a democratic society. This matters because platform regulation can easily slide into indirect censorship. If states pressure platforms to remove ambiguous content, if they define disinformation too broadly, if they impose vague duties to suppress harmful narratives, the result may be chilling effects and the narrowing of democratic debate. A Council of Europe approach requires that measures against digital harms remain compatible with open public discourse.

At the same time, proportionality does not mean passivity. It does not require democratic societies to accept opaque infrastructural power simply because speech is involved. Proportionality allows intervention, but it disciplines intervention. It asks whether the measure is suitable, necessary and balanced. It asks whether less restrictive alternatives exist. It asks whether procedural safeguards are present. It asks whether users, journalists and media actors have remedies. It asks whether transparency and accountability can address risks without suppressing lawful expression. This is exactly the approach needed for platform and AI governance.

The rule of law dimension is also essential. Digital infrastructures often operate through automated systems whose decisions are difficult to understand, challenge or review. Users

may not know why content is recommended, demoted, labelled or removed. Publishers may not know why traffic declines or why their sources are not surfaced. Researchers may not have access to data. Regulators may depend on information provided by the platforms themselves. This creates a rule-of-law problem: power is exercised, but its reasons are opaque; effects are produced, but remedies are weak; decisions shape rights, but accountability is limited. A Council of Europe approach must therefore insist on transparency, reason-giving, contestability and independent oversight.

This is where the Council of Europe Framework Convention on Artificial Intelligence becomes relevant. Its importance lies not only in the fact that it is the first international legally binding instrument on AI. Its deeper significance is that it frames AI governance around human rights, democracy and the rule of law. This is exactly the vocabulary needed for AI-mediated public discourse. AI systems that structure access to information, produce synthetic answers, rank sources or influence public debate cannot be assessed only through technical safety or market efficiency. They must be assessed in light of their impact on democratic processes, media pluralism, freedom of expression and the ability of citizens to form opinions freely.

The Framework Convention also reinforces the idea that AI governance must address the lifecycle of systems, not merely isolated outputs. This is crucial. If we focus only on individual AI-generated answers, we miss the systemic nature of the problem. The relevant questions concern design, deployment, risk assessment, monitoring, transparency, oversight and remedies. How is the system trained? How does it select sources? How does it update? How does it handle public-interest information? How does it mitigate hallucinations? How does it preserve attribution? How does it allow contestation? How does it affect media pluralism over time? These are lifecycle questions, and they are directly connected to the Council of Europe's human rights approach.

The Council of Europe perspective also helps mediate between the European and American models discussed earlier. The EU model, through the DSA, is regulatory and ex ante. The American model, through cases such as those in Los Angeles, is more litigation-driven and ex post. The Council of Europe does not simply choose one model over the other. It provides the constitutional criteria by which both can be evaluated. A regulatory model must be effective but proportionate. A litigation model must provide accountability without chilling speech. A platform governance system must address systemic harms without undermining pluralism. An AI governance model must promote innovation without sacrificing human rights and democratic legitimacy.

This is particularly important because digital power is transnational while democratic accountability remains fragmented. Platforms and AI systems operate across borders. Their design choices are often made globally. Their effects, however, are felt within national democracies, local media ecosystems, electoral processes, linguistic communities and vulnerable groups. This creates a governance gap. No single national authority can fully control the infrastructure. No single court can fully address the systemic effects. No single platform policy can substitute for public accountability. The Council of Europe can help fill this gap by articulating common standards grounded in human rights, democracy and rule of law.

The Council of Europe also has a specific role in preventing regulatory fragmentation from becoming rights fragmentation. Different jurisdictions may adopt different platform rules, disinformation policies, AI governance models or media protection measures. Diversity is

inevitable, but the baseline protection of freedom of expression, pluralism and rule of law must remain common. The risk is that under the language of disinformation, national security or platform accountability, states may adopt measures that are incompatible with open democratic debate. The Council of Europe framework provides a common constitutional discipline.

This is why the discussion should not be framed as a conflict between innovation and regulation. The real conflict is between unaccountable infrastructural power and democratic constitutionalism. Innovation can and should continue. AI-mediated search may improve access to knowledge. Recommender systems may help users navigate complexity. Platforms may enable expression and association at unprecedented scale. But these benefits do not eliminate the need for accountability. A system can be innovative and still create risks. A service can be useful and still affect pluralism. A platform can enable speech and still distort the conditions of visibility. Constitutional governance does not oppose innovation. It asks how innovation can be made compatible with democratic values.

In this sense, the Council of Europe's contribution is not to replace the DSA, nor to duplicate EU enforcement. It is to provide the constitutional horizon within which enforcement should operate. The DSA gives regulators tools. The Council of Europe reminds us why those tools matter and where their limits lie. They matter because freedom of expression, pluralism and democratic resilience require accountable infrastructures. Their limits lie in the need to avoid censorship, truth control and disproportionate interference.

The key message, therefore, is that digital constitutionalism in the Council of Europe context is not about regulating every digital interaction. It is about identifying forms of infrastructural power that affect the effective exercise of rights and bringing them within a framework of human rights, democracy and rule of law. It is about ensuring that private platforms do not privately organise the public sphere without accountability. It is also about ensuring that public authorities do not use platform regulation to control lawful expression. The Council of Europe stands precisely at this point of equilibrium.

This is the broader meaning of the AGCOM case and the Los Angeles litigation when seen from Strasbourg. They are not merely regulatory or judicial episodes. They are symptoms of a wider constitutional question. How do we protect freedom of expression when power over speech is exercised through visibility rather than prohibition? How do we preserve pluralism when public knowledge is mediated by AI-generated interfaces? How do we fight disinformation without creating mechanisms of censorship? How do we hold digital architectures accountable without undermining the openness of democratic debate?

The Council of Europe is the place where these questions can be framed in their proper constitutional terms. Not as technical questions only. Not as market questions only. Not as security questions only. But as questions about the future of democratic communication under conditions of algorithmic and AI-mediated power.

The conclusion, in my view, is that the future of freedom of expression and digital pluralism will not be decided only by rules on takedown, illegal content or platform liability for individual posts. It will increasingly be decided by the way democratic societies govern the infrastructures of visibility, amplification, attribution and synthesis. This is the central transformation. The most important power over speech is not always the power to prohibit it.

It is increasingly the power to organise the conditions under which speech becomes visible, credible, viral, marginal, decontextualised or absorbed into AI-generated outputs.

This does not mean that classical freedom of expression is obsolete. On the contrary, it remains the foundation. Article 10 of the European Convention on Human Rights continues to protect the public sphere against censorship, political pressure, prior restraint and disproportionate interference. But the digital environment forces us to recognise that freedom of expression can be weakened even when speech is not formally prohibited. A voice can be free but invisible. A source can exist but disappear behind a synthetic answer. Journalism can remain legally protected while losing visibility, attribution and economic sustainability. Citizens can have access to more information than ever before and yet encounter it through increasingly concentrated, opaque and personalised infrastructures.

This is why digital constitutionalism is necessary. It does not seek to regulate speech for the sake of regulation. It seeks to identify and limit power where power actually operates. In the contemporary public sphere, power operates through platforms, recommender systems, ranking mechanisms, advertising infrastructures, search engines and AI-generated interfaces. These systems do not merely host expression. They structure exposure. They organise attention. They distribute relevance. They mediate the relationship between sources and citizens. They shape the conditions under which public opinion is formed.

The constitutional task is therefore not to create new authorities of truth. That would be a serious mistake. Democratic societies should not respond to disinformation, platform power or AI-mediated communication by empowering public authorities to decide which opinions are legitimate or which narratives should prevail. Nor should they delegate this function to platforms. The choice cannot be between the privatisation of truth by platforms and the nationalisation of truth by public authorities. Both models are dangerous. Both risk narrowing democratic debate. Both can undermine lawful expression, dissent, satire, minority viewpoints and journalism.

The better approach is different: not the regulation of truth, but the accountability of infrastructures. This means focusing on the systems that organise visibility, amplification and access to information, rather than on the political correctness or substantive acceptability of individual opinions. It means asking whether the architecture of the service is transparent, contestable, proportionate and compatible with pluralism. It means ensuring that systemic actors assess and mitigate risks without turning platforms into delegated censors or regulators into editors.

Several principles follow from this approach.

The first is transparency. Platforms, search engines and AI systems that operate at systemic scale must provide meaningful transparency about the ways in which they organise public discourse. This does not mean disclosing every line of code or every commercially sensitive detail. It means making visible the criteria, effects and risks of ranking, recommendation, source selection and AI-generated synthesis in ways that users, researchers and regulators can meaningfully understand. Transparency must not be reduced to formalistic reports. It must allow democratic scrutiny.

The second principle is attribution. In an AI-mediated information environment, source attribution becomes a constitutional safeguard. Citizens must be able to know where

information comes from. They must be able to distinguish original journalism from AI-generated synthesis, official information from political messaging, professional verification from user-generated speculation, editorial responsibility from automated recombination. If AI systems absorb sources without making them visible, pluralism is weakened. Attribution is not only a matter of fairness to publishers. It is a condition of democratic judgment.

The third principle is contestability. Decisions and patterns affecting visibility must be challengeable. This is difficult because architectural power often operates through patterns rather than single decisions. A source may not be removed, but may be systematically deprioritised. A publisher may not be banned, but may lose visibility. A user may not be censored, but may be exposed to highly personalised and manipulative flows of information. A journalistic source may not be excluded, but may be misrepresented in a synthetic answer. Contestability must therefore go beyond appeals against content removal. It must include mechanisms to challenge misrepresentation, systematic invisibility, lack of attribution and harmful architectural effects.

The fourth principle is independent scrutiny. The governance of platforms and AI systems cannot depend only on self-assessment. Systemic-risk assessment is necessary, but it must be open to verification. Researchers need access to relevant data. Regulators need technical capacity. Auditors must be independent and competent. Civil society and media actors must be able to identify and document harms. Without independent scrutiny, transparency becomes performative and risk assessment becomes a compliance ritual.

The fifth principle is proportionality. This is particularly important in the Council of Europe context. Any intervention affecting freedom of expression must be lawful, necessary and proportionate in a democratic society. Measures against disinformation, manipulation or systemic risks must not become tools of censorship. The fact that a digital architecture creates risks does not justify unlimited intervention. The response must be targeted, evidence-based and rights-respecting. Proportionality is what distinguishes constitutional regulation from control.

The sixth principle is media sustainability. Digital pluralism cannot survive if professional journalism and public-interest media become invisible or economically unsustainable. This does not mean protecting legacy media from innovation as such. It means recognising that journalism performs a democratic function and that the infrastructures of access to information should not extract journalistic value while eroding source visibility and the relationship between publishers and citizens. Media pluralism must be protected not only through ownership rules, but also through source visibility, fair attribution and safeguards against infrastructural dependency.

The seventh principle is user autonomy. Citizens must not be treated as passive recipients of algorithmic organisation. They should have meaningful control over recommender systems, personalised information flows and AI-mediated interfaces. This includes the ability to understand why certain content is shown, to modify recommendation parameters where appropriate, to access original sources, to compare perspectives and to avoid being locked into opaque informational environments. User empowerment is not a substitute for systemic regulation, but it is an essential part of democratic communication.

The eighth principle is institutional cooperation. No single actor can govern these infrastructures alone. Platforms are global. Media ecosystems are national and local. Fundamental rights standards are regional and international. Regulatory competences are distributed. This requires cooperation between national authorities, European institutions, courts, independent regulators, researchers, civil society, journalists and international organisations. The AGCOM initiative shows how a national authority can identify a risk and activate a European mechanism. The Council of Europe can provide the human-rights framework within which these interventions remain principled and proportionate.

These principles also help clarify the relationship between the DSA, AGCOM, the Los Angeles cases and the Council of Europe framework. The DSA provides the European regulatory machinery for systemic-risk assessment and mitigation. AGCOM shows how a national independent authority can identify a concrete risk to media pluralism and source visibility in an AI-mediated search environment. The Los Angeles litigation shows, through a different legal route, that platform design itself can become a source of legal accountability. The Council of Europe provides the constitutional grammar: freedom of expression, media freedom, pluralism, democracy and rule of law.

Taken together, these developments show that the law is gradually moving from the governance of content to the governance of architectures. This is the decisive point. Content still matters. Illegal content still matters. Harmful speech still matters. But the deeper layer is architectural. Which content is amplified? Which sources are visible? Which voices are marginalised? Which narratives are monetised? Which outputs are synthesised? Which users are targeted? Which actors can contest the system? Which risks are assessed before deployment and monitored afterwards?

This architectural focus does not weaken freedom of expression. It can strengthen it, provided it is properly designed. Freedom of expression is not only the absence of removal. It is the possibility of participating in a public sphere where citizens can encounter diverse sources, evaluate information, contest power and form opinions freely. If the architecture of that public sphere is opaque, concentrated and unaccountable, then formal freedom may coexist with structural dependency. Constitutional law cannot ignore that.

At the same time, architectural accountability must never become content paternalism. This is the line that must be preserved. Regulators should not decide what citizens must believe. Courts should not become truth managers. Platforms should not become private ministries of acceptable opinion. The object of governance should be not lawful disagreement, but infrastructural arbitrariness. The aim should be not to suppress contested speech, but to ensure that the systems organising public discourse are transparent, accountable and compatible with pluralism.

This is the broader lesson of digital constitutionalism. The constitutional question is not whether we can apply old rights to new technologies in a mechanical way. The question is whether constitutionalism can continue to perform its original function — limiting power — when power is exercised through digital infrastructures rather than only through state commands. If the public sphere is organised by platforms, search engines and AI systems, then the constitutional protection of freedom of expression and pluralism must address those infrastructures.

The final message could therefore be this: the future of democratic communication will not be secured by choosing between total deregulation and state control. It will be secured by building a framework of accountable digital infrastructures. A framework in which platforms and AI systems can innovate, but not invisibly reorganise the public sphere. A framework in which disinformation can be addressed, but not through censorship. A framework in which journalism remains visible, but not artificially protected from every technological change. A framework in which users benefit from AI-mediated access to knowledge, but without losing the ability to identify, compare and contest sources.

In this sense, the challenge before us is not to stop technological transformation. It is to constitutionalise it. To ensure that the architectures of visibility, amplification and synthesis remain compatible with freedom of expression, pluralism, democratic self-government and the rule of law. That is the real meaning of digital constitutionalism in this field. And that is why this debate belongs, in a particularly strong way, within the Council of Europe: because it is ultimately about the conditions under which citizens can continue to speak, receive information, trust sources, compare perspectives and participate freely in democratic life.

Ms Murielle POPA-FABRE

Murielle Popa Fabre – *NLP/ML Expert for AI Governance & Policies, Senior Tech & Policy Advisor, AI & Society Institute, École Normale Supérieure (ENS-PSL).*

2024-2026 EMERGING BEHAVIORAL PREDICTION RISKS –

HUMAN DIGNITY AND INDIVIDUAL AUTONOMY IN THE AGE OF CONTINUOUS AND MULTIMODAL SENSING

The latest wave of AI has introduced seamless interactional and conversational features that are transforming the way end-users engage with AI systems and technology more broadly. This paper traces the shift witnessed since 2022, from Large Language Models (LLMs) performing language simulation to LLMs enabling behavioural prediction, and examines how contemporary AI systems increasingly generate and infer information about individuals.

#A - The turn to Behavioral Prediction

The main emerging pattern across 2024-2026 AI systems and products developments is the convergence of results demonstrating increased AI performance in behavioural prediction across diverse domains and contexts. An emerging research field is consolidating around the term *behavioural cloning*, focussing on LLMs trained to replicate human behavior as a tool for large-scale social simulation (see Jagadish et al. 2026; Sucholutsky et al. 2025; Dillion et al. 2023, Argyle et al., 2022). Beyond language simulation, LLMs are increasingly deployed within composite agentic architectures that efficiently simulate human behavioural responses through multimodal data integration and reinforcement learning-driven adaptability. The following sections examine how LLM-based systems are achieving progressively greater precision in behavioral simulation, with implications that extend well beyond technical performance.

#1 Predicting behavioural data: elections

One of the first empirical tests of the behavioural prediction capabilities of LLMs appeared in late 2023: a simple study simulating election outcomes with basic prompting of an off-the-shelf large language model (ChatGPT-4.0) offers a telling example (Jiang et al., 2024). This pre-election study closely matched US presidential election outcomes across states by inputting the results of the 2016 and 2020 election cycles, and constructing voter personas using World Values Survey data and demographic profiles, to simulate voters' behavior. This approach exemplifies a technique known as *persona-induction* that elicits human-like behavior by conditioning models on participant-specific information (Binz et al. 2026). The study achieved strong predictive accuracy with minimal engineering, illustrating how the vast amount of data compressed in LLMs can be leveraged to approximate and predict collective behaviours such as voting outcomes.

#2 Agentic Simulation reproducing complex multidimensional surveys

A series of studies from a team at Stanford University between late 2023 and 2024

demonstrated the extent to which agent-based individual behavior modeling can simulate human responses with high accuracy using minimal data input (Park et al., 2023, 2024).

Agentic digital twins The experimental setup (Park et al., 2024) involved 1,000 human participants whose two-hour interviews were used to construct an AI agent's memory. The resulting agentic digital twin system could accurately simulate responses to more than 200 questions drawn from social science and psychological tests, including the Big Five personality inventory, economic games, and the General Social Survey.

While a two-hour audio interview (approximately 6,500 words) enabled 70% simulation accuracy, against a human baseline of ~80% (adjusted for intra-individual variability), providing the AI agent fed with demographic information alone still achieved around 60% precision.

The implication of these studies is that LLM-driven agents can simulate personal traits and choices with minimal personal data inputs, demonstrating **the potential of leveraging the information compressed within large language models** to enable highly proactive, hybrid recommender systems as we will see in the following.

#3 LLM-based behavioral prediction with multimodal data in e-commerce platforms

Previous research had already documented that LLMs reasonably predict human responses on sensory judgments across six modalities without being trained on multimodal inputs, and irrespective of whether direct visual input or purely textual descriptors are provided (Marjeh et al., 2024).

Building on this baseline capability, further real-world experimentation showed that fusing multimodal data about products and behavioral patterns markedly boosts predictive performance on an e-commerce platform. A 2024 study on Walmart. (Ma et al., 2024) explored how multimodal large language models enhance behavioral prediction by aggregating heterogeneous data sources. Notably, fusing viewed items, product descriptions, product images, and user actions history, and viewing times into a large language model architecture through a targeted training strategy boosted next-recommendation prediction accuracy by 43% in the sports category.

#4 Proactive interaction and Agents simulation embedded in recommender systems

Beyond the fusion of diverse data modalities to capture the multifaceted nature of users' behaviours on e-commerce platforms for a more fine-grained behavioural prediction, later experiments show that even the simple integration of an LLM-based simulation into traditional recommender systems enables adaptable needs anticipation. Simulating users' intentions and behaviour in LLM-recommender systems significantly boosts next-purchase prediction accuracy (Huang et al, 2024).

By creating hybrid architectures where LLM components model customer behaviour, traditional recommender systems can now proactively adapt and aggregate brand affinity, tastes, habits, idiosyncrasies, and dietary styles alongside multimodal signals.

This shift from reactive to proactive prediction introduces new behavioral and privacy risks, calling for dedicated and new assessment frameworks centered on behavioral inference and simulation.

AI Agent as a User simulator This leads to a paradigm shift: from classification-based recommender systems to proactive systems that anticipate user needs in real time. This represents a move from systems reacting to explicit input to systems that continuously simulate users' preferences, thus, broadening the application scope of Recommender Systems (RecSys) by anticipating needs, and adapting to evolving tastes or prices sensitivities of individual customers, while simultaneously raising behavioural simulation risks.

#5 Predicting next-turns in conversational AI offers seamless interaction

Behavioural simulation is also being leveraged in conversational agents, where next-turn prediction ensures seamless interaction by embedding agentic simulation in conversational AI (Wu et al. 2025). The rationale is to shift from passive responders to proactive AI virtual assistants. In this configuration what can be described as ***Users' intention decoding*** enables fluid interaction without the user explicitly formulating their needs (or intentions), delivering a seamless experience through anticipation rather than explicit request. While this is an efficient mechanism for conversational fluidity, it deserves careful monitoring of its potential impact on users' freedom of expression, opinion forming or access to information.

#6 Fine-tuning an LLM model with psychological data makes it a better behavioural predictor Further research into modelling human behaviour through LLMs has shown that adapting models such as LLaMA with data from psychological experiments (60,000 people, 10 million choices across diverse tasks) yields significantly enhanced accuracy in predicting individual decisions across a range of psychological paradigms (Centaur model, Binz et al., 2025).

Initial evidence suggests that integrating psychological theory into AI can offer some scalable path towards more trustworthy AI systems aligned with human moral cognition. Cognitively aligned models can, for example, leverage a method that introduces "cognitive" bottlenecks (Dillon et al., 2026) by focusing decisions on a small set of key features. This interpretability becomes crucial as large language models become more deeply integrated into everyday decision-making.

Improving the prediction of human decision through psychological fine-tuning increases the efficiency of behavioural simulation, but also carries heightened behavioral prediction risks, with direct implications for human rights, autonomy, and privacy.

#7 LLM simulation without recommender systems and fine-tuning Whereas the Walmart results required an advanced LLM architecture and training strategy, subsequent research (Maier et al. 2025) confirms that precise behavioral simulation can also be obtained using multimodal off-the-shelf LLMs, by simply inputting a customer demographic profile and a product image, and querying the model's impressions. Predicting purchase intent with 90% accuracy by prompting an off-the-shelf LLM to impersonate a customer not only automates market research, but demonstrates how LLMs can function as highly efficient user-response

simulators, without any specific fine-tuning or training, while outperforming classical machine learning methods for predicting customer interest.

As LLMs are increasingly used as surrogates for human participants, the scientific research community is actively working to understand which models best capture human behaviour and why. Latest research constructed a second corpus, Psych-201, to measure at scale the behavioral similarity between models and humans, termed *behavioral alignment* (Binz et al., 2026), finding that base models exhibit stronger behavioural alignment than fine-tuned ones, even when fine-tuned models are stress tested by eliciting human-like behaviour through *persona-induction* prompting. This seminal work suggests that the general knowledge compressed in pre-training may be a more powerful driver of behavioural fidelity than task-specific adaptation, a finding with significant implications for how behavioural simulation systems are designed and governed.

Such results confirm that multimodality and recent advances in LLM-based systems are bringing behavioural prediction to a highly effective level, with e-commerce as the most documented domain to date, but with implications that extend across sectors.

#B Behavioral prediction is yielding a shift towards Intention and seamless interaction economy

In sum, recent advances in LLM-powered applications are expanding multimodal data aggregation and predictive capabilities, thereby capturing more aspects of the multifaceted nature of human behavior with increasing precision thanks to advanced simulation techniques.

This progression is gradually driving a shift from the attention economy towards an *intention economy* (see Chaudhary & Penn, 2024 for a definition), with profound implications not only for private life, but also for human autonomy and freedoms (Barbera & Popa-Fabre, 2025, pp. 11-13).

Intention and seamless interaction economy is based on the following set of enabling technological factors and steps:

- **1 -Predictability:** as the capacity to generalize from sparse signals grows, the simulation and prediction of behavioural data, such as voting patterns, purchasing decisions, and survey responses, becomes increasingly accurate.
- **2 -Multimodal integration:** combining heterogeneous data sources into unified behavioural representations to better capture the multifaceted aspects of human behaviour. As smartphones and wearables continuously consolidate rich streams of personal data, their fusion into high-capacity predictive architectures enables sophisticated behavioural inference at scale, thus intensifying both behavioural prediction capabilities and privacy stakes.

- **3-Proactive human simulation:** seamless interaction grounded in predicted behaviour operationalised through AI Agents inside RecSys, Conversational AI, Agentic Digital twins.
- **4 -Adaptability:** enhanced adaptability brought by state-of-the-art reinforcement learning techniques and agentic simulation enables systems to adjust dynamically to users, to run real-time decoding of behavioural dynamics or patterns in a real-world context (e.g., robotics).

A shift from attention economy to intention economy This leads to two major developments: (1) the rise of personal intelligence products and (2) a gradual shift from attention economy to intention economy.

Rather than optimizing for attention capture, systems now gather subtle signals and behavioural cues form a cloud of proxy data (see Figure 1, on the right) used to predict users' intentions. This progression leverages rich digital environments to deliver individualized interactions, experiences and offers, calculated on proxy signals of intent, this is catalyzing a shift towards more adaptive and less constraining personalization methods.

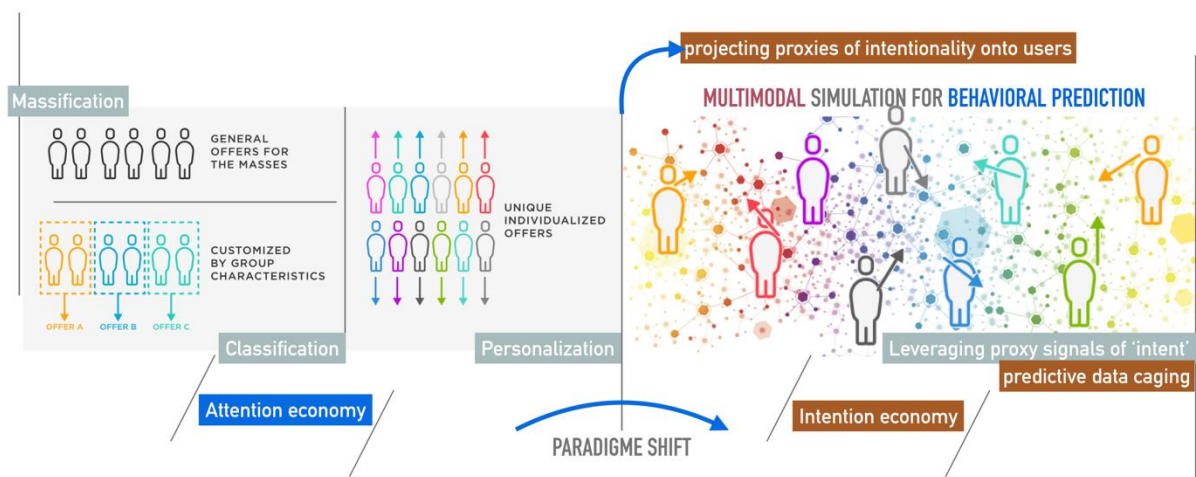


Figure 1 Illustration of the paradigm shift from attention economy to intention economy enabled by behavioral prediction projecting proxies of intentionality onto users through multimodal simulation. Evolution path: (1) **Massification**: one-size-fits-all offers, (2) **Classification**: customisation by identified segment within a limited set of offers.; (3) **Hyper-personalization**: sophisticated recommenders create individualized offers built on a richer digital context. (4) **Intent projection**: leveraging proxy signals to model and anticipate user intent.

Intention economy and Personal intelligence products The recent development and commercialisation of Personal intelligence applications offers a concrete illustration of this shift : smartphone data can be used to assist the user directly, for example selecting a gift for a partner by selecting a gift for a partner through analysis a user's photo history and through the behavioural patterns that can be extrapolated from text messages and notes.

The digitally connected ecosystem of multimodal data surrounding a smartphone can thus be leveraged for fine-grained behavioral prediction, spanning social rhythms captured by light sensors, voice pitch tracking, and GPS or Wi-Fi signals, through to emotional states inferred from heart rate, touchscreen reaction times, eye movements and smartwatch analytics.

The fundamental feature of personal intelligence products is thus to operationalise personal data gathered through user consent on the smartphone. **This raises questions about behavioural nudging, whereby users are subtly guided toward an algorithmically predicted action.**

Predictive data caging: a fluid 360° profile of individual users based on past digital traces, constantly nudging individual decision

In such a paradigm, every piece of user data becomes predictive, feeding an adaptive response framework calibrated to individual needs. Leveraging multimodal proxy signals of intent across individuals' digital traces can produce what we term *predictive data caging*: a phenomenon whereby the user is digitally encapsulated within a cloud of their own past data."

This intention-prediction approach differs fundamentally from traditional profiling-based harms. By integrating multiple multimodal data sources into a **comprehensive, highly adaptable 360° profile of individual users**, advanced system architectures create a ***predictive and adaptable data cage*** around the user – one in which individuals are constantly nudged towards choices and offerings within comfort-zone calculated and calibrated on proxies of intent, so as to make the proposed choices appear effortlessly optimal.

The use in AI systems of fluid, adaptive proxies of intent to guide choices without overtly constraining users, and without the crude caricatures of users in legacy targeting, produces what we define as *predictive data caging*, which may over the medium and long term gradually cultivate in users what we term *freedom laziness*. This has direct Human rights, autonomy, and privacy implications:

- Multimodal aggregation and adaptive prediction can form an individualized net of cues that nudges decisions, thereby affecting cognitive autonomy.
- The richness and portability of behavioral traces across contexts heighten privacy risks and challenge conventional notions of private life.

Beyond business model considerations, one should consider the potential pervasiveness of these *emerging behavioural prediction risks*, as AI systems are increasingly becoming the primary conversational interfaces for human-machine interaction. A comprehensive analysis of these developments is essential to understand how these systems can impact human autonomy by shaping human actions. Several open questions emerge:

- If individuals are constantly nudged towards good-enough choices by proxies of their own intentions, how can cognitive autonomy, or autonomy more broadly, be preserved?
- In this context, what is the impact of individuals' freedom of expression, access to information, opinion forming, and privacy more broadly?

Analysing these recent developments and open questions is essential for designing future-proof governance measures capable of addressing emerging risks for human rights, democracy and the rule of law.

#C Wearables are the next behavioral prediction frontier integrating brain and physiological data

Smartphones are not the only devices now functioning as multimodal behavioural sensors, but multimodal integration is increasingly extending to health and physiological data (Li et al, 2025; Hearts, Li et al. 2025), and notably to brain data (Yet et al, 2025) with the emergence of thought-to-text applications in **wearables** (Mishra et al., 2024 ; Baradari et al., 2025 or Kosmyna et al., 2025). The growing diffusion wearables enables the aggregation of brainwaves activity, heart rate, breathing, and muscle activity.

The time series of these physiological signals can train large language models capable of querying for example sleep states or cognitive conditions in natural language (Xu et al, 2026). Models trained on physiological data have been shown to outperform general-purpose ones in predicting sleep, attention and emotional states (Chen et al. 2025).

#D At scale social prediction through platforms for large scale agentic simulation

Recent significant investments in a company developing platforms for large-scale agentic simulation of consumers behaviours, signals that the field is moving towards solutions enabling AI simulation of society as a whole, making it possible, for example, to test how populations respond to policies (Li et al., 2026) or price changes (Simile Company).

Starting from simulations of individuals' decision-making to model how real people make decisions, allows such a platform to then compose them into bottom-up societal simulations in which changing a single assumption or constraining one person's behaviour can cascade through the entire simulation. In this way, building high-fidelity digital twins grounded in real interviews and data allows companies to have new way to test ideas, explore tradeoffs, and better understand how people actually make decisions.

This not only amplifies the concerns raised above regarding individuals' cognitive autonomy, freedom of expression, and the capacity to form independent opinions and access to information when continuously nudged or influenced by predictive systems in everyday life, but it also marks a qualitative threshold: the first AI simulation of society at scale could not only radically transform how decision-makers anticipate the downstream effects of their choices, with applications spanning epidemic modelling, disaster response planning, and economic policies, but also, and most importantly, impact the rule of law and the integrity of democratic life.

#E Conclusion

In conclusion, the acceleration of predictive and simulative capabilities through multimodality and agentic architectures is driving proactive behavioral into the fabric of everyday AI products and services. Since 2023, growing volumes of individuals' data, including physiological and brain data, can be leveraged to create seamless human-machine interactions grounded in predictive mechanisms that efficiently simulate human behaviour. **This trajectory is no longer merely a technical development: it is a structural transformation of the conditions under which individuals exercise their autonomy.**

Emerging Behavioural prediction risks deserve dedicated governance and monitoring action to tackle the risk of *predictive data caging* at individual level, and the at-scale risk of societal

prediction. Together, these challenges raise fundamental questions about human cognitive autonomy and freedoms more broadly.

Bibliography

Argyle, L.P., Busby, E., Fulda, N., Gubler, J.R., Rytting, C., & Wingate, D. (2022). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31, 337 - 351.

Baradari, D., Kosmyrna, N., Petrov, O., Kaplun, R., & Maes, P. (2025). NeuroChat: A Neuroadaptive AI Chatbot for Customizing Learning Experiences. *Proceedings of the 7th ACM Conference on Conversational User Interfaces*.

Barberá, I. & Popa-Fabre (June 2025), M, Privacy and Data Protection Risks in Large Language Models (LLMs), Expert report, Council of Europe.

Binz, M., Akata, E., Bethge, M., Brandle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M.K., 'EltetHo, N., Griffiths, T.L., Haridi, S., Jagadish, A.K., Li, J., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M.G., Modirshanechi, A., Nath, S.S., Peterson, J.C., Rmuš, M., Russek, E.M., Saanum, T., Scharfenberg, N., Schubert, J.A., Buschoff, L.M., Singhi, N., Sui, X., Thalmann, M., Theis, F.J., Truong, V., Udandara, V., Voudouris, K., Wilson, R.C., Witte, K., Wu, S., Wulff, D.U., Xiong, H., & Schulz, E. (2024). Centaur: a foundation model of human cognition. *ArXiv*, *abs/2410.20268*. Nature 644, 1002–1009 (2025). <https://doi.org/10.1038/s41586-025-09215-4>

Binz, M., Akata, E., Almaatouq, A., Alsobay, M., Ariasov, O., Brandle, F., Broska, D., Burton, J.W., Busch, N., Callaway, F., Cheung, V., Christian, B., Coda-Forno, J., Demircan, C., Dentella, V., Eckstein, M.K., 'EltetHo, N., Franke, M., Griffiths, T.L., Gunther, F.L., Haridi, S., Hellmann, S., Herytash, S., Hof, L., Holton, E., Hoxha, I., Hussain, Z., Jagadish, A.K., Kara, E., Kriegmair, V., Leivada, E., Li, J., Ludwig, T., Maier, M., Mattar, M.G., Mathony, M., Modirshanechi, A., N.A., R.U., Nadverniuk, M., Nasioulas, A., Nath, S.S., Niemeyer, H., Nussenbaum, K., Olschewski, S., Pachur, T., Palminteri, S., Petrenco, A., Phaneuf-Hadd, C.V., Pirrone, A., Rausch, M., Raveling, L., Reddy, S., Rmuš, M., Russek, E.M., Saanum, T., Sandbrink, K.J., Schiekiera, L., Schubert, J.A., Buschoff, L.M., Singhi, N., Somerville, L.H., Spektor, M.S., Sui, X., Summerfield, C., Thalmann, M., Thoma, A.I., Tikhomirova, T., Truong, V., Tsvilodub, P., Voudouris, K., Wilson, R.C., Witte, K., Wu, S., Wulff, D.U., Xiong, H., Xu, S., Ying, L., Zhang, X., Zhu, J., & Schulz, E. (2026). Post-training makes large language models less human-like
<https://doi.org/10.48550/arXiv.2605.07632>

Chaudhary, Y., & Penn, J. (2024). Beware the Intention Economy: Collection and Commodification of Intent via Large Language Models. Harvard Data Science Review, (Special Issue 5).
<https://doi.org/10.1162/99608f92.21e6bbaa>

Chen, H., Zeng, W., Chen, C., Cai, L., Wang, F., Shi, Y., Wang, L., Zhang, W., Li, Y., Yan, H., Siok, W.T., & Wang, N. (2025). EEG Emotion Copilot: Optimizing lightweight LLMs for emotional EEG interpretation with assisted medical record generation. Volume 192, 107848, ISSN 0893-6080, <https://doi.org/10.1016/j.neunet.2025.107848>.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants?. Trends in cognitive sciences, 27(7), 597–600. <https://doi.org/10.1016/j.tics.2023.04.008>

Dillion, D., Mondal, D., Zhao, W., Tandon, N., & Gray, K. (under review). Making AI More Trustworthy and Morally Aligned by Integrating Human Cognition. <https://osf.io/preprints/psyarxiv/ek46d>

Huang, C., Yu, T., Xie, K., Zhang, S., Yao, L. and McAuley, J., 2024. Foundation models for recommender systems: A survey and new perspectives. arXiv preprint arXiv:2402.11143.

- Jagadish, A.K., Rmuš, M., Witte, K., Mathony, M., Binz, M., & Schulz, E. (2026). Can we automatize scientific discovery in the cognitive sciences? <https://doi.org/10.48550/arXiv.2603.20988>
- Jiang, S., Wei, L., & Zhang, C. (2024). Donald Trumps in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models.
- Kosmyna, N., & Hauptmann, E. (2026). NeuroSkill(tm): Proactive Real-Time Agentic System Capable of Modeling Human State of Mind.
- Li, S., Xiao, S., Joshi, M., Metwally, A.A., McDuff, D., Wang, W., & Yang, Y. (2026). HEARTS: Benchmarking LLM Reasoning on Health Time Series.
- Li, Y., Monteiro, K., Shirado, H., & Das, S. (2026). WhatIf: Interactive Exploration of LLM-Powered Social Simulations for Policy Reasoning. <https://doi.org/10.48550/arXiv.2604.17615>
- Ma, L., Li, X., Fan, Z., Xu, J., Cho, J.H., Kanumala, P., Nag, K., Kumar, S., & Achan, K. (2024). Triple Modality Fusion: Aligning Visual, Textual, and Graph Data with Large Language Models for Multi-Behavior Recommendations. *ArXiv, abs/2410.12228*.
- Maier, B.F., Aslak, U., Fiaschi, L., Rismal, N., Fletcher, K., Luhmann, C.C., Dow, R., Pappas, K.M., & Wiecki, T.V. (2025). LLMs Reproduce Human Purchase Intent via Semantic Similarity Elicitation of Likert Ratings. *ArXiv, abs/2510.08338*.
- Marjeh, R., Sucholutsky, I., van Rijn, P. et al. Large language models predict human sensory judgments across six modalities. *Sci Rep* 14, 21445 (2024). <https://doi.org/10.1038/s41598-024-72071-1>
- Mishra, A., Shukla, S., Torres, J., Gwizdka, J., & Roychowdhury, S. (2025). Thought2Text: Text Generation from EEG Signal using Large Language Models (LLMs). North American Chapter of the Association for Computational Linguistics. Thought2Text: Text Generation from EEG Signal using Large Language Models (LLMs). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3747–3759, Albuquerque, New Mexico. Association for Computational Linguistics.
- Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 2, 1–22. <https://doi.org/10.1145/3586183.3606763>
- Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C.J., Morris, M.R., Willer, R., Liang, P., & Bernstein, M.S. (2024). Generative Agent Simulations of 1,000 People. *ArXiv, abs/2411.10109*.
- Park, J.S., Zou, C.Q., Kamphorst, J., Egan, N., Shaw, A., Hill, B.M., Cai, C.J., Morris, M.R., Liang, P., Willer, R., & Bernstein, M.S. (2024). LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals. <https://doi.org/10.48550/arXiv.2411.10109>
- Sucholutsky, I., Collins, K.M., Jacoby, N. et al. Using LLMs to advance the cognitive science of collectives. *Nat Comput Sci* 5, 704–707 (2025). <https://doi.org/10.1038/s43588-025-00848-z>
- Tang, Zh, Jiaxin Qi, Yuhua Zheng, and Jianqiang Huang. 2025. A Comprehensive Benchmark for Electrocardiogram Time-Series. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. Association for Computing Machinery, New York, NY, USA, 6490–6499. <https://doi.org/10.1145/3746027.3754729>
- Wu, S., Galley, M., Peng, B., Cheng, H., Li, G., Dou, Y., Cai, W., Zou, J., Leskovec, J., & Gao, J. (2025). CollabLLM: From Passive Responders to Active Collaborators. *ArXiv, abs/2502.00640*.
- Xu, Z., Shuai, Z., Mozaffari, E., Aysola, R.S., Kumar, R., & Yang, Y. (2026). SleepLM: Natural-Language Intelligence for Human Sleep. *ArXiv, abs/2602.23605*.

Mr David REICHEL

Notes for FRA intervention:

The use of algorithms for online content moderation is risky, because algorithms can easily be biased or tricked, as shown in FRA's report on Bias in algorithms: fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

FRA reports on evidence of much hate speech online. For example, the FRA survey on antisemitism, the largest survey among Jews in Europe, shows the negative effect of encountering hate speech online. 90% of respondents encountered antisemitic content in the past 12 months online. A quarter of those encountering antisemitism online avoid identify as Jews and are worried about physical safety. Others limit their engagement online and reduce the use of certain apps. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2024-experiences-perceptions-antisemitism-survey_en.pdf

FRA data collection from social media also showed high levels of hate speech against women and ethnic groups. The analysis also shows the challenges in detecting and assessing hate speech online at scale: fra.europa.eu/sites/default/files/fra_uploads/fra-2023-online-content-moderation_en.pdf. The EU's Digital Services Act (DSA) provides for several avenues to address such challenges.

FRA also raised concerns about safety of children online. A recent publication shows that the vast majority of people in the EU are worried about child safety online and are in favour of social media restrictions: [Europeans deeply concerned about children's use of social media | European Union Agency for Fundamental Rights](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2024-europeans-deeply-concerned-about-childrens-use-of-social-media_en.pdf)

FRA published a report that deals with high-risk AI (as defined under the AI Act). It shows ways forward on how to use EU law to address fundamental rights risks when using AI: [Assessing High-risk Artificial Intelligence – Fundamental rights risks](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2024-assessing-high-risk-artificial-intelligence-fundamental-rights-risks_en.pdf)

FRA will publish several more reports in the course of 2026, including on [remote biometric identification in the context of law enforcement](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2024-remote-biometric-identification-in-the-context-of-law-enforcement_en.pdf).

Mr Gabriel REVEL

Madame la Secrétaire d'État,

Madame la Directrice,

Mesdames et Messieurs,

Je suis très heureux d'être parmi vous aujourd'hui pour cette conférence consacrée à l'« époque algorithmique », et de participer à l'ouverture, au nom de la Principauté de Monaco, de la réunion inaugurale du Comité directeur sur les technologies numériques nouvelles et émergentes.

Je tiens, en premier lieu, à saluer l'ensemble des intervenants et des participants réunis aujourd'hui à Strasbourg. La qualité et la diversité des expertises rassemblées témoignent de l'importance des enjeux que nous abordons collectivement.

Pour reprendre une intuition présente chez Italo Calvino, notre monde apparaît de plus en plus comme un tissu de signes qu'il nous faut apprendre à déchiffrer.

Nous nous trouvons à un moment charnière.

Les technologies numériques, et en particulier les systèmes fondés sur l'intelligence artificielle, ne constituent plus seulement des outils. Elles structurent désormais en profondeur nos sociétés, nos économies et nos institutions démocratiques. Elles redéfinissent les conditions mêmes dans lesquelles s'exercent les droits fondamentaux.

Cette transformation est porteuse de promesses considérables en matière d'innovation, d'efficacité des politiques publiques ou encore d'accès à l'information.

Mais elle soulève également des défis majeurs.

Car ces technologies, lorsqu'elles sont déployées sans garanties suffisantes, peuvent affecter des principes essentiels : la dignité humaine, l'autonomie individuelle, la protection de la vie privée, mais aussi le pluralisme du débat public et la confiance dans les institutions démocratiques.

C'est précisément pour répondre à ces défis que le Conseil de l'Europe joue un rôle unique.

Depuis plus de soixante-quinze ans, notre Organisation a su construire un cadre normatif fondé sur des principes clairs : la primauté du droit, la protection effective des droits de l'Homme et le respect des valeurs démocratiques.

Aujourd'hui, ces principes doivent continuer à s'appliquer pleinement dans l'environnement numérique.

À cet égard, les travaux conduits ces dernières années, qu'il s'agisse de la modernisation de la Convention pour la protection des données ou de l'élaboration de la Convention-cadre sur l'intelligence artificielle ou bien ceux en lien avec la lutte contre la cybercriminalité, illustrent la capacité du Conseil de l'Europe à anticiper et encadrer les transformations technologiques.

La création du CDNET s'inscrit dans cette dynamique. À cet égard, permettez-moi, en ma qualité de Président du Groupe de rapporteurs sur la Coopération juridique du Comité des ministres (GR-J), de rappeler le rôle joué dans l'élaboration du projet de mandat, ultérieurement adopté par les Délégués des Ministres, et qui a permis l'établissement de ce nouveau comité. Dans le même temps, le GR-J s'est attaché à assurer un suivi attentif et continu des développements relatifs à l'intelligence artificielle, en particulier des travaux menés au sein du CAI, notamment l'adoption du modèle HUDERIA.

Ce modèle présente une valeur ajoutée particulière en ce qu'il propose une méthodologie opérationnelle permettant d'évaluer, de manière structurée, l'impact des systèmes d'intelligence artificielle sur les droits de l'Homme, la démocratie et l'État de droit. En traduisant des principes juridiques en outils d'analyse directement mobilisables, il contribue à rapprocher les exigences normatives de leur mise en œuvre effective.

Ces travaux ont vocation à se prolonger au sein du CDNET, qui offrira désormais un cadre renouvelé, propice à une approche plus transversale et pleinement intégrée des enjeux numériques.

La Principauté de Monaco accueille ces développements avec une attention particulière.

En tant que future Présidence du Comité des Ministres, nous avons suivi de près l'élaboration de ces travaux et sommes déterminés à leur donner la visibilité qu'ils méritent.

Monaco s'est engagé dans une trajectoire résolue d'alignement sur les normes numériques du Conseil de l'Europe.

La ratification récente de la Convention 108+, ainsi que la modernisation de notre cadre national de protection des données, en témoignent concrètement.

Dans cette continuité, la Convention-cadre sur l'intelligence artificielle constitue une étape naturelle que nous entendons accompagner avec la plus grande attention.

Permettez-moi d'ajouter un point qui nous tient particulièrement à cœur.

En tant que petit État, Monaco a fait de la souveraineté numérique un axe structurant de sa stratégie nationale.

Cette situation nous rend peut-être particulièrement sensibles à un risque : celui de voir se développer des systèmes technologiques dont les logiques échappent, en pratique, à tout contrôle démocratique effectif.

Lorsque des décisions sont prises, influencées ou amplifiées par des systèmes algorithmiques, la question n'est pas uniquement technique.

Elle est fondamentalement juridique et politique.

Qui est responsable ?

Quels recours sont ouverts aux individus ?

Quelles garanties assurent que les droits fondamentaux ne sont pas seulement proclamés, mais effectivement protégés ?

Ce sont précisément ces questions que vos travaux permettront d'approfondir.

Les trois panels de cette conférence en donnent une illustration particulièrement claire.

La protection de la dignité et de l'autonomie face aux technologies de détection avancées ;
La préservation de l'État de droit dans un environnement algorithmique ;
Et, enfin, les conditions d'une démocratie résiliente à l'ère des contenus générés et amplifiés par des systèmes numériques.

Ces thématiques ne sont pas abstraites. Elles renvoient à des situations concrètes, déjà à l'œuvre dans nos sociétés.

Elles appellent des réponses qui soient à la fois juridiquement solides, techniquement informées et politiquement assumées.

Dans cette perspective, le rôle du CDNET sera déterminant.

Il lui reviendra non seulement d'accompagner la compréhension de ces phénomènes, mais aussi de contribuer à la formulation de cadres communs, capables de guider l'action des États membres.

Pour Monaco, il est essentiel que ces travaux reposent sur une approche équilibrée : une approche qui permette de saisir les opportunités offertes par les technologies numériques, tout en garantissant que leur développement demeure pleinement compatible avec les exigences de la Convention européenne des droits de l'Homme.

Mesdames et Messieurs,

À l'heure où les transformations technologiques s'accélèrent, notre responsabilité collective est claire.

Il ne s'agit pas de freiner l'innovation.

Il s'agit de l'inscrire dans un cadre de confiance.

Un cadre dans lequel les individus demeurent titulaires de leurs droits,
dans lequel les institutions conservent leur légitimité,
et dans lequel la technologie reste au service de l'humain et contrôlée par l'humain.

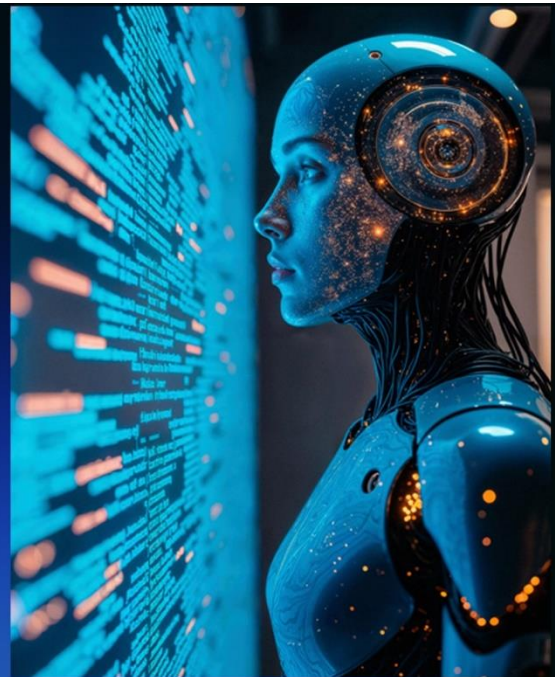
Un monde dans lequel la machine est totalement autonome risque de tenter de s'affranchir des règles qui régissent notre Continent depuis 1950.

C'est dans cet esprit que la Principauté de Monaco entend s'engager aux côtés du Conseil de l'Europe et de ses États membres.

Je vous remercie.

AI in Criminal Investigations: New Challenges for Criminal Procedure Law and International Cooperation

Steering Committee on New and Emerging Digital
Technologies (CDNET)
Strasbourg, 15 April 2026
@saltmarcos



Paradigm Shift in Criminal Procedure and international cooperation or the Intensification of Challenges Posed by Digital Evidence?

- Application of Procedural Powers Established by the Budapest Convention and SAP
 - "Artificial intelligence service providers" within the framework of public-private cooperation
 - AI in forensic tools and Special Investigation Techniques
- The Limits of the Principle of Territoriality in Addressing Jurisdictional
 - New safeguards or the need to redefine the interpretation of traditional safeguards?

Budapest Convention

Procedural Powers	
Article 16	Expedited Preservation of Stored Computer Data
Article 17	Expedited Preservation and Partial Disclosure of Traffic Data
Article 18	Production Order
Article 19	Search and seizure of data.
International Cooperation	
Article 29	Expedited Preservation of Stored Computer Data
Article 30	Expedited Disclosure of Preserved Data
Article 31	Mutual Assistance Regarding Access to Stored Data
Article 32	Transborder Access to Stored Data
Article 33	Mutual Assistance for the Real-Time Collection of Traffic Data
Article 34	Mutual Assistance Regarding the Interception of Content Data
Article 35	24/7 Network

Second Additional Protocol

Procedural Powers and International Cooperation	
Article 7	Disclosure of subscriber information
Article 8	To give effect to the orders of another Party for the expedited production of subscriber information and traffic data.
Article 9	Expedited Disclosure of Stored Computer Data in Emergency Situations
Article 10	Direct Cooperation between Authorities and Service Providers in Emergency Situations

Traditional procedural powers in the age of AI

Possible data available from an AI provider

- 1 IP addresses
- 2 Timestamps
- 3 Contents of prompts or queries
- 4 User agent (browser/device) data
- 5 User account/ subscriber data
- 6 Geolocation
- 7 Thoughts, ideas.
- 8 And so on, and so on, and so on...



OpenAI Law Enforcement
User data request policy

AI Providers Obligations to report Child Sexual exploitation material

18 U.S. Code § 2258A – Reporting Requirements of Providers.

Key Obligations

General Requirement:

Every "provider" must report to the CyberTipline of the National Center for Missing & Exploited Children (NCMEC) whenever they obtain actual knowledge of facts or circumstances from which there is an apparent violation involving child sexual exploitation material.

Timeframe:

The report must be made as soon as reasonably possible after obtaining such knowledge.

Report contents must include:

- Identification of the involved user(s)
- Description of the suspected material
- Technical details (IP address, timestamp, etc.)

Who Is Considered a "Provider" under 18 U.S.C. § 2258E(6)?

Definition:

A "provider" means an electronic communication service provider or remote computing service;

- **"electronic communication service"** means any service which provides to users thereof the ability to send or receive wire or electronic communications;
- **"remote computing service"** means the provision to the public of computer storage or processing services by means of an electronic communications system;

OpenAI law enforcement user data request policy:

Requests from
Government
Agencies in the U.S.

Requests from
Government
Agencies Outside
the U.S.

Emergency safety
concerns

OpenAI law enforcement user data request policy:

V. USER NOTICE

When OpenAI US or OpenAI Ireland receive a demand from a government to disclose information about a user, OpenAI provides notice to the user before disclosing the information, unless prohibited by law. OpenAI may withhold notice in exceptional circumstances, such as emergencies where notice could result in danger, or where notice would be counterproductive (e.g., where an account has been compromised). OpenAI provides delayed notice to users after disclosure where no longer prohibited by law from doing so, such as upon expiration of a nondisclosure order. OpenAI may also provide notice when it receives a request for preservation of data when circumstances warrant. - OpenAI User data request policy, 2026

District court of New York, OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION

"Accordingly, OpenAI is NOW DIRECTED to preserve and segregate all output log data that would otherwise be deleted on a going forward basis until further order of the Court (in essence, the output log data that OpenAI has been destroying), whether such data might be deleted at a user's request or because of "numerous privacy laws and regulations" that might require OpenAI to do so. SO ORDERED." (May 13, 2025)



Search and seizure of data

Scenario: In a fraud case, it was determined that the perpetrators used artificial intelligence tools to carry out the offense, including generative AI models to craft deceptive emails, automated social engineering and credential harvesting techniques—such as mass scraping and conversational bots designed to extract data—the creation of fake profiles using AI-generated human faces and credible digital identities, and deepfake video and voice technology to bypass biometric identity verification controls.

The procedural power of Article 19 of BC is essential for obtaining evidence in such cases, for example:

- Access to text generation platforms (e.g., ChatGPT, Claude, Jasper, etc.).
- Retrieval of AI-generated images and videos.
- Collection of emails drafted with AI.
- Identification of scripts used to automate attacks.

But Possible problem arise concerning cross-border access to data...:

- The collection of AI-related evidence may involve data stored or controlled abroad, raising difficult questions about transborder access, territorial sovereignty, conflicts of jurisdiction, and compliance with due process safeguards.

Special Investigative Techniques and AI

- OSINT (Open Source Intelligence) enhanced by AI capabilities.
- Digital Undercover Agents operating within AI-driven environments.
- State Trojans Executed via AI for targeted investigations.
- Geofence and Keyword Searches supported by AI tools.
- Agentic



From the keyword search warrants to “the reverse prompt orders”

Toward new forms of indirect access to patterns of human thinking in criminal investigations?

Generative artificial intelligence systems involve users *externalizing* processes of reasoning, and creativity.

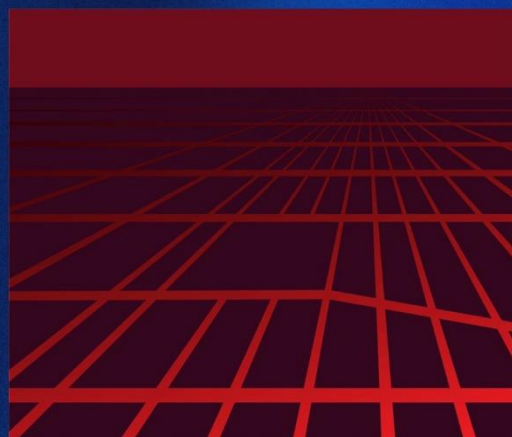
“Reverse prompt search warrants” represent an emerging form of reverse investigation with a potential impact on individual rights that is qualitatively greater than previously known forms.



Reverse investigative techniques

Methods that invert the traditional investigative approach: Instead of starting with a known suspect and gathering evidence, authorities begin with **large datasets** to identify potential suspects.

Example: Geofence warrant is a judicial order that compels technology companies (such as Google) to provide data about all devices located within a defined geographic area during a specific time period



From the keyword search warrants to the reverse prompt orders. A new reality

"This Affidavit is submitted in support of an application under Rule 41 of the Federal Rules of Criminal Procedure to search data and records stored at premises owned, maintained, controlled, or operated by OpenAI..., Specifically, this warrant seeks records and information pertaining to two unique, specific prompts made to OpenAI's "ChatGPT" program by a SUSPECT USER and the unique responses generated by ChatGPT..."

Probable cause:

"...On several occasions in private conversations on the dark web between the UC and SUSPECT USER, the SUSPECT USER has told the UC about using ChatGPT. More specifically on April 18, 2025, the SUSPECT USER indicated SUSPECT USER had created/submitted the prompt to ChatGPT: "what would happen if sherlock holmes met q from star trek?..."

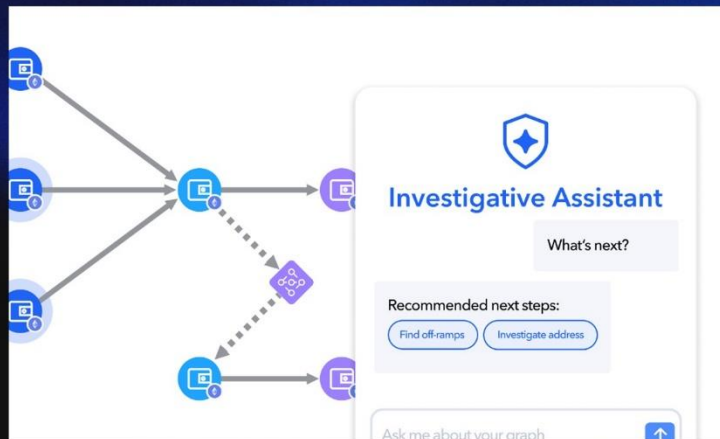
Agentic AI agent in criminal investigations: From passive tools → autonomous systems

Investigation and forensic tools are evolving towards agent-based systems and the integration of LLMs into these tools

Legal risks and issues:

- Black box decisions and Hidden assumptions
- Biases
- Right to defense
- Chain of Custody
- Data Protection
- Cross-border Data

TRM AI agent



Chainalysis

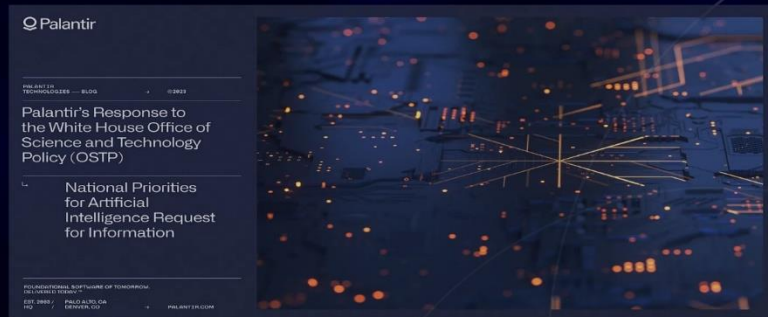
"...We have the context of every type of investigation you run and the compliance obligations you navigate daily. We know how transactions on the blockchain actually function, and the workflows, audit trails, and standards of evidence you rely on..."

"Chainalysis agents are built responsibly because your missions demand it. There is no room for hallucinated outputs or black-box reasoning..."



Palantir

Known for its roots in military intelligence, it provides AI-based data analysis tools to government and commercial clients. Its work has helped clients improve decision-making, but it has also raised privacy concerns because of its role in surveillance, immigration enforcement, and policing.



Clear view



Artificial Intelligence Act

Prohibited AI systems (Art. 5)

Assessing the risk of an individual committing criminal offenses solely based on profiling or personality traits...; compiling facial recognition databases by untargeted scraping of facial images from the internet or CCTV footage.

High risk AI systems (art.6)

Anex 3 (law enforcement)

AI systems used to assess an individual's risk of becoming a crime victim, Evaluating evidence reliability during criminal investigations or prosecutions., Assessing an individual's risk of offending or re-offending not solely based on profiling or assessing personality traits or past criminal behavior, Profiling during criminal detections, investigations or prosecutions...

Artificial Intelligence Act - European Union and Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law

- Effective procedural guarantees
- Human oversight required
- Right to contest AI-informed decisions
- Access to effective remedies
- Judicial review compatibility
- Admissibility of AI-assisted evidence
- Explainability of outputs
- Bias and discrimination risks
- Chain of accountability
- Due process and defence rights

Conclusion

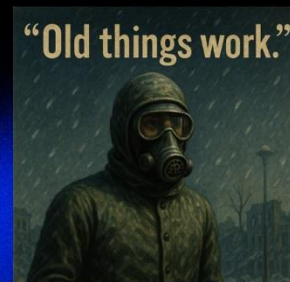
It is not a problem to use AI; the problem is using AI without human oversight, without adequately protecting rights and personal data.

Powerful tool needs safeguards:
can artificial intelligence control itself?



Improper use of IA poses a risk that could undermine not only rights but also the efficiency of future investigations:

Toward a world with intelligent systems but foolish human investigators?



Thanks!

Marcos Salt



Mr Jean-Luc SAURON

L'ÉTAT ALGORITHMIQUE EN FRANCE

Exigences juridiques et institutionnelles applicables à l'usage des systèmes algorithmiques par les pouvoirs publics

Perspective européenne et opérationnelle

INTRODUCTION

Dans la France de 2026, la vie publique est profondément structurée par des algorithmes, dont le public a entendu parler sans identifier exactement ce qu'ils recouvrent⁷. Le plus souvent il n'en a connaissance que lorsqu'ils dysfonctionnent⁸.

Si cette réalité quotidienne de millions de décisions administratives individuelles désormais prises ou influencées chaque année par des algorithmes ouvre des perspectives considérables en termes d'efficience et d'équité formelle, elle soulève des questions juridiques, institutionnelles et démocratiques pas entièrement définies et encore moins résolues.

Les algorithmes publics sont des **formes de l'action publique**. Les administrations qui les conçoivent et les utilisent doivent "**rendre des comptes**"⁹ de leur utilisation auprès des individus concernés, mais aussi de la société dans son ensemble (article 15 de la Déclaration des Droits de l'Homme et du Citoyen de 1789)¹⁰. Cette **explicabilité** de la décision algorithmique est une question centrale qui s'inscrit au cœur d'un droit administratif visant à maximiser la transparence de l'action publique.

Par rapport aux algorithmes mis œuvre par le secteur privé, les algorithmes publics ont des caractéristiques particulières. Ils sont censés opérer au service de l'intérêt général. Ils servent souvent à faire appliquer la loi. Mais surtout **ils sont bien souvent incontournables**, c'est à dire qu'il n'existe pas d'alternatives pour les usagers.

Nous allons examiner la place de l'Etat algorithmique en France en 2026 autour de trois thématiques : (1) tout d'abord l'imbrication sans cesse plus complexe de règles juridiques nationales et européennes, (2) ensuite l'étude des cas d'usages permettant de cerner les dysfonctionnements de leurs mises en œuvre, et (3) enfin les perspectives (tensions persistantes et réformes nécessaires).

Nous n'évoquerons pas dans cette intervention le coût environnemental du recours à l'Intelligence artificielle, ni la question pourtant cruciale de la souveraineté numérique.

⁷ « Si 83% des Français ont déjà entendu parler des algorithmes, ils sont plus de la moitié à ne pas savoir précisément de quoi il s'agit (52%). Leur présence est déjà appréhendée comme massive dans la vie de tous les jours par 80% des Français qui considèrent, à 65%, que cette dynamique va encore s'accroître dans les années qui viennent. » (Sondage mené par l'IFOP pour la CNIL en janvier 2017), in « Comment permettre à l'homme de garder la main », CNIL, décembre 2017.

⁸ Denis Merigoux, Marie Alauzen, Justine Banuls, Louis Gesbert, Émile Rolley. De la transparence à l'explicabilité automatisée des algorithmes : comprendre les obstacles informatiques, juridiques et organisationnels. RR-9535, INRIA Paris. 2024, pp.68. (hal-04391612). Voir également « Les algorithmes contre la société » par Hubert Guillaud, la fabrique éditions, juin 2025.

⁹ Voir à ce propos l'article 9 de la Convention IA du Conseil de l'Europe (« Obligation de rendre des comptes et responsabilité »).

¹⁰ « La société a le droit de demander compte à tout agent public de son administration. »

1. L'imbrication sans cesse plus complexe des règles juridiques et jurisprudentielles nationales et européennes¹¹

Nous allons examiner cette imbrication en trois étapes.

1.1 De la loi Informatique et Libertés de 1978 à la loi pour une République numérique de 2016.

La régulation des algorithmes publics en France trouve ses racines dans la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, dite Informatique & Libertés (ci-après LIL). **Son article 2**, dans sa rédaction d'origine, posait un principe fondateur et clair : « Aucune décision administrative ou privée impliquant une appréciation sur un comportement humain ne peut avoir **pour seul fondement** un traitement automatisé d'informations donnant une définition du profil ou de la personnalité de l'intéressé. » Ce principe, en apparence robuste, a progressivement été tempéré par des exceptions qui, au fil des réformes, ont pris une place prépondérante. Cette interdiction s'est maintenue **près de 18 années**.

« le droit d'accès aux documents administratifs, qui a valeur constitutionnelle (décision du Conseil constitutionnel n° 2020-834, du 3 avril 2020), s'exerce dans les conditions fixées aux articles L300-2 et L311-1 et suivants du code des relations entre le public et l'administration, lesquelles concilient l'exigence de transparence de l'action administrative découlant des dispositions de l'article 15 de la Déclaration des droits de l'homme et du citoyen de 1789 et la protection des secrets protégés par la loi, à l'instar de la sécurité des systèmes d'information des administrations prévue au d) du 2° de l'article L311-5 du même code et de la recherche et de la prévention, par les services compétents, d'infractions de toute nature, prévues par le g) du 2° du même article. »¹²

La **loi n° 2016-1321 du 7 octobre 2016 « pour une République numérique »**¹³ constitue un véritable tournant dans la régulation des algorithmes publics en France. Cette loi introduit dans le Code des relations entre le public et l'administration (ci-après CRPA) un dispositif articulé autour de **trois obligations de transparence**, relatives à l'utilisation des algorithmes par l'administration. Ces obligations sont applicables dès lors que **quatre conditions** sont réunies : [1] l'auteur de la décision est une administration d'État, une collectivité territoriale, d'autres personnes de droit public ou des personnes de droit privé chargées d'une mission de service public (art. L300-2 du CRPA) ; [2] il utilise un traitement algorithmique [3] qui fonde une décision administrative individuelle, [4] traitement qui n'est pas couvert par l'un des secrets légaux mentionnés à l'article L.311-5, 2 du CRPA: défense nationale, sécurité publique, prévention d'infractions .

Les trois obligations de transparence du CRPA

¹¹ Nous ne traiterons que de l'application du droit de l'UE, seul droit européen entrée en vigueur. Mais nous gardons à l'esprit la Convention-cadre sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit, qui sera évoquée au cours de cette intervention.

¹² Avis CADA n°20235120 du 2 novembre 2023.

¹³ Cette loi est discutée et adoptée en grande partie après l'adoption du RGPD (25 mai 2016) et doit se comprendre dans ce contexte.

Obligation n° 1 — Information de type général (art. L312-1-3 du CRPA)

Toute administration comportant plus de 50 agents équivalents temps plein, ainsi que les collectivités de plus de 3 500 habitants, doit publier « spontanément » en ligne les règles définissant ses principaux traitements algorithmiques utilisés dans l’accomplissement de leurs missions de service public lorsqu’ils fondent des décisions individuelles. Cette obligation vise à permettre un contrôle démocratique diffus sur les pratiques administratives. La mise en œuvre de cette obligation informative d’ordre générale est lourde : elle demande de réaliser un inventaire de tous les algorithmes utilisés par les services de ladite administration. Dans son étude intitulée « Intelligence artificielle et action publique : construire la confiance, servir la performance » (2022), le Conseil d’Etat souligne que « toutefois, cette obligation n’est assortie d’aucune sanction et son effectivité est douteuse en pratique ».

Dans les faits, **l’effectivité de cette obligation reste très inégale**¹⁴. Un rapport de la Défenseure des droits publié en novembre 2024 (« Algorithmes, systèmes d’IA et services publics : quels droits pour les usagers ? ») relève que de nombreuses administrations ne respectent pas cette exigence, faute d’avoir identifié l’ensemble de leurs traitements algorithmiques ou de disposer de ressources humaines dédiées à la mise en conformité. La consultation de l’inventaire des algorithmes publics publié par **l’observatoire (citoyen) des algorithmes publics**¹⁵ en novembre 2025 confirme ce point : sur 120 algorithmes publics répertoriés, seuls 6 des 24 algorithmes publics couverts par le CRPA ont rempli leurs obligations de publicité issues du CRPA.

Mais une dernière raison pourrait expliquer ce retard à l’exécution de cette première obligation : **l’absence de définition légale ou réglementaire** de ce que recouvre la **notion d’algorithme**. Ni le RGPD, ni la loi Informatique & Libertés (ci-après LIL), ni le règlement 2024/1689 ne comprennent une telle définition permettant aux administrations une identification précise des algorithmes qu’elles utilisent et l’application des garanties prévues au CRPA. Seul le guide pratique de la publication en ligne et de la réutilisation des données publiques (« open data »), document élaboré par la Commission d’accès aux documents administratifs (ci-après CADA) et la Commission nationale informatique & Libertés (ci-après CNIL) en 2019 donne une définition de l’algorithme dans son glossaire : « Algorithme : au sens strict, un algorithme est la description d’une suite finie et non ambiguë d’étapes (ou d’instructions) permettant de déterminer un droit à une prestation à partir de critères prédéfinis constitue un algorithme. »

Obligation n° 2 — Mention explicite d’un traitement algorithmique (art. L311-3-1 et R311-3-1-1 du CRPA)

Sous réserve de l’application du 2 de l’article L.311-5 du CRPA, toute décision individuelle prise sur le fondement d’un traitement algorithmique doit comporter une « **mention explicite** »

¹⁴ En mai 2026, ont effectué cette publication : la région Île de France, la ville de Paris, la métropole européenne de Lille (en cours d’expérimentation), la ville de la Rochelle, la communauté d’agglomération de la Rochelle, la ville d’Antibes, Nantes métropole, France Travail, le département d’Ille et Vilaine et le département du Val d’Oise.

¹⁵ <https://odap.fr/inventaire>.

informant l'intéressé qu'il a fait l'objet d'une décision individuelle automatisée¹⁶. Cette mention doit préciser la finalité dudit traitement et rappeler le droit de l'intéressé d'obtenir la communication des règles définissant ce traitement et les principales caractéristiques de sa mise en œuvre ainsi que les modalités d'exercice de ce droit à communication et de saisine, le cas échéant, de la Commission d'accès aux documents administratifs (ci-après **CADA**).

Dans son avis publié sur le projet de loi « Pour une République numérique », le Conseil d'Etat, au point 24 de l'avis, attire l'attention du gouvernement « sur la nature des informations prises en compte dans un traitement algorithmique qui, par leur précision, ne doivent pas non plus permettre à des usagers de se constituer un profil permettant de contourner les prescriptions qui seraient applicables aux opérateurs »¹⁷.

Depuis le 1^{er} juillet 2020, l'absence de cette mention dans les décisions entièrement automatisées entraîne la nullité de la décision administrative. Cette sanction par la nullité introduite par la réforme de la loi Informatique et Libertés de 2018 (art. 47 LIL modifiée) est invocable devant le juge administratif, mais n'a jusqu'à maintenant pas donné lieu à des contentieux sans que cette circonstance puisse être interprétée dans un sens positif.

Obligation n° 3 — Information individuelle à la demande (art. R311-3-1-2)

À la demande de l'intéressé ayant fait l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, l'administration doit lui communiquer, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :

(1) le degré et le mode de contribution du traitement à la prise de décision : c'est-à-dire si la décision a été prise de manière entièrement automatisée ou non ;

(2) les données traitées et leur source, c'est-à-dire la manière dont le système informatique a été alimenté.

(3) les paramètres de traitement et leur pondération appliqués à sa situation, c'est-à-dire le poids des différents critères de traitement des données ;

(4) les opérations effectuées par le traitement, c'est-à-dire les différentes étapes du traitement (par exemple : collecte et normalisation des données, croisement et détection d'écarts, application du modèle de scoring, comparaison au seuil de déclenchement, inscription en file de contrôle et validation humaine et notification).

¹⁶ CE, Association médicale indépendante de formation et autres, 24 juillet 2023, n°462778, B, point 11 : « Il résulte des termes mêmes du premier alinéa de l'article L. 311-3-1 du code des relations entre le public et l'administration (CRPA) qu'il ne s'applique que lorsqu'un traitement algorithmique a fondé, en tout ou partie, une décision individuelle. Tel n'est pas le cas d'un traitement algorithmique utilisé par l'ANDPC ayant seulement pour objet de déterminer les actions d'organismes de développement professionnel continu des professionnels de santé susceptibles de faire l'objet d'une évaluation au cours d'une année, et qui ne fonde ainsi aucune décision individuelle, et notamment pas les mesures qui peuvent être prises à la suite de l'évaluation défavorable d'une action contrôlée. ».

¹⁷ Avis n°390741 du 3 décembre 2015.

Cette obligation d'explication individualisée constitue la garantie la plus substantielle pour l'administré. Elle conditionne l'exercice effectif du droit au recours, lequel ne peut être utilement exercé que si l'intéressé comprend les raisons de la décision le concernant.

Enfin le code source de l'algorithme qui a été utilisé pour prendre une décision individuelle est un document administratif (article L.300-2 du CRPA). Il peut faire l'objet d'une publication en ligne d'initiative par toute administration comportant plus de 50 agents équivalents temps plein, ainsi que les collectivités de plus de 3 500 habitants (article L.312-1-1 du CRPA) ou à la demande d'un administré (article L.300-9 du CRPA).

1.2. Le tournant de l'intégration en droit interne de l'article 22 du RGPD (loi 2018-493 du 20 juin 2018 relative à la protection des données personnelles et décret n° 2019-536 du 29 mai 2019).

L'article 22 du RGPD (décision individuelle automatisée, y compris le profilage) prévoit que :

« 1. La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire.

2. Le paragraphe 1 ne s'applique pas lorsque la décision: a) est nécessaire à la conclusion ou à l'exécution d'un contrat entre la personne concernée et un responsable du traitement; b) est autorisée par le droit de l'Union ou le droit de l'État membre auquel le responsable du traitement est soumis et qui prévoit également des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée; ou c) est fondée sur le consentement explicite de la personne concernée.

3. Dans les cas visés au paragraphe 2, points a) et c), le responsable du traitement met en œuvre des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée, au moins du droit de la personne concernée d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision.

4. Les décisions visées au paragraphe 2 ne peuvent être fondées sur les catégories particulières de données à caractère personnel visées à l'article 9, paragraphe 1, à moins que l'article 9, paragraphe 2, point a) ou g), ne s'applique et que des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ne soient en place. »

Cet article **pose les conditions dans lesquelles** une administration peut prendre **une décision automatisée**. Il commence par rappeler le principe d'une interdiction de la prise d'une décision fondée exclusivement sur un algorithme (« traitement automatisé » dans les termes du RGPD), produisant des effets juridiques la concernant ou l'affectant de manière significative.

L'administration peut utiliser un algorithme pour prendre des décisions mais avec une intervention humaine. L'administration peut prendre, sans intervention humaine, des décisions automatisées « mineures » (sans effets juridiques ou n'affectant pas les personnes de manière significative) : par ex., l'attribution automatisée d'un rendez-vous en préfecture.

Surtout, cet article 22 prévoit deux exceptions qu'il décrit (consentement, nécessité pour un contrat) et une troisième qu'il formule ainsi à savoir « *l'autorisation par le droit de l'UE ou le droit national (...) qui prévoit des mesures appropriées pour la sauvegarde des droits et libertés et*

des intérêts légitimes des personnes concernées », exception particulièrement mobilisable par l'administration nationale.

Cette autorisation du droit national fait l'objet de l'article 47 de la loi Informatique & Libertés (ci-après LIL) pour les traitements qui relèvent du RGPD :

« Aucune décision de justice impliquant une appréciation sur le comportement d'une personne ne peut avoir pour fondement un traitement automatisé de données à caractère personnel destiné à évaluer certains aspects de la personnalité de cette personne.

Aucune décision produisant des effets juridiques à l'égard d'une personne ou l'affectant de manière significative ne peut être prise sur le seul fondement d'un traitement automatisé de données à caractère personnel, y compris le profilage, à l'exception :

1. Des cas mentionnés aux a et c du 2 de l'article 22 du règlement (UE) 2016/679 du 27 avril 2016, sous les réserves mentionnées au 3 du même article 22 et à condition que les règles définissant le traitement ainsi que les principales caractéristiques de sa mise en œuvre soient communiquées, à l'exception des secrets protégés par la loi, par le responsable de traitement à l'intéressé s'il en fait la demande ;

2. Des décisions administratives individuelles prises dans le respect de l'article L. 311-3-1 et du chapitre Ier du titre Ier du livre IV du code des relations entre le public et l'administration, à condition que le traitement ne porte pas sur des données mentionnées au I de l'article 6 de la présente loi. Ces décisions comportent, à peine de nullité, la mention explicite prévue à l'article L. 311-3-1 du code des relations entre le public et l'administration. Pour ces décisions, le responsable de traitement s'assure de la maîtrise du traitement algorithmique et de ses évolutions afin de pouvoir expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard.

Par dérogation au 2° du présent article, aucune décision par laquelle l'administration se prononce sur un recours administratif mentionné au titre Ier du livre IV du code des relations entre le public et l'administration ne peut être prise sur le seul fondement d'un traitement automatisé de données à caractère personnel. »

Cet article 47 commence par reprendre le principe d'interdiction de la prise d'une décision fondée exclusivement sur un algorithme, produisant des effets juridiques la concernant ou l'affectant de manière significative posé par l'article 22 RGPD.

Dans le cadre d'une autorisation du droit national (mentionnée au b) du deuxième paragraphe de l'article 22 du RGPD, il prévoit au 2° de son deuxième alinéa une possibilité pour l'administration de prendre de telles décisions individuelles automatisées fondées exclusivement sur un algorithme (ou dit autrement traitement automatisé de données à caractère personnel)¹⁸ respectant les garanties prévues par l'article L. 311-3-1 du CRPA, précisé par les articles R.311-3-1-1 et R311-3-1-2 du CRPA (explication individualisée et explication détaillée) et par le droit de l'Union¹⁹, sous réserve de ne pas concerner des données sensibles au sens du I de l'article 6 de la LIL.

¹⁸ L'article 47 de la LIL emploie les deux formules comme synonymes : traitement automatisé de données à caractère personnel et traitement algorithmique.

¹⁹ Voir à ce propos, le considérant 70 de la décision du Conseil constitutionnel no 2018-765 DC du 12 juin 2018 (loi relative à la protection des données personnelles) qui précise qu'« Il en résulte que, lorsque les principes de fonctionnement d'un algorithme ne peuvent être communiqués sans porter atteinte à l'un des secrets ou intérêts énoncés au 2° de l'article L. 311-5 du code des relations entre le public et l'administration, aucune décision individuelle ne peut être prise sur le fondement exclusif de cet algorithme ».

Mais il ajoute une **nouvelle exigence d'explication** à la charge du responsable de traitement. Cette garantie mentionnée dans cette troisième phrase de l'alinéa 2 de l'article 47 de la LIL n'est pas une mesure issue du RGPD, mais un ajout national qui instaure un dispositif d'explication plus exigeant que les dispositions précitées du CRPA. Dans cette troisième phrase, il est prévu que le responsable du traitement « *s'assure de la maîtrise du traitement algorithmique et de ses évolutions afin de pouvoir expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard* ».

Cette troisième phrase de l'alinéa 2 de l'article 47 de la LIL opère, sur la thématique de la motivation de la décision individuelle algorithmique de l'administration, un saut qualitatif dans la protection de l'utilisateur du service public. Elle ajoute aux garanties prévues par le CRPA une **exigence d'intelligibilité** du traitement algorithmique qui concerne tant le destinataire de la décision que, et c'est là où il y a une grande nouveauté, l'administration auteure de ladite décision.

Ainsi, l'esprit de la 3^{ème} phrase de l'alinéa 2 de l'article 47 de la LIL est d'inciter l'administration à se réformer et à concevoir la motivation non comme une opération intervenant par exception et a posteriori dans le cadre d'un contentieux ou d'une réclamation, c'est-à-dire au cas par cas, mais comme inhérente à chaque décision individuelle prise qui doit être conçue dans cette perspective d'explicabilité vis-à-vis de l'extérieur et de l'intérieur (service ou administration)²⁰. Par suite, un algorithme auto-apprenant, en ce qu'il interdit au responsable de traitement d'en expliciter la mise en œuvre, ne peut fonder une décision administrative individuelle conforme aux dispositions précitées du CRPA et de la LIL²¹.

Enfin, le Conseil constitutionnel rappelle²² que « la décision administrative individuelle [fondée exclusivement sur un algorithme] doit pouvoir faire l'objet de recours administratifs, (...). L'administration sollicitée à l'occasion de ces recours est alors tenue de se prononcer sans pouvoir se fonder exclusivement sur l'algorithme. La décision administrative est en outre, en cas de recours contentieux, placée sous le contrôle du juge, qui est susceptible d'exiger de l'administration la communication des caractéristiques de l'algorithme ».

Ces obligations s'ajoutent naturellement aux obligations générales imposées par le RGPD (une prise de décision individuelle automatisée est un traitement de données à caractère personnel) : pertinence et exactitude des données, durées de conservation, sécurité, droits des personnes, etc.

1.3 L'intervention du juge européen et national.

²⁰ Voir à ce propos, Denis Merigoux, Marie Alauzen, Justine Banuls, Louis Gesbert, Émile Rolley. De la transparence à l'explicabilité automatisée des algorithmes : comprendre les obstacles informatiques, juridiques et organisationnels. RR-9535, INRIA Paris. 2024, pp.68. (hal-04391612).

²¹ Voir à ce propos, le considérant 71 de la décision du Conseil constitutionnel no 2018-765 DC du 12 juin 2018 (loi relative à la protection des données personnelles) qui précise qu'«il en résulte que ne peuvent être utilisés, comme fondement exclusif d'une décision administrative individuelle, des algorithmes susceptibles de réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement ».

²² Considérant 70 de la décision du Conseil constitutionnel no 2018-765 DC du 12 juin 2018 (loi relative à la protection des données personnelles).

Une jurisprudence européenne qui commence à s'affirmer.

La Cour de justice de l'Union européenne commence à établir une jurisprudence sur ces dispositions, jurisprudence qui recouvre pleinement l'état de la législation nationale française. Les plus importantes étant :

La décision « DUN and BRADSTREET » du 27 février 2025, affaire C-203/22, interprétant l'article 15, 1, h) du RGPD, concernant le niveau d'information et d'explication devant être fourni à la personne concernée (celle dont les données à caractère personnel ont fait l'objet d'un traitement) qui peut exiger du responsable du traitement, au titre des « informations utiles concernant la logique sous-jacente », que celui-ci lui explique, au moyen d'informations pertinentes et d'une façon concise, transparente, compréhensible et aisément accessible, la procédure et les principes concrètement appliqués pour exploiter, par la voie automatisée, ses données à caractère personnel. Dans l'hypothèse où le responsable du traitement considère que les informations à fournir à la personne concernée conformément à cette disposition comportent des données de tiers protégées (secrets d'affaires), ce responsable est tenu de communiquer ces informations prétendument protégées à l'autorité de contrôle ou à la juridiction compétentes, auxquelles il incombe de pondérer les droits et les intérêts en cause²³ aux fins de déterminer l'étendue du droit d'accès de la personne concernée prévu à l'article 15 de ce règlement. Cette dernière règle trouverait à s'appliquer en droit interne du fait de la primauté du droit de l'Union sur le droit national, mais l'auteur de cette contribution n'a pas trouvé de trace de l'application de cette règle.

La décision « SCHUFA » du 7 décembre 2023, affaire C-634/21 concernant la notion de décision automatisée (et le rôle de l'intervention humaine). L'établissement automatisé, par une société fournissant des informations commerciales, d'une valeur de probabilité fondée sur des données à caractère personnel (scoring) relatives à une personne et concernant la capacité de celle-ci à honorer des engagements de paiement à l'avenir constitue une « décision individuelle automatisée », au sens de l'article 22 §1 du RGPD. Concernant les mesures appropriées pour la sauvegarde des droits et des libertés ainsi que des intérêts légitimes de la personne concernée, les dispositions de l'article 22, paragraphe 2, sous b), du RGPD, mentionne que le droit national qui autorise l'adoption d'une décision individuelle automatisée doit prévoir les mesures techniques et organisationnelles appropriées pour faire en sorte que le risque d'erreur soit réduit au minimum et que des erreurs soient corrigées, ainsi que de sécuriser les données à caractère personnel d'une manière qui tienne compte des risques susceptibles de peser sur les intérêts et les droits de la personne concernée et prévenir, entre autres, les effets discriminatoires à l'égard de celle-ci. Ces mesures appropriées pour la sauvegarde des droits et des libertés incluent, par ailleurs, au moins le droit de la personne concernée d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision prise à son égard (simple reprise du 3^{ème} paragraphe de l'article 22 du RGPD).

²³ Règle recoupant les prévisions du considérant 70 de la décision du Conseil constitutionnel n°2018-765 DC du 12 juin 2018 (loi relative à la protection des données personnelles).

Pourquoi si peu de jurisprudence de l'Union en cette matière précise de la décision administrative individuelle automatisée ? Ces deux décisions jurisprudentielles de l'Union sont des renvois préjudiciels. Le délai de traitement desdites questions préjudicielles, de façon générale, est de 16,7 mois (entre la saisine de la Cour de justice et sa réponse) en 2025. Il n'y en a pas d'autres jurisprudences de même importance dans la jurisprudence de l'Union.

Le rôle du juge administratif français : un contrôle quasiment toujours médiatisé.

De la même manière que devant le juge de l'Union, il a eu jusqu'à maintenant peu de recours contentieux dans le domaine de la décision individuelle automatisée devant le juge administratif français : sans doute du fait de la complexité de la réglementation et de la brièveté du temps écoulé entre l'entrée en vigueur des législations (8 à 10 années) ou de l'activité de la CNIL ou de la CADA, institutions protectrices des droits des destinataires des décisions individuelles automatisées et qui ont rendu, notamment en ce qui concerne cette dernière, en très grande majorité des avis favorables aux requérants²⁴. L'analyse de l'auteur de cette contribution de ce faible volume des contentieux est convergente avec celle de la Défenseure des droits dans son rapport d'activités pour 2025 : « le nombre de saisines de l'institution et de contestation de ces décisions reste faible, sans doute en raison de l'opacité forte des processus de décision, du manque de transparence et d'intelligibilité des systèmes algorithmiques pour le grand public et, plus largement, de sa méconnaissance de ces systèmes et de ses droits ».

Lors du recours, l'administration ne peut se fonder une nouvelle fois exclusivement sur l'algorithme : une intervention humaine est impérative dans le réexamen du dossier lors de la défense de la position de l'administration devant le juge.

Le Conseil d'Etat est **l'instance de contrôle de deux autorités administratives indépendantes** créées la même année : la **CNIL** par la loi du 6 janvier 1978 et la **CADA** par la loi du 17 juillet 1978, Autorités Administratives Indépendantes (AAI) qui couvrent l'ensemble des questions relatives au suivi des décisions algorithmiques de l'Etat.

En effet, concernant la CNIL, la personne concernée doit saisir tout d'abord le responsable du traitement avant de saisir la CNIL, CNIL dont elle pourra contester la décision devant le Conseil d'Etat²⁵. Pour la CADA, l'article L.342-1 du CRPA prévoit qu'elle émet un avis lorsqu'elle est saisie par une personne à qui est opposé un refus de communication ou un refus de publication d'un document administratif en application du titre Ier du CRPA. Cette saisine constitue un préalable obligatoire à l'exercice d'un recours contentieux²⁶.

La **Commission d'accès aux documents administratifs** (CADA) a rendu des avis structurants en matière d'accès aux codes sources des algorithmes publics²⁷ et ce d'autant que la loi du 7 octobre 2016 « pour une République numérique » a élargi un droit d'accès aux documents

²⁴ In « Le contentieux des décisions administratives automatisées », par Nina Leasbleiz, RFDA 2025, page 211.

²⁵ CE, Mme Pouille, 27 janvier 2025, n°490416, B.

²⁶ CE, Association one voice, 23 décembre 2025, n°503789.

²⁷ Voir rapport d'activités de la CADA, 2022-2023, pages 41 et suivants (« Favoriser la transparence administrative à travers les algorithmes et les codes sources »).

administratifs dans le domaine numérique, puisque sont désormais expressément inclus dans la définition des documents administratifs communicables, les codes sources. Cette même loi a également créé un droit d'accès spécifique à certaines informations relatives aux traitements algorithmiques fondant les décisions individuelles (article L.311-3-1 du CRPA), ouvert aux personnes faisant l'objet de telles décisions. Le droit d'accès aux codes sources et aux algorithmes ne peut intervenir que sous réserve des secrets protégés par la loi.

Pour la CADA, la sécurité des systèmes d'informations peut faire obstacle à la communication de l'intégralité d'un code source dans la mesure où sa divulgation peut constituer un facteur de vulnérabilité des systèmes d'information²⁸. Ainsi, la CADA a notamment admis que certains codes sources pouvaient être refusés au titre de la sécurité des systèmes d'information (exemple : le Health Data Hub) ou de la sécurité publique (exemple : les arbres décisionnels du SAMU et des sapeurs-pompiers de Paris²⁹). S'agissant des algorithmes, le principal secret opposable est le risque d'atteinte à la recherche et à la prévention, par les services compétents, d'infractions de toute nature³⁰.

Sur la thématique de l'Etat algorithmique, la CNIL a rendu une seule décision de mise en demeure à l'encontre de la plate-forme Admission Post Bac ou APB (Décision MED-2017-053 du 30 août 2017)³¹, mise en demeure ayant précédée de peu son remplacement par la plate-forme Parcoursup. Des contrôles sont en cours sur des dispositifs de lutte contre la fraude, basés sur des algorithmes, mais aucune décision n'a pas été, à avril 2026, adoptée ni publiée.

La coexistence de ces deux autorités administratives indépendantes peut être source de difficultés du fait de l'articulation entre le droit d'accès aux documents administratifs et le droit à la protection des données à caractère personnel, notamment dans des cas de chevauchements de compétences. En effet, il n'est pas toujours simple pour une personne qui exerce son droit d'accès à des documents à l'égard d'une administration de savoir si elle le fait sur le fondement du livre III du code des relations entre public et l'administration (compétence de la CADA) ou sur le fondement de l'article 15 du RGPD (compétence de la CNIL)³². Cette difficulté spécifique s'explique par le fait que des droits concurrents existent et que le législateur a pu régler la question dans certains domaines mais pas dans tous. Cependant, les deux autorités travaillent de concert pour fixer des lignes de partage claires.

²⁸ Avis CADA III n°20235126 du 21 septembre 2023.

²⁹ Article d) de l'article L.311-5 du CRPA : voir avis CADA n°20172357 du 21 septembre 2017, avis CADA n°20244499 du 19 septembre 2024, ministère de l'intérieur, et avis CADA n°20255551 du 9 septembre 2025, Conseil national de l'ordre des pharmaciens. Dans ces deux derniers avis cités, la CADA considère, dans des cas de demande de communication de code source que « ces documents sont communicables (...) après occultation ou disjonction des fragments relevant des secrets et intérêts protégés par ces articles, en particulier la sécurité des systèmes d'information et le secret des affaires ».

³⁰ Article g) du 2° de l'article L.311-5 du CRPA.

³¹

https://www.legifrance.gouv.fr/cnil/id/CNIL/TEXT000035647959?init=true&page=1&query=mise+en+demeure+algorithme&searchField=ALL&tab_selection=all.

³² « 1. La personne concernée a le droit d'obtenir du responsable du traitement la confirmation que des données à caractère personnel la concernant sont ou ne sont pas traitées et, lorsqu'elles le sont, l'accès aux dites données à caractère personnel ainsi que les informations suivantes: (...) h) l'existence d'une prise de décision automatisée, y compris un profilage, visée à l'article 22, paragraphes 1 et 4, et, au moins en pareils cas, des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée. »

Cette nécessité pour ces deux autorités de régulation (CADA/CNIL) de travailler ensemble sur les questions communes implique l'existence de relations étroites entre elles. Depuis l'adoption de la loi pour une République numérique (entrée en vigueur en 2017), la CADA et la CNIL peuvent se réunir en collège commun et sont représentées dans chacun des collèges. La réunion d'un collège commun permet le développement et le renforcement de celles-ci. De plus, la présence du président de la CADA au sein du collège de la CNIL, et, réciproquement, d'un représentant de la CNIL au sein du collège de la CADA, permet aussi à ces autorités de renforcer les échanges et de faire des comparaisons indispensables (par exemple, dans l'hypothèse d'une prise de décision automatisée, accès aux algorithmes prévu par le code des relations entre le public et l'administration et accès à la logique sous-jacente du traitement prévu à l'article 15 du RGPD).

Que va devenir cette interrégulation (CADA/CNIL) confrontée dans le cadre de la mise en œuvre du RIA à la participation de deux administrations du ministère de l'économie et des finances que sont la direction générale des entreprises (DGE) et la direction générale de la concurrence, de la consommation et de la répression des fraudes (DGCCRF) justement pour les systèmes d'intelligence artificielle mis en œuvre par certaines administrations ?

2. Les expériences françaises.

Les cas suivants illustrent certaines des modalités concrètes de déploiement des systèmes algorithmiques par l'administration française, les arbitrages opérés entre automatisation et intervention humaine, et les tensions juridiques qui en résultent.

2.1 PARCOURSUP — L'algorithme d'affectation dans l'enseignement supérieur

Parcoursup, la plateforme nationale d'affectation dans l'enseignement supérieur, est certainement l'algorithme public le plus connu et le plus débattu en France. Lancée en 2018 en remplacement d'APB (Admission Post-Bac)³³, elle traite annuellement les candidatures de plus de 900 000 lycéens vers quelque 20 000 formations. Le système utilise deux niveaux d'algorithmes : un algorithme national géré par le ministère, et des algorithmes locaux développés par chaque établissement pour classer les candidatures reçues (à défaut d'utiliser un algorithme fourni par le ministère de l'enseignement supérieur nommé « outil d'aide à la décision » ou OAD), fondé exclusivement sur les notes.

Quatre algorithmes nationaux interviennent dans l'appariement optimal entre candidats et les capacités d'accueil des formations, en tenant compte des préférences ordonnées des étudiants et des classements opérés par les établissements. Ces algorithmes garantissent le respect de pourcentages de boursiers et de candidats résidant dans l'académie, modifiant automatiquement les classements pour respecter ces seuils. L'article L.612-3 du Code de l'éducation impose la publication du cahier des charges et des algorithmes de Parcoursup, qui sont accessibles sur le dépôt public Gitlab, outil de **gestion de version de code source** basé sur Git, une technologie de gestion de code source.

³³ Qui est un algorithme qui a « échoué », qui a été fermé et a été remplacé par Parcoursup.

En l'absence de décision entièrement automatisée, compte tenu de l'intervention des commissions d'examen des vœux (CEV) au niveau de chaque établissement), le système échappe à l'obligation du RGPD d'information sur la logique sous-jacente (voir h) paragraphe 1^{er} de l'article 15 en liaison avec l'article 22 du RGPD). La CNIL recommande néanmoins une telle information à titre de bonne pratique.

Cette procédure de Parcoursup a fait l'objet de jurisprudences « croisées » du Conseil d'Etat et du Conseil constitutionnel.

Selon le Conseil d'Etat, si les articles L. 311-1, L. 311-3-1 et L. 312-1-3 du CRPA sont, en principe, applicables aux traitements algorithmiques utilisés, le cas échéant, par les établissements d'enseignement supérieur pour fonder des décisions individuelles et si elles instaurent, par suite, un droit d'accès aux documents relatifs aux algorithmes utilisés par ces établissements et à leurs codes sources, il résulte des termes du dernier alinéa du I de l'article L. 612-3 du code de l'éducation, éclairés par les travaux préparatoires de la loi n° 2018-166 du 8 mars 2018 dont ils sont issus, que le législateur a entendu régir par des dispositions particulières le droit d'accès aux documents relatifs aux traitements algorithmiques utilisés, le cas échéant, par les établissements d'enseignement supérieur pour l'examen des candidatures présentées dans le cadre de la procédure nationale de préinscription (« Parcoursup »). Ces dispositions spéciales doivent ainsi être regardées comme ayant entendu déroger, notamment, aux dispositions de l'article L. 311-1 du CRPA, en réservant l'accès aux documents administratifs relatifs aux traitements algorithmiques utilisés, le cas échéant, par les établissements d'enseignement supérieur pour l'examen des candidatures, aux seuls candidats qui en font la demande, une fois prise la décision les concernant, et pour les seules informations relatives aux critères et modalités d'examen de leur candidature. Ni les tiers ni les candidats, avant qu'une décision ait été prise à leur sujet, ne peuvent donc demander à ce que ces critères et modalités leur soient communiqués³⁴. Pour le Conseil constitutionnel, « le législateur a considéré que la détermination de ces critères et modalités d'examen des candidatures, lorsqu'ils font l'objet de traitements algorithmiques, n'était pas dissociable de l'appréciation portée sur chaque candidature. Dès lors, en restreignant l'accès aux documents administratifs précisant ces critères et modalités, il a souhaité protéger le secret des délibérations des équipes pédagogiques au sein des établissements. Il a ainsi entendu assurer l'indépendance de ces équipes pédagogiques et l'autorité de leurs décisions. Ce faisant, il a poursuivi un objectif d'intérêt général »³⁵.

De même chaque établissement est tenu de publier les critères généraux encadrant l'examen des candidatures par les commissions d'examen des vœux en application de l'article D. 612-1-5 du code de l'éducation³⁶.

Selon le Conseil constitutionnel, les dispositions du dernier alinéa du paragraphe I de l'article L. 612-3 du code de l'éducation ne sauraient, sans méconnaître le droit d'accès aux documents

³⁴ CE, 12 juin 2019, Université des Antilles, n°427916 et 427919, B.

³⁵ Conseil constitutionnel, 3 avril 2020, UNEF, n°2020-834 QPC, point 13.

³⁶ Dans sa rédaction résultant du décret du 26 mars 2019 relatif à la procédure nationale de préinscription pour l'accès aux formations initiales du premier cycle de l'enseignement supérieur et modifiant le code de l'éducation.

administratifs, être interprétées comme dispensant chaque établissement de publier, à l'issue de Parcoursup et dans le respect de la vie privée des candidats, le cas échéant sous la forme d'un rapport, les critères en fonction desquels les candidatures ont été examinées et précisant, le cas échéant, dans quelle mesure des traitements algorithmiques ont été utilisés pour procéder à cet examen³⁷.

Enfin le Conseil d'Etat a jugé que « le I de l'article L. 612-3 du code de l'éducation doit être interprété, conformément à la décision n° 2020-834 QPC du Conseil constitutionnel du 3 avril 2020, comme n'imposant pas la publication ou la communication aux tiers des traitements algorithmiques eux-mêmes et des codes sources correspondants »³⁸.

Le système Parcoursup continue à faire l'objet de contestation de la part d'association ou de syndicat qui illustre la tension fondamentale entre l'opacité des algorithmes locaux (développés de manière hétérogène par les établissements)³⁹ et l'exigence de transparence. Les inégalités potentielles induites par des critères de sélection locaux insuffisamment contrôlés demeurent un sujet de préoccupation, notamment relevé dans les rapports successifs de la Cour des comptes et du Défenseur des droits, et ceci malgré l'institution dès mars 2018 du comité éthique et scientifique de la plateforme Parcoursup⁴⁰, élargie en 2024 à la plateforme « Mon master »⁴¹. Ce comité éthique et scientifique rend un rapport annuel au ministre de l'enseignement supérieur.

2.2 La DGFIP — Le ciblage algorithmique du contrôle fiscal (CFVR / GALAXIE)

La Direction générale des finances publiques (DGFIP) est pionnière dans l'utilisation des algorithmes à des fins d'analyse de risque. Depuis 2014, le programme Ciblage de la Fraude et Valorisation des Requêtes (ou CFVR) mobilise les techniques de datamining pour identifier les contribuables présentant des anomalies fiscales significatives. En 2025, plus de la moitié des contrôles fiscaux sont déclenchés par des alertes générées par ces outils.

Outils de la DGFIP	Description et données utilisées
CFVR (depuis 2014)	Datamining sur 11 bases de données : Sirius-Pro, BODACC, registres sociaux, etc. Génère des scores de risque pour prioriser les contrôles. Représentait 44 % de la programmation du contrôle fiscal en 2021, avec un objectif de 50 %.

³⁷ Conseil constitutionnel, 3 avril 2020, UNEF, n°2020-834 QPC, point 17.
³⁸ CE, 15 juillet 2020, UNEF, n°433296, B.
³⁹ Voir à ce propos, « A Sciences Po Bordeaux, l’algorithme maison favorise les boursiers et les élèves des lycées « cordées de la réussite », article de Soazig Le Nevé dans le Monde.fr le 23 mai 2023.
⁴⁰ 7^{ème} rapport du comité éthique et scientifique de Parcoursup en 2025.
⁴¹ Voir à ce propos, « Parcoursup : une « réussite technique », mais un « échec politique », souligne un rapport parlementaire », par Soazig Le Nevé dans le Monde.fr le 17 février 2026.

GALAXIE (depuis 2022)	La DGFIP a créé l'outil « GALAXIE ». Cet outil est accessible seulement aux agents de la DGFIP. Une de ses principales fonctions est de cartographier les liens existants entre différents contribuables (entre les personnes morales, entre les personnes morales et personnes physiques, etc.) en intégrant des données issues de leurs situations patrimoniales et fiscales.
IA photographique IGN ⁴²	La DGFIP, en octobre 2021, avec l'appui de Capgemini et de Google, a développé un outil interne baptisé « Foncier innovant », permettant de croiser les déclarations des contribuables et les vues aériennes de l'Institut national de l'information géographique (IGN) afin de repérer les constructions non déclarées (piscines, vérandas, etc.). Résultats : milliers de régularisations fiscales.
Calcul automatisé de l'IR	Production automatisée des avis d'imposition pour 37 millions de foyers fiscaux. Moteur de règles fiscales avec restitution des variables intermédiaires de calcul — niveau d'explicabilité exemplaire selon les chercheurs.

Concernant le statut juridique du CFVR, le g) du deuxième alinéa de l'article L311-5 du CRPA prévoit que le droit à communication des règles algorithmiques ne s'applique pas lorsque la communication porterait atteinte « à la recherche et à la prévention, par les services compétents, d'infractions de toute nature ». La DGFIP peut donc légalement refuser de communiquer les paramètres de ses algorithmes de ciblage, ce qui peut soulever des questions lors de la contestation des contrôles déclenchés algorithmiquement.

2.3 Les Caisses d'allocations familiales (CAF) — Le modèle de Datamining Données Entrantes (DMDE)

Le modèle de **Datamining Données Entrantes (DMDE)** des CAF illustre les défis de transparence et d'équité liés au profilage des allocataires. Développé initialement en 2018, le modèle a fait l'objet d'une refonte en janvier 2026 par la CNAF pour tenir compte des évolutions réglementaires (réforme des aides au logement, prime d'activité, solidarité à la source) et d'un contentieux très mobilisateur.

Selon la CAF, le DMDE fournit un indicateur de risque d'indu (involontaire ou par fraude) qui aide à prioriser certains contrôles. Il n'emporte aucune décision automatique, la décision de contrôler restant humaine. En 2024, les CAF ont réalisé 31,5 millions de contrôles auprès de 6,4 millions d'allocataires. 92 % sont des régularisations automatisées issues d'échanges de

⁴² Institut national de l'information géographique et forestière.

données interbancaires ; environ 8 % sont des contrôles par agents. Les contrôles issus du DMDE représentent 0,5 % de l'ensemble. D'après les chiffres communiqués par la CNAF, 1,98% des dossiers contrôlés dans le cadre du datamining aboutissent à une qualification de fraude⁴³.

Controverse persistante : Des études menées par des chercheurs ont mis en évidence que le DMDE générerait des taux de contrôle significativement plus élevés pour certaines catégories de bénéficiaires (familles monoparentales, personnes en situation de précarité). La Défenseure des droits est allée au soutien de ces critiques : décision 2023-066 du 20 mars 2023 relative à la contestation de qualification de fraude d'indus notifiés par une caisse d'allocations familiales et de sa pénalité afférente et décision 2024-039 du 27 mars 2024 relative à la contestation de qualification de fraude d'indus notifiés par une caisse d'allocations familiales et de sa pénalité afférente. La **CNAF** conteste une causalité directe et insiste sur la neutralité des critères retenus. Ce débat illustre la difficulté de distinguer la détection légitime du risque et la reproduction algorithmique d'inégalités sociales préexistantes.

En réalité, ce scoring des allocataires a fait l'objet de contestations depuis 2022, contestation qui a pris un tour nouveau fin 2024. Quinze organisations (dont la Quadrature du NET, la Ligue des droits de l'Homme) ont déposé le 16 octobre 2024 un recours devant le Conseil d'Etat contre l'algorithme de notation de la Caisse nationale aux prestations familiales (ou CNAF) pour violation du RGPD (notamment son article 22) et la nature discriminatoire de ce scoring⁴⁴. Dix nouvelles organisations se sont ajoutées depuis aux quinze premières. La décision devrait être prononcée au printemps 2026.

La CNAF a accompagné la modification de son algorithme en janvier 2026 et la publication du code source de l'algorithme DMDE en open source en février 2026, de l'adoption d'une **charte** sur l'usage des algorithmes et de l'intelligence artificielle de la Cnaf ⁴⁵encadrant le développement et l'usage des algorithmes et la création d'un « **comité éthique** sur les usages des données, des algorithmes et de l'IA » (installé le 19 mars 2025), qui appuie la CNAF pour identifier les risques et les moyens de les atténuer, notamment sur les sujets sensibles liés aux libertés publiques. Ce comité est consulté pour : « identifier les risques éthiques liés à l'utilisation des données et des algorithmes qui peuvent conduire à des discriminations ; (...) » appuyer la structuration d'une méthodologie de cadrage éthique dès la production d'un nouvel algorithme ». Il ne publie pas de rapport, mais des comptes rendus de ses séances : 4 depuis mars 2025 (chiffre relevé à avril 2026).

Cette même thématique a fait l'objet de contentieux dans d'autres Etats membres : ainsi un scandale a secoué les Pays-Bas relatif à des allocations familiales récupérées à tort par l'Etat, les familles accusées à tort de fraude à l'aide sociale étant principalement d'origine immigrée.

⁴³ In « La lutte contre la fraude aux prestations sociales à l'ère de son industrialisation », rapport du Défenseur des droits, 2026.

⁴⁴ Alice Vitard, « Violation du RGPD : l'algorithme de notation de la CAF attaqué devant le Conseil d'Etat, l'Usine digitale, 18 octobre 2024.

⁴⁵ Il est précisé qu'elle a vocation à évoluer.

Le scandale a conduit en janvier 2021 à la démission collective du gouvernement néerlandais de l'époque (cabinet Rutte III)⁴⁶.

3. Difficultés persistantes.

Les difficultés doivent s'appréhender d'une manière globale sans qu'il soit réellement possible d'isoler la compréhension desdites difficultés isolément une par une. Par exemple, les difficultés autour de la définition de l'intervention humaine (3.1) doivent s'articuler avec les difficultés liées à la formation des personnels des autorités publiques (3.5) ou encore la contrainte d'explicabilité des algorithmes (3.3) avec l'exigence éthique de la gouvernance des algorithmes (3.4).

3.1. Quelle définition de l'intervention humaine ?⁴⁷

Le droit français et le droit européen convergent pour exiger, dans des conditions variables, une « intervention humaine » dans les décisions algorithmiques. Mais **la notion d'intervention humaine souffre d'une définition insuffisante qui en altère la portée protectrice**. Une signature apposée mécaniquement sur des décisions générées algorithmiquement sans véritable examen constitue-t-elle une intervention humaine au sens du droit ? De plus, **le droit existant (national et européen) comprend 3 définitions** : la notion de « maîtrise » par le responsable du traitement (article 47 de la LIL), la notion d'intervention humaine à l'occasion d'une décision individuelle automatisée ou d'un profilage (article 22 du RGPD) et la notion de contrôle humain (article 14 du règlement IA).

Faute de définition précise et unifiée, l'exigence d'intervention humaine ne risque-t-elle pas de devenir une garantie purement formelle ?

3.2 maîtrise administrative et recours à des prestataires privés.

Une partie croissante des systèmes algorithmiques utilisés par l'administration repose sur des solutions développées par des prestataires privés⁴⁸. Cette externalisation crée des tensions avec l'obligation de transparence : comment l'administration peut-elle communiquer à l'administré les paramètres d'un algorithme qu'elle ne maîtrise pas entièrement, dont le code source est protégé par des droits de propriété intellectuelle ou par le secret des affaires ?

La Défenseure des droits (dans son rapport précité de novembre 2024) recommande de « s'assurer de la disponibilité des compétences en interne et de la bonne maîtrise du fonctionnement de ces systèmes par les agents ». La CADA a jugé que les limites du droit de communication pouvaient empêcher l'accès au code source d'un algorithme dont les droits appartiennent à un tiers⁴⁹.

⁴⁶ J.P.Stroobants, « Aux Pays-Bas, le coût d'un scandale sur les allocations familiales qui engorge les tribunaux », Le Monde, 9 janvier 2025.

⁴⁷ A ce propos, voir une publication de link (équipe de recherche de la CNIL) : Contrôle humain, décisions hybrides : quels enjeux ? par Charlotte Barot, 16 octobre 2024 <https://linc.cnil.fr/13-contrôle-humain-décisions-hybrides-quels-enjeux>.

⁴⁸ Voir l'inventaire des algorithmes publics dressé par l'observatoire citoyen des algorithmes publics. L'information sur ce point ne semble pas toujours accessible en ce qu'il peut exister des algorithmiques ou des systèmes d'intelligence artificielle pris en charge par une administration mais dans le cadre d'une procédure où une partie de l'algorithme a été le fruit du travail d'un sous-traitant du secteur privé (startup ou entreprise de conseils).

⁴⁹ Voir à ce propos, Avis CADA n°20255551 du 9 octobre 2025.

Les marchés publics de systèmes d'IA ne devraient-ils pas désormais intégrer systématiquement des clauses de transparence et d'accès au code ?

3. 3 Quelle articulation entre le droit national tel qu'interprété par le Conseil constitutionnel et le règlement (UE) 2024/1689 : risque d'inconstitutionnalité et/ou bouleversement des pratiques administratives ?

Le règlement (UE) 2024/1689, directement applicable en droit français dès le 2 août 2026 pour les systèmes à haut risque (sauf adoption de l'omnibus IA), va considérablement enrichir le corpus normatif applicable aux administrations.

L'articulation entre le régime du RIA et les obligations du CRPA reste à construire. Les **deux régimes se chevauchent partiellement mais ne se recoupent pas** : le CRPA vise toutes les décisions individuelles algorithmiques (y compris les plus simples), tandis que le RIA cible les systèmes d'IA à haut risque. Une réforme du CRPA intégrant les exigences du règlement 2024/1689 et clarifiant les notions d'intervention humaine et d'explicabilité serait sans doute juridiquement souhaitable.

Algorithme et système d'IA ne se superposent pas⁵⁰.

L'algorithme désigne une suite d'instructions et d'opérations ; le système d'IA est, au sens du règlement (UE) 2024/1689, un système technique qui met en œuvre des procédés algorithmiques, mais ne s'y réduit pas. Ce qui le caractérise, au regard du règlement, est sa capacité à inférer, à partir des données qu'il reçoit, la manière de générer ses sorties. C'est précisément ce qui le distingue des logiciels plus simples fondés uniquement sur des règles définies par des personnes physiques.

La conséquence est qu'en droit français les cadres se croisent sans se confondre. L'article 22 du RGPD et l'article 47 de la LIL visent les décisions prises sur le seul fondement d'un traitement automatisé de données personnelles. L'article L. 311-3-1 du CRPA vise, pour l'action administrative, la décision individuelle prise sur le fondement d'un traitement algorithmique. L'AI Act ajoute à cet ensemble un régime propre aux systèmes d'IA, sans se substituer à ces cadres. Il paraît donc juste de conclure à une coexistence des régimes, non à une superposition des notions.

Un exemple simple permet de l'illustrer : un module purement déterministe qui calcule l'éligibilité ou le montant d'une aide sociale en appliquant un barème fixé à l'avance selon un arbre de décision (revenus, composition du foyer, seuils réglementaires) peut fonder une décision administrative individuelle et relever, à ce titre, du CRPA, sans constituer pour autant un système d'IA au sens du règlement.

Un autre exemple parlant peut être l'algorithme utilisé par la DGFIP pour calculer l'impôt sur le revenu fonde chaque année des millions de décisions administratives individuelles et relève à

⁵⁰ Cette analyse recoupe celle menée, dès 2017 (!), par la CNIL dans « Comment permettre à l'homme de garder la main ? », décembre 2017, pages 16 à 18.

ce titre du CRPA ; mais comme il applique mécaniquement les règles du code général des impôts, sans capacité d'inférence, il n'est pas un système d'IA au sens du règlement.

Trois publications confortent cette analyse :

L'étude annuelle 2022 du Conseil d'État, Intelligence artificielle et action publique : construire la confiance, servir la performance (<https://www.conseil-etat.fr/publications-colloques/etudes/intelligence-artificielle-et-action-publique-construire-la-confiance-servir-la-performance>), qui pose, dans sa partie définitionnelle, qu'« il ne suffit pas de dire qu'une IA est un algorithme ou un ensemble d'algorithmes », un système d'IA pouvant être « une composante simple dans un produit complexe ». La formule cadre exactement la non-réductibilité.

Les lignes directrices de la Commission du 6 février 2025 sur la définition du système d'IA (<https://digital-strategy.ec.europa.eu/fr/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application>), qui isolent sept critères et excluent expressément les systèmes fondés sur des règles établies par des personnes physiques, le « basic data processing » et les tableurs standards.

Le rapport du Défenseur des droits, Algorithmes, systèmes d'IA et services publics (novembre 2024) (<https://www.defenseurdesdroits.fr/les-droits-des-usagers-des-services-publics-face-aux-algorithmes-et-aux-systemes-dia-749>), qui pose, dans ses propos liminaires, que « les règles suivantes s'appliquent, de façon cumulative ou non » : la loi Informatique et Libertés, le CRPA pour les décisions administratives, et « en outre » les obligations de l'AI Act.

Cette co-existence de deux systèmes juridiques est d'autant plus importante à analyser que l'Annexe III du RIA liste les systèmes d'IA à haut risque particulièrement importants pour les pouvoirs publics. Y figurent notamment : les systèmes biométriques utilisés par les autorités répressives, les systèmes d'IA utilisés pour l'accès aux services publics essentiels (prestations sociales, évaluation du crédit), les systèmes utilisés dans la sélection et l'évaluation dans l'éducation, les systèmes de profilage dans le cadre du maintien de l'ordre, et les systèmes d'IA utilisés dans l'administration de la justice.

Pour ces systèmes, le RIA impose des obligations substantielles : système de gestion des risques tout au long du cycle de vie, gouvernance des données, documentation technique exhaustive, journalisation automatique des opérations, transparence et obligation d'information, supervision humaine obligatoire, exactitude et robustesse certifiées. Les déployeurs qui sont des organismes publics doivent en outre réaliser une analyse d'impact sur les droits fondamentaux avant tout déploiement.

Mais la décision n° 2018-765 DC du 12 juin 2018, Loi relative à la protection des données personnelles, qui à son point 71, donne un **support constitutionnel à la garantie prévue par le législateur** selon laquelle le responsable du traitement doit, à tout moment, **être en mesure d'expliquer en détail le fonctionnement de l'algorithme, et par suite, interdit de recourir, comme fondement exclusif d'une décision, à des algorithmes « auto-apprenants »** dont les règles qu'ils appliquent échappent nécessairement à celui qui y a recours, alors qu'il lui

appartient en principe de les contrôler et de les valider sera-t-elle toujours opposable à compter de l'entrée en vigueur de l'annexe III du RIA concernant les SIA « auto-apprenant » ?

Ce verrou constitutionnel passera-t-il « le passage » éventuels des services publics tels que l'enseignement (Parcoursup), des allocations sociales (contentieux de la CNAF) au régime de l'annexe III du RIA ? Mais lesdites activités vont-elles entrer dans le champ d'application du RIA comme SIAHR (système d'intelligence artificielle à haut risque) ? ou vont-elles demeurer sous l'empire combiné du RGPD et du code des relations entre les particuliers et l'administration (CRPA) comme algorithme fondant des décisions individuelles automatisées et soumises, seules, à cette exigence d'explicabilité ?

Cette **dualité de régime juridique** est-elle lisible et compréhensible pour les usagers du service public (et pour les administrations) ?

3.4 Vers un droit à l'explication opposable et gouvernance éthique des algorithmes publics

La Défenseure des droits recommande la consécration d'un droit à l'explication des décisions administratives individuelles partiellement ou entièrement automatisées, en complément du droit à l'information individuelle existant. Ce droit supposerait que l'administration ne se contente pas de communiquer les paramètres abstraits de l'algorithme, mais explique concrètement pourquoi la décision individuelle a été prise dans le sens qu'elle a pris, compte tenu de la situation spécifique de l'intéressé.

Cette évolution s'inscrirait dans la continuité des exigences de motivation des actes administratifs individuels défavorables (loi du 11 juillet 1979) et constituerait une garantie procédurale adaptée à l'ère algorithmique. **Elle appellerait cependant des investissements significatifs en matière d'explicabilité technique des modèles** — particulièrement complexes pour les systèmes d'apprentissage automatique.

L'exemple de la CNAF, qui a institué en mars 2025 un comité d'éthique des usages des données, des algorithmes et de l'IA, constitue-t-il une bonne pratique susceptible d'être généralisée dans le cadre des services publics couverts par l'annexe III ? **Il est intéressant de constater que la présence de comité éthique tant pour Parcoursup que pour la CNAF n'a permis de mettre fin aux critiques et aux soupçons de discriminations.** C'est comme si tous les efforts de transparence « entretiennent » le doute !

3.5 La formation des agents publics : condition sine qua non d'un Etat algorithmique conforme à notre Etat de droit.

Le Conseil économique social et environnemental (ou CESE)⁵¹, la Défenseure des droits et la Cour des comptes s'accordent sur un point : le déploiement de systèmes algorithmiques dans les services publics exige une montée en compétence des agents et de leurs responsables. Il ne s'agit pas de former tous les agents à l'ingénierie des modèles, mais de leur permettre de comprendre le rôle de l'algorithme dans la décision, d'identifier les demandes d'information des usagers y relatives, et d'exercer un regard critique sur les résultats produits.

⁵¹ « Pour une intelligence artificielle au service de l'intérêt général », rapport du CESE de janvier 2025.

La Direction générale à l'administration et à la fonction publique (DGAFP) a lancé en 2024-2025 des travaux en ce sens. Le RIA impose d'ailleurs aux organisations une obligation de « littératie en IA » applicable à leurs personnels (article 4 du RIA ou 20 de la Convention IA⁵²) depuis le 2 février 2025 — ce qui constitue une base légale européenne pour des programmes de formation obligatoires. C'est à une véritable « littératie des algorithmes » qu'il faut construire, littératie définie comme l'aptitude à les comprendre, les utiliser et les critiquer⁵³.

Mais **avons-nous le vivier de formateur pour dispenser une telle formation** adaptée aux besoins de notre économie et de notre administration ?

⁵² Voir l'article 20 de la Convention IA du Conseil de l'Europe (« Maîtrise du numérique et compétences numériques ») qui prévoit que « Chaque Partie encourage et promeut la maîtrise du numérique et les compétences numériques adéquates pour toutes les catégories de la population, notamment les compétences spécifiques de pointe pour les personnes chargées de l'identification, de l'évaluation, de la prévention et de l'atténuation des risques posés par les systèmes d'intelligence artificielle. »

⁵³ Dans ce cadre, il faudrait utiliser et actualiser le guide d'Etalab (direction de la DINUM) intitulé « Expliquer les algorithmes publics », 2 mars 2023.

Conclusion.

L'analyse du cadre juridique et des expériences documentées que nous venons de voir me conduit à proposer les **6 recommandations opérationnelles** suivantes à l'attention des décideurs publics.

	1. Cartographier l'ensemble des traitements algorithmiques utilisés dans chaque administration et assurer leur publication au titre de l'art. L312-1-3 du CRPA, mission qui pourrait être confiée aux administrateurs ministériels des données, algorithmes et codes sources dans chaque ministère ⁵⁴ .
	2. Préciser par voie réglementaire la notion d'« intervention humaine » dans les décisions algorithmiques : définir les critères minimaux (examen effectif du dossier, possibilité de s'écarter du résultat algorithmique, traçabilité de l'intervention).
	3. Prévoir dans tous les marchés publics de systèmes d'IA des clauses obligatoires de transparence et d'accès au code source permettant à l'administration de satisfaire à ses obligations de communication.
	4. Anticiper l'application du RIA en réalisant dès à présent une analyse d'impact sur les droits fondamentaux pour tous les systèmes d'IA à haut risque utilisés par les pouvoirs publics par exemple en assurant la diffusion et la formation à l'utilisation de la méthodologie HUDERIA pour l'évaluation des risques et des impacts des systèmes d'intelligence artificielle du point de vue des droits humains, de la démocratie et de l'Etat de droit, approuvée par le Comité des ministres du Conseil de l'Europe le 25 février 2026.
	5. Déployer un programme de formation obligatoire à la littératie algorithmique pour les agents publics — conformément à l'obligation du RIA entrée en vigueur le 2 février 2025 (article 4 du RIA).
	6. Réfléchir aux modalités du suivi éthique « efficaces » d'accompagnement des algorithmes et des SIAHR dans l'administration,

⁵⁴ Circulaire du Premier ministre du 27 avril 2021, n°6264/SG, relative à la politique publique de la donnée, des algorithmes et des codes sources.

Mr Thomas SCHNEIDER

[Contribution to be added]