FACTA UNIVERSITATIS
Series: **Law and Politics** Vol. 24, N° 1, 2026, pp. 23-36
https://doi.org/10.22190/FULP251226003M

Original Scientific Paper

RETHINKING HUMAN RIGHTS PROTECTION IN THE AGE OF AI: FROM EX-POST REMEDIES TO RIGHTS-BY-DESIGN

UDC 004.8:340.131(4-672EU)
004.738.5:341.231.14(4-672EU)
342.7:004.7 (4-672EU)

Janko Munjić

Appellate Court in Kragujevac, Republic of Serbia
Faculty of Law, University of Kragujevac, Republic of Serbia

ORCID iD: Janko Munjić https://orcid.org/0009-0004-8649-2233

Abstract. *As artificial intelligence rapidly becomes a mediating but often invisible factor in decision-making across many areas that are central to the realisation of human rights, ranging from social welfare to the judiciary, legal protection still predominantly operates through an ex-post logic, reacting only after harm has occurred. This article identifies the functional defects of that model in contemporary algorithmic systems, where decision-making opacity, automation, and fragmented responsibility complicate the identification of the responsible actor, proof of causation, and the effective contestation of outcomes, while simultaneously enabling diffuse and systemic harms, particularly through discrimination and the intensification of surveillance. Through the lens of the European normative framework, and especially the EU Artificial Intelligence Act, the paper examines the emerging shift towards preventive risk governance and conceptualises "rights-by-design" as a necessary architectural evolution of human rights protection in algorithmic environments. The author warns that this approach must remain tied to measurable protective outcomes and effective accountability mechanisms. Otherwise, it risks collapsing into technocratic formalism and bureaucratic simulations of compliance.*

Key words: *artificial intelligence, human rights protection, ex-post remedies, rights-by-design, EU Artificial Intelligence Act, fundamental rights impact assessment.*

1. INTRODUCTION

Artificial intelligence (AI) is no longer just a new source of discrete, occasional violations of human rights but also an increasingly structural mode of decision-making that reorganises how power is exercised, how risks are distributed, and how individuals are classified and treated

Received: December 26th, 2025 / Accepted 9th March 2026
Corresponding author: Janko Munjić, Senior Judicial Assistant, Criminal Division, Appellate Court in Kragujevac, PhD Candidate, Criminal Law Department, Faculty of Law, University of Kragujevac, Republic of Serbia | janko.munjic@kg.ap.sud.rs

across the economy and public administration (Rodrigues, 2020: 1). In practice, the human rights challenge is not only that AI can be biased or opaque but also that its deployment tends to be scalable, automated, and embedded into ordinary workflows, which ultimately changes the speed, reach, and reversibility of harm (Shany, 2025: 23). Human rights bodies and networks increasingly emphasise that AI risks are not limited to single mistakes but include structural and collective dimensions, including discriminatory patterns, chilling effects, and diffuse harms that are difficult to attribute to a single decision-maker (ENNHRI, 2024).

The traditional model of human rights protection is largely reactive as it presumes that rights violations materialise in identifiable events, that affected individuals can recognise the harm, and that courts or oversight bodies can provide an effective remedy after the fact. That architecture struggles when AI systems operate as black boxes, when meaningful contestation depends on technical access and expertise, and when the harm is experienced as a pattern rather than a single actionable incident (Quintavalla, Temperman, 2023: 449). Even where transparency and explanation are legally required, explainability may fail to translate into meaningful understanding for affected individuals, and transparency may only partially illuminate internal processes, leaving a gap between formal rights and the real-world ability to challenge automated outcomes (Quintavalla, Temperman, 2023: 70). This gap is particularly visible in high-impact settings, including health, education, social welfare, and justice, where AI can shape eligibility, risk assessment, or access to services (FOC, 2025: 1).[1] This broader shift toward digitally mediated institutional decision-making has been recognised in regional scholarship as creating new pressure points for effective rights protection and procedural guarantees (Živkovska, Prželska, Lalevska, 2022: 13). It is also visible in the mechanics of remedies because safeguards that appear robust on paper can fail if they are bolted on late, treated as optional, or undermined by institutional incentives that favour speed and standardisation over careful review (Rodrigues, 2020: 4). In that context, an exclusively ex-post posture risks becoming a form of "aftercare" that arrives too late, particularly where harms scale quickly, where affected groups are unaware that an AI system shaped their outcome, or where the burden of proof is practically insurmountable without system access (Quintavalla, Temperman, 2023: 449).

Against that background, the key question is whether the law is undergoing a methodological shift from reactive enforcement toward proactive protection, and European governance developments are often presented as an early indicator of that shift. Shany (2025: 35-36) argues that the EU is increasingly relying on regulatory techniques that work through design-stage obligations and systemic risk controls, including product-style requirements and platform governance tools, rather than relying only on courts and *post factum* enforcement; the orientation matters because it moves human rights protection upstream, into the conditions under which systems are built, procured, evaluated, deployed, and monitored. The paper conceptualises that upstream approach as *rights-by-design*. This notion implies embedding human rights constraints and safeguards throughout the AI lifecycle rather than treating them as external checks added after deployment (FOC, 2025:1). It also entails translating rights into operational requirements, such as meaningful human review, bias mitigation mechanisms, transparency about capabilities and limitations,

[1] AI is already used by public authorities for allocating services and forms of assistance, and increasingly also within the judiciary and law-enforcement settings in which errors, bias, or opacity can translate directly into rights-impacts (Prlja, Gasmi, & Korać, 2022: 8-9). In the judicial domain, the use of AI tools by judges and court staff can jeopardise fair trial guarantees, including the right to an independent and impartial tribunal and equality of arms, which makes operational safeguards non-negotiable (Matić Bošković, 2024: 481).

privacy protections across the lifecycle, and accountability structures that allocate responsibility among developers, deployers, and users (DCO, 2025: 50-51).[2]

Methodologically, the paper uses normative legal analysis of human rights standards and duties, combined with a functional analysis of protection mechanisms across the AI lifecycle, and a limited comparative perspective that uses the EU's emerging approach as a reference point against more classical ex-post protection models.[3] The core claim developed here is that the concept of *rights-by-design* is not simply an ethics slogan; it implies a functional reorganisation of protection that aims to preserve the point of human rights law under conditions of automation, opacity, and scale. The conceptual move is from viewing human rights as primarily judicially vindicated entitlements to viewing them also as design constraints and governance duties that must be internalised by institutions and private actors across procurement, development, deployment, and monitoring (DCO, 2025: 50). This is consistent with the observation that private actors play central roles across the lifecycle and should take responsibility to identify, mitigate, and prevent adverse human rights impacts linked to their activities (FOC, 2025: 2).

2. Traditional Human Rights Protection and the Ex-post Paradigm

Traditional human rights protection is built around an ex-post logic. Namely, rights are declared as entitlements; thus, if an entitlement is violated, the core institutional promise is that the individual can challenge the violation through a remedy that is capable of producing redress. In that sense, access to justice is not only a procedural convenience but also a right in itself and an enabling tool that makes other rights practically real (FRA, 2016: 3-4).

2.1. Ex-post remedies in international human rights law

In international human rights law, the relief architecture is largely remedial, complaint-driven, and backward-looking. The system presupposes that an alleged violation has already occurred, or is occurring, and that protection is operationalised through procedures that can establish whether a violation took place and provide an effective response. At the European level, Article 13 of the ECHR frames this remedial promise as a national-level entitlement to a remedy for arguable claims of the Convention violations, while leaving States a degree of discretion as to the form of remedy (FRA, 2016: 92-93).[4] At a structural level, the right to a remedy is defined by effectiveness, in the sense that a remedy must allow

[2] The technical literature mapped through a human rights lens similarly emphasises that many violations are preventable at early stages, particularly through careful requirements setting and specifications, where human rights limitations can be built into the system's design parameters (Quintavalla, Temperman, 2023: 57).

[3] The original contribution is a structured account of the shift from remedy-centred protection to lifecycle-centred protection, plus a concrete evaluation frame for rights-by-design that links doctrinal commitments to implementable governance requirements. Moreover, the paper examines three research questions. First, why do predominantly ex-post models of human rights protection become functionally insufficient in AI-mediated governance, even where the substantive rights catalogue remains strong? Second, does the emerging EU regulatory toolkit represent a methodological turn toward upstream protection, and what does that turn look like in operational terms? Third, under what conditions can rights-by-design become a credible standard of human rights protection rather than a rhetorical label, and how should its success be evaluated?

[4] The ex-post paradigm is also tied to the *principle of subsidiarity*, because domestic institutions, especially courts, are meant to be the first place where people seek redress. For that reason, applicants must exhaust domestic remedies before turning to international bodies, which ties international scrutiny to whether national channels for complaint and redress actually work (FRA, 2016: 95).

the competent domestic authorities to deal with the substance of the relevant Convention complaint and to grant appropriate relief, while also being available and sufficient, sufficiently certain in practice, and effective in practice as well as in law (CoE Committee of Ministers, 2013: 12), which reflects the logic of the ex-post model that frames protection as a response to concrete allegations through procedures capable of producing legally meaningful relief, e.g., cessation, annulment, compensation, or another form of redress.[5]

2.2. Judicial review and individualised harm

Judicial review sits at the centre of the ex-post model, turning protection into a process where an individual can bring a claim, the decision or conduct is examined, and a binding outcome can follow; even when remedies are not strictly judicial, the model still rests on adjudicative reasoning through fact finding, applying legal standards, and a case specific determination of responsibility and relief.[6] The individualised character of the model matters in two ways. First, remedies are typically triggered by a complainant who can point to a specific interference with a right and who can place that interference within an existing legal framework. Second, the assessment of whether a remedy is "effective" is normally evaluated through its capacity to address the complainant's situation. In the ECHR context, the relevant question is whether there is a remedy that can enforce the substance of the Convention rights and provide relief for an arguable claim, not whether the legal system has abstract supervisory tools (FRA, 2016: 92-93).

As a result, the ex-post model is built around individual harm and the way that harm is framed in legal terms, i.e. a remedial focus that also surfaces in recent analyses of artificial intelligence in adjudicative settings. Examining AI in adjudication through Article 6(1) ECHR, Patricheev (2025: 121-122) analyses how AI-assisted and fully autonomous models interact with core fair-trial safeguards, notably equality of arms, the right to a public hearing, and the duty to give reasoned decisions, with particular attention to opacity and informational imbalance. This analysis is grounded in how these safeguards operate in concrete proceedings and treats procedural backstops such as review and appeal as key protections where autonomous adjudication is contemplated, even though the analysis ultimately argues for strict ex-ante regulation and regulatory foresight to preserve fair-trial standards. More generally, human rights litigation assumes that the claimant can demonstrate standing, identify a contested act or omission, and articulate the nature of the interference in a form that legal institutions can evaluate. In this respect, the "chain of justice" framing is instructive as it highlights that access to justice spans multiple stages, so that weaknesses

[5] This is not to suggest that human rights protection is exclusively ex-post. The Convention system has long recognised preventive dimensions through positive obligations and due diligence duties requiring States to take reasonable measures to avert foreseeable violations. The point here is functional rather than doctrinal because, even where preventive duties exist, the dominant operational pathway of protection still tends to be complaint-driven and remedial, triggered after deployment, after harm becomes legible, and after responsibility can be attributed through evidence. Accordingly, this functional dominance becomes a weak fit for AI systems, where scale, opacity, and diffusion of agency undermine identifiability, attribution, and contestability. This preventive strand is well-established in the Court's positive-obligations case law, which treats foreseeable rights-risks as triggering duties of reasonable preventive action. *See*: ECtHR case *Osman v. the United Kingdom* [GC], Application no. 23452/94, Judgment of 28 October 1998; ECtHR case *Öneryıldız v. Turkey* [GC], Application no. 48939/99, Judgment of 30 November 2004; ECtHR case *Budayeva and Others v. Russia*, Applications nos. 15339/02 et al., Judgment of 20 March 2008.

[6] This is why access to justice is treated as a cross-cutting condition for human rights protection, extending beyond entry to a courthouse to cover the full chain of justice, from awareness and access to information about rights, through the actions of justice institutions including law enforcement authorities, to the application of appropriate legal remedies (Nastić, 2025: 237).

at any point in the chain can undermine the practical possibility of obtaining judicial protection (Nastić, 2025: 236).

2.3. Structural assumptions of the ex-post model

The dominance of the ex-post model stems from historical institutional design, since courts, complaint mechanisms, and remedial procedures developed for harms tied to identifiable acts, traceable decision makers, and contestable reasons; that legacy still shapes the assumptions of traditional human rights protection. The model assumes that the claimant can articulate an arguable claim and identify the acts or omissions complained of as a rights-relevant interference so that a national authority can examine the substance of the complaint and, if appropriate, grant redress (CoE Committee of Ministers, 2013: 12-13). The model also assumes that there is a competent national authority with powers and guarantees sufficient to decide the claim and provide relief, which is why international review places weight on the existence and effectiveness of domestic remedies and on the exhaustion requirement (CoE Committee of Ministers, 2013: 13; FRA, 2016: 95). A third assumption is contestability because the right of access to court must be effective by providing individuals with a clear, practical opportunity to challenge acts that interfere with their right (Nastić, 2025: 237). These assumptions are what make the traditional model work when harms are discrete, traceable, and open to challenge, and what make it falter when causation is diffuse, responsibility is spread across actors or systems, or decision grounds are hard to see and contest; this is exactly where algorithmic settings strain it by changing how easily a measure can be identified, responsibility assigned, and reasons tested.[7]

3. Artificial Intelligence and the Crisis of Ex-post Human Rights Protection

The limits of the ex-post paradigm become most visible once AI systems are understood as socio-technical infrastructures rather than isolated decision tools. Shany (2025) argues that AI does not merely create new categories of rights violations but also challenges the underlying assumptions of existing human rights frameworks by reshaping how power, risk, and decision-making are organised across society. In practice, AI-supported decision-making can be deployed at speed and scale, fed by continuous data collection, and insulated by technical and organizational opacity; these features shift human rights problems from identifiable unlawful acts to patterns of risk, cumulative disadvantage, and governance failures, which ex-post remedies were not designed to capture (Lila, 2024: 11-12). Three

[7] To be clear, human rights law is not exclusively ex-post since it contains preventive dimensions through positive obligations to secure the effective enjoyment of rights, due diligence duties to prevent foreseeable harm, and procedural duties to put safeguards in place before interferences occur. This paper does not argue that prevention is missing but that protection still works mainly through remedies and complaints; institutions usually identify violations only after deployment, once affected people can describe the harm and the facts are clear enough for adjudication and proof. Hence, this ex-post model is a weak fit for AI systems because scale, opacity, and dispersed responsibility make it harder to identify the relevant measure, assign responsibility, and challenge that measure through existing remedial mechanisms.

stress-tests illustrate this mismatch: welfare fraud analytics,[8] biometric surveillance in public space,[9] and risk tools in criminal justice.[10]

3.1. Scale, automation, and algorithmic opacity

AI systems extract and process data at volumes that make broad monitoring feasible in ways that used to be too costly or too wide to run (Lila, 2024: 12). Yet, many ex-post remedies still rely on a claimant being able to pinpoint a specific decision, reconstruct its reasoning, and tie it to a concrete rights interference. When decisions are routine, repeated, and spread across time and platforms, that burden grows because it becomes harder to identify the legally relevant moment of interference (NIST, 2023: 42).

Opacity worsens the scale problem because it is not just technical but also institutional, given that many platforms route interaction through automated messages that leave no real way to engage the provider, limiting access to information and rights while putting profit first (Mudrinić, 2022: 97-98). A recurring concern is that algorithmic outputs can be explained only in narrow terms while institutional users treat the system as authoritative, which limits effective challenges. Even where transparency duties exist, black box dynamics can keep affected people from knowing what factors drove a result, weakening ordinary legal routes for review (Gordon, 2023: 19; Lila, 2024: 11). Risk frameworks themselves increasingly treat explainability, interpretability, and transparency as governance conditions that enable documentation, audit, and oversight, which is a recognition that ex-post review is structurally weakened without such

[8] Let's assume that a welfare agency uses an ML model trained on past enforcement data to flag fraud risk or rank social assistance claims, and the model ends up disproportionately targeting applicants from certain neighbourhoods and vulnerable communities, as a result of which people face denials or delays backed by generic explanations; whereas each decision looks minor and formally reviewable, the discriminatory effect only shows up across thousands of routine cases while the real proof sits in data pipelines and continuous model updates that claimants cannot realistically access. This is exactly the kind of welfare fraud analytics pattern that courts have already treated as raising serious Convention-level concerns, including privacy safeguards and structurally indirect discrimination. *See*: District Court of the Hague case *NJCM c.s. v. The Netherlands (SyRI)*, ECLI:NL:RBDHA:2020:1878, Judgment of 5 February 2020; ECtHR case *D.H. and Others v. the Czech Republic* [GC], Application no. 57325/00, Judgment of 13 November 2007; ECtHR case *Stec and Others v. the United Kingdom* [GC], Applications nos. 65731/01 and 65900/01, Judgment of 12 April 2006.

[9] Let's consider a case where a city deploys facial recognition or other biometric tools in public spaces to spot threats or manage crowds; even without a single adverse individual decision, the system still reshapes public life by raising the baseline risk of identification, tracking, and later profiling, creating anticipatory and collective harms like a chilling effect on protest, association, and everyday movement, while ex-post remedies usually fail because people cannot prove they were scanned, what inferences were made, or how the data may be reused. It aligns with courts' settled view that biometric and large-scale surveillance demands especially clear legality and robust safeguards precisely because affected persons can rarely reconstruct or contest the interference after the fact. *See:* ECtHR case *S. and Marper v. the United Kingdom* [GC], Applications nos. 30562/04 and 30566/04, Judgment of 4 December 2008; ECtHR case *Big Brother Watch and Others v. the United Kingdom* [GC], Applications nos. 58170/13, 62322/14 and 24960/15, Judgment of 25 May 2021; Court of Appeal (England and Wales*), R (Bridges) v Chief Constable of South Wales Police and others* [2020] EWCA Civ 1058, Judgment of 11 August 2020.

[10] Criminal courts and probation services use risk assessment tools or AI decision support to recommend detention, bail conditions, sentencing severity, or supervision intensity, and the output is framed as mere "guidance" while the human decision-maker remains formally responsible. In practice, it fragments accountability across vendor, agency, and court, and allows small shifts in risk scores to systematically tighten liberty restrictions for certain groups even though errors are hard to contest because training data, feature choices, and calibration are opaque; thus, appeals (when they exist) arrive only after liberty has already been curtailed and rarely reach the upstream model governance that produced the pattern, precisely the concern courts have flagged by insisting on caution, limits on opacity, and demonstrated validity across affected groups. *See*: Supreme Court of Wisconsin, *State v. Loomis*, 2016 WI 68, 881 N.W.2d 749 (2016); Supreme Court of Canada, *Ewert v. Canada*, 2018 SCC 30 (2018).

properties (NIST, 2023: 16-17). National human rights institutions have similarly identified algorithmic opacity, limited transparency, and information asymmetries as core obstacles to effective remedies, noting that affected individuals are often unable to understand or challenge AI-driven decisions even when formal rights exist (ENNHRI, 2024). A further implication is evidentiary, given that complex models, training data, and iterative updates can scatter the record relevant to legality across technical layers and organizational boundaries. The NIST notes that AI scale and complexity can involve extremely large numbers of decision points and risks that extend beyond a single enterprise, which helps explain why case-by-case reconstruction is often difficult in practice (NIST, 2023: 38-39).

3.2. Diffuse and systemic harm in AI-driven decision-making

A second pressure point is how harm changes shape, considering that many AI-related impacts are not a single individual injury but diffuse or systemic effects that build across populations and over time. It is clearly demonstrated by surveillance-enabled AI; continuous tracking, large-scale data aggregation, and automated identification can affect privacy and autonomy without any single incident by changing the basic conditions of social life (Lila, 2024: 12). The ex-post model is a weak fit because it assumes that a claimant can show a personal, individual interference, while AI can shape outcomes for people who never interact with the system through decisions based on its outputs, making standing, proof, and causation harder, especially in discrimination and social sorting where disadvantage shows up as a pattern spread across many similarly placed people (NIST, 2023: 37; Gordon, 2023: 51-53). Chilling effects show the mismatch; as Gordon (2023: 54) links broad state surveillance to pressures that deter people from voicing critical views, rights harms often appear as self-censorship and behavioral change rather than a discrete violation that can be easily litigated. In such settings, ex-post litigation may arrive after the systemic effect has already reshaped conduct, and the claimant may be unable to demonstrate individualized harm with the clarity typically expected by courts. Human rights monitoring bodies recognise this gap, and ENNHRI (2024) notes that many AI harms accumulate as cumulative, group-based, or structural effects, including discriminatory patterns and chilling effects, which are hard to address through individual complaints and traditional remedies.

3.3. Accountability gaps and the fragmentation of responsibility

Despite the visible impacts on human rights, accountability may be hard to allocate. A persistent difficulty is that responsibility may be unclear across the AI lifecycle, including uncertainty about whether and how to attribute outcomes to system developers, deployers, operators, or the system's autonomous behavior (Lila, 2024: 11). This fragmentation is not incidental; it reflects how AI systems are built and deployed through procurement, third-party components, and layered governance. The NIST's AI RMF is explicit that risk management should begin at plan and design level, and be performed throughout the AI system lifecycle, which frames accountability as distributed and continuous rather than episodic (NIST, 2023: 4). It also emphasizes that AI risk mapping should cover all components including third-party software and data, which directly speaks to the structural difficulty of locating a single accountable decision-maker after harms materialize (NIST, 2023: 26). The framework further notes that technologies acquired from third parties may be complex or opaque and that risk tolerances may not align between vendors and deploying organizations, a dynamic that predictably complicates both legal responsibility and practical access to evidence (NIST, 2023:

36-37). These accountability gaps deepen the problem for ex-post remedies because, when affected people are unable to obtain meaningful explanations, fail to map which actor controlled which part of the pipeline, and find it unfeasible to isolate a legally relevant act within a chain of socio-technical decisions, judicial review shifts from correcting a wrong to coping with uncertainty, as the primary route through which individuals vindicate rights and obtain effective relief (Human Rights Committee, 2004: 6). This is one reason why contemporary governance approaches increasingly treat transparency and accountability as risks to be actively examined and documented, rather than assuming that post-hoc dispute resolution will reliably surface them (NIST, 2023: 29).

4. The European Shift Towards Preventive Human Rights Protection

The European response to AI risks increasingly treats fundamental rights not only as standards for post hoc adjudication but also as design and governance constraints that must shape systems before they affect people. The EU AI Act is emblematic because it structures obligations around a prevention logic that operates across the full lifecycle of certain AI systems, from development and deployment to monitoring and corrective action.[11] This is a methodological shift in human rights protection because it seeks to reduce the probability and severity of rights' infringements upstream, instead of relying predominantly on downstream litigation and individual remedies once harm has already materialised.[12]

4.1. Human rights as a regulatory objective in the EU AI Act

A first indicator of the EU shift is that the AI Act positions fundamental rights as an explicit regulatory objective that shapes explicit prohibitions, risk management duties, and operational requirements. At the level of outright prohibitions, the AI Act bans several AI practices that are difficult to control through classical ex-post mechanisms because they can operate at scale and without transparent traceability.[13] For high-risk AI systems, the AI Act sets a lifecycle risk management framework that treats rights protection as a governance goal rather than a side-effect of sector compliance and requires a risk management system that is established, maintained, reviewed, and updated throughout the system's life, including identifying and analysing reasonably foreseeable risks to health, safety, or fundamental rights, assessing risks under intended use and reasonably foreseeable misuse, and taking targeted measures to address the identified risks.[14]

---

[11] 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). *Official Journal of the EU,* L series (12 July 2024).

[12] AI Act, Art. 9(2).

[13] These include manipulative or deceptive techniques that materially distort behaviour, exploitative use of vulnerabilities linked to age, disability, or socio-economic situation, social scoring that triggers unjustified or disproportionate adverse treatment, and certain forms of predictive criminal risk assessment based solely on profiling or personality assessment. AI Act, Art. 5(1)(a)-(d). This preventive approach also shows up in the rules on remote biometric identification, where the Act pairs substantive limits with ex-ante safeguards and, even where a use is allowed, requires prior authorisation, strict time and place limits, and a fundamental rights impact assessment before the technology is deployed. See AI Act, Art. 5(2)-(3).

[14] See AI Act, Art. 9(1)-(2) and Art. 9(2)(a)-(d). This preventive turn is reinforced by the rule that the relevant risks are those that can be reasonably mitigated or eliminated through development and design choices or through adequate technical information, which treats rights protection as something built into systems and their deployment context rather than left for litigation after the fact. Cf. AI Act, Art. 9(3). The human oversight

4.2. Fundamental rights impact assessments as an ex-ante safeguard

A second indicator of the European preventive turn is the institutionalisation of impact assessment techniques as a safeguard that is triggered prior to deployment. The EU AI Act introduces a dedicated Fundamental Rights Impact Assessment (FRIA), which is required before deploying certain high-risk AI systems by deployers that are public bodies or private entities providing public services, as well as for certain systems listed in Annex III contexts.[15] Two elements are especially important for the human-rights protection methodology. First, the FRIA shifts the analytical lens from individualised harm to categories and groups likely to be affected in context, which is closer to how AI-related discrimination, exclusion, and chilling effects often operate in practice.[16] Second, the FRIA links risk identification to organisational response capacities, including governance arrangements and complaint pathways, which anticipates the remedy problem by requiring deployers to design internal escalation and response mechanisms before deployment.[17] Furthermore, the AI Act links the FRIA to existing European impact assessments, especially DPIAs (Data Protection Impact Assessments); it also stipulates that, where a DPIA already meets the relevant duties (envisaged in Art. 35 GDPR), the FRIA should build on it rather than repeat it.[18] This points to rights-focused risk review becoming a standard part of deployment decisions instead of being triggered mainly by litigation.[19] Lila (2024: 31) also treats DPIA-style assessments as useful for AI because many deployments involve sensitive data, profiling risks, and ongoing audits and compliance checks. The AI Act also addresses implementation by asking the AI Office to develop a questionnaire template, including an automated tool, to help deployers meet their FRIA duties,[20] while the aim is to make rights' assessments repeatable at scale, in direct response to the fact that individual litigation cannot keep pace with system wide automated decision making.[21]

4.3. Risk-based regulation and preventive governance

The broader European shift is not only about adding a single assessment requirement but also about implementing a risk-based regulatory structure that treats prevention as the central logic of governance. The AI Act's lifecycle risk management obligations are designed as iterative processes that integrate post-market signals and require ongoing review and updating, which aligns more closely with dynamic technological systems than with static compliance models.[22] This risk-based approach builds prevention into a set of

---

requirement supports this preventive approach by ensuring that people can spot, understand, and intervene when rights risks arise, so protection rests on keeping high risk systems under human control through procedural choices rather than mainly on ex-post court review. Cf. AI Act, Art. 14.

[15] See AI Act, Art. 27(1). The FRIA duty is a concrete checklist that requires deployers to describe how the system will be used, the period and frequency of use, the people and groups likely to be affected, the specific risks to them, the human oversight measures, and what steps they will take if risks occur, including internal governance and complaint mechanisms. Cf. AI Act, Art. 27(1)(a)-(f).

[16] AI Act, Art. 27(1)(c)-(d).

[17] AI Act, Art. 27(1)(f).

[18] AI Act, Art. 27(4).

[19] *Ibid*.

[20] AI Act, Art. 27(5).

[21] *Ibid.*

[22] AI Act, Art. 9(2)(c).

linked design and governance duties.[23] This architecture connects naturally with the "rights-by-design" concept, understood as the embedding of human rights principles into the development, procurement, and deployment lifecycle, rather than treating rights as external constraints invoked only after harm occurs. In policy terms, rights-by-design recommendations typically include integrating human rights review processes into product development, aligning governance practices with human rights principles, and establishing oversight and accountability structures that can respond to rights risks in a timely manner (DCO, 2025: 9-10). The value of this approach is not that it replaces courts but that it redistributes the burden of rights' protection by shifting part of it into the organisations that build and deploy systems, where design choices and operational controls can prevent harm more effectively than after-the-fact remedies (DCO, 2025: 10-11).[24]

5. Rights-by-design as a New Paradigm of Human Rights Protection

Europe's preventive turn embeds rights into system design and governance because, in algorithmic settings, harms often arise before courts, regulators, or even affected persons become aware of those harms. If system design sets the limits of what can happen, rights protection must constrain how systems are built and run up front, not only provide remedies after the fact.

5.1. Conceptualising rights-by-design in algorithmic environments

The rights-by-design framework means turning human rights norms into concrete design constraints, governance duties, and assurance mechanisms that apply across the full AI lifecycle, from development to deployment and monitoring in AI use (FOC, 2025: 1; DCO, 2025: 9). Rights-by-design is framed as a practical approach for mainstreaming human rights culture in the digital age through concrete tools and institutional practices (Loder, 2022: 390). Rights protection is not treated as a purely ex-post exercise but as an operational requirement that must be built into how systems are specified, procured, tested, deployed, and continuously supervised (DCO, 2025: 9).[25] Rights-by-design is also defined

[23] Article 9(4) of the AI Act ties risk identification to targeted measures that reduce risks while coordinating compliance across requirements; Article 14 adds human oversight as a control layer to spot and correct harmful outputs before they become rights violations; Article 5(1)-(3) adds prohibitions and strict conditions for practices whose rights risks are seen as unacceptable or not controllable in practice.

[24] Concurrently, a preventive governance model should be read as a functional complement to the ex-post paradigm, not as its normative substitute. The EU shift is a layered approach that aims to make rights protection workable at scale despite opacity and dispersed responsibility by requiring documented, auditable governance steps that can be improved over time, including risk mitigations that must be achievable through design or adequate technical information and a structured assessment of fundamental rights impacts before deployment in sensitive settings. See: AI Act, Art. 9(3) and Art. 27(1).

[25] A useful way to conceptualise rights-by-design is as a translation pipeline with three linked steps. First, there is normative translation, where affected rights are mapped in the concrete context and converted into system-level requirements, such as explainability calibrated to rights impact, privacy protections that go beyond narrow data protection, and bias detection and mitigation (DCO, 2025: 53; Rodrigues, 2020: 1). Second, there is institutional embedding, where responsibility and decision authority are allocated inside organisations and across the supply chain, including internal review structures with real influence and routes for stakeholder feedback and grievance resolution (DCO, 2025: 52). Third, there is operational assurance, where organisations generate evidence that safeguards work in practice through continuous monitoring, periodic audits that examine real-world rights impacts, public reporting, and remediation pathways that can be triggered when risks materialise (DCO, 2025: 52–53).

by what it is not;[26] it is broader than isolated "fairness-by-design" or "privacy-by-design" because the relevant bundle of rights in AI settings often includes transparency, non-discrimination, privacy, and contestability together (Rodrigues, 2020: 1).[27]

5.2. Normative implications for human rights architecture

In AI-mediated settings, the meaning of effective protection has to change because a remedies-only model can let systemic harm run for too long, considering that the harm is spread out, hard to detect, and difficult to litigate at scale. Rights-by-design treats prevention as part of human rights protection when automated decision making is deployed widely (FOC, 2025: 1; DCO, 2025: 9). In practice, it entails routine ex-ante assessment and ongoing governance across a system's life, including required human rights impact assessments in high-risk contexts and oversight bodies with clear human rights mandates (DCO, 2025: 9). A second implication is that responsibility shifts across the AI lifecycle because rights-by-design spreads duties across designers, developers, integrators, procurers, and users, with governance arrangements that extend through the supply chain (DCO, 2025: 52).[28] A third implication is proceduralisation and evidentiary discipline in rights practice because rights-by-design increases the importance of documented processes that generate verifiable traces of how rights risks were identified, mitigated, and monitored.[29] In short, rights-by-design nudges human rights protection toward an architecture where prevention and accountability operate together, with ex-post remedies preserved as a backstop rather than the whole system.

5.3. Limits and risks of rights-by-design approaches

Rights-by-design is promising but it has predictable failure modes that have to be confronted if the paradigm is to remain credible. One risk is technocratisation and legitimacy loss.[30] The

---

[26] It is not a move toward treating AI systems as rights-bearing subjects because, even where technology is central to a socially sensitive function, the normatively relevant harm is typically mediated through the system to human or collective interests, and anthropomorphic framing can mislead governance choices and dilute a human-centred focus on rights-holders (Munjić, 2025: 299-300). It is not the idea that rights can be fully coded into systems because rights remain legal standards whose application depends on context, proportionality, and evidence. It is not a replacement for judicial remedies because even strong preventive governance must remain accountable through access to redress.

[27] Rights-by-design is not the same as ethics-by-design; it is not a compliance checklist but a method for translating legal rights into concrete design and governance constraints and for producing verifiable evidence that safeguards reduce rights-relevant risks in practice. By contrast, ethics-by-design usually relies on voluntary principles and self-assessment. For instance, the ALTAI (Assessment List for Trustworthy AI) is a self-assessment tool but it does not prove that a system meets the listed requirements (AI HLEG, 2020: 2-3). Although these tools may be useful, without enforceable duties, independent oversight, and checks tied to real-world impacts, they have clear limits. The rights-by-design framework goes further by grounding requirements in law, naming who is responsible, and requiring monitoring, audits, and information disclosure aimed at human rights outcomes rather than box-ticking (DCO, 2025: 52-53).

[28] This responds to how AI is built and delivered, since key choices about data, model design, testing, and monitoring are often made well before deployment, sometimes by different actors in different jurisdictions, and it therefore calls for shared assessment methods and common governance expectations that work across borders and organisational boundaries (DCO, 2025: 50).

[29] This includes continuous monitoring, periodic audits designed to evaluate real-world human rights impacts, public reporting, and grievance channels that make it possible to detect and address harms during operation, not only after litigation begins (DCO, 2025: 52-53).

[30] If rights protection becomes dominated by technical and compliance elites, affected communities can be reduced to symbolic consultation, while "what counts" becomes whatever can be measured most easily. A serious rights-by-design approach therefore needs participatory governance and meaningful stakeholder involvement as a legitimacy safeguard, not a decorative add-on (DCO, 2025: 52-53), and this includes meaningful public

second risk is box-ticking.[31] The third risk is rights-washing and accountability dilution because organisations can market "rights-preserving AI" while shifting responsibility onto ethics teams, consultants, or auditors, in ways that blur who is actually accountable for operational decisions.[32] More broadly, the digital domain can enable new forms of coercion and intervention that are less visible and harder to attribute, which is a useful reminder that 'by-design' governance can also be geopolitically instrumentalised unless anchored in robust accountability and oversight (Dakić, 2025: 103-104). The way out is not to abandon rights-by-design but to set minimum success conditions that separate substantive protection from paperwork.[33] As a minimal success test, rights-by-design should be able to demonstrate: (i) broad and systematic impact assessment across the lifecycle; (ii) independent oversight with a real mandate and powers; and (iii) accessible avenues of effective redress for those affected (CoE Commissioner for Human Rights, 2023: 11; 25; 29-30).[34]

6. CONCLUSION

AI does not merely add new categories of human rights risk, it reshapes how decisions are made and how harms emerge, often at scale, through opacity, and across distributed chains of actors, which makes a purely ex-post, court-centred model of protection increasingly fragile. The European shift is significant because it treats fundamental rights as an upstream governance objective and moves protection into the lifecycle of AI systems through risk management duties, impact assessments, and structured oversight that aim to prevent harm before it crystallises. This article has argued that rights-by-design is the most coherent way to understand that shift, not as an ethics slogan or a compliance accessory but as a reorganisation of protection which translates rights into design constraints, embeds responsibility across organisations and supply chains, and requires evidence that safeguards work in practice through monitoring, auditing, transparency, and accessible remediation pathways. At the same time, prevention cannot replace adjudication, and rights-by-design will fail if it becomes a technocratic box-ticking or a branding exercise that dilutes responsibility. Therefore, the credible path forward is layered protection, where upstream

consultations with affected groups and relevant stakeholders as part of the governance architecture (Council of Europe Commissioner for Human Rights, 2023: 19-20).

[31] Where "by design" gets reduced to a few technical controls like encryption, anonymisation, or consent prompts and the social and governance questions are left aside, so assessment turns into paperwork that can be completed even as harms persist, especially when organisations rely on self-assessment tools that explicitly disclaim any guarantee of compliance (Shwartz Altshuler et al, 2025: 5; AI HLEG, 2020: 2). The way out is to treat assessment as evidence generation that is independently checked and audited against real-world effects, not merely technical conformity (DCO, 2025: 52-53 and Rodrigues, 2020: 3).

[32] A simple way to detect whether rights-by-design is real (rather than performative) is to apply three success criteria: (i) evidence of measurable risk-reduction against a specified rights-risk; (ii) independent verification (audit or certification) that does not rely on vendor self-reporting; and (iii) an accessible grievance channel that can trigger meaningful human review and remedy.

[33] The judiciary is a stress-test for "objectivity" claims. Even there, AI can at best increase predictability and consistency but it cannot guarantee impartiality because it is ultimately a product of human choices and may reproduce human prejudice (Avramović, Jovanov, 2023: 162-163).

[34] In this regard, a rights-by-design programme should count as successful only if three criteria are met: 1) it produces evidence of risk reduction in practice, shown through monitoring that tracks rights relevant outcomes over time and across affected groups, not only policy compliance; 2) it is independently assured, meaning that claims about safeguards can be verified through external audit or oversight with access to documentation and the ability to trigger corrective action; and 3) it preserves contestability and remediation, meaning that affected persons have accessible channels to challenge outcomes, obtain timely responses, and receive meaningful remedies, and that complaints feed back into updates of the system and its governance.

controls reduce systemic risk, while downstream remedies remain strong, contestable, and effective for individuals and groups who are affected because, in the age of AI, rights risk becoming ineffective unless they are engineered to operate under AI conditions.

REFERENCES

Academic sources

Avramović, D. S., & Jovanov, I. D. (2023). Sudijska (ne)pristrasnost i veštačka inteligencija [Judicial (im)partiality and artificial intelligence], *Strani pravni život*, Vol. LXVII, No. 2, 2023, pp. 162-163.

Dakić, D. (2025). Međunarodno javno pravo u digitalnoj (sf)eri: sajber sankcije i održivi razvoj [Public international law in the digital (sph)era: Cyber sanctions and sustainable development]. In: *Zbornik radova "Savremeno pravo u eri digitalizacije i održivog razvoja"*, Kragujevac, pp. 103-104.

Gordon, J. S. (2023). *The Impact of Artificial Intelligence on Human Rights*. Palgrave Macmillan, pp. 19-54.

Lila, A. (2024). *The Effects of Artificial Intelligence on Human Rights: Kosovo Case*. Institute for Technology and Society, in cooperation with Metamorphosis Foundation for Internet and Society, Prishtina, pp. 11-31.

Loder, L. (2022). *Rights by Design: Mainstreaming Human Rights Information, Education and Culture*, PhD thesis, p. 390.

Matić Bošković, M. M. (2024). Uticaj veštačke inteligencije na obavljanje profesija u pravosuđu [The impact of artificial intelligence on the performance of professions in the judiciary], *Sociološki pregled*, Vol. LVIII, No. 3, 2024, p. 481.

Mudrinić, Lj. M. (2022). Posledice razvoja novih tehnologija na ljudska prava sa aspekta međunarodnog prava [Consequences of the development of new technologies on human rights from the aspect of international law]. *Megatrend revija*, 19 (2), 2022, pp. 93-116.

Munjić, J. (2025). Robots as victims? Examining criminal law's boundaries in the digital age. *Alternative Law Journal*, 50(4), 2025, pp. 299-303.

Nastić, M. (2025). Access to Justice in the Digital Age. *TEME*, Vol. XLIX, No. 2, 2025, pp. 236-237.

Patricheev, I. (2025). To regulate or not to regulate: Assessing the necessity of regulating AI systems in adjudication in light of Article 6(1) ECHR. *Facta Universitatis: Law and Politics*, Vol.23 (2) 2025, pp.121-122.

Prlja, D., Gasmi, G., & Korać, P. (2022). *Ljudska prava i veštačka inteligencija* [Human rights and artificial intelligence], Belgrade: Institute for comparative law, pp. 8-9.

Rodrigues, R. (2020). *Legal and human rights issues of AI: gaps, challenges and vulnerabilities*. *Journal of Responsible Technology*, Vol. 4, December 2020, Article 100005, pp. 1-4.

Shwartz Altshuler, T., Aridor Hershkovitz, R., Mikulinsky, N., & Müller, K. (2025). Governing phygital spaces: Human rights by design meets speculative design, *Internet Policy Review*, 14(4), 2025, p. 5.

Quintavalla, A., & Temperman, J. (2023). *Artificial Intelligence and Human Rights*. Oxford University Press, pp. 57-449.

Živkovska, R., Prželska, T., & Lalevska, T. (2022). The status of the pledge creditor during the realization of the right of pledge in the Macedonian law. In: *Protection of Human Rights and Freedoms in Light of International and National Standards* (Thematic Conference Proceedings), Faculty of Law, University of Priština in Kosovska Mitrovica, p. 13.

Legal documents

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). *Official Journal of the EU*, L series (12 July 2024).

Online sources

European Network of National Human Rights Institutions (ENNHRI), (2024). *Key Human Rights Challenges*, Retrieved 22 December 2025, from https://ennhri.org/ai-resource/key-human-rights-challenges/

Freedom Online Coalition (FOC), (2025). *Joint Statement on Artificial Intelligence and Human Rights*. Freedom Online Coalition, Retrieved 20 December 2025, from https://freedomonlinecoalition.com/joint-statement-on-ai-and-human-rights-2025/

Soft law documents

AI HLEG (High-Level Expert Group on Artificial Intelligence (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office of the EU, Luxembourg.

Council of Europe Committee of Ministers/CoE CM (2013). *Guide to good practice in respect of domestic remedies*. Council of Europe, Strasbourg.

Council of Europe Commissioner for Human Rights/CoE CHR (2023). *Human rights by design: future-proofing human rights protection in the era of AI*, Council of Europe.
Digital Cooperation Organization/DCO (2025). *Rights by Design: Embedding Human Rights Principles in AI Systems*. Digital Cooperation Organization.
European Union Agency for Fundamental Rights/FRA (2016). *Handbook on European law relating to access to justice*. Publications Office of the European Union, Luxembourg.
National Institute of Standards and Technology/NIST (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, Gaithersburg, MD.
Shany, Y., (2025). *The Need for and Feasibility of an International AI Bill of Human Rights*. White paper, Institute for Ethics in AI, University of Oxford, 2025, pp. 23-36.
UN Human Rights Committee, General Comment No. 31 (2004).

Court decisions

Court of Appeal (England and Wales*), R (Bridges) v Chief Constable of South Wales Police and others* [2020] EWCA Civ 1058, Judgment of 11 August 2020.
District Court of The Hague, *NJCM c.s. v. The Netherlands (SyRI)*, ECLI:NL:RBDHA:2020:1878, Judgment of 5 February 2020;
European Court of Human Rights, *Osman v. the United Kingdom* [GC], Application no. 23452/94, Judgment of 28 October 1998.
European Court of Human Rights, *Öneryıldız v. Turkey* [GC], Application no. 48939/99, Judgment of 30 November 2004.
European Court of Human Rights, *Budayeva and Others v. Russia*, Applications nos. 15339/02 et al., Judgment of 20 March 2008.
European Court of Human Rights, *D.H. and Others v. the Czech Republic* [GC], Application no. 57325/00, Judgment of 13 November 2007.
European Court of Human Rights, *Stec and Others v. the United Kingdom* [GC], Applications nos. 65731/01 and 65900/01, Judgment of 12 April 2006.
European Court of Human Rights, *S. and Marper v. the United Kingdom* [GC], Applications nos. 30562/04 and 30566/04, Judgment of 4 December 2008.
European Court of Human Rights, *Big Brother Watch and Others v. the United Kingdom* [GC], Applications nos. 58170/13, 62322/14 and 24960/15, Judgment of 25 May 2021.
Supreme Court of Wisconsin, *State v. Loomis*, 2016 WI 68, 881 N.W.2d 749 (2016)
Supreme Court of Canada, *Ewert v. Canada*, 2018 SCC 30 (2018)

PREISPITIVANJE ZAŠTITE LJUDSKIH PRAVA U ERI VEŠTAČKE INTELIGENCIJE: OD EX-POST PRAVNIH LEKOVA KA PREVENTIVNOJ ZAŠTITI PO DIZAJNU

Dok veštačka inteligencija ubrzano preuzima ulogu posrednog, često nevidljivog faktora u odlučivanju u sferama presudnim za ostvarivanje ljudskih prava, od socijalne zaštite do pravosuđa, pravni poredak i dalje dominantno počiva na ex-post logici, odnosno na reagovanju tek nakon što šteta nastane. Autor u radu ukazuje na funkcionalne defekte takvog modela u uslovima savremenih algoritamskih sistema, gde netransparentnost odlučivanja, automatizacija i fragmentisana odgovornost otežavaju identifikovanje odgovornog subjekta, dokazivanje kauzaliteta i delotvorno osporavanje odluka, dok istovremeno omogućavaju difuzne i sistemske povrede, naročito kroz diskriminaciju i intenziviranje nadzora. Kroz prizmu evropskog normativnog okvira, a posebno Uredbe o veštačkoj inteligenciji, autor u radu razmatra prelazak na preventivno upravljanje rizicima i konceptualizuje „prava po dizajnu" (rights-by-design) kao nužnu arhitektonsku evoluciju zaštite ljudskih prava u algoritamskim okruženjima. Istovremeno se upozorava da ovaj pristup mora ostati vezan za merljive zaštitne ishode i delotvorne mehanizme odgovornosti, kako se ne bi sveo na tehnokratski formalizam i birokratsko simuliranje usklađenosti.

Ključne reči: *veštačka inteligencija, zaštita ljudskih prava, ex-post pravni lekovi, prava po dizajnu, Uredba EU o veštačkoj inteligenciji, procena uticaja na osnovna ljudska prava.*