Athens Journal of Law

Quarterly Academic Periodical,
Volume 12, Issue 3, July 2026
URL: https://www.athensjournals.gr/ajl
Email: journals@atiner.gr
e-ISSN: 2407-9685 DOI: 10.30958/ajl

Front Pages

Athens Journal of Law

Published by the Athens Institute

****

***

Athens Institute is an Athens-based World Association of Academics and Researchers based in Athens. Athens Institute is an independent and non-profit Association with a Mission to become a forum where Academics and Researchers from all over the world can meet in Athens, exchange ideas on their research and discuss future developments in their disciplines, as well as engage with professionals from other fields. Athens was chosen because of its long history of academic gatherings, which go back thousands of years to Plato's Academy and Aristotle's Lyceum. Both these historic places are within walking distance from Athens Institute's downtown offices. Since antiquity, Athens was an open city. In the words of Pericles, Athens"…is open to the world, we never expel a foreigner from learning or seeing". ("Pericles' Funeral Oration", in Thucydides, The History of the Peloponnesian War). It is ATINER's mission to revive the glory of Ancient Athens by inviting the World Academic Community to the city, to learn from each other in an environment of freedom and respect for other people's opinions and beliefs. After all, the free expression of one's opinion formed the basis for the development of democracy, and Athens was its cradle. As it turned out, the Golden Age of Athens was in fact, the Golden Age of the Western Civilization. Education and (Re)searching for the 'truth' are the pillars of any free (democratic) society. This is the reason why Education and Research are the two core words in Athens Institute's name.

****

The ***Athens Journal of Law (AJL)*** is an Open Access quarterly double-blind peer reviewed journal and considers papers from all areas of law. Many of the papers published in this journal have been presented at the various conferences sponsored by the Business, Economics and Law Division of the Athens Institute. All papers are subject to Athens Institute's Publication Ethical Policy and Statement.

The Athens Journal of Law
ISSN NUMBER: 2407-9685 - DOI: 10.30958/ajl
Volume 12, Issue 3, July 2026
Download the entire issue (PDF)

Athens Journal of Law
Editorial and Reviewers' Board

President's Message

All Athens Institute's publications including its e-journals are open access without any costs (submission, processing, publishing, open access paid by authors, open access paid by readers etc.) and is independent of presentations at any of the many small events (conferences, symposiums, forums, colloquiums, courses, roundtable discussions) organized by Athens Institute throughout the year and entail significant costs of participating. The intellectual property rights of the submitting papers remain with the author. Before you submit, please make sure your paper meets the basic academic standards, which includes proper English. Some articles will be selected from the numerous papers that have been presented at the various annual international academic conferences organized by the different divisions and units of the Athens Institute for Education and Research. The plethora of papers presented every year will enable the editorial board of each journal to select the best, and in so doing produce a top-quality academic journal. In addition to papers presented, Athens Institute will encourage the independent submission of papers to be evaluated for publication.

The current issue is the third of the twelfth volume of the *Athens Journal of Law (AJL)*, published by the *Business and Law Division of Athens Institute*.

Gregory T. Papanikos
President
Athens Institute

Social and Educational Program

The Social Program Emphasizes the Educational Aspect of the Academic Meetings of Atiner.

- Greek Night Entertainment (This is the official dinner of the conference)
- Athens Sightseeing: Old and New-An Educational Urban Walk
- Social Dinner
- Mycenae Visit
- Exploration of the Aegean Islands
- Delphi Visit
- Ancient Corinth and Cape Sounion
    - More information can be found here: https://www.atiner.gr/social-program

Conference Fees

Conference fees vary from 400€ to 2000€
Details can be found at: https://www.atiner.gr/fees

Athens Institute for Education and Research

A World Association of Academics and Researchers

14th Annual International Conference on Business, Law & Economics
3-8 May 2027, Athens, Greece

The Business, Economics and Law Division (BLRD) of Athens Institute is organizing its **14th Annual International Conference on Business, Law & Economics, 3-8 May 2027, Athens, Greece,** sponsored by the Athens Journal of Business & Economics and the Athens Journal of Law. In the past, the six units of BLRD have organized more than 50 annual international conferences on accounting, finance, management, marketing, law and economics. This annual international conference offers an opportunity for cross disciplinary presentations on all aspects of business, law and economics. This annual international conference offers an opportunity for cross disciplinary presentations on all aspects of business, law and economics. Please submit an abstract (email only) to: atiner@atiner.gr, using the abstract submission form (https://www.atiner.gr/2027/FORM-BLE.doc)

Important Dates

- Abstract Submission: **6 October 2026**
- Acceptance of Abstract: 4 Weeks after Submission
- Submission of Paper: **5 April 2027**

Academic Member Responsible for the Conference

- **Dr. Gregory T. Papanikos, President, Athens Institute.**
- **Dr. Chris Sakellariou, Vice President of Finance, Athens Institute & Associate Professor of Economics (Retired), Nanyang Technological University, Singapore.**
- **Dr. Anica Hunjet, Deputy Director, Business and Law Division, Athens Institute & Vice Rector, University North, Croatia.**
- **Dr. Jorge Emilio Núñez, Head, Law Unit, Athens Institute & Reader, Manchester Metropolitan University, UK.**

Social and Educational Program

The Social Program Emphasizes the Educational Aspect of the Academic Meetings of Atiner.

- Greek Night Entertainment (This is the official dinner of the conference)
- Athens Sightseeing: Old and New-An Educational Urban Walk
- Social Dinner
- Mycenae Visit
- Exploration of the Aegean Islands
- Delphi Visit
- Ancient Corinth and Cape Sounion

More information can be found here: https://www.atiner.gr/social-program

Conference Fees

Conference fees vary from 400€ to 2000€
Details can be found at: https://www.atiner.gr/fees

Athens Journal of Law - Volume 12, Issue 3, July 2026 – Pages 149-168

Regulatory Impact Analysis in Brazilian Health Surveillance: An Examination of Anvisa's Methodological Practices

*By Juliano Heinen**

This article analyses the decision-making architecture of the National Health Surveillance Agency (ANVISA) from the perspective of the OECD's Cost-Benefit Analysis (CBA) guidelines. Through a systematic documentary analysis of the Regulatory Impact Analysis Reports (RIA) produced between 2022 and 2025, it is demonstrated that the agency operates under a technical-regulatory rationale that favours procedural and qualitative metrics over the economic valuation of social welfare. The systematic absence of indicators such as Net Present Value (NPV), Social Internal Rate of Return (SIRR), and the monetisation of externalities reveals a structural divergence between institutional practice and the scientific standard of allocative efficiency advocated nationally and internationally. It is argued that "scientific uncertainty," repeatedly invoked as a methodological barrier, is treated by the agency as a justification for quantitative inertia, when it could be the subject of stochastic modelling and rigorous sensitivity analysis. The article concludes that ANVISA could broaden its analytical criteria to incorporate the cost-benefit analysis provided for in Article 7, II, of Decree No. 10,411/2020, bringing it closer to the regulatory evidence paradigm required by multilateral organisations and cutting-edge specialist literature.

Keywords: *Regulatory Impact Analysis; ANVISA; cost-benefit analysis; allocative efficiency; health regulation; OECD.*

Introduction

Health regulation occupies a unique position in economic regulation theory. Unlike other regulated sectors, where risks are predominantly concentrated in the economic sphere, the regulation of health products and services deals with potentially catastrophic externalities, involving risks to life, physical integrity and systemic confidence in the markets for medicines and medical technologies. For this reason, the law and economics literature often identifies health regulation as one of the paradigmatic cases in which state intervention finds strong theoretical justification, especially in the face of market failures related to information asymmetry, scientific uncertainty, and negative externalities.

Paradoxically, it is in this very field that the costs of regulation tend to be less visible and more difficult to challenge politically. While regulatory benefits are often presented as diffuse and projected into the future–in the form of avoided risks or prevented harm– regulatory costs manifest themselves immediately and in a concentrated manner, affecting specific economic agents and, often, the very

*Professor in the Master's and Doctoral Program at the Pontifical Catholic University of Rio Grande do Sul., Brazil.

dynamics of technological innovation. This temporal and distributive misalignment between costs and benefits generates institutional incentives for regulatory expansion without a corresponding systematic assessment of its economic efficiency, a phenomenon identified in the literature as "pro-regulation bias."

It is precisely in this context that Regulatory Impact Analysis (RIA) emerges as an institutional instrument designed to discipline regulatory decision-making. As widely disseminated by the international guidelines of the Organisation for Economic Co-operation and Development (OECD), analysed below, RIA seeks to introduce analytical rationality, decision-making transparency and technical *accountability* into the regulatory process, mainly through the application of structured methods for assessing the social costs and benefits of regulatory alternatives. At the heart of this methodological toolkit is Cost-Benefit Analysis (CBA), whose purpose is to systematically compare the expected impacts of regulatory options, converting them, whenever possible, into comparable monetary metrics discounted to present value.

In Brazil, the formal institutionalisation of RIA was consolidated with Decree No. 10,411/2020, which regulated its preparation within the scope of federal regulatory agencies, implementing the provisions of Article 6 of Law No. 13,848/2019 (Regulatory Agencies Law) and Article 5 of Law No. 13,874/2019 (Economic Freedom Law). This regulatory framework established procedures, criteria, and minimum steps for the preparation of RIA reports, including the identification of the regulatory problem, the analysis of alternatives, and the assessment of expected impacts.

Among Brazilian regulatory agencies, the National Health Surveillance Agency (ANVISA) is a particularly relevant case for empirical analysis. Due to the breadth and sensitivity of its field of activity – which involves medicines, food, medical devices, and health technologies – ANVISA produces extensive RIA reports that are technically structured and formally compliant with current regulatory requirements. However, a fundamental question remains: to what extent do these reports effectively incorporate the methodological standards of economic efficiency found in the international literature on regulatory analysis?

This article argues that there is a structural deficit of economic efficiency in ANVISA's HIA practice. In analytical terms, it is argued that the agency shows a high degree of formal compliance with the procedural requirements of RIA, but systematically avoids applying its most demanding methodological component: the monetary quantification of the social costs and benefits of regulatory alternatives. This methodological gap compromises RIA's ability to fulfil its central function of disciplining regulatory decision-making through explicit efficiency criteria.

The analysis suggests that this institutional resistance to economic quantification is often justified by two recurring narratives. The first invokes scientific uncertainty, arguing that the absence of robust evidence would make reliable measurement of regulatory impacts unfeasible. The second maintains the moral incommensurability of public health, according to which attempting to monetise benefits related to the protection of health or human life would be normatively inappropriate. Although these objections have philosophical and methodological relevance, the international regulatory analysis literature has developed tools precisely to address these difficulties, such as methods of statistical value of life, sensitivity analysis, and scenario evaluation.

Given this context, the objective of this article is to critically examine the extent to which the RIA reports produced by ANVISA incorporate the essential methodological elements of cost-benefit analysis as established by OECD guidelines and contemporary literature. In doing so, the study seeks to contribute to the debate on the institutional quality of regulatory governance in Brazil, offering an empirical assessment of RIA practice in one of the most sensitive sectors of regulatory administration.

Against this backdrop, three research questions guide the present inquiry. First, to what extent do ANVISA's RIA reports incorporate the essential elements of cost-benefit analysis as defined by OECD guidelines and contemporary regulatory scholarship? Second, what methodological approaches predominate in these reports, and how does the agency justify its methodological choices? Third, what institutional, epistemic, or normative factors may account for the systematic divergence observed between the agency's formal compliance with RIA procedural requirements and the substantive application of economic valuation methods?

To achieve this objective, the article is structured as follows. Section 1 presents the theoretical and dogmatic foundations of Regulatory Impact Analysis (RIA) and Regulatory Outcome Assessment (ROA), defining their concepts and functions in the regulatory cycle. Section 2 develops the theoretical framework of the research, discussing the fundamentals of cost-benefit analysis in light of OECD guidelines and the law and economics literature. Section 3 describes the documentary research methodology adopted to examine ANVISA's RIA reports. Section 4 presents the results of the empirical analysis, organised into three critical axes. Finally, Section 5 discusses the institutional implications of the findings and proposes ways to improve the AIR methodology in the context of Brazilian health regulation.

Theoretical and Dogmatic Framework of RIA

Regulatory Impact Analysis (RIA) was already being practised in Brazil by some regulatory agencies, which stipulated the need to prepare this document in their internal regulations or in the administrative rules for sectoral standardisation. An example of this is the Internal Regulations of the ANS (National Health Agency), according to Normative Resolution No. 197/2009 and Ordinance No. 354/2006. ANATEL, in Article 62, sole paragraph, of its Internal Regulations (Resolution No. 61212013) determined that the RIA should be prepared prior to the issuance of regulatory acts. Similarly, ANEEL (National Electric Energy Agency) issued, at the time, Organisation Standard No. 40/2013 (Article 2), which requires the use of AIR in its regulatory procedures. ANCINE (National Film Agency) has long been formulating AIRs in the process of standardising the sector. ANVISA (National Health Surveillance Agency), depending on the subject or complexity of the rule to be issued, provided for Regulatory Impact Analysis on three levels (*e.g.* AIR level 1, 2 or 3). For example, level 3 would be applied to high-impact regulatory proposals involving high costs for those affected by the rule or a large budgetary contribution (Palma, 2014).

With the enactment of Law No. 13,848/19 (*General Law on Regulatory Agencies*), the adoption and proposed amendments to normative acts of general interest to economic agents, consumers or users of the services provided shall,

under the terms of the regulation, be preceded by a Regulatory Impact Analysis (RIA), which shall contain information and data on the possible effects of the normative act (Art. 6). Even before the enactment of this legislation, the *Economic Freedom Law* (Rule No. 13,874/19) already determined that proposals for the enactment and amendment of normative acts of general interest to economic agents or users of the services provided, enacted by a federal public administration body or entity, including autonomous agencies and public foundations, shall be preceded by a regulatory impact analysis (Article 5). This document must contain information and data on the possible effects of the normative act in order to verify the proportionality (*i.e.,* reasonableness) of its economic impact.

In other words, the RIA is nothing more than an administrative procedure that aims to analyse the burdens and benefits of regulating a given sector, establishing a rational decision on the subject, revealing, to the maximum extent possible, the externalities of regulating a given subject or sector. Thus, this administrative procedure becomes mandatory when considering regulation by such independent agencies. And, in cases where the AIR is not carried out, "[...] at least a technical note or equivalent document that has substantiated the proposed decision must be made available." (Paragraph 5 of Article 6 of Rule No. 13,848/19).

The institutional origins of RIA trace back to the United States, where the Administrative Procedure Act of 1946 first required agencies to state the reasons for their regulatory choices and to hear affected parties. Concerns about regulatory inflation subsequently prompted President Gerald Ford, in 1974, to issue Executive Order No. 11,821, mandating inflationary impact assessments of new regulations. In 1981, President Ronald Reagan substantially expanded this requirement through Executive Order No. 12,291, which made cost-benefit analysis the central evaluative criterion for significant regulations and established the Office of Information and Regulatory Affairs (OIRA) as the central oversight body (Radaelli & Francesco, 2020; Papanicolas, Woskie & Jha, 2018). In the United Kingdom, impact analysis emerged from a distinct political context. The mid-1980s deregulation agenda sought primarily to reduce compliance burdens on businesses, giving rise to the Cost Compliance Assessment (CCA) method – a tool designed to estimate the direct "compliance costs" imposed by proposed rules on regulated entities (Aragão, 2010).

To this end, the results can be obtained from a series of metrics. Many of these are disseminated *by the OECD (Organisation for Economic Co-operation and Development*) to to conduct an analysis of regulatory impacts (OECD, 2008). To this end, it is suggested that the problem be defined first and that its resolution can be justified by state action. It is clear that administrative legality must allow for state action. In this regard, the RIA should demonstrate that the costs of regulation are lower than its benefits. This assessment must be transparent, clear and accessible. It is therefore extremely important to allow everyone to participate in this process.

In addition, the aforementioned document should be responsible for projecting the effects of regulation, especially in terms of national development, protection of the vulnerable, harm or benefit to competition, preservation of fundamental rights, etc. With regard to the OECD guidelines, the organisation recommends the integration of RIA from the early stages of policy formulation (OECD, 2012). RIA is defined as a tool that provides decision-makers with valuable empirical data (OECD, 2007). Without this

comprehensive framework, government action may suffer from flaws due to a poor understanding of the problem (OECD, 2007).

The OECD's recommendations for Brazil suggested standardising RIA to strengthen governance for growth (OECD, 2007). The organisation proposes that the tool clearly define problems and identify the *trade-offs* of different approaches (OECD, 2012). In Brazil, this inspired the PRO-REG Programme, aimed at institutional capacity building.

An important institutional guideline is the creation of a central regulatory quality oversight body (OECD, 2008). This body should oversee the use of RIA in other institutions and provide technical expertise (OECD, 2008). The OECD emphasises that implementation requires a change in administrative culture, often through experimental phases (OECD, 2008).

Public participation is an emphatic guideline of the OECD and should be systematically incorporated through consultations and hearings (OECD, 2015). Stakeholder engagement gathers information dispersed throughout society that the government is unaware of. This ensures greater adherence by those regulated and democratic legitimacy for the standard (MENDONÇA, 2014).

Finally, the OECD recommends conducting periodic "retrospective reviews" of existing regulations (OECD, 2015). This process allows for the identification of obsolete rules or those that negatively impact the economy, enabling their repeal (OECD, 2015). This practice ensures that the regulatory system remains efficient and less costly over time (OECD, 2015).

Thus, when AIR weighs the various interests and considers the possibilities of being effective, it tends to significantly reduce the judicialisation of the regulation under analysis. This proactive and preventive action can be very valuable in achieving legal certainty for the regulated sector. In the United States, there are a series of *Executive Orders* that regulate the RIA procedure (*e.g.,* No. 12,886/1993) and how regulation should be carried out (No. 13,777/2017 and 13,771/2017). These are binding documents, including for independent regulatory authorities.

In other words, AIR allows for concerted or dialogical regulation, insofar as the State builds the normative policy in question together with the affected sectors or with citizens. Or, at the very least, it allows any individual to control the determining factors set out in the impact analysis. The question that always remains is: to what extent does the population have real and effective power to interfere in decision-making processes related to the provision of public services? We will answer this question in the following topic.

Cost-benefit Analysis (CBA) as a Scientific Standard for Regulation

Fundamentals of Cost-Benefit Analysis and OECD Guidelines

Cost-benefit analysis (CBA) has its roots in welfare economics and, more specifically, in the Kaldor-Hicks criterion, according to which a policy is efficient if the gains of the winners are large enough to compensate, at least hypothetically, for the losses of the losers (EPA, 2010). The pillars of CBA rest on the principles of microeconomics: maximisation, equilibrium and efficiency (Cooter & Ulen, 2012). after

all, it is assumed that individuals are rational maximisers of their satisfaction, making choices that minimise costs and increase benefits (Posner, 2014).

In the public sphere, the document aims to highlight the matrix for the allocation of scarce resources, with a view to minimising social costs (Cooter & Ulen, 2012). Therefore, the ethical basis of the technique seeks to overcome the rigidity of the Superior Pareto criterion, which requires that no one be harmed by a change in *the status quo* (Adler & Posner, 2006). In its place, the Kaldor-Hicks criterion, mentioned above, or potential Pareto test, is adopted, because this metric considers that a decision is efficient if the beneficiaries can, theoretically, compensate those who are harmed, generating a net gain in social welfare (Adler & Posner, 2006).

In legal terms, Richard Posner places CBA in the context of legal pragmatism, where wealth maximisation is seen as a goal to be achieved by law (Posner, 1993). Thus, this pragmatism manages to prioritise the empirical method and consideration of the systemic consequences of decisions (Mendonça, 2014), with economic rationality serving as a criterion for assessing the reasonableness of state interventions (Posner, 2010).

A crucial foundation of CBA is its ability to circumvent the harmful effects of cognitive biases, such as biases and heuristics (Kahneman, 2012). Due to these distortions, public decisions may focus on marginal problems, ignoring more serious and risks (Sunstein, 2002), which is crucial to the regulatory issues addressed by ANVISA. Note that quantifying impacts avoids the so-called "simultaneous system of paranoia and negligence" in public management (Sunstein, 2002), a pathology in the public decision-making process characterised by a disproportionate state reaction to certain risks to the detriment of others. This phenomenon occurs when public management is influenced by cognitive biases, such as biases and heuristics, which distort the perception of the probabilities and consequences of certain events.

Additionally, CBA is justified by the strengthening of the democratic regime and transparency (Sunstein, 2002), because it allows for the accountability of government officials by requiring that the assumptions behind decisions be made explicit (Sachs, 2015). In this way, the population can participate in the evaluation process through a well- informed deliberative democracy (Sunstein, 2002).

As for the essential requirements, the procedure requires an "explicit evaluation," overcoming merely implicit convictions of managers (Sen, 2004), so that it is imperative to clearly identify the objective of state action and alternative options, including the cost of inertia. This rationalist transparency allows society to understand the reasons behind the government's choice (Sen, 2004).

Identifying the ownership of affected individuals is another critical procedural requirement (Beck, 2015). The asymmetry between those who make the decision and those who bear the risks must be considered (Beck, 2015). This mapping is essential to assess distributional impacts and ensure that the policy does not disproportionately burden certain segments of the population – an issue related to vaccination policies, drug dispensing, form of dispensing, etc.

Another essential item, perhaps the most central to the institute, is the detailed cataloguing of impacts, which should include a *risk trade-off analysis* (Graham & Wiener, 1995). After all, it is recognised that minimising one risk can unintentionally exacerbate another secondary risk. For example, banning a toxic substance may lead to the use of a substitute whose effects are still unknown or worse (Revesz & Livermore, 2008).

Monetisation is the distinctive feature of CBA, requiring the conversion of benefits and costs to a common monetary metric scale (Adler & Posner, 2006). Note that when market prices do not exist, methodologies such as "revealed preferences" or "stated preferences" (contingent valuation) are used. These techniques estimate how much society is willing to pay for risk reduction (Adler, 2012), which is very relevant to ANVISA's AIRs.

Another central requirement is the establishment of the "statistical value of life" (VSL), as diagnosed by Sunstein (2014, p. 12): the calculation is based on evidence of individual choices involving trade-offs between risk and money (Viscusi & Aldy, 2003). Although ethically debated, VSL is necessary to prioritise health and safety policies in an impersonal and rational manner (Sunstein, 2014). For impacts that occur at different times, it is essential to use a "social discount rate" (Boardman et al., 2006), bearing in mind that it converts future values into present values, reflecting society's time preference (Cole, 2012). The choice of rate must balance economic efficiency with the moral obligation to future generations.

But that is not all: the qualitative assessment of unquantifiable goods, such as dignity and justice, is part of the requirements of the procedure (Sunstein, 2014). In such cases, *breakeven analysis* helps to identify the minimum value of these goods necessary to justify the costs. This prevents subjective and necessarily protected values from being ignored simply because of the lack of a monetary metric (Sunstein, 2014).

The process ends with a "sensitivity analysis", which investigates the degree of uncertainty in the estimates (Boardman et al., 2006). Here, alternative variables should be tested to ensure that the result is not the result of arbitrary assumptions (Boardman et al., 2006). Finally, the final result is consolidated in a Regulatory Impact Analysis Report (RIA).

The OECD, in its reference document *Cost-Benefit Analysis and the Environment: Further Developments and Policy Use* (2018), establishes that a methodologically robust CBA should include: (i) a precise definition of the counterfactual scenario (*baseline*); (ii) the identification and valuation of all relevant externalities, including non-market ones; (iii) the use of the Value of a Statistical Life (VSL) to monetise impacts on mortality; (iv) the application of a Social Discount Rate for calculating the NPV; and (v) the performance of Sensitivity Analysis to test the robustness of estimates in the face of parametric uncertainty (OECD, 2018). These requirements are not mere academic technocracy. They respond to a fundamental democratic concern: if citizens cannot compare, in common terms, how much a regulation costs them versus how much it benefits them, political deliberation on regulatory choices loses its intersubjective rationality. Monetisation is not a mercantile reduction of human values – it is, on the contrary, a requirement for transparency about the *trade-offs* that any regulation inevitably imposes.

When RIA obscures this analysis, there is a choice not to account for the benefits of protective regulations, which paradoxically weakens the very justification for regulatory intervention. The literature on *behavioural regulation* also shows that regulations that avoid monetisation tend to create a *pro-status quo* or pro-regulation bias, depending on how the alternatives are framed (Thaler & Sunstein, 2008). In the context of ANVISA, this bias manifests itself in the asymmetry between the meticulous measurement of internal administrative costs – via the Standard Cost Model – and the complete absence

of valuation of the corresponding social benefits.

It is important to acknowledge, however, that cost-benefit analysis is not without its critics. Scholars in legal theory and ethics have raised legitimate concerns about the use of the statistical value of life, noting that monetary proxies for human wellbeing inevitably rest on normative assumptions that are themselves contestable (Adler & Posner, 2006, p. 78). Distributional effects pose a further challenge: standard CBA aggregates gains and losses across individuals without regard to who bears the costs and who receives the benefits, potentially obscuring regressive impacts (Revesz & Livermore, 2008). These limitations do not undermine the case for CBA, because they do, however, counsel in favour of complementing monetary analysis with distributional impact assessments and explicit sensitivity testing of key normative assumptions.

It is equally necessary to situate ANVISA's methodological choices within their institutional context. The agency operates under a legal framework that explicitly assigns priority to the Precautionary Principle, and its technical staff may face genuine constraints regarding access to reliable epidemiological data, econometric training, and dedicated resources for economic modelling. The gap documented in this study may therefore reflect, at least in part, structural limitations of capacity rather than deliberate resistance to quantitative analysis. Addressing this gap will require not only methodological guidance but also institutional investments in data infrastructure and human capital.

Precaution and Efficiency: False Antinomy

Although ANVISA has consolidated Regulatory Impact Analysis (RIA) as a pillar of its governance, there is a methodological gap in relation to OECD guidelines. The agency often adopts a qualitative or multi-criteria analysis approach which, although legally robust, could be further improved by translating health risks into comparable economic variables.

Note that this is the logic of the OECD: regulation must be justified by a Cost-Benefit Analysis (CBA) that minimises "precautionary paralysis". At ANVISA, the Precautionary Principle is often interpreted in a "strong" way, in the sense of Sunstein (2005), where scientific uncertainty justifies prohibition or severe restriction without proper quantification of opportunity costs or trade-offs.

In other words, the proposed integration – that uncertainty is not the opposite of efficiency, but its object – challenges the bureaucratic culture of Brazilian health regulation. For ANVISA to achieve the standard of economic efficiency expected in global markets, the transition from qualitative to stochastic is imperative. If ANVISA adopted CBA in accordance with OECD standards, the discipline of comparison would further reduce the uncertainties of the effects of regulation, because the use of mathematical models allows the regulator to decide not only "whether" to regulate, but also "how much" the adopted safety margin costs.

The operationalisation of this transition relies on statistical simulation techniques. Rather than yielding a single deterministic figure, stochastic modelling allows the Net Present Value (NPV) of a health regulation to be expressed as a probability distribution, converting scientific uncertainty into confidence intervals. Uncertainty in key parameters (such as the magnitude of morbidity reductions on the benefit side, or

the cost of regulatory compliance) is thus rendered analytically tractable, rather than treated as a justification for abandoning quantification altogether.

The adoption of this technical rigour by ANVISA would make it possible to overcome the false dichotomy between public health and economics, considering that, by using the OECD standard, the agency would not be "monetising life", but rather ensuring that society's scarce resources are allocated where they generate the greatest net social gain, i.e., save more lives. But that is not all: with balance, it is possible to prevent fear of the unknown from generating regulations that, in attempting to protect, end up stifling innovation and access to new health technologies.

Data Analysis Methodology

Nature of the Research

This research adopts a qualitative approach of an interpretative nature, anchored in the method of critical documentary analysis (Cellard, 1997; Sá-Silva, Almeida & Guindani, 2009). Between 2022 and 2025, ANVISA produced twenty- five AIR reports, which is the total production made available by the agency's portal during that period (BraziL, 2026). No progress was made in previous years because the context of the pandemic caused by the Covid-19 virus naturally altered the Agency's mode of operation, and in many cases, a CBA could not even be required.

The analytical procedure was structured around a seven-item methodological checklist applied uniformly to each report: (i) explicit definition of the regulatory problem; (ii) identification and comparative assessment of regulatory alternatives; (iii) application of cost-benefit analysis or explicit monetisation of impacts; (iv) use of quantitative efficiency indicators such as Net Present Value (NPV) or Social Internal Rate of Return (SIRR); (v) use of the statistical value of life (VSL) or equivalent willingness-to-pay estimates; (vi) performance of sensitivity or scenario analysis; and (vii) use of outcome indicators (e.g., variations in morbidity, mortality, or market prices) rather than purely process-based outputs. For each item, presence or absence was coded as a binary variable, and the overall methodological profile of each report was mapped accordingly.

The replicability of this analysis is assured by the public availability of all twenty-five RIA reports on ANVISA's official portal (Brazil, 2026). Consistent application of the checklist across reports of varying thematic scope required interpretive judgements in borderline cases, which are acknowledged as a limitation of single-coder documentary research. Where ambiguity arose regarding the classification of a given methodological element, the more conservative interpretation (recording absence rather than partial presence) was adopted, so as not to overstate the degree of CBA incorporation.

The sample of 25 reports is distributed as follows, considering their completion or official publication dates:

2022 (4 reports)

1. Donation of Food with Health Safety (April 2022).
2. Food and Packaging Regularisation (May 2022).

3. Electronic Smoking Devices (ESDs) (June 2022).
4. Modernisation of the Framework for Novel Foods and Ingredients (2022).

2023 (6 reports)
1. Control of Degradation Products in Medicines (March 2023).
2. Administrative Resources (General Resource Management - GGREC) (July 2023).
3. Cosmetovigilance (July 2023).
4. Risk Classification for Economic Activities (August 2023).
5. Health Control of Travellers (September 2023).
6. Cannabis Products for Medicinal Purposes (November 2023).

2024 (10 reports)
1. Health Control of Aircraft and Airports (2024).
2. Clinical Research with Medicines and Biological Products (March 2024).
3. Organisation of the SNVS and Decentralisation (RDC 560/2021) (April 2024).
4. Prosecution and Judgement of Health Violations (PAS) (May 2024).
5. Imported Goods and Products (RDC 81/2008) (May 2024).
6. Designation and Selection of Reference Medicines (2024).
7. Regularisation of Companies in Ports, Airports and Borders (AFE/BPA) (2024).
8. Health Control of Ports and Vessels (2024).
9. Prioritisation and Special Procedure for Medicinal Product Analysis (November 2024).
10. Health Requirements for Dental Care (November 2024).

2025 (5 reports)
1. Prices of New Products (CMED) (March 2025).
2. Handcrafted Cosmetics (August 2025).
3. Good Manufacturing Practices (GMP) for Food (2025).
4. Infection Prevention and Control (IRAS) (November 2025).
5. Fractionation of Hygiene and Cosmetic Products (December 2025).

Table 1. *Methodological Characteristics of ANVISA's RIA Reports (2022–2025)*

RIA Report (Year)	CBA/Mon.	Quant. Ind.	Sens. Analysis
Food Donation / Safety (2022)	No	No	No
Food & Packaging (2022)	No	No	No
Electronic Smoking Devices (2022)	No	No	Partial
Novel Foods & Ingredients (2022)	No	No	No
Degradation Products / Medicines (2023)	No	No	No
Administrative Resources GGREC (2023)	No	No	No
Cosmetovigilance (2023)	No	No	No
Risk Classification / Econ. Activities (2023)	No	No	No
Health Control of Travellers (2023)	No	No	No
Cannabis / Medicinal Purposes (2023)	No	No	No
Health Control of Aircraft (2024)	No	No	No
Clinical Research / Medicines (2024)	No	No	No
SNVS Organisation / Decentralisation (2024)	No	No	No
Health Violations / PAS (2024)	No	No	Partial
Imported Goods / RDC 81 (2024)	No	No	No
Reference Medicines (2024)	No	No	No
Companies in Ports / Airports (2024)	No	No	No
Health Control of Ports & Vessels (2024)	No	No	No
Priority Procedure / Medicines (2024)	No	No	No
Dental Care Requirements (2024)	No	No	No
Prices of New Products / CMED (2025)	No	No	No
Handcrafted Cosmetics (2025)	No	No	No

GMP for Food (2025)	No	No	No
Infection Prevention/IRAS (2025)	No	No	No
Fractionation / Hygiene Products (2025)	No	No	No

Note: CBA/Mon. = cost-benefit analysis or explicit monetisation; Quant. Ind. = quantitative efficiency indicators (NPV, SIRR); Sens. = sensitivity or scenario analysis. "Partial" indicates limited qualitative scenario discussion without full stochastic modelling.

Through the analysis of twenty-five Regulatory Impact Analysis (RIA) Reports completed between 2022 and 2025, the study identifies a methodological hegemony of Multicriteria Analysis (MCA) and the systematic neglect of economic valuation. It is argued that the agency uses the "complexity of the issue" and the "unavailability of data" as rhetorical shields to avoid the monetisation of externalities, distancing itself from the OECD guidelines and Decree No. 10,411/2020.

Results obtained from the Analysis of the AIRs surveyed during the Period

Contemporary regulatory governance requires that state intervention be justified not only by technical imperatives, but also by a full demonstration of its allocative efficiency. In Brazil, Decree No. 10,411/2020 established the AIR as the gold standard for this accountability (ANVISA, 2024). However, systematic analysis of ANVISA's documentary *corpus* reveals a different reality. The pattern identified is the almost universal replacement of Cost- Benefit Analysis (CBA) by Multi-Criteria Analysis (MCA), generally operationalised by the *Analytic Hierarchy Process* (AHP) method (ANVISA, 2025). Although AHP is a valid tool for dealing with subjective criteria, the agency uses it as an absolute substitute for economic valuation, rather than as a complement to it (ANVISA, 2025). In the AIR Report on *Cannabis* products, the agency justifies this choice by stating that quantitative methodologies would bring a "reductionist view of a complex issue," focusing on economic aspects that "would not adequately reflect reality" (ANVISA, 2024). This ideological resistance to monetisation ignores the fact that CBA is, by definition, the instrument that allows the marginal social utility of different public policies to be compared.

The absence of indicators such as Net Present Value (NPV) and Social Internal Rate of Return (SIRR) is not an accidental technical flaw, but a recurring methodological choice. In the report on infection control (IRAS), it is noted that the criterion of "health safety" received an overwhelming weight of 65.99%, while "costs to the regulated sector" were weighted at only 4.39% (ANVISA, 2025). This disproportion demonstrates that economic impact is treated as a mere procedural accessory, rather than an essential component of social welfare. The agency admits that the monetisation of health benefits is "highly complex," opting for descriptive qualitative comparison (ANVISA, 2024). The incorporation of the method would reveal whether the billion-dollar investment required for infection control is higher or lower than the social cost of antimicrobial resistance, estimated at 100 billion dollars globally (ANVISA, 2025). Without CBA, the agency decides in the "monetary dark," without knowing whether the marginal weight assigned to safety justifies the

increased cost of health services.

In the AIR Report on *Cannabis* products, ANVISA opted for a qualitative approach, arguing that monetisation would bring a "[...] reductionist view (...)" (ANVISA, 2024). However, it is understood that the method could reveal the cost of judicialisation to the SUS compared to the cost of more permissive regulation. By monetising patient waiting times and legal costs, CBA would demonstrate that the agency's qualitative "prudence" has a real and high price for the treasury and families.

The argument of "unavailability of data" is systematically invoked to avoid stochastic modelling. In the report on the review of prices for new products, the agency claims that "[...] no data were available that would allow the application of quantitative methodologies (...)" (ANVISA, 2025). Even in areas of post-market monitoring, such as Cosmetovigilance, the agency explicitly states that "[...] it is not appropriate to quantify and/or monetise the benefits (...)" due to uncertainty about future prices and behavioural responses (ANVISA, 2023). This position is corroborated in the report on clinical research, which states that monetisation would make the analysis "[...] even more complex (...)" (ANVISA, 2024). The concrete benefit of migrating to CBA would be transparency regarding the social *trade-off*: society would have clarity on whether the gain in health security justifies, for example, the loss of competitiveness or the increased cost of access to new technologies.

In summary, ANVISA operates under a technical-regulatory rationale that favours process over allocative efficiency. To align itself with Article 7, II, of Decree No. 10,411/2020 and the OECD evidence paradigm, the agency must transcend the purely qualitative use of AMC. The inclusion of CBA would allow scientific uncertainty to be addressed through statistical modelling rather than analytical inertia. As Sunstein (2014) aptly summarises, regulation that avoids quantification is not more prudent, it is merely less transparent. Health regulation cannot be an end in itself; it must demonstrate, with verifiable evidence, that the value it generates for society outweighs the burden it imposes on the economy and innovation.

Table 2. *Synthetic Comparison of the Three Critical Analytical Axes identified in ANVISA's RIA Corpus (2022–2025)*

Critical Axis	Pattern Observed	Illustrative Reports
Scientific uncertainty as barrier	Data unavailability or epistemic complexity invoked to justify non-monetisation; stochastic alternatives not considered.	CMED Price Review (2025); Cosmetovigilance (2023); Clinical Research (2024)
MCA as substitute for CBA	AHP-based multi-criteria scoring replaces economic valuation; weights assigned without monetary comparability.	Cannabis Products (2023); IRAS Infection Control (2025); Dental Care (2024)
Asymmetric cost–benefit treatment	Internal administrative costs quantified via Standard Cost Model; social benefits remain exclusively qualitative.	Traveller Control (2023); Aircraft & Airports (2024); PAS (2024)

Note: Illustrative reports are cited for each pattern; the pattern itself was observed across the full corpus. Page references appear in the individual report citations in the text above.

Implications and Reform Agenda

The Hidden Cost of Quantitative Inertia

We have seen that ANVISA has standardised the use of Multi-Criteria Analysis, especially the *Analytic Hierarchy Process* (AHP) method (ANVISA, 2024; ANVISA, 2025). In the report on Good Practices for the Prevention and Control of IRAS, it is noted that the criterion of "Health Safety" dominates the overall score, while the costs for the regulated sector are minimised (ANVISA, 2025). The benefit of migrating to CBA would be the requirement for a sensitivity analysis and stochastic modelling that would test the robustness of these choices in the face of uncertainty, rather than treating it as a justification for not doing so.

The inclusion of CBA, as recommended by the OECD, would transform AIR reports from mere procedural documents into instruments of social transparency. The concrete benefit would lie in the valuation of externalities: In the analysis of "cosmetovigilance" , it would be possible to quantify the social gain in reducing adverse events against the compliance cost of micro-enterprises (ANVISA, 2023). In the report on "traveller control", it would be relevant to demonstrate how to balance the operational cost of enforcement with the economic impact of potential outbreaks at airports and ports (ANVISA, 2023).

Systematic resistance to full CBA has specific distributional and political consequences. When ANVISA decides not to monetise the benefits of a regulation, it removes from public debate the information needed to compare that

regulation with alternatives, structurally favouring the maintenance of *the* regulatory *status quo*. Furthermore, the absence of rigorous CBA weakens the agency's position in the face of legal challenges. The "cost consideration" test (*Michigan v. EPA*, 135 S.Ct. 2699, 2015) requires agencies in the US to demonstrate that they have weighed the economic impacts of their choices – a precedent whose normative weight grows as Brazil advances in its process of joining the OECD. The 2023 *Regulatory Policy Outlook* points out that the quality of RIA is a central criterion for evaluating national regulatory systems, and the absence of systematic monetisation of benefits is explicitly identified as a methodological deficiency (OECD, 2023).

Suggested Methodological Reforms

Based on the analysis undertaken, and in line with OECD guidelines and the best literature on regulatory quality, four priority methodological reforms are proposed:

(a) Incorporation of VVE into AIRs on mortality impact: ANVISA could adopt a reference VVE, estimated by revealed preference studies in the Brazilian labour market or transferred from international studies adjusted for GDP per capita – *benefit transfer* methodology (Rosenberger & Loomis, 2017)[1].
(b) Introduction of stochastic models for uncertainty analysis: Monte Carlo simulations or scenario analyses with probability distributions allow confidence intervals to be derived for the regulatory NPV, making uncertainty manageable without eliminating the discipline of comparison.
(c) Expansion of the scope of AIRs to *outcome* indicators: these documents should evolve from monitoring *outputs* to *outcome* indicators (variation in morbidity and mortality, effective patient access, price variation in the regulated sector), systematically integrating the capabilities of DATASUS, VIGITEL, and INCA into the agency's assessments.
(d) Institutionalisation of methodological review by external peers? Following the model of the US *Office of Information and Regulatory Affairs* (OIRA) or the British *Regulatory Policy Committee*, Brazil should create an independent review mechanism for AIRs and ARRs, with the technical capacity to assess the robustness of methodological choices and recommend corrections before the publication of standards.

Conclusions

ANVISA's regulatory output, examined through the content of twenty-five AIR reports, reveals an agency that has mastered the formal process of impact analysis, producing technically structured qualitative analyses. However, the documents analysed could be strengthened by essential methodological cores, notably Cost-Benefit Analysis (CBA), whose incorporation could substantially raise the

[1]The *benefit transfer* methodology, as described by Rosenberger and Loomis (2017), is an economic technique for estimating the value of environmental goods or ecosystem services in unstudied locations by transferring results from pre-existing primary studies.

quality and legitimacy of regulatory decisions. It was found that governance based on Multi-Criteria Analysis (MCA) and the internal Standard Cost Model formally meets national regulatory requirements, but is insufficient to provide the allocative transparency required by international best practices and Brazil's accession process to the OECD.

Three patterns emerged from the analysis. First, the systematic invocation of scientific uncertainty as a barrier to quantification, when stochastic models and sensitivity analyses are precisely designed tools to make it manageable. Second, there was a lack of analytical differentiation between internal operational efficiency (understood as the reduction of the agency's own administrative costs) and social allocative efficiency, which presupposes the maximisation of net collective welfare. Third, treating public health as an intrinsically immeasurable value produces a particularly serious analytical paradox: an agency unable to calculate the value of the benefits it produces also cannot demonstrate, in a transparent and verifiable manner, that these benefits outweigh the costs it imposes on society.

Overcoming this *deficit* does not require abandoning the principles of precaution or public health protection, nor does it imply the indiscriminate monetisation of values. On the contrary, it is argued that these principles should be supported by the best available evidence: an agency that quantifies the benefits of its regulations is not reducing human life to a market value, but rather building the most solid and transparent argument possible for the legitimacy of its interventions. At the current stage of Brazil's accession to the OECD, the adoption of full CBA is not a technocratic concession, but rather an imperative of democratic *accountability*. Without it, there is a risk of making health protection invisible in its foundations, incalculable in its effects and, therefore, immune to rational challenge by both its beneficiaries and its critics.

Several limitations of this study merit explicit acknowledgement. The analysis relies exclusively on publicly available RIA reports and does not draw on internal agency documents, staff interviews, or budgetary data that might illuminate the institutional factors underlying the patterns observed. As previously mentioned, the analysis relies exclusively on publicly available Regulatory Impact Assessment (RIA) reports and does not utilize internal agency documents, employee interviews, or budgetary data. Therefore, the coding of methodological elements was carried out using a binary and relatively objective approach for most checklist items, introducing the possibility of interpretive bias, which does not detract from, but rather enhances, the merit of the results. The final, publicly available outcome of a broader regulatory process can be explored in future research, which could be achieved through multi-coder designs.

For Brazilian health regulation to reach the standard of evidence required by international literature and OECD membership criteria, it is imperative that "scientific uncertainty" cease to be the end of the analysis and become its object. Imperfect quantification, transparent in its limits and methodologically explicit in its premises, is infinitely more valuable to regulatory democracy than the calculated silence of strategic imprecision. In this sense, this research contributes to the debate on the institutional quality and scientific nature of regulatory governance by empirically demonstrating that ANVISA has the technical capacity for formal impact analysis and could mobilise the economic valuation instruments

that would give the decision-making process the analytical robustness required by the contemporary scientific paradigm. Future research could investigate the institutional, budgetary, and technical training factors that explain this gap, as well as assess whether specific reforms in the agency's organisational structure – such as the creation of a unit specialising in economic modelling – would be sufficient to bring ANVISA's practice closer to the OECD standard, or whether broader legislative reforms in the regulatory framework of AIR in Brazil would be necessary.

References

Adler, M. D., & Posner, E. A. (2006). *New foundations of cost-benefit analysis*. Harvard University Press.

Agência Nacional de Vigilância Sanitária (ANVISA). (2025). *Relatório de AIR sobre boas práticas de prevenção e controle das IRAS em serviços de saúde*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2025). *Relatório de AIR sobre a revisão dos critérios para definição de preços de produtos novos (CMED)*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR: produtos de cannabis para fins medicinais*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2023). *Relatório de AIR sobre cosmetovigilância*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2023). *Relatório de AIR sobre as diretrizes para classificação de risco para as atividades econômicas*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre controle sanitário de aeronaves e aeroportos*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2023). *Relatório de AIR sobre o controle sanitário de viajantes*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2022). *Relatório de AIR sobre procedimentos para regularização de alimentos e embalagens*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2023). *Relatório de AIR sobre controle de produtos de degradação em medicamentos*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2023). *Relatório de AIR sobre análise e deliberação dos recursos administrativos (GGREC)*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre priorização e procedimento especial de análise de medicamentos*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre indicação e eleição de medicamentos de referência*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre regularização de empresas em PAF (AFE e BPA)*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre requisitos sanitários para serviços de assistência odontológica*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre controle sanitário de portos e embarcações*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2025). *Relatório de AIR sobre fracionamento de produtos de higiene e cosméticos*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre procedimentos para autuação e julgamento de infrações sanitárias (PAS)*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2022). *Relatório de AIR sobre doação de alimentos com segurança sanitária*. Anvisa.

Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre*

revisão do regulamento de bens e produtos importados (RDC 81/2008). Anvisa.
Agência Nacional de Vigilância Sanitária (ANVISA). (2025). *Relatório de AIR sobre cosméticos produzidos de maneira artesanal*. Anvisa.
Agência Nacional de Vigilância Sanitária (ANVISA). (2022). *Relatório de AIR sobre novos alimentos e novos ingredientes*. Anvisa.
Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório de AIR sobre pesquisa clínica com medicamentos e produtos biológicos*. Anvisa.
Agência Nacional de Vigilância Sanitária (ANVISA). (2022). *Relatório final de AIR sobre dispositivos eletrônicos para fumar*. Anvisa.
Agência Nacional de Vigilância Sanitária (ANVISA). (2024). *Relatório final de AIR sobre organização do SNVS e descentralização (RDC 560/2021)*. Anvisa.
Agência Nacional de Vigilância Sanitária (ANVISA). (2025). *Relatório parcial de AIR obre BPF para industrializadores de alimentos*. Anvisa.
Aragão, A. S. de. (2010). Análise de impacto regulatório. *Revista de Direito Público da Economia*, (32), 9–15.
Beck, U. (2015). *Pioneer in cosmopolitan sociology and risk society*. Springer.
Boardman, A. E., et al. (2006). *Cost-benefit analysis: Concepts and practice* (4th ed.). Prentice Hall.
Brasil. Agência Nacional de Vigilância Sanitária. (2026). *Análises de impacto regulatório*. https://www.gov.br/anvisa/pt-br/assuntos/regulamentacao/air/analises-de-impacto-regulatorio
Brasil. (2020). Decreto nº 10.411, de 30 de junho de 2020. *Diário Oficial da União*.
Cellard, A. (1997). A análise documental. In J. Poupart et al., *A pesquisa qualitativa* (pp. 295–316). Vozes.
Cole, D. H. (2012). Law, politics, and cost-benefit analysis. *Alabama Law Review*, 64, 55–89.
Cooter, R., & Ulen, T. (2012). *Law and economics* (6th ed.). Pearson.
EPA. (2010). *Guidelines and specifications for preparing economic analyses*. U.S. Environmental Protection Agency.
Gico Júnior, I. (2014). Introdução ao direito e economia. In L. B. Tim (Ed.), *Direito e economia no Brasil* (2nd ed., pp. 1–33). Atlas.
Goldman, D. P., et al. (2018). The value of medical innovation in the United States: 1983–2012. *RAND Health Quarterly*, 8(1).
Graham, J. D., & Wiener, J. B. (1995). *Risk versus risk*. Harvard University Press.
Kahneman, D. (2012). *Rápido e devagar: Duas formas de pensar*. Objetiva.
Mendonça, J. V. (2014). *Direito constitucional econômico*. Fórum.
OECD. (2008). *Building an institutional framework for regulatory impact analysis*. http://www.oecd.org/regreform/regulatory-policy/40984990.pdf
OECD. (2018). *Cost-benefit analysis and the environment*. OECD Publishing.
OECD. (2023). *OECD regulatory policy outlook 2023*. OECD Publishing.
OECD. (2009). *Regulatory impact analysis*. OECD Publishing.
OECD. (2012). *Recommendation of the council on regulatory policy and governance*.
OECD. (2015). *Regulatory policy outlook*.
OECD. (2007). *Relatório sobre a reforma regulatória – Brasil*.
Palma, J. B. de. (2014). Processo administrativo normativo na regulação. *Revista de Direito Administrativo Contemporâneo*, 12.
Papanicolas, I., Woskie, L. R., & Jha, A. K. (2018). Health care spending. *JAMA*, 319(10), 1024–1039.
Posner, R. A. (2010). *Direito, pragmatismo e democracia*. Forense.
Posner, R. A. (2014). *Economic analysis of law* (9th ed.). Wolters Kluwer.
Posner, R. A. (1993). *The problems of jurisprudence*. Harvard University Press.

Radaelli, C. M., & De Francesco, F. (2007). *Regulatory impact assessment*. http://regulation.upf.edu/ecpr-07-papers/cradaelli.pdf
Revesz, R. L., & Livermore, M. A. (2008). *Retaking rationality*. Oxford University Press.
Rosenberger, R. S., & Loomis, J. B. (2017). Benefit transfer. In P. A. Champ et al. (Eds.), *A primer on nonmarket valuation* (pp. 431–462). Springer.
Sachs, J. D. (2015). *The age of sustainable development*. Columbia University Press.
Sá-Silva, J. R., Almeida, C. D., & Guindani, J. F. (2009). Pesquisa documental. *Revista Brasileira de História & Ciências Sociais*, 1(1), 1–15.
Sen, A. (2004). *Rationality and freedom*. Harvard University Press.
Sunstein, C. R. (2005). *Laws of fear*. Cambridge University Press.
Sunstein, C. R. (2014). *The cost-benefit revolution*. MIT Press.
Sunstein, C. R. (2002). *The cost-benefit state*. ABA Publishing.
Sunstein, C. R. (2014). *Valuing life*. University of Chicago Press.
Sunstein, C. R. (2007). *Worst-case scenarios*. Harvard University Press.
Thaler, R. H., & Sunstein, C. R. (2008). *Nudge*. Yale University Press.
Viscusi, W. K., & Aldy, J. E. (2003). The value of a statistical life. *Journal of Risk and Uncertainty*, 27(1), 5–76.

Athens Journal of Law - Volume 12, Issue 3, July 2026 – Pages 169-188

Online Suicide and Self-harming Communications: Providing Protection for Children under Online Regulation?

By Daniel Fenwick[*]

Online content promoting suicide and self-harm has been found by a number of researchers to be a major and rising concern for users aged 13-17, on the basis of finding a correlation between suicidal or self-harming behaviour in under-18s and visiting sites depicting images of self-harm and suicide, accompanied by text advocating or promoting such activity. Ofcom has found that age is a risk factor: the risk of encountering and being influenced by such content increases with the age of children and teenagers. Further, those with existing mental health challenges may be more likely to engage with this content, and self-harming behaviour or suicidal ideation may therefore increase. Children and teenagers, the Ofcom Report pointed out, are also more susceptible to a contagion effect than adults: they are more likely to imitate behaviours encountered online and give in to impulse. The Online Safety Act 2023 (OSA) in the UK was introduced partly to address the harm to under-18s created by the presence and influence of self-harm and suicide sites/communications, by creating a regulatory regime covering in-scope online services, including such sites and also social media platforms such as Facebook, Instagram, X, TikTok, on which such content may appear, and be promoted by recommender algorithms. This article will critically examine the efficacy of the OSA in relation specifically to regulation via content moderation and other methods of protecting children from harm caused by encountering such content online, concluding that the OSA regime in a range of respects is proving to be ineffective. This article will focus most closely on the legal but harmful scheme under the OSA aimed at under-18s, not adults, in relation to curbing the harm caused by content encouraging children to engage in self harm or suicide.

Introduction

Online services have the capacity to enrich children's lives in myriad ways so long as they have the 'resources and support in their physical and social environments to navigate digital spaces safely'.[1] This is especially true in terms of enhancing their opportunities to exercise freedom of expression, not only in relation to their personal interactions. That is in part because the internet provides them with the technological architecture to circumvent various barriers to communication,[2] which for children can include parental control or oversight.[3] When accompanied with adequate support and supervision, it can, therefore, increase their opportunities to freely communicate – often instantaneously – with others, including friends and

[*]Senior Lecturer, Northumbria University, UK.

[1]OECD (2025) para 2.2.

[2]Fenwick & Coe (2025), p. 747.

[3]E.g. Higson-Bliss & Street (2026) for an argument on this point in respect of the use of virtual private networks (VPNs).

family from anywhere in the world, at any time of day.[4]

However, such freedom of communication also brings with it a range of online harms, especially affecting children, in particular, online content promoting suicide and self-harm. Such content has been found by a number of researchers to be a major and rising concern for users aged 13-17, on the basis of finding a correlation between suicidal or self-harming behaviour in under-18s and visiting sites/pages presenting text and images promoting self-harm and suicide accessible by children.[5] In particular, content promoting self-harm and suicide has been found by Ofcom to be a particular concern in relation to that age group; its 2025 Report found that 30 per cent of persons within that age group had seen or heard of primary priority content including self-harm and suicide.[6] Ofcom has found that age is a risk factor: the risk of encountering such content increases with the age of children and teenagers.[7] Ofcom has also found that a number of deaths in the UK have been linked with online material where detailed information is shared on methods of suicide and self-harm, or where suicidal and self-harm behaviours are actively encouraged.[8] It further found that children may encounter such content accidentally or have it recommended to them by algorithms. Its 2024 Report suggested that four per cent of UK internet users have seen online content promoting suicide in the last month, and children are more likely to see it than adults.

The Online Safety Act 2023 (OSA) was introduced in order to address a range of online harms created by content on various online services, such as Facebook, Instagram, Snapchat, *X,* TikTok, but in particular those affecting children, one of the main reasons given for its introduction.[9] To further that aim the OSA includes offences linked to suicide and self-harm as 'priority illegal content' for regulatory purposes, and also creates a legal but harmful scheme aimed only at under-18s, not adults, which specifically singles out sites and processes encouraging self-harm or suicide in that age group. This article will focus most closely on that legal but harmful scheme in relation to curbing the harm caused by sites and processes, such as recommender algorithms, encouraging children directly or indirectly to engage in self-harm or suicide.

Ofcom was installed as the online regulator, and accorded various powers in order to enforce the safety duties discussed below, imposed on the services. It can fine regulated services up to £18 million, or 10 per cent of annual global turnover, whichever is higher, if they fail in their duty of care (Sched. 13, para. 4). Ofcom has the power to block non-compliant services from being accessed in the UK since ss.144-148 provide for 'business disruption measures' that allow it to apply for a variety of 'restriction orders' if the regulated service has failed to meet certain conditions relevant to the restriction sought. That could include requesting third-party companies to stop providing services or facilitating access to the non-compliant

[4]Ofcom (2024), pp. 17-18.

[5]The Molly Rose Foundation (2025) commissioned research which found that around 1.5 million young people in the UK may be exposed to high-risk online content on a weekly basis. It demonstrated that this content is frequently delivered through algorithmic recommendations rather than active searches, exacerbating the risks created for vulnerable users.

[6]See Ofcom (2025b), p. 85.

[7]Ofcom (2025c), p. 50: 'Research indicates that the risk of encountering this content online increases with children's age'.

[8]Ofcom (2025d).

[9]OSA, s1(3)(b)(i).

platform, meaning that it would be erased from search results, app stores, or links on social media posts. S.110 creates criminal offences, pursuant to s.109, for named senior managers of in-scope services in respect of failures to provide the requisite information needed to determine compliance (Part 7 chp. 4; see also ss.111-113). The problem is that Ofcom is under-resourced when compared with the companies it is attempting to regulate. Further, Ofcom appears already to be adopting a cautious approach to the demands placed on the companies and deployment of the sanctions at its command.[10]

While seeking to address online harms, the OSA also attempts to preserve free expression online,[11] while creating a scheme intended to protect users from online harms, especially those affecting children. A key concern is with the ability of this model of regulation to navigate a path between fostering such expression while curbing such harms. The primary provisions that preserve and protect free expression are section 22 (for user-to-user services) and section 33 (for search services). These provisions sections mandate that platforms have a statutory duty to prioritize and protect users' rights to freedom of expression within the law, protected in the UK by Article 10 of the European Convention on Human Rights, received into domestic law by the Human Rights Act 1998 (HRA). Thus these sections place the services, as private bodies, in a position similar to that of public bodies which are subject to Article 10 due to s6 of the HRA. Specifically, these provisions require providers to avoid over-reach by ensuring that their automated filters and takedown rules do not disproportionately censor lawful speech.

Therefore, in discussing the provisions of the OSA below, intended to protect children from online content promoting suicide or self-harm, the services need to take free expression into account, especially when confronting lawful content (section 4 below). This article sets out to critically examine the efficacy of the OSA in relation specifically to regulation via content moderation and other methods of protecting children from harm caused by encountering such sites and content, concluding that the regime in a range of respects is proving to be ineffective. Thus, it is not fully protecting children from the online harms discussed here, but there is also little evidence as yet that Ofcom is taking steps to address company over-reach, adversely affecting free expression, in terms of deploying their own terms of service in ways that at times loeadss to content removal even where that is not required by the OSA.

Harms caused by Suicide and Self-harm-related Content Online

The perception that online communications were contributing to serious self-harm and suicide among children was a key reason for the introduction of the OSA.[12] The Molly Rose Foundation, whose recommendations informed the OSA, argue that the exposure of children to self-harm and suicide content on various sites and social media is not merely incidental or transient, but is rather a pervasive aspect of the business model of the platforms.[13] This is so due to methods inbuilt by

[10]See: Children's Commissioner (2025).
[11]See e.g.: Bhagwat and Weinstein (2021); Post (2006); Greenawalt (1989); Blasi (1977). While these works concern free speech off-line, their messages are also applicable to online expression.
[12]House of Lords, House of Commons (2021).
[13]Molly Rose Foundation (2025).

design of encouraging user continued engagement and also partly because digital media, unlike traditional media, has limited editorial control over content production and dissemination, and therefore allows harmful user-generated content to spread.[14] Poorly moderated services, including social media are demonstrably capable of immersing vulnerable children in a culture of self-harm and suicide,[15] and of contributing to pro-self-harm and pro-suicide ideation and motivation, as discussed below.[16] Both of the latter were evident from the inquest into the suicide of Molly Russell, a fourteen year-old who killed herself after becoming immersed in suicide and self-harm content on Instagram.[17] The Coroner in the inquest that followed found that social media had significantly contributed to Molly's suicide, singling out Instagram's recommendation-algorithm as a substantial factor in confirming her depressive suicidal ideation.[18] While a causal link between any specific images and engaging in serious self-harm was not found, it was established that Instagram's algorithm contributed to Molly becoming overwhelmed with thoughts of ending her life.[19]

Molly Russell's case provides an effective illustration of the harms of online platforms. Molly began posting on Instagram when she was twelve, contrary to the terms of service,[20] as Instagram lacked age-verification screening at the time,[21] and by the time of her suicide had amassed hundreds of hours browsing, commenting and messaging on the platform. While maintaining an outward appearance of normality to her family, she had, in the six months prior to her suicide, begun to withdraw from offline social relationships and spent a significant amount of her free time on Instagram and other online platforms.[22] This time was, unbeknownst to her family and offline friends, devoted in part to accessing unnuanced pro-self harm and suicide content,[23] which formed thirteen percent of the over sixteen thousand pieces of Instagram content that she had engaged with in the six months prior to her suicide.[24] Dramatisation of suicide was a prominent feature of this material, including graphic videos of suicide and self-harm set to music and TV-media.[25] Molly had also periodically attempted to contact social media influencers who raised suicide and self-harm themes in their content.[26] While much of the material Molly accessed was contrary to Instagram's guidelines, the latter material was considered to fall within them, a position that was criticised by the judge at the subsequent inquest on the basis that a user of Molly's age would struggle to contextualise such stories.[27]

Ian Russell, Molly's father, campaigned successfully for the encouragement

[14]Ibid; see also Crawford (2023).
[15]Molly Rose Foundation (2025).
[16]Marchant et al. (2017); see also Crawford (2023).
[17]Molly Rose Foundation (2025), p. 3.
[18]Ibid, p. 12. Coroner Andrew Walker ruled that Molly "died from an act of self-harm whilst suffering from depression and the negative effects of on-line content".
[19]Ibid, p. 12.
[20]Milmo (2022).
[21]North London Coroner's Service (2022), p.5.
[22]Milmo (2022).
[23]North London Coroner's Service (2022), p.4.
[24]Molly Rose Foundation (2025), p. 12.
[25]Milmo (2022).
[26]Ibid; see also Milmo (2022).
[27]Milmo (2022).

of self-harm offence that would become s184 OSA 2023, discussed in the next section,[28] and her case was cited by Chris Philp, then Minister for Technology and the Digital Economy, when discussing duties regarding suicide and self-harm content.[29] The Online Harms White Paper[30] and report of the Joint Committee on the Draft Online Safety Bill[31] similarly stated that prevention of harms in cases such as Molly's is a fundamental goal of the legislation. The Joint Committee in particular found, by reference to Molly's case, that suicide and self-harm material on websites was "far more dangerous when served up automatically, proactively, and repeatedly by the recommender systems of platforms popular with young people".[32] A consensus emerged in early literature on this subject concerning the 'contagion effect' whereby children and teenagers encounter harmful behaviours seen online and then upload content emulating such behaviours.[33] Studies have demonstrated a correlation between suicidal or self-harming behaviour in under-18s and the contagion effect phenomenon.[34]

While the contagion effect provides an adequate starting point, studies of suicidal and self-harm ideation provide a more nuanced understanding of how these harmful patterns of behaviour arise on user-to-user services, including social media platforms. There is now an established body of research concerning social media's facilitation of self-harm and suicidal ideation in children,[35] and the role of crude recommender-algorithms in this process is now better understood.[36] One area of developing study is 'suicide-sensationalism' which concerns the particular contribution to users' suicidal ideation of video and image-sharing media. This phenomenon, which is also observable in traditional media,[37] is associated with the engagement-excitement-compulsion cycle typical of social media platforms.[38] Digital platforms encourage users' engagement by creating a cycle of consumption and production of media using varied stimuli that can become compulsive, which – when a crude recommender algorithm is operating – results in the promotion of depressive, suicide and self-harm content. Similar, social media systems operate social reward systems, such as 'view' and 'like' counters, that – when applied crudely – capture depressive, suicide and self-harm content.[39]

It is now recognised that, without effective regulation, visceral depressive posting on social media, including acts of self-harm, is likely to be promoted, while posting discouraging such acts and encouraging cognition and other healthy strategies to manage suicidality and depression, is less likely to be as high profile.[40] For certain

[28]House of Lords, House of Commons (2021).
[29]Digital, Culture, Media and Sport Committee (2022), Q307.
[30]Department of Culture Media and Sport & Home Office (2020), p.19.
[31]House of Lords, House of Commons (2021).
[32]Ibid, p.95.
[33]Marchant et al. (2017), p. 22; Bell & Westoby (2025), p. 4.
[34]See eg Sedgwick et al. (2019).
[35]Marchant et al. (2017).
[36]Molly Rose Foundation (2025), p. 7; House of Lords, House of Commons (2021), para 322.
[37]Thom (2011).
[38]Liu et al. (2020).
[39]Ibid.
[40]Molly Rose Foundation (2025).

users who are most susceptible, the effect of such unregulated algorithms is to create a spiral whereby intensely bleak and depressive content that complements ideation is normalised.[41] Additionally, when users upload and share their experiences this can reinforce sense-memories of previous acts of self-harm or attempted suicide.[42] Poorly moderated social media is therefore capable of both spreading and *distilling* harmful, depressive content for users who engage in it, and thus of creating not only an unrelentingly negative online suicide and self harm 'culture' but also one that is algorithmically tailored towards the extreme.[43] It should be noted that this spreading and distilling effect extends beyond the image/video itself to comments left by users.[44]

In addition to emotive content that normalises suicidality and depressive feelings in victims, unregulated social media content can encompass practical and emotional support for those who are actively contemplating self-harm or suicide.[45] Digital media-use by suicide and self-harm attempters is characterised by practical content, such as discussion of methods of self-harm, as well as content designed to motivate contemplating users to implement such methods. Individuals who are contemplating or practicing have been found to upload, comment on and repost planned or completed acts of self-harm or attempted suicide.[46] One recent study identifies a three-stage process for attempters: firstly, selection and retention of practical information and emotive quotes and slogans; secondly, reproduction of such information in comments and reposts and, finally, posting designed to increase attempters' resolve to complete the self-harm or suicide attempt, such as discussing a specific action plan.[47] The latter stage has the strongest correlation with an attempt,[48] in contrast to more discursive posting. This suggests a flaw in the approach to suicide and self harm content in current OfCom guidelines that are weighted towards graphic text/images associated with normalising the emotional basis of suicide rather than the latter content which is most proximately associated with self harm and suicide.[49] The harms associated with the reproduction stage are also poorly reflected in current guidance, which encourages companies to treat such posting merely as a content violation, rather than as an opportunity for intervention, such as referral to protective content, such as the Samaritans.[50] It should be noted that a nuanced approach to such content is required, however, as the sharing of experiences by survivors in a manner that provides no practical or emotional support to those actively contemplating has been shown to correlate with a diminution in self harm and suicide attempts.[51]

Another problem with current guidance is that the relationship between suicide and self-harm and related online harms is poorly understood.[52] For example, child

[41]Marchant et al. (2017), p. 14.
[42]Liu et al. (2020), p. 8.
[43]Bell & Westoby (2025), pp.4-5, 7-8.
[44]Ibid.
[45]O'Connor & Kirtley (2018).
[46]Liu et al. (2020), p.7; Sueki (2015).
[47]Liu et al. (2020), p.8.
[48]Ibid.
[49]Ofcom (2025a), p. 42.
[50]Ofcom (2025a), p. 42.
[51]Liu et al. (2020), p.1
[52]Ofcom (2025a), p. 42.

engagement with mental health information undergoes a similar toxic spiral to the one detailed above for depressive suicide and self-harm ideation and often runs alongside specifically self-harm and suicide-related posting.[53] Such engagement may involve discussion of mental health topics, particularly depression and anxiety, as well as discussion of self-destructive behaviours, such as sexting and alcohol or drug misuse.[54] Users engaging in such behaviour, who are not already involved in posting self-harm and suicide related media, have been shown to become more susceptible to suicide or self-harm ideation.[55]

In addition to the self-harm, suicide and mental health-related online content, there are two significant background factors that facilitate ideation and attempts: immaturity and isolation. As regards the former, suicide and self-harm content on various sites and social media platforms has a greater impact on underage or emotionally immature children.[56] Posting about depressive topics is capable of having a particularly harmful impact on such users.[57] Furthermore, content addressing such content that may otherwise be beneficial for mature children and adults, such as stories of recovery, may be harmful for such users. In relation to the other factor, isolation, immersion in digital media has been demonstrated to retard the development of pro-social behaviours and to encourage anti-social ones.[58] The impact of such immersion on the sharp increase in children and teenagers, reporting that they suffer from loneliness, is well documented.[59] Loneliness is closely associated with higher than average engagement with online services including social media and is a significant factor contributing to suicidal ideation and attempt.[60] Relatedly, algorithmic processes can undermine the development of strategies for successful social interactions concerning complex depressive topics, so that a user who appears socially well adjusted, may – in relation to certain such topics – be further discouraged from discussing them with off-line friends or family.[61]

In addition to the harmful patterns of digital media-use detailed above, the positive impact of protective discourse facilitated by online activities should be recognised.[62] For example, targeted interventions by mental health workers working with platform moderators, have been shown to be capable of countering suicide and self harm ideation.[63] Similarly, more discursive platforms, such as suicide and self harm forums, on which the typical discourse is in a longer, text-based format, have also been shown to be beneficial, particularly when combined with effective moderation.[64] Such messaging has been shown to be capable of fostering "hope, recovery, and well-being".[65] The

[53]Molly Rose Foundation (2025).
[54]Liu et al. (2020), p.6.
[55]Ibid.
[56]Gannon et al. (2025).
[57]Garg & Singh (2025), p. 51.
[58]Ibid.; see also Uhls et al. (2014).
[59]Yang & Crespi (2025), pp. 7, 27.
[60]Mikuska et al. (2020); Yang & Crespi (2025), p. 7.
[61]Molly Rose Foundation (2025).
[62]European Parliament (2023), p.7.
[63]Bell & Westoby (2025), p. 7.
[64]Ibid, p. 7.
[65]Ibid, p. 7.

potential utility of permitting users to continue to post such material was raised by Adam Mosseri, the CEO of Instagram, in response to Molly Russell's case.[66] However, the policing of the distinction between protective suicide or self-harm discourse in the digital media, and harmful sensationalist content, requires effective moderation and a detailed understanding of user-behaviour. On image-based platforms that are more susceptible to sensationalism, such as Instagram, this challenge is particularly acute. Moderators typically have limited knowledge or direct experience of the mental health context of self-harm or suicidal ideation and attempt related content on their platforms, especially if the moderation is partially or wholly automated. Therefore, even when a platform has guidelines in place and has curated its algorithm to avoid promoting dangerous content, the actual experience of the platform for users is not dissimilar from the relatively unregulated one that of previous generations of users, such as Molly Russell.

Illegal Content Duties

The OSA imposes illegal content duties – which came into force in March 2025 - on services that are within the scope of the Act; that includes so-called user-to-user services. It is irrelevant that a company running the services is based outside the UK so long as it has links to the UK due to the number of UK users (s4 OSA). As of March 2025, Ofcom has the power to issue fines of up to £18 million or 10% of a company's global revenue, as mentioned above, if companies fail to discharge their illegal content duties, and can also seek court orders to block access to illegal sites within the UK.[67] The illegal content Codes are also in force.[68]

'Illegal content' is defined in s59(2) as content amounting to a 'relevant offence';[69] if no such offence applies to particular content, even if it could be harmful to children, it either falls entirely outside the scheme, or, depending on its nature, it could be covered by the legal but harmful provisions applying only to under-18s, not adults, discussed in Section 4. Under section 192 services must find illegality if they have 'reasonable grounds to infer' that the elements of the relevant offence are made out – in this case, offences relating to aiding/encouraging suicide and self-harm. That means that the tech company's moderators, usually automated systems, must consider whether, as a reasonable inference, based on the nature of the content on the platform, the *actus reus* and *mens rea* elements of the relevant offence appear to be present.[70] Moderators must also determine whether a defence to the offence, if any, appears to be present.[71] If one or more of the elements of the offence in question do not appear to be present – ie the content, even if potentially harmful to children, does not appear to fall within the offence in question, and/or a defence does appear to be present, the content must be deemed legal and therefore no response is needed to it from the service in terms of the illegal content duties. If

[66]Marsh & Waterson (2019).
[67]See the Introduction for further detail as to enforcement methods.
[68]See: Ofcom (2025e); Ofcom (2025f).
[69]S.59(4), (5).
[70]Section 192(6)(b) OSA.
[71]Section 192(5), (6)(a).

the content, however, is deemed to be illegal and is also listed as priority illegal content (PIC) in Schedule 7 OSA it should be prevented from appearing on the service or, if that safeguard fails, the period of time for which it is present should be minimised (ss10 and 27 OSA).

For the purposes of this article the relevant offences are, firstly, the new offence of encouraging serious self-harm introduced under the OSA s184. It covers content encouraging or assisting serious self-harm with intent to do so.[72] S184 covers a range of online content, going well beyond self-harm or suicide sites, and it 'does not matter whether the content of the communication or publication is created by the defendant (so for example, in the online context, the offence under this section may be committed by forwarding another person's direct message or sharing another person's post)' (s184(7)). It also covers (s184(8)) communications consisting of or including a hyperlink to other content, where the content can be accessed directly via the hyperlink. But service providers themselves fall outside s184 (sub-section 10). Secondly, there is also the offence of encouraging or assisting suicide *if* intending to so encourage or assist suicide or an attempt at suicide; that was already an offence pre-OSA under s2 Suicide Act 1961 and s13 Criminal Justice Act (Northern Ireland) 1966. These offences are deemed by the OSA to cover priority illegal content – PIC. Therefore, content falling within those offences would count as PIC for the purposes – in theory - of triggering the relevant, quite stringent duties of service providers.

But there are obvious problems with this illegal content scheme relating to suicide and self-harm sites or other inducements via online content to commit self-harm or suicide. Clearly, tech company moderators, usually automated systems, are not well equipped to detect the *mens rea* elements of these offences since usually they are reliant on detecting illegality based on the content of a communication alone. They may be able to identify content based on the *actus reus* of these offences, but services are likely to be finding leeway currently to maintain that the *mens rea* element was not clearly present or could not be identified. In relation to s.184, in some very clear-cut instances of encouragement to self-harm that element of intent could be inferred to be present, from the content, but in less obvious but still persuasive instance it might appear to an automated moderation system that the intention element was absent.

A service could also argue that the communication in question was merely exploring the issue of self-harm or suicide and reporting on it: the provider might concede that negligence or recklessness as to encouraging the acts in question might be found, but not intention. Or a service provider could argue that the self-harm depicted was not clearly 'serious' enough to fall within the s184 offence. The result, therefore, in a number of instances, is likely to be that content of this nature, although identified in various studies as harmful to under-18s – and sometimes

[72]'A person (D) commits an offence if—(a) D does a relevant act capable of encouraging or assisting the serious self-harm of another person, and (b) D's act was intended to encourage or assist the serious self-harm of another person'. The 'relevant act' includes 'sends, transmits or publishes a communication by electronic means'. Serious self-harm amounts to GBH within the meaning of the Offences Against the Person Act 1861, and in Scotland, severe injury. Cumulative acts of self-harm can reach that threshold, in combination.

most harmful to younger teenagers – could remain available to them online.

Aside from the problems of identifying illegality discussed, enforcement is also complex due to jurisdictional issues, since many of these sites are hosted outside the UK and also use anonymity tools. However, the fact that user-to-user services and search services are outside the UK does not necessarily mean that they are not regulated under the OSA.[73] While some such sites covered by the OSA have blocked UK users, they may still be accessible via VPNs, leading to calls from organisations representing the families of victims for faster and more decisive action from Ofcom.[74] Some services are banning content relating to suicide or self harm under their own terms of service, even where the content is probably not illegal. Facebook (Meta), for example, states that it explicitly prohibits material that encourages or promotes suicide, self-injury, or eating disorders under its Community Standards. While the platform allows users to discuss these topics to raise awareness or seek support, it states that it removes content that encourages, provides instructions for, or celebrates these acts. Clearly, however, these claims are sometimes open to doubt: the child-protective safety tools in question may not be as effective as the company in question claims.[75]

Further, it is already apparent that some services are not complying with this scheme, or not complying fully, in relation to curbing online encouragement to commit suicide or to self-harm, although following pressure from Ofcom, some such sites have "voluntarily" restricted access for UK users to comply with the Act via geo-blocking. In April 2025, Ofcom launched its first investigation under the OSA into an unnamed, US-hosted pro-suicide forum linked to over 50 UK deaths. In February 2026 Ofcom found provisionally that the site was breaching its duties in relation to the site.[76] Ofcom found: 'Last year, the forum implemented a 'geoblock' in response to our enforcement proceedings against it, to restrict access by people with UK IP addresses. However, after a period of monitoring the service, we became concerned that the block was ineffective and/or was not consistently maintained, and continued to a provisional breach decision as a result'.

An alternative possibility arises of addressing the problems arising due to the nature of the sites in question, or other posts in various respects encouraging self-harm or promoting suicide. If, due to the problem of identifying the *mens rea* elements of the offence in question, for the purpose of triggering the illegal content

[73]They are regulated if they have links with the UK. This means that the service has a significant number of UK users, *or* UK users form one of the, or the only, target market(s) for the service (OSA, s.4(5)(a)-(b)) *or* the service is capable of being used in the UK by individuals, *and* there are reasonable grounds to believe that there is a material risk of significant harm to individuals in the UK due to the user-generated content of the service or the search content of the service (whichever is applicable) (OSA, s.4(6)(a)-(b)(i),(ii)).

[74]In particular, the Molly Rose Foundation (2025) has repeatedly been 'highly critical of Ofcom's deeply unambitious approach to implementing the Online Safety Act'. It stated: 'These findings [see note 5 above] both justify and increase our concern, with Ofcom's current set of measures poorly placed to respond to the scale at which children were being exposed to harmful content and the potential effects of cumulative harm'.

[75]Research has found that child-protective safety tools, such as age verification and content filters, often fall short of company claims due to ineffective enforcement, algorithmic failures, and rapid technological changes. See.

[76]On 27 February 2026 (Ofcom (2026a)).

duties, those duties are evaded by some companies, duties arising under the legal but harmful provisions applying to under-18s could be relied on instead. Since they do not rely on identifying content as *illegal*, it appears at face value that these provisions could, to an extent, address this problem, although the problems of enforcement discussed would still arise. The application of the legal but harmful provisions are discussed and criticised in the next section, below.

The 'Legal but Harmful' Scheme applying to Children

The legal but harmful provisions divide content into primary priority content (PPC) and priority content (PC); the duties are more stringent in relation to primary priority content (ss12 and 29).[77] Such content is covered by s61 OSA. Sub-sections 61(3) and (4) cover content aimed at under-18s which 'encourages, promotes or provides instructions' for suicide or for an act of deliberate self-injury. As of July 2025, when those aspects of the OSA came into force, platforms became legally required to use age assurance to prevent children from accessing this content. Services likely to be accessed by children must – if the scheme is taken at face value - take strict, proactive measures to prevent users under 18 from encountering this "primary priority content". Duties in relation to PPC include enforcing highly effective age assurance (verification or estimation) in order to block access, or prohibiting such content entirely from appearing in their terms of service (s12(3)(a)).[78] Services must clearly define in their terms how they prevent children from accessing this content and enforce these policies consistently.[79] In relation to PC services only have a duty to 'protect' children from encountering the content. The codes (Ofcom's Children's Safety Codes of Practice, published April 2025) set out how platforms can reduce toxic algorithms which can recommend harmful content to children without them seeking it out. This includes ensuring that algorithms do not operate in a way that harms children, by guiding them to content related to suicide and self-harm.

But s61 is badly drafted: it appears that, in order to fall within the scope of the provisions of s61, the content, under s.61(6), must consist of 'text only' or consist of text accompanied by 'identifying content which consists only of text', a GIF, 'emoji or other symbol'. If those words were taken at face value they could be found by moderators to mean that sites/content encouraging suicide or serious self-harm, consisting only of images with no accompanying text, or of both images and text, are excluded from the primary priority content category. That would be a very

[77]Services have a duty (s12)(3)): to operate a service using proportionate systems and processes designed to—

(a) prevent children of any age from encountering, by means of the service, primary priority content that is harmful to children;

(b) protect children in age groups judged to be at risk of harm from other content that is harmful to children (or from a particular kind of such content) from encountering it by means of the service.

[78]See: Ofcom (2025g) p.57: "The provider should include the following in the terms of service: a) provisions specifying how children in the United Kingdom are to be protected from content that is harmful to children."

[79]Ibid, p.57-58.

strange situation, given that images may be more compelling and emotive than text but might equally or more probably create the effects of contagion. The relevant research tends to focus more emphatically on images as opposed to text.[80] There is therefore an apparent mismatch between research into links between encountering online content relating to suicide or self-harm and engaging in such behaviours which does not appear to focus mainly on text,[81] and the designations of content as 'primary priority content' (PPC) or 'priority content' (PC), given that the duties arising in relation to content in the former category are more likely to prevent under-18s encountering it. However, the wording of s61(6) is misleading and should be re-drafted. Section 236(1) of the Act defines content as 'anything communicated by means of an internet service, whether publicly or privately. That includes text *and* images.

Thus a communication, due to the poor drafting of s61(6), might be found by automated systems, depending on their programming, to fall outside ss61(3) and (4). That might also arise on the basis that the content is not found to encourage, promote or provide instructions for suicide or deliberate self-injury, on the basis that the content is indirect and subtle; it could be viewed as merely discussing these issues, bearing in mind that automated systems as moderators may fail to detect more indirect forms of persuasion. Moreover, s61 speaks of 'content'; it does not mention recommender algorithms guiding children towards such content.

If for one of these reasons content was found by automated moderation systems to fall out of the higher PPC category, it could still, in some instances, be caught by the relevant priority content (PC) provisions, under s62. However, the provisions in question under s62 are not as precise as might have been expected and provide further scope for the tech companies to find that the content fell outside the categories, requiring, therefore, no response in terms of removal of the images and other content. The harms in question include under s62(6)(b) depicting 'the real or realistic serious injury of a person in graphic detail'. Obviously, such images may depict a range of injuries that are *not* self-inflicted; therefore, this provision is not closely linked to the problem of encouraging self-harm. A service might further argue that the detail was not graphic enough to be covered.

Section 62(9) further covers 'content which encourages a person to ingest, inject, inhale or in any other way self-administer: (a) 'a physically harmful substance'; (b) 'a substance in such a quantity as to be physically harmful'. Section 62(9) therefore covers encouragement to cause harm to oneself by ingesting something harmful, but *not* by other means, such as by wrist-cutting. Common forms of self-injury include cutting, severe scratching, burning, and banging or hitting; most individuals who self-injure have used more than one method. But encouragement to use these common forms lies outside s62(9).

Section 62 therefore does not expressly cover sites or content generally promoting suicide or self-harm, unless the specific content falls within s62(6) or (9). That omission may have arisen on the basis of the presence of the new offence mentioned above in the OSA, s184, of encouraging or assisting serious self-harm with intent to do so. But if that offence or the offences in relation to encouraging or aiding suicide were not deemed applicable, for the reasons discussed above, the provider would not be affected by the

[80]See: Susi et al. (2023).
[81]Ibid.

illegal content provisions. Given the narrow wording of s62(6) or (9), the duties arising in relation to legal but harmful priority content might be susceptible to avoidance by the service as well.

If relevant content falls, or appears to fall, just outside the scope of the relevant offences, it would not, as discussed, count as 'illegal content'. But, as discussed, such material may, for various reasons, also fall outside the two categories of 'primary priority content' or 'priority content'. It could still possibly be covered as 'non-designated content that is harmful to children',[82] and therefore would be covered by the less demanding age-group dependent duties.[83] Section 234(2) defines harm as "physical or psychological harm", while section 234(3) explains that harm includes harm arising from the nature of the content, the fact of its dissemination, or the manner of its dissemination. Section 234(4) includes cumulative harm arising from repeated content encounters. On that basis content linked to encouraging suicide or self-harm that is not deemed to be primary priority content but is still harmful, on the bases discussed in Section 2, could fall within these provisions and children should be protected from it if an age group deemed to be at risk, under s12(9)(c). However, those provisions leave some loopholes open that services may be able to exploit. Children need only be protected from the content, not prevented from encountering it. The likelihood that it would be removed swiftly from a service is probably low, if the content is not found to be within the primary priority or priority category, and could in any event be disregarded if targeting certain age groups deemed to be at low risk of harm. In some instances, of discussions of suicide or self-harm, including more subtle inducements to engage in that behaviour, automated systems might not pick up the more nuanced albeit harmful content, finding that it was not covered at all. Further, services only need to use 'proportionate' means to address content harmful to children; smaller companies might argue that resource constraints mean that at present they need to focus on addressing PIC or PPC; or they might argue that at present they are technologically unable to detect PC or unspecified harmful content.

For all the reasons discussed, there are very clear gaps in this highly significant aspect of the OSA scheme – since especially harmful content is at stake - for the protection of children online via addressing legal but harmful content. There is also, clearly, the question of Ofcom's willingness or ability to use its available sanctions against non-compliant service providers, which is open to doubt, partly because the scheme in general relies heavily on enforcement by a regulator that is under-funded as compared with the services it is regulating.

Ofcom has shown some awareness of these problems; in June 2025 it published the 'Additional Safety Measures' consultation,[84] which sets out proposals to ask platforms to go further to keep users safe. These include proposals that some service providers should assess whether proactive technology to detect certain kinds of content is available and meets specific criteria. This includes technology to detect illegal suicide content, and suicide and self-harm content which is harmful to children. Ofcom states: 'Where such tools exist, they should use them'. They should also, Ofcom recommends, enable real-time reporting of livestreams showing imminent

[82]OSA, s.60(4) which refers to s.60(2)(c).
[83]It falls within OSA, s.12(3) (user-to-user services) or s.29(3) (search services); see also s12(9)(c).
[84]Ofcom (2026b).

harm, and ensure human moderators are available when livestreaming is active, and design and operate their recommender systems so that content likely to be certain kinds of priority illegal content, (including illegal suicide-related content) is excluded from users' feeds. Ofcom also proposed the expanded use of proactive technologies, such as hash matching to block known illegal images, and automated tools to detect harms, including illegal suicide content. It also proposed to address repeat offending through new user sanctions.

The Samaritans responded, commenting that content related to suicide or self-harm should be treated as in the highest risk category for content, requiring the swiftest response from the companies. They also pointed out: 'Panel members had highlighted that the most valuable improvement to the reporting system would be better feedback on what actions have been taken in response to a report, or, where no action is taken, a clear explanation of why. This would help users understand how decisions are made and reinforce confidence in the platform's moderation processes'. They also criticized the use of automated systems alone to flag and take down content related to suicide and self-harm on the basis that such systems 'cannot capture the emotional tone, urgency, or complexity of such situations. Human moderators should therefore be supported by trained mental health or crisis support staff in cases where potential suicide or self-harm risk is identified'.[85] However, whether from Ofcom or the Samaritans organisation, these are only proposals at present and they are not at present backed up by statutory powers.

Conclusions

The Online Safety Act 2023 was introduced amid claims that it creates a world-leading child-protective model. But the interrogation of the OSA in terms of the harms with which this article is concerned that it is claimed to address indicates that practice as between the various companies is very variable, and that not all are fully compliant with the scheme opposing online encouragement of suicide and serious self-harm in under-18s.[86]

The leeway created, as discussed, for evasion of the illegal content duty in relation to encouraging suicide and self-harm also, it is concluded, undermines the duties pertaining to similar legal but harmful content in relation to under-18s, which could potentially cover certain postings or sites, where they were found to fall outside the illegal content duty. It was found above that relying on the relevant existing offences to identify illegal content for content removal purposes is open to criticism since the mens rea elements may not appear to be present in instances of more subtle encouragement to commit suicide or to self-harm. So there is a case for simply designating the content that should be removed, as occurs under the legal but harmful scheme, which obviously does not rely on existing offences. Or, preferably, it should be prevented from appearing at all – without involving such reliance.

[85]Ofcom (2026c).

[86]Those findings are backed by recent research; see eg, Rahman-Jones & McMahon (2025). "The testing, by child safety groups and cyber researchers, found 30 out of 47 safety tools for teens on Instagram were "substantially ineffective or no longer exist".

Further, some content encouraging suicide or self-harm may not be found to be illegal as discussed, but may also fall outside s61(3) or (4) OSA. If so, the legal but harmful scheme only requires that children should be protected from it if it then falls within s62 as PC only. But ss12 and 29 do not mandate that PC should be prevented from appearing on the service, and does not specify that it is only allowed to appear on the service for a minimal amount of time – as is the case for PPC. Even if some services do remove PC content at some point, relating to depictions of serious self-injury of a person in graphic detail or content encouraging ingestion of a harmful substances, such content removal is in any event not very effective in terms of protecting children from the online harms discussed here. Apart from the possibility of evasion of those sub-sections by some services on the basis that their automated systems did not find that they applied, once children have viewed the content, the psychological damage may well have been done, even if the content is subsequently removed.

This article concludes with some recommendations for development and reform of the OSA regime in relation to online encouragement of suicide and self-harm in children. Alternatives to content removal are however in general unsuitable for use in relation to children since they rely on an evaluation of risk by the child themselves. They include: user Empowerment Tools (Opt-in Filters); instead of platforms blanket-removing content, they are required to provide users with tools to control their own experience. This allows users to filter out content they do not wish to see. To achieve this, platforms can provide "mute," "block," or "hide" functionalities that empower users to customize their feed. These include options to mute keywords, hide non-verified accounts, or opt-out of certain algorithmic recommendations. However, this alternative is not of great value in relation to the content with which this article is concerned for obvious reasons, in relation to children; the same can be said of labelling in the sense of applying warning labels to content that has been flagged as possibly but not clearly illegal (due to the difficulty of identifying the presence of the *mens rea* elements); since users can click through to view it, its efficacy is diminished in the context with which this article is concerned since a significant number of users are still likely to view it.

A further possibility that could be somewhat more effective is to engage in demotion/reduced reach, meaning that services reduce the visibility of content via algorithms (for example, News 11.1.25: shadowbanning or limiting the reach in feeds). As a concomitant to this possibility Ofcom could place a greater focus on the ways that algorithms promote or amplify harmful content. So doing could run alongside content removal by "de-prioritising" it in user feeds. One possibility – of pertinence in the context covered by this article - would be to amend the OSA to provide that if an automated system flags the content as satisfying the actus reus of one of the offences discussed above in Section 3, it should then be de-prioritised, even if not removed. Under Article 25 Digital Services Act (DSA), platforms cannot use, design, or operate interfaces that manipulate, deceive, or substantially impair a user's ability to make free and informed choices. Such 'dark patterns' are understood in the DSA's Recitals to be techniques that 'materially distort or impair... the ability of recipients... to make autonomous and informed choices'. This article has mentioned the addictive features of the platforms, but also algorithmic recommenders guiding children to content linked to suicide or self-harm. Expressly banning addictive features and such recommender

algorithms, due to an OSA amendment, echoing the DSA model to an extent, would clearly be a significant reform, especially in the context covered by this article.

This article has argued that the OSA scheme intended to protect children from content advocating suicide or self-harm is flawed in a range of respects. In a few years time it may become more readily apparent, as Ofcom concludes various investigations into company practice, that, while well-intentioned to an extent, the scheme is largely unable to deliver on the promises in relation to such protection which accompanied its introduction. The government launched an investigation in March 2026 into the child-protective measures stemming from the OSA; its findings may also be of value in future in judging the efficacy of the current scheme.[87] The recommendations for future reforms of the OSA made here, and those recently made by Ofcom, discussed above, could have some effect in ameliorating the position under the OSA scheme, as it currently stands. But at present the findings of this article concur with the calls for the strengthening of the OSA in relation to protection of children from the online harms discussed, coming from, among others, Molly Russel's father after she killed herself, having visited a number of suicide-promoting sites.[88] The UK government has stated recently that it intends to bring forward legislation in late 2026 to ban under-16s from social media, covering certain large social media sites, which may be in force by Spring 2027. But even if that occurs, it would not address all the harms discussed here because, leaving aside the likelihood that some children will be able to circumvent the ban, it would not catch all the sites promoting suicide or self-harming in children, since they are not social media platforms.[89]

References

Bhagwat, A., & Weinstein, J. (2021) Freedom of Expression and Democracy in Stone A., & Schauer F. (eds) *The Oxford Handbook of Freedom of Speech* (Oxford University Press, 2021).

Bell, J., & Westoby C. (2025) Public and mental health professionals perspectives on social media and suicide exposure. *BMC Public Health* 25:1380-89.

Blasi, V. (1977) The checking value in First Amendment theory *American Bar Foundation Research Journal* 521-649.

Crawford, A. (2023). *Molly Russell: Tech firms still failing after teenager's death, says father,* BBC News, https://www.bbc.co.uk/news/uk-67556756.

Children's Commissioner (2025). *Statement from the Children's Commissioner on Ofcom's new Children's Safety Codes*, https://www.childrenscommissioner.gov.uk/ news-and-blogs/statement-from-the-childrens-commissioner-on-ofcoms-new-childre ns-safety-codes/.

Department of Culture Media and Sport & Home Office (2020). *Online Harms White Paper*, https://www.gov.uk/government/consultations/online-harms-white-paper/online-har ms-white-paper.

Department of Science, Innovation and Technology (2026). *Growing up in the online world: a*

[87]Department of Science, Innovation and Technology (2026).

[88]See: Kuenssberg (2025).

[89]Snapchat, TikTok, YouTube, Instagram, Facebook and X are to be covered, but the ban will probably cover some other large platforms deemed social media ones.

national conversation CP 1528, https://www.gov.uk/government/consultati ons/gro wing-up-in-the-online-world-a-national-consultation/growing-up-in-the-online-world -a-national-conversation.

Digital, Culture, Media and Sport Committee (2022). *Oral evidence: Online safety and online harms*, HC 620, https://committees.parliament.uk/event/6804/formal-meeting-oral-evi dence-session/.

European Parliament (2023). *The influence of social media on the development of children and young people*, https://www.europarl.europa.eu/RegData/etudes/STUD/2023/73 3109/IPOL_STU(2023)733109_EN.pdf .

Fenwick, H., & Coe, P. (2025). The Online Safety Act: effectively navigating tensions between online expression fostering democratic aims and creation of online anti-democratic harms? *Northern Ireland Legal Quarterly* 76(4):741-793.

Gannon, J., O'Hanlon, F. & Conroy, N. (2025). TikTok, teens and mental health: quantifying algorithmic exposure to harmful content. *Irish Journal of Medical Science*, https://doi. org/10.1007/s11845-025-04244-4.

Garg, S., Singh, S. (2025). The Role of Piaget's Stages of Development in Understanding the Cognitive Growth of Children in the Digital Age. *International Journal of Research Culture Society* 9(4):48-55.

Greenawalt, K. (1989). Free speech justifications. *Columbia Law Review* 89(1): 119-115.

Higson-Bliss, L., & Street, L. (2026). *Education, Not Exclusion: Rethinking How We Protect Young People Online*, Inforrm, https://inforrm.org/2026/04/30/education-not-exclusion-re thinking-how-we-protect-young-people-online-laura-higson-bliss-and-louisa-street/.

House of Lords, House of Commons (2021). *Joint Committee on the Draft Online Safety Bill*, HL Paper 129, HC 609, https://committees.parliament.uk/publications/8206/do cuments/84092/default/.

Kuenssberg, L. (2025). *Molly Russell's dad warns UK 'going backwards' on online safety and urges PM to act*, BBC News, https://www.bbc.co.uk/news/articles/cp3j5kp8501o.

Liu, X., Huang, J., Yu, N., Li, Q., Zhu, T. (2020). Mediation Effect of Suicide-Related Social Media Use Behaviors on the Association Between Suicidal Ideation and Suicide Attempt: Cross-Sectional Questionnaire Study. *Journal of Medical Internet Research* 22(4):e14940.

Marchant, A., Hawton, K., Stewart, A., Montgomery, P., Singaravelu, V., Lloyd, K., et al. (2017). A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: The good, the bad and the unknown. *PLoS ONE* 12(8):1-26.

Marsh, S & Waterson, J. (2019). *Instagram bans 'graphic' self-harm images after Molly Russell's death*. The Guardian, https://www.theguardian.com/technology/2019/feb/07/instagram-ba ns-graphic-self-harm-images-after-molly-russells-death.

Mikuska, J., Smahel, D., Dedkova, L., Staksrud, E., Mascheroni, G., Milosevic, T. (2020). Social relational factors of excessive internet use in four European countries'. *International Journal of Public Health* 65:1289–1297.

Milmo, D. (2022). *'The bleakest of worlds': how Molly Russell fell into a vortex of despair on social media.*

Molly Rose Foundation (2025). *Pervasive-by-design* https://mollyrosefoundation.org/wp-con tent/uploads/2025/08/proof3_PervasivebyDesign.pdf.

North London Coroner's Service (2022). *Molly Russell - Prevention of future deaths report*. https://www.judiciary.uk/wp-content/uploads/2022/10/Molly-Russell-Preven tion-of-future-deaths-report-2022-0315_Published.pdf.

O'Connor, R., Kirtley O. (2018). The integrated motivational-volitional model of suicidal behaviour. *Philosophical Transactions of the Royal Society of London B* 373:1-10.

OECD (2025). *How's life for children in the digital age?*, https://www.oecd.org/en/publications/how-s-life-for-children-in-the-digital-age_0854b900-en.html.

Ofcom (2024). *Children and Parents: Media Use and Attitudes Report.* https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/children/children-media-use-and-attitudes-2024/childrens-media-literacy-report-2024.pdf?v=368229.

Ofcom (2025a). *Illegal content Codes of Practice for user-to-user services*, https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/statement-protecting-people-from-illegal-harms-online.

Ofcom (2025b). Online Nation. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2025/online-nations-report-2025.pdf?v=409837.

Ofcom (2025c). *Consultation: Protecting children from harms online*, https://www.ofcom.org.uk/online-safety/protecting-children/protecting-children-from-harms-online.

Ofcom (2025d). *Protecting people online from online suicide and self-harm material,* https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/protecting-people-from-online-suicide-and-self-harm-material.

Ofcom (2025e). *Illegal content codes of practice for search services.* https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/illegal-content-codes-of-practice-for-search-services-24-feb.pdf?v=391888.

Ofcom (2025f). *Illegal content codes of practice for user-to-user services.* https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/illegal-content-codes-of-practice-for-user-to-user-services-24-feb.pdf?v=391889.

Ofcom (2025g). *Protection of Children Code of Practice for user-to-user services.* https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/statement-protecting-children-from-harms-online/main-document/protection-of-children-code-of-practice-for-user-to-user-services.pdf?v=399754.

Ofcom (2026a). *Ofcom provisionally finds suicide forum in breach of Online Safety Act.* https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/ofcom-provisionally-finds-suicide-forum-in-breach-of-online-safety-act.

Ofcom (2026b). *Consultation: Online Safety - Additional Safety Measures*. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/online-safety-additional-safety-mea sures.

Ofcom 2026c). *Consultation on Online Safety: Additional Safety Measures: Response by the Samaritans.* https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/consultation-online-safety---additional-safety-measures/responses/samaritans.pdf?v=409230.

Post, R. (2006). Democracy and Equality. *The Annals of the American Academy of Political and Social Science* 603(1): 24-36.

Rayman-Jones, I., & McMahon, L. (2025). *Instagram teen accounts still show suicide content, study claims*. BBC News, https://www.bbc.com/news/articles/ce32w7we01eo

Sedgwick, R., Epstein, S., Dutta, R., Ougrin, D. (2019). Social media, internet use and suicide attempts in adolescents. *Current Opinion in Psychiatry* 32(6):534-541.

Stoilova, M., Bulger, M., & Livingstone, S. (2024). Do parental control tools fulfil family expectations for child protection? A rapid evidence review of the contexts and outcomes of use. *Journal of Children and Media* 18(1):29–49.

Sueki, H. (2015). The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan. *Journal of Affective Disorders* 170:155-160.

Susi, K., Glover-Ford, F., Stewart, A., Bevis, R., Hawton, K. (2023). Research Review: Viewing self-harm images on the internet and social media platforms: systematic review of the

impact and associated psychological mechanisms. *Journal of Child Psychology and Psychiatry* 64(8):1115-1139.

Thom, K. (2011). Suicide online: Portrayal of website-related suicide by the New Zealand media. *New Media & Society* 13(8) 1356-1370.

Uhls, Y., Michikyan, M., Morris, J., Garcia, D., Small, G., Zgourou, E., Greenfield, P. (2014). Five days at outdoor education camp without screens improves preteen skills with nonverbal emotion cues. *Computers in Human Behavior* 39:387-392.

Yang, N., Crespi, B. (2025). I tweet, therefore I am: a systematic review on social media use and disorders of the social brain. *BMC Psychiatry* 29(95):1-37 https://doi.org/10.1186/s12888-025-06528-6.

Athens Journal of Law - Volume 12, Issue 3, July 2026 – Pages 189-208

Administrative Responsibility for the use of AI in Public Administration – A Theoretical Analysis

*By Elena Emilia Ștefan**

Nowadays, it has become common for modern means of communication to be used in interactions with public administration, through which individuals can schedule visits to the premises of public authorities or, conversely, send or request documents remotely, within the virtual space. The interplay between new technologies and the public and private environments determines the need for a firm regulatory framework that clearly defines rights and obligations, limits, and liability - a regulatory framework which must, however, be subject to ethical scrutiny. In this context, the proposed aim of the study is to document the issue of liability in the context of the use of artificial intelligence systems in public administration, based on legislation, legal doctrine and administrative practice, through an examination of both national and comparative law. Therefore, the proposed topic of liability is highly relevant and generates widespread interest not only in the public sector but also in the private one and it should consistently remain in the attention of the scientific community, especially since humanity is facing new challenges in reinterpreting the law and adapting it to a new type of personality, non-human, namely the artificial intelligence. The findings of the study underline the increasing importance of establishing a solid regulatory framework for AI liability, at the international level, from a legal and ethical perspective, as a foundation for the development of innovation in this field.

Keywords: *administrative act, administrative liability, AI Act, risk assessment, ANAF (National Agency for Fiscal Administration).*

Introduction

In recent years, the proliferation of new technologies has generated new ethical challenges that must be addressed at a regulatory level. Field-specific literature is increasingly engaged in analysing the connection between law, ethics and business. Artificial intelligence has led to normative approaches that extend beyond the level of states. For example, at the EU level, the law on artificial intelligence, AI ACT[1] applies to multiple states, which are in turn obliged to implement it at the national level.

Thus, we can no longer speak merely about the possibility of damage being caused using artificial intelligence as long as various situations that have caused prejudice are now being publicised internationally, bringing to the forefront the issue of liability for automated decisions, such as autonomous vehicles but also the issue of tax scoring. However, "international factors cannot be disregarded in the contemporary context, in

*Associate Professor, Faculty of Law, "Nicolae Titulescu" University of Bucharest, Romania.

[1]Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/1144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, OJ L 2024/1689, 12.7.2024.

which the life of every human community is closely linked to the fate of humanity as a whole[2]".

The French Tribunal des Conflits established on 8 February 1873, through a cornerstone decision (s.n. Arrêt Blanco), that administrative liability is governed by the rules of public law rather than those of civil law. In French legal doctrine, it has been noted that the Blanco decision determined that: "*for damage caused to individuals by the acts of persons performing certain public services, the principles established by the Civil Code for relations between individuals cannot be applied; this liability is neither general nor absolute. It is subject to special rules which vary according to the needs of the service and the necessity of reconciling the rights of the State with private law*"[3]. Likewise, another author has assessed that in French law "the Blanco Decision of 1873 delivered by the Tribunal des Conflits (the court empowered to resolve jurisdictional conflicts between administrative courts and ordinary courts) is at the origin of a genuine jurisprudential revolution, radically changing the orientation on the issue of *state liability for damages"*[4].

The 153rd anniversary of the famous Blanco decision was commemorated on 8 February 2026, providing the impetus for the present research analysis, the curiosity to examine a new form of administrative liability in relation to artificial intelligence. From this perspective - of theoretical discourse - extrapolating the concept of administrative liability from the Blanco decision to the implications of the use of artificial intelligence in today's public services, such an analysis is both compelling and capable of attracting the interest of the global scientific community.

Nowadays, the issue of liability for damages caused by AI has become scientifically challenging. An analysis of the legal doctrine reveals differences in the regulatory philosophy between the legal instruments and the way the legislator perceives artificial intelligence from one continent to another, such as, for example, Europe or the United States of America. As regards Europe, the member states are obliged to harmonize their legislation with that of the Union. Otherwise, liability intervenes and a sanction can be applied[5]. In this regard, as shown "the Court of Justice first established the principle of Member States' liability for infringements of European Union law in the 1991 Francovich judgment[6]".

The aim of the study is to discover a potential solution for liability in case damage is caused by the use of artificial intelligence. The paper proposes several research questions related to the topic under review, such as: "*is administrative liability for damages created by AI regulated in a uniform manner at a global level?*", "*what is the vision of EU and United States legislation regarding AI*?", "*in the case of an automated decision that causes damages, who is responsible: the human or the machine?*".

The proposed objectives of the study are:

1. Brief documentation of the normative framework which regulates liability in the context of artificial intelligence.

[2]See Muraru (coord.), Muraru, Bărbățeanu & Big (2020) at 2.
[3]See Chapus (1988) at 978 apud Vedinaș (2009) at 288.
[4]See Apostol Tofan (2024) at 366.
[5]See an interesting paper on sanctions: Nunez (2025).
[6]See Popescu (2011) at 214.

2. Identification and clarification at a general level of the concepts of administrative act, automated decisions, administrative liability.
3. Identifying vulnerabilities or risks of using AI in the administrative decision-making system - case study.

This paper starts from the hypothesis that existing legal studies have focused on many topics, such as the principle of good administration with an emphasis on transparency and motivation for AI decisions.[7]At the same time, among the topics addressed in recent years by researchers are: automated decisions, the issue of tax scoring, issues related to personal data protection, and many others.

In order to achieve the research objective, from a methodological point of view, the paper has several sections, organized in a chronological order as follows. Section I – *The Administrative Act and Administrative Liability, Fundamental Concepts of Administrative Law* (where the administrative act and automated decisions, as well as administrative liability were theoretically analysed in the context of AI). Section II is dedicated to *the analysis of the regulatory framework applicable to artificial intelligence from the perspective of automated decision-making and liability* (the Union and American plan). Section III is dedicated to *a case study* on the administrative act issued based on a risk score. The final part of the paper is dedicated to the conclusions.

The empirical data were collected by accessing several databases in which relevant studies were identified, indexed in internationally recognized databases. The identification of documentary sources was carried out in Romanian, French and English by using keywords such as: liability, artificial intelligence, automated decisions, liability for defective products, etc. As the number of bibliographic sources increased, the initial research plan was gradually adapted until it reached a final form that also included the research of several online resources from various geographical areas. The key contributions of the paper consist of: the interdisciplinary analysis of the subject, the extension of the documentation to the American jurisdiction and the identification of a case study from Romania regarding risk analysis based on tax algorithms. Thus, the proposed theme encourages reflection because it brings information from comparative law about the regulation of artificial intelligence but also about the issue of liability for the use of AI in administration, the analysis being interdisciplinary.

At the same time, the empirical data collected were interpreted according to research methods specific to law, which revealed the vulnerabilities associated with the use of AI systems in the decision-making process. In this sense, the data were filtered using the logical-deductive method that supported the process of discussions and conclusions.

[7]See Pedro (2023) at 159.

The Administrative Act and Administrative Liability, Fundamental Concepts of Administrative Law

Several Theoretical Considerations on the Administrative Act

The unilateral manifestation of the administration's will to which it appeals when placed in a position to make a decision is the administrative act. When damage occurs in connection with an administrative act or with the poor functioning of public services, the guilty are held accountable.

The unilateral expression of will of the public administration, resorted to whenever it is required to make a decision, constitutes the administrative act. When damage occurs in connection with an administrative act or with the improper functioning of public services, those responsible are held accountable. In administrative law, as a general principle, pursuant to art. 52 of the revised Constitution of Romania, even the state may incur liability for damages caused to individuals, the legal remedy through which such liability may be pursued being the administrative litigation action. Furthermore, judicial control is imperative in order to ensure compliance with the principle of legality in administrative action. Or, as the legal doctrine has stated, "in any system of administrative law, the court exercises control over the factual and discretionary determinations made by the administration."[8]

In Romania, pursuant to Law No. 554/2004[9] on Administrative Litigation, art. 2 para. (1) letter c), the administrative act is defined as: *unilateral act of an individual or normative nature, issued by a public authority for the purpose of executing or organising the execution of the law, giving rise to, amending or extinguishing legal relations*. Field-specific literature[10] has adopted this definition and integrated it into administrative law courses. At the same time, in the Romanian legal system the legislator also considers the administrative-fiscal act which, according to art. 46 paragraph (1) of the Fiscal Procedure Code[11] "is issued in writing, on paper support or in electronic form". Whether materialized in physical or electronic form and regardless of whether it is a pure administrative act or a fiscal administrative act, both forms have one common element: they must comply with the principle of legality.

Automated Decisions-Making

Within state activity, "the most important activity of public administration entities is the adoption of administrative decisions that have either negative or positive consequences for natural or legal persons[12]". In common language, automated decisions are administrative acts issued based on algorithms. Essentially, automated decision-making involves a complex process that does not ignore the fact that it is closely linked to technology, and this leads to the imperative requirement of compliance with the legality of the issuance procedure. At the same time, just as, in the case of the traditional

[8]See Craig (2015) at 477.
[9]Administrative Litigation Law No. 554/2004, Official Gazette No. 1154 of 7 December 2004.
[10]For example see Săraru (2022) at 72.
[11]Law No. 217/2015 on the Fiscal Procedure Code, Official Gazette No. 547 of 23 July 2025.
[12]ee Bareikyte (2021) at 65.

administrative act, legislation imposes the obligation to provide reasons (in fact and in law), we consider that this obligation must also apply to automated decisions. From a normative perspective, the administration is obliged to provide reasons for its decisions, and this derives from the right to good administration, as provided for in art. 41 of the Charter of Fundamental Rights of the European Union[13].

According to the *Principles of Administrative Law concerning Relations between Individuals and Public Authorities* contained in the manual developed by the European Commission, "algorithmic decision making", or "algorithmic decision-making system", means a process of making a decision with the support of automated means. "It usually involves the use of automated reasoning to aid or replace a decision-making process that would otherwise be performed by humans. It does not necessarily involve the use of artificial intelligence but generally involves the collection and processing of data[14]". Therefore, according to the perspective presented in that manual, automated decisions are made by a machine and not by a human being, as a result of a complex process involving the use of data.

Regarding the frequency of this method of making decisions, it has been observed that "automated decision-making systems (ADM) have been increasingly utilized by both private and public entities across the world to reduce errors by humans, increase efficiency, and make more consistent decisions"[15]. Also, from a conceptual perspective "Within the umbrella term of part-ADM, the role of the automated system's output in the process and how humans use it largely differs. In automated triage, the system classifies a new case or application based on the automated assessment; the human can get a case assigned or be required to take follow-up actions[16]".

According to Malgieri[17], "*The French Law[18] regulates automated decision-making in a different manner considering three different cases: (1) automated decisions in the judicial field; (2) administrative automated and semi-automated decisions and (3) all other kinds of automated decisions with legal effects or significant effects on individuals". (...) "For administrative decisions three is a difference between semi-automated decisions and fully automated decisions. Fully automated decisions are prevented within the administrative appeal (....)*[19]".

In the case law of the French Constitutional Council with reference to automated decisions, it was appreciated that: "an exclusive basis form an individual administrative decision, algorithms likely to revise by themselves the rules to

[13]Charter of Fundamental Rights of the European Union (2012/C 326/02), OJ C 326/391, 26.10.12, https://eur-lex.europa.eu/legal-content/RO/TXT/PDF/?uri=CELEX:12012P/TXT

[14]Council of Europe (2024). The Administration and You, A handbook, 3rd edition, *Principles of administrative law concerning relations between individuals and public authorities,* ISBN 978-92-871-9460-2, p.*7,* https://rm.coe.int/handbook-the-administration-and-you-3rd-edition-005924-gbr-web/1680b04d3f.

[15]See Bantekas & Bratsiakou (2026).

[16]See Palmiotto (2024) at 210-236.

[17]See Malgieri (2019).

[18]Loi°2018-493 du 20 juin 2018 relative a la protection des donnees personnelles, JORF n° 0141 du 21 juin 2018, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037085952

[19]See Malgieri (2019).

which they apply cannot be used, without the oversight and validation of the data processor[20]" (para.71).

Moreover, automated decisions are increasingly used in many areas of activity. In this sense, "Examples include the "visa streaming algorithm" and automatic detection of "sham marriages" in the UK, the risk-assessment tool used in the Netherlands to screen employment sponsorships, the EU-funded project iBorder Ctrl, and the automated case-management system for the EU settlement scheme. In automated evidence, the system provides information or an expert assessment humans use to prove a fact relevant to the decision[21]".

Administrative Liability in the Context of AI – General Considerations

Regarding legal liability, according to university courses on the general theory of law, the conditions required for engaging liability are[22]: unlawful conduct, result of unlawful conduct, causal relationship, fault of the perpetrator. From the perspective of administrative law, when analysing administrative liability, the following elements are important: the unlawful act, the fault, the causal relationship and the damage[23]. In traditional administrative law, administrative liability is classified into objective liability and fault-based liability.

According to the Romanian Administrative Code[24], the following forms of administrative liability exist:

- *objective liability* that is incurred regardless of the guilt of the public authority, and this includes: the patrimonial liability of the state for damages caused by judicial errors as well as the exclusive patrimonial liability of public authorities for the limits of public service;
- *subjective liability* which implies the fault of the responsible public authority and includes: the administrative-patrimonial liability of public authorities and institutions, respectively of public officials for damages caused by administrative acts as well as the joint and several administrative-patrimonial liability for damages caused in connection with the valorisation of public goods and services.

Analysing Romanian law, as mentioned above, we can say that there is: fault-based administrative liability, no-fault administrative liability, as well as joint and several administrative liability of the administration.

Currently, as states toy with the idea of introducing AI into administration and damage inevitably occurs, the question arises as to whether the conditions for administrative liability still apply from a legal standpoint. From this perspective,

[20]Decision no 218-765 DC of 12 June 2018, https://www.conseil-constitutionnel.fr/en/decision/2018/2018765DC.htm
[21]See Palmiotto (2024).
[22]See Bădescu (2025) at 373-375.
[23]See Ștefan (2013) at 240-268.
[24]Government Emergency Ordinance No. 57/2019 on the Administrative Code, Official Gazette No. 555 of 5 July 2019.

we consider that the topic of administrative liability in the context of AI use is interesting because several problems revealing vulnerabilities and risks must be clarified, namely: who is to blame if an AI system creates harm? In this regard, who is to blame, man or machine? Does the lack of human oversight over an AI system in administration constitute fault or risk?

A recent study on algorithmic accountability noted that "the challenges arising from algorithm use give rise to deficits that strike at the heart of accountability processes: compounded informational problems, the absence of adequate explanation or justification of algorithm functioning, and ensuing difficulties with diagnosing failure and securing redress[25]".

Fault-Based Liability for Damages caused by Automated Administrative Decisions

A valuable recent study highlighted the major legal challenges in the context of analysing administrative liability for AI-based decisions, namely: "how do we define fault in a digital context?

1. Is a programming error considered administrative fault?
2. Does the deployment of an untested or unsupervised system constitute negligence?[26]".

This form of liability was considered to arise when "the damage arises when harm can be traced to an error attributable to the administrative authority, whether that error lies in its legal acts or its material operation[27]". According to this author, fault denotes "a service fault (faute de service), a defect imputed to the public service as an institution, irrespective of the individual staff involved, arising from its failure to discharge the functions it is legally bound to perform in a proper manner"[28]. Another study supports to the idea that "the application of the concept of fault-based unlawful acts is difficult to implement effectively due to the autonomous and non-transparent nature of AI"[29].

It was assessed that, "When the damage flows from intelligent automated administrative decisions, liability consists in the administration's obligation to indemnify individuals for injury resulting from decisions rendered by automated systems that rely on artificial-intelligence techniques or algorithms, with no direct human intervention, whenever that injury is traceable to: a software error, a malfunction in the automated processing, or the misuse of the intelligent systems through which the decision is generated[30]".

From the perspective of the nature of the damage that can be satisfied in cases of fault-based liability, we believe that compensation for both material damage and moral damages may be awarded by a court of law.

[25]See Busuioc (2021) at 825-836.
[26]See Hamid (2025).
[27]See Khader (2025) at 620.
[28]Ibidem at 621.
[29]See Pelupessy (2026).
[30]See Khader (2025) at 616.

As a general perception, we consider that in cases involving the liability of the administration, the notion of fault may also be understood to include situations in which there is no human oversight of AI systems and such systems are capable of causing damage, as well as situations in which fault may be interpreted as a lack of transparency in decision-making and, implicitly, a lack of reasoning. We particularly refer to scoring mechanisms (tax risk analysis) that form the basis for the adoption of an administrative act, as will also be illustrated in the case study presented in the final part of this paper.

No-Fault Liability of the Public Authority for Damages caused by Automated Decisions

This form of liability is based on the idea of risk. In this case, we are talking about the existence of a public service that by its very nature contains the risk of causing certain harm to the beneficiaries. Consequently, beneficiaries have the obligation to demonstrate before the courts the causal link between the automated decision and the damage suffered. Furthermore, the causal relationship must demonstrate the existence of a public service which, by its nature, contains the risk of causing certain harm. Similar to fault-based liability for damages caused by automated administrative decisions and in the case of no-fault liability of the authority, we support the idea of obtaining in the court of law compensation for material damages as well as moral damages.

Moreover, it has been argued that "the risks associated with the absence of a human in control are projected into risks of accountability[31]". Among the categories of risks, we note that a serious impediment to attracting liability lies in the particular nature of AI systems' operation which complicates the burden of proof regarding causality, as the relevant data are excessively technical and therefore inaccessible to an ordinary person lacking specialised expertise.

Joint and Several Liability of the Public Authority

An analysis of the relevant field-specific literature reveals that, in addressing the issue of administrative liability arising from the use of artificial intelligence, certain authors advocate for the liability of AI system developers. Thus, it has been argued that "developers should be held accountable for the algorithmic design and functionality of their GenAI systems. This requires clear disclosures of their AI's modelling and reasoning process, enabling scrutiny for potential bias or flaws in the algorithms. However, liability allocation networks must avoid overly punitive developer liability frameworks[32]". We do not concur with this doctrinal position, as we consider that, within the context of artificial intelligence usage, fault cannot be attributed exclusively to a single actor; rather, there exists a chain of culprits. Consequently, given the current stage of legislation, we consider that joint liability could arise on the part of the public administration alongside other participants, such as the software programmer responsible for the system underlying the operation of the AI system, the implementer, and others involved in the process.

[31]See Mureddu, Paciaroni, Pavelka, Pemberton & Remotti (2025).
[32]See Socol de la Osa & Remolina (2024).

In conclusion, based on the above assertions, we consider that it remains difficult, from the perspective of liability, to formulate a definitive answer regarding fault or what fault is and to whom it may be attributed when an AI system causes damage. The answer depends on the nature of the liability involved, whether objective or subjective. However, under no circumstances, at least at the present stage of legislative development, is it possible to hold a machine itself liable.

The Regulatory Framework Applicable to Artificial Intelligence from the Perspective of Automated Decision-Making and Liability

The EU Regulatory Approach

The delineation of the legal regime governing liability for the use of artificial intelligence in public administration, particularly with regard to the identification of applicable legal instruments, is currently difficult, given the absence of a unified legal framework at Union level. Nevertheless, we consider that among the most relevant instruments applicable to the subject under analysis are the GDPR[33]and the AI Act[34], to which various other regulations and directives may be added. In this respect, it is worth noting that "a Regulation has general applicability[35]". Scholars rightly point out that: "(...) automated decision-making (ADM) and AI-supported decision-making create new dilemmas, especially in relation to accountability, data protection, and general principles of administrative law[36]".

Regarding the AI Act and GDPR, "Unlike the GDPR, which sets requirements for solely automated decisions, the AI Act primarily concerns AI systems that pose an unacceptable or high risk, considering AI-driven decision making as a potential source of risk"[37]. Moreover, "Even if the legislation does not provide a definition of AI decision-making, the role of AI systems in influencing decisions is a core concept in the classification rules for high-risk AI set in Article 6 of the AI Act[38]".

In addition, European liability law is also shaped by Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024[39] on liability for defective products and repealing Council Directive 85/374/EEC. This represents the common law on liability for damages and applies to the case of no-fault liability for manufacturers in the European Union. In essence, this regulatory act states that when there is a defective product that causes harm to a consumer,

[33]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/CE (General Data Protection Regulation), OJ L 119/1, 4.5.2016.
[34]We understand that the aim is not to provide an extensive development of the identified legal instruments, but rather to select, at a general level, the most relevant information for the proposed research topic.
[35]See Fuerea (2010) at 141.
[36]See Jakubek – Lalik (2024) at 109**.**
[37]See Palmiotto (2024).
[38]Ibidem.
[39]Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC, JO L, 2024/2853, 18.11.2024.

the manufacturer may be held liable. From our perspective, we advance the idea that the respective directive could also be applied to AI software, given the lack of a special normative act regulating the legal regime of liability applicable to the use of artificial intelligence. This interpretation emerges from the provisions of art. 4 of Directive (EU) 2024/2853 which provides that "products" include any movable item, even where integrated into another movable or immovable item or interconnected with it, and which encompasses electricity, digital production files, raw materials, and software. Consequently, we consider the possibility of holding the public authority and the manufacturer jointly and severally liable for the implementation of a defective product (AI system) in the administration.

It should be noted that, from a chronological perspective, following the adoption of the AI Act, there were notable intentions to improve the legislation, materialising in the form of a proposed Directive[40]on the adaptation of non-contractual civil liability rules to artificial intelligence. The Preamble of this proposal stated that "the current liability rules, in particular those based on fault, are not suitable for addressing claims for damages caused by AI-based products and services (...)". The legal doctrine considered in this sense that: "The EU's proposed AI Liability Directive partly alleviates the burden of proof by adopting the 'presumption of causality', noting precisely that it is increasingly more onerous for individuals, due to reasons of lack of transparency, complexity and autonomy of AI systems, to demonstrate harm or call them into account[41]". However, the proposal for a directive on liability was withdrawn on 6 December 2025 and at the time of drafting this study there is no other regulatory proposal in this field of liability.

Therefore, at present, two key Union instruments remain relevant in relation to the legal regime of liability, namely the AI Act and Directive (EU) 2024/2853 on liability for defective products. There is a close relationship between these two instruments, as under the Directive's framework, presumptions of defect may arise from non-compliance with the requirements laid down in the AI Act.

At the same time, one of the key elements underpinning the legality of algorithmic decision-making is transparency. *Per a contrario*, in the context of artificial intelligence use, a lack of transparency may result in the infringement of certain rights, such as the right to be informed and the right to receive a statement of reasons for administrative acts, thereby potentially leading to the occurrence of damage. In this regard, according to Regulation EC no. 1049/2001 of the European Parliament and of the Council of 30 May 2001 regarding public access to European Parliament, Council and Commission documents[42] "any citizen of the Union and any natural or legal person residing or having its registered office in a Member State shall have a right of access to documents of the institutions, subject to the principles, conditions and limits defined in this Regulation" [Article 2(1)].

---

[40]COM/2022/496 final - Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual liability rules to artificial intelligence (AI Liability Directive), 2022/0303(COD), Brussels, 28.9.2022.

[41]See Teo (2025) at 2265-2280.

[42]Regulation (EC) No. 1049/2001 of the European Parliament and of the Council of 30 May 2001 regarding public access to European Parliament, Council and Commission documents, OJ L 145/43, 31.05.2001, pp.43-48.

Essentially, it seems that in Europe, the emphasis regarding AI is on the protection of fundamental human rights, on its operation in a safe environment and under conditions of transparency, while any breach of the relevant legal requirements gives rise to liability.

Discussions on Administrative Liability for Artificial Intelligence from the United States Perspective

In the American jurisdiction there is no federal law on liability for the use of AI in administration. For this reason, at the level of the legislation of the states forming the federation, certain responses may be identified with regard to the applicable regulatory framework in this area. Therefore, this section first outlines the applicable normative framework governing artificial intelligence in the United States, followed by an analysis of the legislation relevant to liability arising from the use of AI in public administration. In this context, a general overview may be obtained of the manner in which the American legislator approaches the use of emerging technologies in public life, and consequently of the legal issues that may arise in the event of harm.

As Donovan observes: "The United States' approach to AI regulation remains fragmented, hindered by shifting political priorities and the absence of a cohesive federal framework[43]".

Regarding legal instruments governing the use of artificial intelligence in public administration, two key developments may be identified in recent years. In this context, reference should also be made to the United Nations Resolution[44] adopted in 2024 on the promotion of safe, secure and trustworthy AI systems, which contains provisions relating to accountability in the use of artificial intelligence. Chronologically, on 30 October 2023, the Biden Administration issued Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (...)[45]", which seeks to ensure that AI systems within government are designed and deployed in ways that respect human rights and prevent harm[46]".

Upon the change of administration, as appears from public information, in his first days in office, President Donald Trump[47]signed Executive Order 14179 of 23 January 2025 - *Removing barriers to American leadership in artificial intelligence*[48].

Moreover, at the state level there is an interesting vision on new technologies, with an emphasis on innovation. Specialists point out: "America's AI Action Plan[49] (released July 2025) remains a key roadmap for federal 'innovation-first' actions[50]". Reading the

[43]See Donovan (2025).
[44]Available online https://docs.un.org/en/A/78/L.49. *Excerpt,* Point 6 letter K) emphasizes: "Promoting transparency, predictability, reliability and understandability throughout the life cycle of artificial intelligence systems that make or support decisions impacting end-users, including providing notice and explanations (....) human decision making alternatives or effective redress and *accountability for those adversely impacted by automated decisions of artificial intelligence systems*".
[45]https://www.congress.gov/crs-product/R47843
[46]See Mureddu, Paciaroni, Pavelka, Pemberton & Remotti (2025).
[47]https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf
[48]Federal Register 90 (20) 8741, https://www.govinfo.gov/content/pkg/FR-2025-01-31/pdf/2025-02172.pdf
[49]https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf
[50]https://www.softwareimprovementgroup.com/blog/us-ai-legislation-overview/

document, it appears that the American Action Plan is based on three pillars: Pillar I - *Accelerate AI Innovation*; Pillar II - *Build American AI Infrastructure* and Pillar III - *Lead in International AI Diplomacy and Security*. Simultaneously, "The Action Plans' objective is to articulate policy recommendations that this Administration can deliver for the American people to achieve the President's vision of global AI dominance[51]".

By contrast, an analysis of legislation at the federal states level, reveals possible answers regarding the legal regime of liability for the use of AI in local government. In this respect, according to the legal doctrine, "In the United States, the absence of a federal law on AI has led to a patchwork of state-level laws (e.g., in California and Colorado) and federal agency guidance (…) with new executive orders driving infrastructure localization and AI safety standards[52]".

One of the most significant examples is represented by the legislation adopted in the State of Colorado. It seems that "Colorado made history on May 17, 2024 when Governor Polis signed into law the Colorado Artificial Intelligence Act ("CAIA"), the first law in the United States to comprehensively regulate the development and deployment of high-risk artificial intelligence ('AI') systems"[53]. The field-specific literature emphasized: "(…) CAIA will require companies doing business in Colorado to meet stringent compliance and oversight requirements intended to prevent algorithmic discrimination in AI systems considered to be 'high risk'"[54]. Although the law was supposed to come into force on 1 February 2026, it seems that its implementation has been postponed. Public information shows that "Colorado Governor Jared Polis recently signed Senate Bill 25B-004 into law, which delays the enforcement date of the Colorado Artificial Intelligence Act ("CAIA") from February 1, 2026, to June 30, 2026. SB 25B-004 does not amend the substantive requirements of the CAIA[55]".

According to experts, "Colorado is not the only state entering the AI regulatory space with trepidation. In 2024, the Connecticut state legislature killed a similar bill to CAIA under threat of veto by the governor. In 2025, the Virginia legislature passed a less-stringent version of CAIA, but it was vetoed by the governor, citing similar concerns as Governor Polis in his CAIA signing statement"[56].

At the same time, legislation concerning artificial intelligence has also been adopted in the State of Texas through the Texas Responsible AI Governance Act (TRAIGA), which entered into force on 1 January 2026. It was stated about TRAIGA that, "Under the new legislation, state government agencies must disclose to consumers when they are interacting with an AI system, regardless of whether such interaction appears obvious. This transparency requirement represents a fundamental shift toward algorithmic accountability in government operations and ensures that

[51]https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf
[52]See Kumar (2026).
[53]https://fpf.org/blog/a-first-for-ai-a-close-look-at-the-colorado-ai-act/
[54]https://www.reedsmith.com/our-insights/blogs/technology-law-dispatch/102lu0u/colorados-hesitation-in-pioneering-ai-regulation-mirrors-similar-hesitation-na/
[55]https://www.hunton.com/privacy-and-cybersecurity-law-blog/enforcement-of-colorado-ai-act-delayed-until-june-2026
[56]https://www.reedsmith.com/our-insights/blogs/technology-law-dispatch/102lu0u/colorados-hesitation-in-pioneering-ai-regulation-mirrors-similar-hesitation-na/

citizens are aware when automated systems are processing their information or making decisions that may affect them[57]".

Considering the above, although no unified federal legal framework currently exists concerning liability for the use of artificial intelligence in public administration, significant legislative efforts may nevertheless be observed at the level of the individual federal states. In conclusion, "While global initiatives like the UN resolution and the EU AI Act provide structured oversight, America oscillates between prioritizing risk mitigation and innovation with each administration[58]".

Discussions on the Administrative Act issued based on a Risk Score - Case Study

Between 2004-2026, as a result of the proliferation of new technologies and digitalization, there is a considerable impact that they have on the administrative decision-making processes. In this context, there is the possibility of issuing an administrative act by using an algorithm. However, "digitalization brings with it new issues of responsibility, legal liability, limits and respect for citizens' rights and freedoms [59]" while "the legal norm requires acceptance and compliance with the prescribed conduct"[60].

From a legal point of view, the administrative act has traditionally been regarded as the will of the administration. Therefore, even when issued in a modern electronic form, the administrative act should still reflect the will of the public authority. But when the administrative decision is made through an automated process, does it reflect the internal will of the administration, so that in the event of damage, it can lead to legal liability?

Theoretically, we believe that the answer to this question is affirmative. The administration may adopt a decision based on an evaluation process founded upon a score obtained through an automated procedure; however, this may render the decision-making process vulnerable. For example, in Romania there is a risk analysis that tax authorities carry out through which economic operators are assigned to different risk categories on the basis of a risk score.

According to tax legislation[61], a risk assessment is carried out according to criteria provided for in Order no. 417/2025[62] on the adoption of criteria for assessing

[57]See Othman (2025).
[58]See Donovan (2025).
[59]See Ștefan (2024) at 565.
[60]See Hegheș N. (2022) at 153.
[61]Joint Order of the President of the National Agency for Fiscal Administration and the President of the Romanian Customs Authority No. 1.826/2.372/2025 amending and supplementing the annex to the Order of the President of the National Tax Administration Agency and the President of the Romanian Customs Authority No. 417/1,204/2025 on the approval of the criteria for assessing tax risk in order to determine economic operators presenting a high tax risk, provided for in art. 375 paragraph (1^1) and in art. 435 paragraph (3^1) of Law No. 227/2015 on the Fiscal Code, Official Gazette No. 708 of 30 July 2025.
[62]Joint Order of the President of the National Agency for Fiscal Administration and the President of the Romanian Customs Authority No. 417/2025 on the adoption of the criteria for assessing the fiscal risk that presents a high fiscal risk, provided for in art. 375 paragraph (1^1) and in art. 435

fiscal risk. In essence, based on certain criteria, an analysis is carried out which results in an official score generated by algorithms. Following the analysis, the economic operator is classified into one of three risk classes: low, medium or high risk.

As a result of the risk analysis carried out by the National Agency for Fiscal Administration (ANAF) and the classification into a particular risk class, legal consequences may arise. Verifications (ANAF controls) can be carried out. The outcome of such verification is recorded in a Verification Report which subsequently forms the basis for the issuance of an administrative act such as, where appropriate, a tax decision or a decision to terminate the documentary verification procedure. Thus, a fiscal scoring system, the reasoning of which is not disclosed to the taxpayer, generates a chain of effects in relation to that taxpayer, the procedure ultimately culminating in the issuance of an administrative act.

In this context, may one speak of the intervention of a form of administrative liability?

The answer to this question can be provided by the Romanian Administrative Code, which regulates, under Article 573, administrative-patrimonial liability as a form of administrative liability. This consists in "*the obligation of the State or, as the case may be, of the administrative-territorial units to compensate for damage caused to a natural or legal person through any judicial error, deficiencies in the public service, an unlawful administrative act, or the unjustified refusal of the public administration to resolve a request concerning a right recognised by law or a legitimate interest"*.

At the same time, regarding the conditions of administrative-patrimonial liability, according to art. 577 of the Administrative Code, the cumulative conditions under which administrative-patrimonial liability may be engaged are: a) the contested administrative act is illegal; b) the illegal administrative act causes material or moral damages; c) the existence of a causal relationship between the illegal act and the damage; d) the existence of the fault of the public authority and/or its staff.

If these general conditions of liability are applied to the case under examination, several observations may be made.

With regard to the first condition, namely that *the administrative act must be illegal*, we consider this requirement to be satisfied through the defective nature of the issuance procedure, insofar as the obligation to provide reasons for the administrative act has not been fulfilled. Furthermore, the lack of transparency in the issuance of the administrative act may result in harm to the taxpayer through the infringement of the right to be informed, since the taxpayer does not have the opportunity to know the criteria for issuing the document, the procedure being devoid of transparency and offering no effective means of challenge.

There is *a causal relationship between the administrative act and the damage* because the risk analysis that generates a risk score automatically triggers the initiation of an ANAF control and based on the Verification Report an administrative act is issued. In relation to the causality report, we estimate that it may be difficult for the injured person to demonstrate that there is a malfunction, an error in the algorithm used by ANAF due to the lack of knowledge of the strictly technical conditions in

---

paragraph (3^1) of Law No. 227/2015 on the Fiscal Code, Official Gazette No. 262 of 26 March 2025.

which it operates. Rather, we can state that in administrative practice, the opaque nature of the AI system may be an impediment to demonstrating causality.

At the same time, we appreciate that *the condition of fault* is also met in this case by the existence of a defect in the public system that enabled the operation of the AI system itself. We consider that based on general data manually input by a human operator, an operation that may itself be defective or incomplete, the automatic system analyses and generates a risk score. This outcome, in accordance with the applicable legislation, is the basis for the decision to initiate an ANAF audit. And therefore, if the AI system is defective, it is evident that the final outcome does not reflect reality, thereby causing harm and consequently lacking legality.

Therefore, we consider that the answer to the question is affirmative: the administration may be held liable. On the one hand, the requirement to justify the administrative act is only met to a small extent, namely the application of the criteria provided for in the regulatory act (legal requirements). On the other hand, the only possibility for the taxpayer is to accept the control that is triggered and possibly to challenge, if necessary, a possible onerous administrative act that could be issued (a tax decision). The existence of the control can be found on the ANAF website which periodically announces the subject of the controls[63]carried out.

Extending the theoretical analysis to the practical, applied level, we have examined the portal of the Romanian courts, the administrative and fiscal litigation sections (www.just.ro) to identify potential litigations concerning liability actions in the context under discussion. Our intention was to identify litigations in which ANAF appeared as a party in proceedings challenging such administrative acts allegedly causing harm. At the time of drafting this study it was not possible to identify in the national plan any case law before the competent administrative and fiscal courts in relation to the issues outlined above. From our perspective, this may have an explanation, meaning that Romanian legislation is still at an early stage regarding the ANAF control procedure based on an automatic risk assessment, which has the potential to end with a sanctioning administrative act. Moreover, "such an approach raises serious compatibility issues with Law no. 554/2004, because the administrative act is issued by a public authority, the documents that were the basis for its issuance, respectively those contained in the administrative file, are the result of artificial intelligence as a tool[64]". The lack of transparency of the decision-making process, combined with the lack of proper reasoning, may constitute sufficient legal grounds for initiating an administrative litigation action to hold the public authority accountable.

Moreover, it should be noted that the Romanian legislator has recently proposed the suspension, until the end of 2026, of the taxpayer's right to request classification into a risk category or subcategory[65]. This development raises genuine concerns regarding the impossibility of ascertaining one's risk classification, as well as the broader lack of decisional transparency. In such circumstances, it may be argued that a potential situation of abuse of power on the part of the public authorities arises.

[63]https://static.anaf.ro/static/3/Anaf/20260223111817_com%20183%20verificari%20antifrauda%2023%20februarie%202026.pdf

[64]See Crețu (2026).

[65]https://www.digi24.ro/stiri/economie/guvernul-vrea-sa-suspende-pana-la-finalul-lui-2026-dreptul-de-a-solicita-de-la-anaf-comunicarea-clasei-de-risc-fiscal-ce-motive-invoca-3649205

Consequently, we consider that it may be possible to speak of an extended form of administrative liability on the part of the public authority, namely the National Agency for Fiscal Administration (ANAF), for the issuance of an administrative act based on an algorithmic system, in breach of the principle of transparency and in the absence of genuine reasoning of the act.

Conclusions

Following the documentation of the issue of liability for the use of AI in the decision-making process, several conclusions can be drawn, the research objective being thus considered fulfilled. In essence, with regard to liability in the context of AI, the legal doctrine remains in a formative stage. From the analysis of the legislation, a conceptual delimitation has been achieved between the administrative act, automated decision-making, and administrative liability.

Regarding the applicable regulatory framework, administrative liability for damages caused by AI is not uniformly regulated at international level. The most important EU legislative acts are the AI Act and the Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 regarding liability for defective products (...). Also, there are notable differences between Union and United States legislation on artificial intelligence. On the one hand, at the Union level, the AI Act and the GDPR are insufficient to outline a general legal regime applicable to liability for the use of AI in administration, which has led to legislative efforts aimed at introducing a dedicated directive on liability, albeit without concrete results to date.

Although there is the Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products (...) it cannot provide all the answers for liability cases because it only regulates the no-fault liability for defective products of manufacturers in the European Union. However, we have advanced the idea that in this case, we see the *possibility of attracting joint and several liability of the public authority and the manufacturer for the implementation in the administration of a defective product (AI system).*

On the other hand, with regard to the United States of America, the documentary analysis reveals differences in the regulation of artificial intelligence and, implicitly, of liability, depending on whether one refers to the federal level or to the level of individual states. The study has highlighted efforts undertaken by certain states to improve the relevant legal framework, such as, for example, the Colorado Artificial Intelligence Act (CAIA).

With regard to the types of administrative liability for the use of artificial intelligence in public administration, certain limitations of the present research have been identified, insofar as the analysis has focused on two main forms of liability, namely objective liability and subjective liability, respectively fault-based liability and strict (no-fault) liability, on the basis of which several conceptual directions have been advanced. Nevertheless, it is considered that, at the current stage of legislative development, a firm answer regarding administrative liability for the use of AI may prove difficult in practice.

With respect to the question of whether the lack of human supervision over an AI system in public administration constitutes fault or risk, the case study suggests that, although it is difficult to classify within a strict conceptual category, it may rather be regarded as fault. We consider that fault is due to the existence of an error within the public system which permitted the operation of a defective AI system, thereby generating the risk of harm and, implicitly, the potential engagement of administrative liability. Furthermore, in this context, it has been argued that extended administrative liability may arise in cases of fault, particularly where there is a lack of human oversight of AI systems or a lack of transparency.

From a comparative law perspective, it has been observed that there are various approaches to liability for the use of artificial intelligence in public administration, depending on the value systems, historical development, and legal traditions of each state. While at Union level the regulatory framework on AI places emphasis on the protection of fundamental human rights and the maintenance of a democratic and secure environment, in the United States the primary focus is placed on innovation.

In conclusion, on the basis of the extensive documentary analysis conducted, it appears that we are currently witnessing a reconsideration of liability regimes in the context of artificial intelligence use, particularly in relation to the potential harm that may be caused by algorithms used within public services.

At the same time, from both an ethical and legal perspective, it is not considered possible to hold a machine legally liable for its actions, but rather the human actor behind it.

References

Bădescu, M. (2025). *Teoria generală a dreptului. Tratat [General Theory of Law. Treaty]*. Bucharest: Hamangiu Publishing House.

Bantekas, I., & Bratsiakou, V. (2026). 'Automated Decision-Making Systems and Black Box Challenges under European Union Administrative Law', *Fordham International Law Journal* 49 (1). https://ir.lawnet.fordham.edu/ilj/vol49/iss1/1

Bareikyte, S. (2021). 'Toolbox of the public administration entity. Intersection of the principle of legality and administrative discretion in exercising the revocation of an administrative decision: the case of Lithuania', in *Bratislava Law Review* 5 (2):65. DOI:10.46282/blr.2021.5.2.255

Busuioc, M. (2021), 'Accountable Artificial Intelligence: Holding Algorithms to Account'. *Public Admin Rev*, 81 (5): 825-836. https://doi.org/10.1111/puar.13293.

Chapus, R. (1988). *Droit administratif.* Paris: Montchrestien, tome II.

Craig, P. (2015). *UK, EU and Global Administrative Law. Foundations and Challenges,* Cambridge University Press.

Crețu, D. (2026). '*Legea contenciosului administrativ în era inteligenței artificiale. Despre nașterea unor noi tipologii de litigii administrative' [Administrative litigation law in the era of artificial intelligence. On the birth on new typologies of administrative litigation]*. https://www.juridice.ro/813905/legea-contenciosului-administrativ-in-era-inteligentei -artificiale-despre-nasterea-unor-noi-tipologii-de-litigii-administrative.html

Donovan, J. (2025). 'The shortcomings of United States artificial intelligence regulations: an international comparison', *Fordham International Law Journal.* https://potato-flamingo-p93z.squarespace.com/iljblog/k73pddfmra65ecd-fxc5k-bdazp-npp5j-l8be9-kbndm?rq=Donovan

Fuerea, A. (2010). *Manualul Uniunii Europene [The European Union Handbook*]. 6th edition revised and supplemented. Bucharest: Universul Juridic Publishing House.

Hegheș, N.E. (2022*)*. 'The non - retroactivity of new legal norms - fundamental principle of law. Exceptions', in *International Journal of Legal and Social Order* (1):153-160. https://ijlso.ccdsara.ro/index.php/international-journal-of-legal-a/article/view/74/60

Khader, S. (2025). 'Administrative Liability for Harm Arising from Intelligent Automated Administrative Decisions', *Zaouli,* 4 (10): 601-629, ISSN: 2788-9343. https://www.revue-zaouli.com/wp-content/uploads/2025/10/19-Somia-KHADER_Zaouli-n%C2%B010-Vol.-4-Aout-2025-1.pdf

Kumar, A. (2026). 'Legal and Regulatory Frameworks Governing Generative AI for Enterprises'. In: Singh, S., et al. GenAI and LLMs for Beyond 5G Networks. *Springer*, Cham. https://doi.org/10.1007/978-3-032-06418-9_3

Jakubek – Lalik, L. (2024). 'The Challenges of AI in administrative law and the need for specific legal remedies: analysis of polish regulation and practice', *Central European Public Administration Review* 22 (2):109. DOI: https://doi.org/10.17573/cepar.2024.2.05.

Muraru I. (coord), Muraru A., Bărbățeanu V., & Big D., (2020). *Drept constituțional și instituții politice. Caiet de seminar [Constitutional law and political institutions. Seminar booklet]*. Bucharest: C.H. Beck Publishing House.

Mureddu F., Paciaroni A., Pavelka T., Pemberton A., & Remotti L.A. (2025). Rights and responsibilities: legal and ethical considerations in adopting local digital twin technology. In Raes L., Ruston McAleer S., Croket I., Kogut P., Brynskov M., Lefever S., (eds) Decide Better. *Springer*, Cham. https://doi.org/10.1007/978-3-031-81451-8_11

Malgieri, G. (2019). 'Automated-decision making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations', C*omputer Law & Security Review,* 35 (5) 105327*,* https://www.sciencedirect.com/science/article/pii/S0267364918303753#cit_105

Nunez, J.E. (2025). *'A multidimensional view on sanctions'. In: Bersier, N., Bezemek C., Schauer F., (eds) Sanctions: An essential element of law? Law and Philosophy Library,* vol.149 Springer, Cham. https://doi.org/10.1007/978-3-031-88512-9_7

Othman, A. (2025). '*The Texas Responsible AI Governance Act: Pioneering State-Level Artificial Intelligence Regulation in the United States*'. DOI: 10.13140/RG.2.2.14596. 85125

Hamid, A.M.A. (2025). 'Administrative liability for damages caused by artificial intelligence systems in public services: an analytical study in light of the principles of legality and transparency', *Humanities and Social Sciences* 13 (4). https://www.sciencepublishinggroup.com/article/10.11648/j.hss.20251304.21

Socol de la Osa D.U., & Remolina, N. (2024). 'Artificial Intelligence at the bench: Legal and ethical challenges of informing-or misinforming-judicial decision making through generative AI', *Data & Policy,* Vol.6: e59, Cambridge University Press. Doi:10.1017/dap.2024.53

Popescu, R.M. (2011). *Introducere în dreptul Uniunii Europene [Introduction to European Union Law]*. Bucharest: Universul Juridic Publishing House.

Pelupessy, E. (2026). 'Civil responsibility in artificial intelligence-driven decision making', *JHK*, 3 (2). https://doi.org/10.61942/jhk.v3i2.544

Pedro, R. (2023). 'Artificial intelligence on public sector in Portugal: first legal approach', *Juridical Tribune* 13 (2):159. https://www.tribunajuridica.eu/arhiva/An13v2/1.%20Ricardo%20Pedro.pdf

Palmiotto, F. (2024) 'When is a decision automated? A taxonomy for a Fundamental Rights Analysis', *German Law Journal,* 25(2):210-236. doi:10.1017/glj.2023.112

Săraru, C.S. (2022). *Tratat de contencios administrativ [Administrative Litigation Treaty]*. Bucharest: Universul Juridic Publishing House.

Vedinaş, V. (2009). *Drept administrativ*, 5th edition revised and updated. Bucharest: Universul Juridic Publishing House.
Teo, S.A. (2025). 'Artificial Intelligence and it's 'slow violence' to human rights', *AI Ethics,* Vol.5, pp.2265-2280. doi.org/10.1007/s43681-024-00547-x
Apostol Tofan, D. (2024). *Drept administrativ [Administrative Law]*, vol. II, 5th edition, Bucharest: C.H.Beck Publishing House.
Ștefan, E.E., (2013). *Răspunderea juridică. Privire specială asupra răspunderii în dreptul administrativ [Legal liability. A special look at liability in the administrative law]*. Bucharest: Pro Universitaria Publishing House.
Ștefan, E.E. (2024). 'Integrity and Transparency in the Work of Public Authorities. Aspects of Comparative Public Law', in Juridical Tribune – *Review of Comparative and International Law* 14 (4):565. DOI: 10.62768/TBJ/2024/14/4/03

Legal Instruments

Charter of Fundamental Rights of the European Union (2012/C 326/02), OJ C 326/391, 26.10.12, https://eur-lex.europa.eu/legal-content/RO/TXT/PDF/?uri=CELEX:12012P/TXT
Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC, OJ L, 18.11.2024.
COM/2022/496 final - Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual liability rules to artificial intelligence (AI Liability Directive), 2022/0303(COD), Brussels, 28.9.2022.
Regulation (EC) 1049/2001 of the European Parliament and of the Council of 30 May 2001 regarding public access to European Parliament, Council and Commission documents, OJ L 145/43, 31.05.2001.
Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/CE (General Data Protection Regulation), OJ L 119/1, 4.5.2016.
Regulation (EU) 2024/1689 of The European Parliament European and of The Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) no. 300/2008, (EU) no. 167/2013, (EU) no. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/1144 and Directives 2014/90/UE, (EU) 2016/797 and (EU) 2020/1828, OJ L 2024/1689, 12.7.2024.
Loi°2018-493 du 20 juin 2018 relative a la protection des donnees personnelles, JORF n° 0141 du 21 juin 2018, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037085952
Administrative Litigation Law No. 554/2004, Official Gazette No. 1154 of 7 December 2004.
Law No. 217/2015 on the Fiscal Procedure Code, Official Gazette No. 547 of 23 July 2025.
Government Emergency Ordinance No. 57/2019 on the Administrative Code, Official Gazette No. 555 of 5 July 2019.
Joint Order No. 1.826/2.372/2025 of the President of the National Agency for Fiscal Administration and the President of the Romanian Customs Authority for the amendment and supplementation of the appendix to Order No. 417/1.204/2025 of the President of the National Agency for Fiscal Administration and the President of the Romanian Customs Authority on the approval of criteria for assessing fiscal risk in order to determine economic operators presenting a high fiscal risk, that are referred to in art. 375 para.(1^1) and art. 435 para. (3^1) of Law no. 227/2015 on the Fiscal Code, Official Gazette No. 708 of 30 July 2025.

Joint Order No. 417/2025 of the President of the National Agency for Fiscal Administration and the President of the Romanian Customs Authority on the approval of criteria for assessing fiscal risk in order to determine economic operators presenting a high fiscal risk, that are referred to in art. 375 para. (1^1) and of art. 435 para. (3^1) of Law No. 227/2015 on the Fiscal Code, Official Gazette No. 262 of 26 March 2025.

Online Sources

Council of Europe (2024). *The Administration and You*, A handbook, 3rd edition, *Principles of administrative law concerning relations between individuals and public authorities,* ISBN 978-92-871-9460-2. https://rm.coe.int/handbook-the-administration-and-you-3rd-edition-005924-gbr-web/1680b04d3f.

Decision no 218- 765 DC of 12 June 2018, https://www.conseil-constitutionnel.fr/en/decision/2018/2018765DC.htm

https://www.congress.gov/crs-product/R47843

https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

https://www.digi24.ro/stiri/economie/guvernul-vrea-sa-suspende-pana-la-finalul-lui-2026-dreptul-de-a-solicita-de-la-anaf-comunicarea-clasei-de-risc-fiscal-ce-motive-invoca-3649205

https://static.anaf.ro/static/3/Anaf/20260223111817_com%20183%20verificari%20antifrauda%2023%20februarie%202026.pdf

https://www.govinfo.gov/content/pkg/FR-2025-01-31/pdf/2025-02172.pdf

https://www.softwareimprovementgroup.com/blog/us-ai-legislation-overview/

https://fpf.org/blog/a-first-for-ai-a-close-look-at-the-colorado-ai-act/

https://www.reedsmith.com/our-insights/blogs/technology-law-dispatch/102lu0u/colorados-hesitation-in-pioneering-a-i-regulation-mirrors-similar-hesitation-na/

https://www.hunton.com/privacy-and-cybersecurity-law-blog/enforcement-of-colorado-ai-act-delayed-until-june-2026

https://docs.un.org/en/A/78/L.49

Athens Journal of Law - Volume 12, Issue 3, July 2026 – Pages 209-224

Ensuring Justice and Non-Discrimination in Automated Decision-Making: A Fundamental Rights Perspective

*By Doina Popescu Ljungholm**

I will start with a simple question: what happens to fundamental rights when an algorithm, not a person, decides whether you receive social benefits, a loan, or even the right to enter a country? This paper examines exactly that problem, in the context of automated decision-making (ADM) systems within the European Union. In short, I argue that the current legal safeguards primarily Article 22 of the GDPR and the new AI Act contain structural gaps that hit the most vulnerable the hardest: ethnic minorities, migrants, and persons with disabilities. Why? Because these systems are opaque, the right to be heard becomes an empty formality, and enforcement authorities lack both the technical training and the resources to verify how complex algorithms actually work. Drawing on CJEU and ECtHR case law, plus two landmark cases the Dutch SyRI system and the UK Home Office visa-streaming tool I argue for three institutional reforms: extending mandatory Fundamental Rights Impact Assessments to all high-risk commercial deployments, defining clearly what "meaningful human oversight" really means, and, not least, equipping supervisory authorities with the resources they need to conduct genuine technical audits.

Keywords: *artificial intelligence, automated decision-making, non-discrimination, fundamental rights, EU law*

Introduction

The deployment of automated decision-making (ADM) systems in domains of fundamental importance welfare allocation, credit assessment, border control, and employment raises a precise and pressing legal question: whether existing EU law provides adequate protection for the individuals subject to algorithmically generated determinations that affect their rights and interests. When a municipal algorithm classifies a resident as a probable welfare fraudster without her knowledge, or when a credit-scoring system declines an application without any human having reviewed the file, the issue at stake is not merely one of administrative inefficiency. It is one of accountability, opacity, and the structural capacity of legal frameworks to respond to a form of governance that is at once pervasive and resistant to conventional oversight. This paper is concerned with exactly that structural resistance.

Here is the central argument I want to make. The existing EU legal framework, for all its stated ambitions, is simply not built well enough to protect fundamental rights in the ADM context. Take the GDPR[1] and its prohibition of "solely automated" decisions under Article 22. That threshold dissolves the moment a deployer hires

*Associate Professor, National University of Science and Technology Politehnica Bucharest, Romania.

[1]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation – GDPR), OJ L 119, 4.5.2016, p. 1.)

someone to rubber-stamp whatever the algorithm has already decided an empty formality. The AI Act[2] is more ambitious in its regulatory reach, but its heaviest obligations fall on system providers not on the commercial deployers whose systems most directly affect people's lives. What we end up with is a framework that speaks the language of human dignity and non-discrimination with considerable sophistication, yet offers the people most at risk legal protections that are, in practice, difficult to invoke and even harder to enforce.

The Charter of Fundamental Rights of the European Union[3] gives me the normative benchmark for the critique that follows. Articles 1, 21, and 47 human dignity, non-discrimination, effective remedy are primary EU law. Any secondary instrument that falls structurally short of those standards is, to the extent of that shortfall, legally deficient. And I argue that both the GDPR and the AI Act exhibit precisely that shortfall and that closing the gap requires institutional redesign, not just tinkering with legislation.

Who bears the costs of these deficiencies matters enormously. Ethnic minorities, migrants, persons with disabilities, and those living in poverty are overrepresented among those subject to algorithmic decision-making in welfare, employment, border control, and housing. In each of these domains, a wrong decision carries consequences of a completely different order of severity from, say, an incorrectly recommended playlist. Hildebrandt has argued convincingly, in my view that automated systems do not just passively reflect social realities; they actively generate them, embedding the biases of historical data into decisions about individual futures.[4] Rouvroy's concept of "algorithmic governmentality"[5] gives this dynamic its most precise theoretical formulation: what is produced is a mode of governance that operates as if it were making no normative choices and is therefore accountable for none of the choices it makes.

The accountability problem gets even worse when we add what Katyal calls the privatisation of consequential decisions,[6] and what Pasquale captured in that wonderful image of the "black box":[7] proprietary systems whose logic is inaccessible not only to the individuals they affect, but also to the courts asked to review their outputs, and to the regulators charged with overseeing their deployment. When you cannot know

---

[2]Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act), OJ L, 12.7.2024. The Act entered into force on 1 August 2024 and applies progressively until 2 August 2027.

[3]Charter of Fundamental Rights of the European Union, OJ C 326, 26.10.2012, p. 391. By virtue of Article 6(1) TEU, the Charter has the same legal value as the Treaties.

[4]Mireille Hildebrandt, Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology (Edward Elgar, 2015) 56–58

[5]Antoinette Rouvroy, 'The end(s) of critique: Data behaviourism versus due process' in Mireille Hildebrandt and Katja de Vries (eds), Privacy, Due Process and the Computational Turn (Routledge, 2013) 143–167. The concept of 'algorithmic governmentality' draws on Foucault's analysis of governmental rationality to describe a mode of rule that operates through the automated production of norms from data, bypassing discursive justification and individual subjectivity.

[6]Sonia Katyal, 'Private Accountability in the Age of Artificial Intelligence' (2019) 66 UCLA Law Review 54, 59–63.

[7]Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (Harvard University Press, 2015) 3–8.

the basis of a decision, the right to be heard becomes formal, not real – a procedural entitlement without substantive content.

There is also a specific discrimination problem that deserves separate attention. Classical anti-discrimination law is built on identifying a protected characteristic and showing differential treatment on that ground. Algorithmic systems frequently produce discriminatory outcomes through a different mechanism: proxy variables – postcode, browsing history, purchasing patterns that are neutral on their face but correlate statistically with race, disability, or social class. Gerards and Zuiderveen Borgesius have shown just how comprehensively this mechanism evades existing legal categories,[8] and Eubanks has documented the human consequences of this phenomenon with an ethnographic precision that legal scholarship rarely achieves.[9]

The Court of Justice of the European Union has not been idle, to be fair. The SCHUFA judgment of December 2023[10] settled a significant interpretive dispute by holding that automated credit-scoring falls within Article 22 of the GDPR where it functions as a determinative input to a third party's decision whatever the formal structure of that decision might be. The Dun & Bradstreet Austria ruling of February 2025[11] moved the transparency question forward, establishing that trade secrecy cannot serve as a blanket defence against a data subject's right to understand how an automated decision was reached. Both rulings matter. But each operates at the level of individual challenge – after the harm has occurred, in cases where the affected person had the resources, knowledge, and stamina to litigate. The ECtHR Grand Chamber in D.H. and Others v Czech Republic[12] recognised, in a different context, that discrimination of a structural kind demands structural responses. That recognition has not yet found its way into the operational design of EU AI governance.

What follows maps the legal framework, evaluates how adequate it is for vulnerable groups, and proposes three institutional reforms: extending mandatory Fundamental Rights Impact Assessments to commercial ADM deployments; defining meaningful human oversight in substantive operational terms; and resourcing supervisory authorities for genuine technical audit.

---

[8]Janneke Gerards and Frederik Zuiderveen Borgesius, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence' (2020) 29 Information & Communications Technology Law 303, 305–310.

[9]Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor (St. Martin's Press, 2018) 11–14.

[10]CJEU, Case C-634/21, OQ v Land Hessen (SCHUFA Holding AG), ECLI:EU:C:2023:957, Judgment of 7 December 2023. The Court held that an automated credit-score constitutes an 'automated individual decision' within Article 22(1) GDPR where third parties determinatively rely on it.

[11]CJEU, Case C-203/22, Dun & Bradstreet Austria GmbH v CK, ECLI:EU:C:2025:137, Judgment of 27 February 2025. The Court held that national law cannot categorically exclude access to explanations in favour of trade secrets; controllers must provide the 'procedure and principles actually applied' in an intelligible and accessible form.

[12]ECtHR, D.H. and Others v Czech Republic [GC], Application No. 57325/00, Judgment of 13 November 2007, paras 175–180. Indirect discrimination producing disproportionately adverse effects on a group falls within Article 14 ECHR; statistical evidence may establish a prima facie case

Literature Review

Scholarship on ADM and fundamental rights has grown into a genuinely interdisciplinary conversation – one that spans computer science, legal theory, political philosophy, and empirical social research. The conversation is productive but uneven. Technical findings on algorithmic harm accumulate faster than the normative frameworks needed to evaluate them. Doctrinal legal analysis of GDPR and AI Act provisions proceeds, with some exceptions, without adequate engagement with how deployed systems actually behave. This review identifies three bodies of work directly relevant to the argument developed below and closes with an account of the gaps that give this paper its reason for being.

The Right to Explanation: From Promise to Paradox

Perhaps the most consequential single contribution to the legal-technical literature on ADM is Wachter, Mittelstadt and Russell's 2018 paper.[13] What made it consequential was its willingness to say plainly what the GDPR does and does not provide: a right to meaningful information about the logic involved in automated decisions, certainly but not a right to an explanation in any sense that would allow an affected individual to understand, challenge, or replicate the decision. A controller satisfies Article 15(1)(h) by communicating the general factors that influenced an output, without disclosing how those factors were weighted or combined. The remedy the authors propose counterfactual explanations that describe the minimum change to the input that would have produced a different result is elegant precisely because it threads the needle between transparency and commercial confidentiality. It is also, as the CJEU's Dun & Bradstreet Austria ruling suggests, increasingly influential on judicial thinking, though without explicit acknowledgment.

The same authors' 2021 paper carries a harder message.[14] The major statistical fairness metrics demographic parity, equalised odds, individual fairness are mutually incompatible when base rates differ across groups. This is a mathematical result, but its legal implications are severe: any system optimised to satisfy one fairness definition necessarily violates another, and EU non-discrimination law gives us no guidance on which definition should prevail. The AI Act inherits this silence. Its high-risk system requirements mandate bias monitoring and testing, but leave open the question of what fairness metric the monitoring is designed to detect. Without a principled answer to that question, the regulatory obligation is formally present and substantively empty. Edwards and Veale put the point from a different angle:[15] demanding explanations of discriminatory decisions does not make those decisions less discriminatory. It relocates the problem from the system to the individual, who

[13]Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31(2) Harvard Journal of Law and Technology 841–887.

[14]Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI' (2021) 41 Computer Law & Security Review 105567.

[15]Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 18, 20–24.

must now obtain, interpret, and legally operationalise an explanation a burden that is, for the populations most affected, prohibitive in practice.

Algorithmic Discrimination: Mapping the Harm

Zuiderveen Borgesius's Council of Europe study remains the most comprehensive mapping of algorithmic discrimination risk across sectors.[16] Its core finding is so direct it is worth preserving in plain language: anti-discrimination law cannot, by itself, adequately address algorithmic bias, because the causal structure of that bias does not match the causal structure anti-discrimination law was built to address. Machine learning systems trained on historically discriminatory data acquire proxies – postcode, purchasing behaviour, web browsing patterns that reproduce discriminatory effects without replicating discriminatory intent or even using protected categories as inputs. The law, as it stands, regulates inputs; the harm occurs through outputs. Closing that gap requires different law, not better algorithms. The computer science tradition's foundational concept of individual fairness[17] similar individuals should be treated similarly offers a partial conceptual bridge, but the operationalisation of "similarity" involves normative judgments that cannot be read off from technical specifications and that AI Act provisions on bias testing currently leave entirely unresolved.

The Regulatory Literature: Advances and Structural Limits

Two recent papers in the Athens Journal of Law take up precisely the structural tensions that this paper analyses. Sarra's examination of the relationship between AI Act Article 14 and GDPR Article 22[18] identifies what may prove to be the most consequential unintended consequence of the new architecture. The AI Act's mandatory human oversight requirement was designed as a safeguard against fully automated decision-making. But "solely automated" is the threshold for Article 22 GDPR's protections. A formally inserted human reviewer even one whose involvement is perfunctory removes a decision from Article 22's scope.

The human oversight requirement, intended as protection, operates as an exemption. Sarra's proposed re-interpretation of Article 22 focusing on the substantive independence of human review from the algorithmic output, rather than on its formal presence is, in my assessment, the most practically workable solution currently on offer. Boura's regulatory analysis[19]identifies a different structural choice: the sandbox model that allows providers to test AI systems in real-world conditions with reduced compliance obligations places the cost of technological development on the individuals who interact with those systems – frequently in welfare, healthcare, and criminal justice contexts, exactly those settings where the individuals concerned are

[16]Frederik Zuiderveen Borgesius, Discrimination, Artificial Intelligence, and Algorithmic Decision-Making (Council of Europe, Directorate General of Democracy, 2018) 13–18.
[17]Cynthia Dwork and others, 'Fairness Through Awareness' (Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ACM, 2012) 214–226.
[18]Claudio Sarra, 'Artificial Intelligence in Decision-making: A Test of Consistency between the EU AI Act and the General Data Protection Regulation' (2025) 11(1) Athens Journal of Law 45–62.
[19]Marta Boura, 'The Digital Regulatory Framework through EU AI Act: The Regulatory Sandboxes' Approach' (2024) 10(3) Athens Journal of Law 385–398.

most vulnerable and least able to refuse. Convention 108[20] provides the international floor: its provisions on automated decision-making operate regardless of whether the decision is "solely" automated, offering a degree of protection that EU secondary legislation has not yet consistently reached.

Gaps and the Contribution of This Paper

Reading this literature as a whole, three gaps become apparent. The Charter of Fundamental Rights legally binding primary law, against which all EU secondary legislation may be measured and found wanting appears in the scholarship largely as background aspiration rather than operative legal standard. The category of "vulnerable groups" is invoked regularly but examined concretely in relatively few legal studies; the human reality behind the doctrinal categories remains underexplored. And the field has produced substantial critique of existing law alongside limited constructive synthesis: what would a rights-compliant ADM framework actually look like in institutional detail? The following sections attempt, if not to answer that question definitively, at least to make it less avoidable.

Methodology

This study employs a legal-doctrinal methodology, which is the standard approach in European Union fundamental rights research when the object of inquiry is the interpretation, coherence, and application of legal norms. The core of the analysis consists of a systematic examination of primary EU law (the Charter of Fundamental Rights), secondary legislation (GDPR and AI Act), and the interpretive case law of the CJEU and the ECtHR. To test how these legal provisions actually work in reality, I use two comparative case studies the Dutch SyRI welfare-fraud detection system and the UK Home Office visa-streaming tool selected because each reveals, with unusual documentary clarity, a specific mechanism through which legal gaps translate into concrete rights violations. The selection of these two cases requires a brief methodological note. Both SyRI and the UK streaming tool are exceptional in one important respect: each attracted litigation and investigative scrutiny that produced a documentary record sufficient to reconstruct the system's operation, its legal basis, and the response of public authorities when challenged. Most ADM deployments in high-risk domains do not generate comparable documentation, precisely because opacity is structurally incentivised. The analytical value of these cases lies not in their representativeness they are, in that sense, unusual but in what they make visible: SyRI exposes the consequences of permitting systemic opacity and placing the burden of challenge on individuals who are never informed that they have been assessed; the streaming tool demonstrates what occurs when existing anti-discrimination law is formally in force but no pre-deployment mechanism requires compliance to be verified. Together, they instantiate two distinct failure modes of the regulatory architecture analysed in Section 4, and they provide the empirical basis for the

[20]Council of Europe, Convention 108+ on the Protection of Individuals with regard to Automatic Processing of Personal Data (modernised version, CETS No. 223, 2018), Articles 8 and 9.

reforms proposed in Section 6. The analysis is normative in orientation: it does not merely describe the current legal architecture but evaluates it against the benchmark of the Charter's guarantees of human dignity, non-discrimination, and effective remedy.

The EU Legal Framework: Architecture and Fault Lines

The EU legal framework governing ADM is layered: primary law establishes the normative standard, secondary legislation attempts to operationalise it, and judicial interpretation fills or tries to fill the gaps that drafting leaves behind. What I call "fault lines" are the points where these layers fail to connect: where the protection promised at one level evaporates at another, and where that evaporation consistently works to the benefit of the deploying institution rather than the affected individual. That consistency, I would argue, is not coincidental.

The GDPR: A Safeguard with Built-In Escape Routes

Article 22 of the GDPR[21] was a legislative first the first binding EU provision to address the risks of consequential automated decisions directly. Its limitations were visible from the start. The "solely automated" threshold is the most obvious: any formal human participation in the decision chain, however perfunctory, removes the case from Article 22's scope. The Article 29 Working Party's guidance[22] attempted to give "meaningful" human review some content, but stopped short of providing operational criteria that supervisory authorities could enforce against deployers unwilling to comply. A second escape route runs through Article 22(2)(a), which excepts decisions necessary for a contract a formulation broad enough to cover most credit, insurance, and employment ADM. The third problem is structural rather than definitional: Article 22 is reactive. It gives you a right to contest a decision already made, but imposes no obligation on controllers to assess, before deploying a system, whether it will produce discriminatory results. The harm arrives before the law is available to address it.

The AI Act: Real Advances, Conspicuous Exclusions

The AI Act represents a genuine step beyond the GDPR's data-protection-centred approach. Its risk-based architecture[23] subjects high-risk AI systems to pre-market conformity assessment, technical documentation requirements, and human oversight obligations. For certain deployers, Article 27 mandates a Fundamental

[21]GDPR, Articles 22(1)–(3). The 'solely automated' threshold has attracted persistent criticism precisely because it is circumvented so easily by nominal human participation. Article 22(2)(a) additionally excepts decisions necessary for a contract, which in practice covers most commercial ADM.

[22]Article 29 Working Party (now EDPB), Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, WP251rev.01 (revised 6 February 2018). The WP29 insisted human review must be 'meaningful' rather than perfunctory, but stopped short of specifying criteria that supervisory authorities could operationalise against deployers.

[23]AI Act, Article 6 and Annex III. High-risk systems include those deployed in biometric identification, education, employment, essential services, law enforcement, migration, and the administration of justice.

Rights Impact Assessment before deployment.[24] Article 5's prohibitions on social scoring by public authorities and on systems that exploit group vulnerabilities[25] are categorical: no proportionality balancing is permitted. These are genuine advances. Their limitation lies in their scope. The FRIA obligation under Article 27 applies only to public-body deployers. Commercial deployers in credit, insurance, recruitment, and platform governance the sectors where algorithmic discrimination against vulnerable groups is best documented fall entirely outside its reach. The human oversight requirements of Article 14 govern system design rather than the quality of review actually practised. There is no mechanism to verify that the human reviewer inserted into a high-risk AI system's decision chain does anything more than approve the algorithmic output. The resulting regulation is, as I argue, most protective where protection is least needed.

The Charter and the Courts: Filling the Gaps Cautiously

Judicial interpretation has begun to give the Charter's guarantees operative content in the ADM context, though progress is sector-specific and reactive. The SCHUFA judgment[26] clarified the scope of Article 22 GDPR in a way the legislature had left ambiguous: an automated score that functions as the determinative input to a third party's decision is itself an automated decision for Article 22 purposes, regardless of how the subsequent decision is formally structured. The Dun & Bradstreet Austria ruling[27] addressed the transparency question: commercial secrecy is not a sufficient reason to deny a data subject access to a meaningful explanation of how an automated decision was reached. The Ligue des droits humains judgment[28]extended the analysis to border-control profiling, holding that systematic automated risk classification constitutes a Charter-level interference requiring necessity and proportionality justification. Each of these rulings closes one escape route. None of them, however, operates upstream of the harm. They provide legal recourse after a wrong decision has been made and in the rare case successfully challenged. The ECtHR's D.H. judgment[29] points toward a more demanding model: structural discrimination, demonstrated statistically, requires systemic institutional responses. That model has not yet been translated into the operative architecture of EU secondary law on AI.

[24]AI Act, Article 27(1). The FRIA obligation is confined to public-body deployers and private entities providing public services — a formulation that leaves commercial deployers in credit, insurance, recruitment, and platform governance outside its mandatory scope.

[25]AI Act, Article 5(1)(c) and (e). These prohibitions are absolute: no proportionality balancing is available. They cover AI systems that exploit vulnerabilities of specific groups and social scoring by public authorities that produces detrimental treatment.

[26]CJEU, Case C-634/21, SCHUFA (n 10), paras 50–63.

[27]CJEU, Case C-203/22, Dun & Bradstreet Austria (n 11), para 71.

[28]CJEU, Case C-817/19, Ligue des droits humains v Conseil des ministres, ECLI:EU:C:2022:491, Judgment of 21 June 2022. Systematic automated PNR profiling was held to constitute a serious interference with Articles 7 and 8 of the Charter, requiring strict necessity and proportionality even where no individual decision is formally automated.

[29]ECtHR, D.H. and Others v Czech Republic (n 12), para 175.

Case Studies: When Law Meets Algorithm

The two cases examined here were chosen for what they reveal rather than for what they represent. Both are exceptional each attracted litigation and public attention that most ADM deployments do not. What makes them analytically useful is that each exposes, in documentary detail, a specific failure mode of the regulatory architecture described in Section 4. SyRI shows the consequences of permitting opacity and placing the enforcement burden on individuals who do not even know they have been harmed. The UK streaming tool shows what happens when the law clearly prohibits discriminatory conduct but imposes no obligation to check for compliance before deployment.

SyRI: The Algorithm the State Refused to Explain

Between 2014 and 2020, the Dutch government ran a system called SyRI Systeem Risico Indicatie which drew on data from tax authorities, municipal housing records, employment registers, and immigration files to generate risk scores identifying individuals as potential welfare fraudsters. The system operated without the knowledge of those it processed. People who received a risk score were not notified; they had no opportunity to see the score, understand how it had been calculated, or challenge it before investigators arrived. They were, to use the language of administrative law, the objects of a consequential classification made entirely behind their backs.In February 2020, the District Court of The Hague ruled SyRI unlawful.[30] The court grounded its ruling in Article 8 ECHR – the right to private and family life rather than in the GDPR, which had been in force for two years by that point and whose enforcement architecture had produced no action in six years of SyRI's operation. The court's central finding was that SyRI legislation failed the Convention's "quality of law" requirement: it was insufficiently clear about which data combinations could justify a fraud risk conclusion. More strikingly, the Dutch state had refused to disclose how the algorithm worked, even to the court itself. Van Bekkum and Zuiderveen Borgesius note the judgment's limited immediate effect:[31] the Dutch government subsequently developed successor systems under different legal bases.

Rachovitsa and Johann identify the deeper wound: "intentional opacity" designing a system so that its operation cannot be externally verified transforms the right to an effective remedy from a substantive guarantee into a procedural formality.[32]Both the ECHR and the GDPR, correctly applied, would have prohibited

[30]District Court of The Hague, NJCM et al. v The Dutch State, ECLI:NL:RBDHA:2020:1878, Judgment of 5 February 2020, paras 6.7–6.9. The state's refusal to disclose the algorithm's logic even to the court prevented judicial verification of discriminatory operation, which the court identified as independently problematic.

[31]Marvin van Bekkum and Frederik Zuiderveen Borgesius, 'Digital Welfare Fraud Detection and the Dutch SyRI Judgment' (2021) 23 European Journal of Social Security 1, 5–8. The authors argue persuasively that without a general doctrine of algorithmic accountability, individual court victories tend to produce system redesign rather than structural reform.

[32]Adamantia Rachovitsa and Niclas Johann, 'The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case' (2022) 22(2) Human Rights Law Review ngac010. The authors' concept of 'intentional opacity' designing a system so that its operation

SyRI. The failure was one of enforcement, and it was systemic: it reflected the structural position of a regulatory architecture that places the burden of challenge on the person who does not know she has been wronged.

The UK Streaming Tool: What Existing Law Required – and Was Never Applied

The UK Home Office's visa-streaming algorithm ran from 2015 to 2020. Over those five years, it assessed virtually every visa application by assigning a traffic-light risk rating green, amber, red that determined the intensity of scrutiny an application received and, consequently, its probability of refusal. The Home Office admitted, when pressed, that nationality was a primary variable: applicants from countries the algorithm designated "suspect" received higher risk scores, longer processing times, and substantially higher refusal rates. This is direct discrimination on grounds of national origin. The Equality Act 2010, section 13, unambiguously prohibits it. The Public Sector Equality Duty under section 149 required the Home Office to have due regard to the need to eliminate such discrimination before adopting the system. GDPR Article 35, operative from May 2018, required a Data Protection Impact Assessment for any processing likely to produce high risks to individuals' rights and freedoms.

None of these obligations were discharged before deployment. When JCWI and Foxglove filed for judicial review in June 2020,[33] they were asserting rights that had been in force for between two and ten years. The Home Office discontinued the tool before a full hearing, committing to conduct the equality and data protection assessments that should have preceded the system's original deployment.[34] No binding legal determination of the tool's unlawfulness was ever made. This is the most important aspect of the case, and the one most consistently overlooked when it is cited as an example of algorithmic accountability in action. It is, equally, an example of a public authority deploying a discriminatory system in confident knowledge that no enforcement mechanism would require prior justification and being proved right. The AI Act's ex ante FRIA requirement, had it existed and applied, would have forced the question before the harm was done. It did not exist; it now does; it applies from August 2026; it does not apply to commercial deployers; it offers no retroactive protection to anyone processed by any algorithm before that date.

---

cannot be externally verified is a valuable analytical contribution that this paper extends to the enforcement context.

[33]JCWI and Foxglove v Secretary of State for the Home Department, judicial review filed June 2020. The claim alleged direct racial discrimination under the Equality Act 2010, s. 13, and breach of the Public Sector Equality Duty under s. 149.

[34]Home Office letter to JCWI, 3 August 2020, discontinuing the Streaming Tool 'pending a redesign of the process'. The commitment to conduct Equality Impact Assessments and Data Protection Impact Assessments standard obligations already operative under the Equality Act 2010 and the GDPR — was presented as a new undertaking, which speaks for itself.

Discussion: Three Reforms and Their Discontents

A pattern runs through the legal framework analysis and the case studies alike, and it is not ambiguous. The existing architecture fails in a consistent direction: it protects the deploying institution and exposes the individual whose rights are at stake. SyRI ran for six years – not because the law permitted it, as the court confirmed, but because enforcement mechanisms placed the burden of challenge on people who did not know they had been harmed. The streaming tool ran for five years encoding direct discrimination not because the law permitted it, but because no authority required an equality assessment before deployment. Article 22 GDPR is circumvented daily through nominal human insertion not by legislative intent, but because no one has defined what meaningful review actually requires. The word for a pattern of failure that consistently advantages the powerful over the vulnerable, and that is predictable from the design of the system producing it, is structural.

Three reforms are necessary. I offer them here as minimum conditions not a complete programme for closing the gap between formal compliance and substantive rights protection.

ADM Risk Mitigation for the Protection of Fundamental Rights: Existing and Prospective Mechanisms

Before setting out the three proposed institutional reforms, it is necessary to map the risk mitigation mechanisms already available under EU law and to identify where each one falls structurally short. This mapping provides the analytical basis for the reforms that follow and establishes that the proposals are not alternatives to existing instruments but responses to their demonstrated inadequacies.

The primary existing mitigation mechanism is the Data Protection Impact Assessment (DPIA) required under GDPR Article 35 for processing likely to result in high risk to individuals' rights and freedoms. The DPIA obligation is, in principle, a pre-deployment instrument: it requires controllers to identify and assess risks before a system goes live. In practice, as the UK streaming tool case illustrates, it is routinely omitted without triggering enforcement action. Its limitation is structural: the DPIA framework is internally assessed, controller-conducted, and subject to no independent technical verification requirement. A controller may satisfy Article 35 on paper without any substantive engagement with the algorithmic processes at issue.

The AI Act introduces two significant mitigation mechanisms for high-risk systems: the conformity assessment under Article 43, which requires technical documentation, risk management procedures, and bias testing before market placement; and the Fundamental Rights Impact Assessment (FRIA) under Article 27, which requires deployers to identify the foreseeable impact of a high-risk AI system on fundamental rights. Both mechanisms represent genuine advances. The conformity assessment, in particular, establishes ex ante obligations that did not exist under the GDPR. Their limitation, as noted in Section 4.2 above, is one of scope: the FRIA is mandatory only for public-body deployers, and conformity assessments in most high-risk categories are conducted by providers rather than independently verified by third parties.

A third mechanism operates at the level of individual redress: the right to contest automated decisions under GDPR Article 22, and the right to an effective remedy under Charter Article 47. These are reactive instruments they operate after a harm has occurred and require the affected individual to initiate proceedings. As the SyRI case demonstrates, where a system is designed to be opaque and individuals are not informed that they have been assessed, these rights are practically unavailable to precisely those who most need them. Convention 108+ provides a floor that operates independently of the "solely automated" threshold, but its enforcement mechanism, at the level of national supervisory authorities, reproduces the resource and capacity constraints that limit GDPR enforcement. It is against this background of mechanisms that are real but structurally insufficient — that the three reforms proposed below are advanced.

Three Proposed Reforms

First reform: extend the mandatory Fundamental Rights Impact Assessment (FRIA) obligation to all high-risk commercial ADM deployments, with standardised methodology, independent verification, and mandatory public disclosure of results. The predictable objection is cost and innovation burden. It should be answered directly: pharmaceutical products, financial instruments, and major infrastructure all require pre-deployment risk assessment as a condition of market access.

The AI systems that determine access to credit, housing, employment, and welfare affect people's lives as consequentially as those products. The question worth asking is why, until 2024, they did not. It is, however, important to be clear-eyed about the institutional and political constraints that a reform of this scope would face. Legislative inertia within EU co-decision procedures, combined with the divergent interests of Member States whose domestic AI industries would bear the compliance costs, creates a structural tendency toward diluted obligations and extended phase-in periods. Industry lobbying by technology firms and financial sector actors who have invested substantially in ADM infrastructure and whose competitive position would be affected by mandatory pre-deployment assessment represents a further source of resistance that should be anticipated rather than discounted. The appropriate response is not to soften the proposal but to design the implementation architecture in a way that distributes costs equitably, allows for proportionate compliance frameworks for smaller operators, and builds in review mechanisms that allow the methodology to be refined as technical audit practice matures.

Second reform: address the human oversight problem Sarra identifies.[35] The AI Act requires that high-risk systems be designed to permit effective oversight. It does not define what effective oversight requires of the human reviewer. A system that presents the reviewer with a dashboard showing an algorithmic recommendation, without providing access to the underlying data, without institutional protection for a reviewer who chooses to override the recommendation, and without a requirement that the substance of the review not merely its outcome be recorded, satisfies the Act's letter and defeats its purpose. A minimum operational standard for meaningful review

[35]Sarra (n 18) 58–60. Sarra's proposal — that Article 22 GDPR be re-read to focus on the substantive independence of human review from the algorithmic output, rather than on formal participation — is one of the more practically workable suggestions in recent literature on this point.

is achievable; it requires political will to impose it on deploying organisations that have strong commercial reasons to prefer the current arrangement.

Third reform: resource supervisory authorities for genuine technical audit. Colonna documented in 2019 that no EU data protection authority had successfully audited a complex machine-learning system and imposed sanctions for Article 22 violations.[36] The AI Act creates new market surveillance authorities but does not ensure they have the technical staff, the financial resources, or – crucially – the institutional independence from government needed to audit AI systems deployed by public authorities. An enforcement body that cannot interrogate a training dataset, reproduce an algorithmic decision, or identify the point in the data pipeline where discriminatory proxies were learned cannot hold anyone accountable. Legislation without enforcement is not law; it is aspiration with a statutory citation.

A further structural issue in this third reform concerns the relationship between the AI Act's newly established market surveillance authorities and the national data protection authorities (DPAs) that already exercise enforcement jurisdiction under the GDPR. The AI Act does not resolve this relationship with adequate clarity. Where an AI system constitutes both a high-risk AI system under Annex III of the AI Act and involves automated processing of personal data under the GDPR which will be the case for most high-risk ADM deployments in welfare, credit, and border control both the market surveillance authority and the competent DPA have arguable jurisdiction.

The potential for overlap is significant: a single algorithmic deployment could in principle be subject to conformity assessment oversight by one authority and Article 22 GDPR enforcement by another, with neither authority having comprehensive visibility over the system's operation.

The reform proposed here therefore encompasses not only the resourcing question but also the coordination architecture: a statutory duty of cooperation between market surveillance authorities and DPAs, with clearly assigned lead competence for different enforcement functions, is a necessary complement to any resourcing uplift.Without it, the fragmentation of oversight reproduces at the institutional level the same accountability gaps that characterise the regulatory framework at the legislative level.

Conclusions

The cases and doctrinal analysis presented in this paper converge on a finding that is both specific and systemic. The Dutch welfare claimant whose risk score was generated by a system she was never informed of, the visa applicant whose file was pre-classified as high-risk on grounds of nationality before any official had reviewed it, and the loan applicant whose refusal was produced by a model whose logic was withheld as a commercial secret these are not isolated administrative failures. They are specific instances of a structural relationship between algorithmic power and legal accountability that the current EU framework has not adequately resolved. What the

[36]Liane Colonna, 'Automated Decision-Making, Profiling, and the GDPR' (2019) 35 Computer Law & Security Review 397, 402–404. Writing before the AI Act's adoption, Colonna was already identifying enforcement capacity as the critical weakness; the subsequent years have done little to address her diagnosis.

analysis reveals is not that the law is absent, but that the mechanisms through which it is supposed to operate are systematically inadequate to the task assigned to them.

The failure this paper has traced is architectural rather than merely legislative. Both the GDPR and the AI Act express genuine commitments to human dignity and non-discrimination; the problem lies in the gap between those commitments and the operative mechanisms through which they are supposed to be realised. The human-in-the-loop loophole converts a substantive safeguard into a procedural formality. The commercial exclusion from the FRIA obligation exempts from the most demanding requirements precisely the deployments that are most consequential for vulnerable people. The enforcement gap between what the law requires and what supervisory authorities can verify is one that deployers rationally exploit, because the cost of non-compliance falls on individuals rather than institutions.

The Charter of Fundamental Rights demands non-discrimination and human dignity with the same legal force as the Treaties. A regulatory architecture that cannot deliver those guarantees in practice is, to that extent, in breach of primary EU law – not as a matter of political rhetoric but as a matter of legal analysis. The task ahead is institutional: building the mechanisms through which commitments already made are actually kept. It would, however, be analytically incomplete to advocate for these reforms without acknowledging the practical challenges their implementation would face. The extension of mandatory FRIA obligations to commercial deployers will encounter resistance not only from industry actors but from Member States whose administrations benefit from the current permissive architecture. The definition of meaningful human oversight in operationally enforceable terms requires the development of audit standards and professional capacity that does not yet exist at scale across EU supervisory authorities.

The resourcing and coordination reforms proposed for supervisory authorities will require sustained budgetary commitment from Member States that have, historically, been reluctant to fund data protection enforcement at the level the GDPR's ambitions require. None of these challenges is insuperable, but none is trivial. The implementation of each proposed reform will require not only legislative amendment but sustained political will, administrative investment, and the development of technical expertise within public institutions that currently lack it. Acknowledging these constraints is not a reason to lower the normative ambition of the proposals; it is a reason to design the transition architecture with realism and to build in the review mechanisms that allow reforms to be calibrated as implementation experience accumulates. That task is feasible, and it is urgent. For those whose rights have already been affected by the inadequacies the current framework has failed to address, the case for prompt and serious action needs no further elaboration.

References

Academic Works

Boura, M. (2024). The Digital Regulatory Framework through EU AI Act: The Regulatory Sandboxes' Approach. *Athens Journal of Law* 10(3): 385–398.

Colonna, L. (2019). Automated Decision-Making, Profiling, and the GDPR. *Computer Law & Security Review* 35(4): 397–410.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). F*airness Through Awareness.* In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. New York: ACM, 214–226.

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review* 16: 18–84.

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

Gerards, J., & Zuiderveen Borgesius, F. (2020). Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence. *Information & Communications Technology Law* 29(3): 303–333.

Hildebrandt, M. (2015). *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*. Cheltenham: Edward Elgar.

Katyal, S. (2019). Private Accountability in the Age of Artificial Intelligence. *UCLA Law Review* 66: 54–141.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Rachovitsa, A., & Johann, N. (2022). The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case. *Human Rights Law Review* 22(2): ngac010.

Rouvroy, A. (2013). The end(s) of critique: Data behaviourism versus due process. In Hildebrandt M and de Vries K (eds) *Privacy, Due Process and the Computational Turn*. London: Routledge, 143–167.

Sarra, C. (2025). Artificial Intelligence in Decision-making: A Test of Consistency between the EU AI Act and the General Data Protection Regulation. *Athens Journal of Law* 11(1): 45–62.

Van Bekkum, M., & Zuiderveen Borgesius, F. (2021). Digital Welfare Fraud Detection and the Dutch SyRI Judgment. *European Journal of Social Security* 23(1): 1–18.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology* 31(2): 841–887.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI. *Computer Law & Security Review* 41: 105567.

Zuiderveen Borgesius, F. (2018). *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*. Strasbourg: Council of Europe, Directorate General of Democracy.

Legislation and Legal Instruments

Article 29 Working Party (2018) Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679, WP251rev.01 (revised 6 February 2018). Brussels: European Commission.

Charter of Fundamental Rights of the European Union (2012) OJ C 326, 26.10.2012, p. 391.
Council of Europe (2018) Convention 108+ for the Protection of Individuals with regard to Automatic Processing of Personal Data (modernised version, CETS No. 223). Strasbourg: Council of Europe.
European Parliament and Council (2016) Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation — GDPR). OJ L 119, 4.5.2016, p. 1.
European Parliament and Council (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). OJ L, 12.7.2024.
United Kingdom (2010) Equality Act 2010, c. 15. London: Her Majesty's Stationery Office.

Case Law

Court of Justice of the European Union, Case C-817/19, *Ligue des droits humains v Conseil des ministres*, ECLI:EU:C:2022:491, Judgment of 21 June 2022.
Court of Justice of the European Union, Case C-634/21, *OQ v Land Hessen (SCHUFA Holding AG)*, ECLI:EU:C:2023:957, Judgment of 7 December 2023.
Court of Justice of the European Union, Case C-203/22, *Dun & Bradstreet Austria GmbH v CK*, ECLI:EU:C:2025:137, Judgment of 27 February 2025.
European Court of Human Rights, *D.H. and Others v Czech Republic* [GC], Application No. 57325/00, Judgment of 13 November 2007. Reports of Judgments and Decisions 2007-IV.
Netherlands, District Court of The Hague, *NJCM et al. v The Dutch State*, ECLI:NL:RBDHA: 2020:1878, Judgment of 5 February 2020.
United Kingdom, High Court of Justice, *JCWI and Foxglove v Secretary of State for the Home Department*, judicial review filed June 2020; discontinued by consent August 2020.