Daniele Amoroso

Human oversight as a procedural safeguard under the Council of Europe AI Framework Convention

(doi: 10.1422/120418)

Sistemi intelligenti (ISSN 1120-9550)
Fascicolo 1, aprile 2026

DANIELE AMOROSO

HUMAN OVERSIGHT AS A PROCEDURAL SAFEGUARD UNDER THE COUNCIL OF EUROPE AI FRAMEWORK CONVENTION

1. Introduction

A few months ago, Guglielmo invited me to deliver two lectures at the 63rd ISODARCO Course in Volterra. Due to a misunderstanding (the details of which I will spare the reader), I had to prepare my second presentation in less than two hours. With little time and even less clarity, I chose to speak on the recently adopted Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (the AI Framework Convention). It was, admittedly, not my finest hour. Yet, in his characteristically generous and reassuring way, Guglielmo congratulated me on the presentation, adding that he always learns something from me.

That was, of course, a lovely lie from a caring friend. But it gave me the idea for this article. In what follows, I will return to the AI Framework Convention, focusing on a question that has featured in several papers I have had the privilege of co-authoring with Guglielmo: whether the Convention can be interpreted as establishing a legal obligation to ensure human oversight of AI systems when their outputs may affect the enjoyment of human rights.

If one were to look solely at the text of the Convention, the question would be easily answered, with a plain and resounding "no". Yet such an answer would tell only half the story, if not even less. As will be shown, such requirement was included – albeit in varying formulations – in several drafts of the Convention and the notion of human oversight finally found its way at least into the Explanatory Report. Against this backdrop, one may legitimately wonder whether – and to what extent – human oversight still plays a normative role within the Convention.

In this respect, it is crucial to distinguish the notion of "human oversight" from the general "oversight requirement" set forth in Article 8 of the Convention. As clarified in the Explanatory Report, "oversight" in Article 8 refers broadly to systemic and institutional supervision of AI, such as independent audits, review boards, or authorities monitoring

compliance[1]. "Human oversight", by contrast, more specifically refers to direct human involvement in reviewing, validating, or overruling individual AI-generated outputs, viz. it presupposes that a human is "in" or "on the loop" when decisions affecting individuals are made. While institutional oversight mechanisms may include elements of human oversight, they extend well beyond it[2].

The contribution will be structured as follows. Section 2 briefly retraces the rise and fall of the human oversight requirement during the Convention's drafting history. Section 3 examines the legal value of its inclusion in the Explanatory Report, in light of Articles 31 and 32 of the Vienna Convention on the Law of Treaties. Section 4 argues that, even in the absence of an express obligation in the Convention, a requirement of human oversight may nonetheless be inferred from the broader logic of human rights protection, as reflected in the doctrine of positive State obligations. Finally, Section 5 identifies several practical considerations that regulators should bear in mind when designing human-AI interactions, in order to ensure that human oversight genuinely contributes to the protection of fundamental rights.

2. The human oversight requirement in the drafting history of the AI Framework Convention

The idea of a human oversight requirement featured prominently throughout most of the process culminating with the adoption of the AI Framework Convention, starting with the Feasibility Study published by the Ad hoc Committee on Artificial Intelligence (CAHAI) on 17 December 2020, where "the necessity to ensure sufficient human control and oversight" is mentioned among the "essential principles that are relevant for the protection of human rights" and that "are currently not explicitly legally assured"[3]. Almost a year later, the same point was raised in the final paper on the "Possible elements of a legal framework on artificial intelligence", which emphasized that, in relation to the use of AI systems in the public sector, the future Convention should have included, "as a minimum, [...] a mandatory right to human review of decisions taken or informed by an AI system [...], and an obligation for public authorities to implement adequate human review for processes which are informed or supported by AI systems"[4].

[1] Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 5 September 2024, paras. 63-65 ("Explanatory Report").

[2] *Ibid*., para. 65.

[3] Ad hoc Committee on Artificial Intelligence, Feasibility Study, 17 December 202, CAHAI(2020)23, para. 85.

[4] Ad hoc Committee on Artificial Intelligence, Possible elements of a legal frame-

It is no surprise, therefore, that the Zero draft, proposed by the Chair of the Committee on Artificial Intelligence (CAI) and released on 30 June 2022[5], included the "human oversight" requirement among the "procedural safeguards" set forth by draft Article 7. Notably, a veritable "right to human review" of decisions substantially informed or taken by AI systems whenever they were capable of "affecting human rights and fundamental freedoms, legal rights and interests" was mentioned at para. 2; para. 4, on the other hand, established "the right to know that one is interacting with an artificial intelligence system rather than with a human" and the corresponding obligation to provide, "where appropriate, [...], for the option of interacting with a human in addition to or instead of an artificial intelligence system". Similar provisions were included in Article 20 of the Revised Zero Draft of 6 January 2023 as "additional procedural safeguards"[6]. Therefore, the Chair understood the human oversight requirement as comprising two distinct prongs: first, a right to human review, and second, a right to human interaction.

In the Consolidated Working Draft, circulated on 7 July 2023 following the first reading of the Revised Zero Draft within the Committee[7], the former prong of "human oversight" was no longer included among the procedural safeguards. Article 14, para. 2, however, retained the second prong, i.e. the right to know that one is interacting with an AI system and the obligation to ensure, where appropriate, the option of interacting with a human. By December 2023, after the second reading, this second element had been dropped too. While the right to be notified that one is interacting with an AI system was kept, the obligation to provide for the possibility to interact with a human disappeared. The "importance of human review/oversight" was merely recognized in draft Article 15, albeit in brackets, as a foundation for the obligation to envision procedural safeguards.

Not even this attenuated language survived the third and final reading. Article 15 of the Convention now only establishes the Parties' duty to ensure the availability of "effective procedural guarantees, safeguards and rights" whenever an artificial intelligence system significantly impacts upon the enjoyment of human rights. To find an explicit mention

work on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law, 3 December 2021, CAHAI(2021)09rev, para. 34.

[5] Committee on Artificial Intelligence, Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, 30 June 2022, CAI(2022)07.

[6] Committee on Artificial Intelligence, Revised Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, 6 January 2023, CAI(2023)01.

[7] Committee on Artificial Intelligence, Consolidated Working Draft of the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, 7 July 2023, CAI(2023)18.

of "human oversight," one must turn to the Commentary to Article 15 in the Explanatory Report submitted to the Committee of Ministers of the Council of Europe. The relevant part of the commentary reads as follows:

> Where an artificial intelligence system substantially informs or takes decisions impacting on human rights, effective procedural guarantees should, for instance, include human oversight, including ex ante or ex post review of the decision by humans. Where appropriate, such human oversight measures should guarantee that the artificial intelligence system is subject to built-in operational constraints that cannot be overridden by the system itself and is responsive to the human operator, and that the natural persons to whom human oversight has been assigned have the necessary competence, training and authority to carry out that role[8].

Human oversight is also mentioned – albeit briefly – in the commentary to Article 8 ("Transparency and oversight"), where a footnote recalls that "[t]he Drafters noted the important link between this concept and that of human determination and human agency"[9].

The legal value of these texts, as well as the interpretative cues one may draw therefrom, will be expounded on later in this contribution. Here it will just be noted that the CAI's deliberation not to include any reference to "human oversight" in the text of the Convention ran counter to the expectations of many, even within the Council of Europe. The need for an "explicit safeguard" of "the right to human review of automated decisions", for instance, was voiced by the Commissioner for Human Rights in a statement released while the negotiations were approaching the final stage[10]. In a similar view, the Parliamentary Assembly, right after the finalization of the text by the CAI, unanimously issued an Opinion urging the Committee of Ministers to (re)introduce a reference to "human review" in Article 15, para. 1[11].

What happened, then, to human oversight, review, interaction? Who "killed" them? The answer cannot be found in the *travaux preparatoires*, as the Drafting Group met behind closed doors – a deviation from the negotiating practice of the Council of Europe, and one that was, quite expectedly, strongly criticized[12]. According to some sources, the language of what eventually became Article 15 was progressively watered down at the insistence of non-member States actively involved in the negotiations – chief among them, the United States, which reportedly opposed

[8] Explanatory Report, para. 103.

[9] Explanatory Report, para. 65, footnote 5.

[10] Commissioner for Human Rights, AI instrument of the Council of Europe should be firmly based on human rights, 13 March 2024.

[11] Parliamentary Assembly, Draft framework convention on artificial intelligence, human rights, democracy and the rule of law, 18 April 2024, Opinion 303 (2024), para. 9.6.

[12] Hickock, Rotenberg, Caunes (2023).

any binding language on human involvement in AI decision-making[13]. As critics noted, this concession was part and parcel of a wider political trade-off aimed at securing signatures from non-member States, such as the United States, Canada, Japan, and Israel, with a view to positioning the Framework Convention as the first global AI treaty, at the cost of weakening its normative strength and legal scope[14].

3. The legal value of the reference to human oversight in the Explanatory Report

Any inquiry into whether the AI Framework Convention entails a legal duty to ensure human oversight over AI systems must begin with some basic premises. First, the importance of human oversight was consistently acknowledged by the drafters, albeit in evolving formulations. Second, notwithstanding this recognition, a decision was taken – at a late stage in the negotiations – to remove any reference to human oversight (or cognate concepts) from the text of the Convention. Third, in seeking to clarify the otherwise vague expression "effective procedural guarantees, safeguards and rights" in Article 15, the Explanatory Report singles out one example only: human oversight. Fourth, the Report also highlights the connection between human oversight and individual autonomy – framed as "human determination and human agency" – which stands as one of the Convention's foundational principles[15].

Taken together, these elements support the view that, in the Drafters' minds, human oversight was the main procedural safeguard in situations where AI systems may affect the enjoyment of human rights. This does not say much, however, as to whether human oversight is a *mandatory* procedural safeguard under Article 15 or it is just *one possible way* to implement the obligations arising from that provision. To answer this question, one should preliminarily establish what interpretative value should be accorded to the Explanatory Report.

From an international law perspective, the search for an answer begins with the standard playbook for treaty interpretation: Articles 31 and 32 of the Vienna Convention on the Law of Treaties, which set out – respectively – the general rule and the supplementary means of interpretation. Under the general rule under Article 31, "[a] treaty shall be interpreted in good faith in accordance with the ordinary meaning to be given to the terms of the treaty in their context and in the light of its object and purpose"; supplementary means of interpretation under

[13] Bertuzzi (2024).

[14] Dubrocard (2024).

[15] See Article 7 of the Convention (Human dignity and individual autonomy). On the nexus between individual autonomy and human agency, see Explanatory Report, paras. 53 and 55.

Article 32, which include the "preparatory work of the treaty" and "the circumstances of its conclusion," can solely be relied on to confirm the meaning reached through the general rule or to determine the meaning when applying the general rule yields ambiguous, obscure, or manifestly unreasonable results.

According to leading commentators of the Vienna Convention, explanatory reports may – under certain conditions – form part of the "context" under Article 31[16], either as an "agreement relating to the treaty which was made between all the parties in connection with the conclusion of the treaty"[17] or as an "instrument" made "in connection with the conclusion of the treaty and accepted by the other parties as an instrument related to the treaty"[18]. This is especially so when the report has been prepared by governmental experts (as opposed to independent ones), and its adoption has gone uncontested by the other Parties[19].

On its face, the Explanatory Report appears to match this description[20]. On the one hand, the Report was prepared by the Drafting Group, which included governmental experts representing the 46 member states of the Council of Europe and other potential parties; on the other hand, it seems to have been approved without reservation or formal dissent. This might suggest that the Explanatory Report constitutes an "authentic" or "authoritative" interpretation of the Convention – namely, one issued by those with "the power to modify or suppress it"[21] – and thus capable of determining its meaning with legally binding effect[22].

That conclusion, however, is complicated by the Report's own disclaimer. In its opening section, the Drafters state that the Report "does not constitute an instrument providing an authoritative interpretation of the text of the Framework Convention"[23].The authors of the Report, hence, ruled out – in the clearest terms – any intent to bind future interpretations of the Convention by the Parties, clarifying instead that the Report's purpose is simply to "facilitate the understanding of [its] provisions"[24].

In light of these considerations, a more defensible view is that the Explanatory Report constitutes a "supplementary means of interpretation"

[16] Sinclair (1984), p. 129; Villiger (2009), p. 403; Dörr (2018), p. 591.

[17] Dörr (2018), p. 591.

[18] Villiger (2009), p. 403.

[19] *Ibid*.

[20] Significantly enough, Sir Ian Sinclair cited the Council of Europe's practice of publishing Explanatory Reports prepared by governmental experts involved in treaty negotiations as a prime example of documents that form part of a treaty's "context." Sinclair (1984), pp. 129-130.

[21] Permanent Court of International Justice, Question of Jaworzina (Polish-Czechoslovakian Frontier), PCIJ Ser B No 8 (1923), p 37.

[22] Villiger (2009), p. 429; Dörr (2018), p. 570.

[23] Explanatory Report, para. II.

[24] *Ibid*.

under Article 32 VCLT, which may be referred to in case of ambiguity or obscurity[25]. On that score, it is hard to deny that the broad and unqualified language of Article 15 is, in effect, *in search of meaning*. In this respect, the Explanatory Report's reference to "human oversight" may undoubtedly flesh out the provision. Still, this stops short of establishing a legal obligation. The Report uses the modal "should"[26], which plainly signals that human oversight is the preferred approach to implementing the safeguards envisaged in Article 15 – perhaps even a paradigmatic example thereof – but not a mandatory one.

4. Human Oversight as a Positive Obligation under Human Rights Law

Although the Convention does not explicitly impose a human oversight requirement, it remains to be seen whether such a requirement can be inferred from the broader human rights framework that the Convention is intended to implement in the AI domain.

Throughout the drafting process, human oversight was consistently included among the "procedural safeguards", which are now regulated by Article 15. In other words, the human involvement in the decision-making process has been conceptualized as a fail-safe mechanism, designed to prevent AI systems from producing outcomes that unlawfully interfere with the enjoyment of human rights[27].

The need for human fail-safe is rooted in the way AI systems operate and, more importantly, in the ways they fail. According to several studies, and as many of us experience in everyday interactions with AI, even the most advanced AI systems remain prone to unpredictable and counterintuitive errors that are sometimes unfathomable even to their own developers. These errors include contextual misinterpretations and perceptual distortions[28]; moreover, in generative contexts, AI systems may fabricate information entirely and present it with unwavering confidence ("hallucinations")[29]; finally, they are vulnerable to intentional

[25] Polakiewicz (1999), p. 27. An exception to this conclusion arises when the Committee of Ministers expressly endorses the Explanatory Report. This occurred in the case of the Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, whose Explanatory Report was, for that reason, explicitly deemed to form part "of the context in which the meaning of certain terms used in the Convention is to be ascertained". See Explanatory Report to the Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, 10 October 2018, para. 6.

[26] Explanatory Report to the AI Framework Convention, para. 103.

[27] On this issue see, also for further references, Amoroso, Tamburrini (2021), pp. 252-253.

[28] See e.g. Lepori, Mozer, Ghandeharioun (2025).

[29] See e.g. Rathkopf (2025).

manipulations, or "adversarial attacks", whereby small perturbations to the input data result in wildly incorrect outputs[30]. When these kinds of mistakes occur in areas where rights and freedoms are at stake, a human rights issue necessarily arises.

Crucially, in most cases, these failures can be prevented, or their consequences mitigated, by the presence of a human in the decision-making loop. This is not to idealize human judgment, of course. Human agents are also fallible, and often egregiously so. They may act on bias, fatigue, misinformation, when not on bad faith. Yet, humans do not fail in the same way as machines. The difference between the two lies in what AI researchers describe as the "semantic gap": unlike human agents, AI systems do not understand the meaning of the inputs they receive or the outputs they generate[31]. They merely detect statistical correlations within vast datasets and identify patterns that may escape human perception. However, they cannot grasp what they are actually dealing with, nor the consequences of the decisions they are taking or suggesting. It is precisely this asymmetry that justifies the qualification of human oversight as a critical safeguard in the quality of the decision-making processes. Its function is to introduce a layer of judgment that can flag and address failures that AI systems themselves are ill-equipped to assess.

The EU AI Act openly embraces this logic in Article 14, which sets the normative goal of the human oversight requirement as "prevent[ing] or minimis[ing] the risks to health, safety, or fundamental rights"[32]. Still, understanding human oversight as a risk mitigation tool also clarifies the basis for affirming its mandatory character under human rights law and, by extension, under the AI Framework Convention.

Indeed, the European Court of Human Rights has consistently held that States are under a positive obligation to adopt effective regulatory frameworks aimed at minimizing foreseeable risks to the enjoyment of human rights[33]. Similar views are reflected in the interpretative practice of other human rights bodies, which emphasize that States are bound to adopt legislative, administrative, or other general measures to shield individuals from the consequences of potentially dangerous activities[34]. While these positive obligations have traditionally applied to risks arising from hazardous industrial activities or environmental threats, their preventive rationale naturally extends to the AI domain: where rights are

[30] See e.g. Jha (2025).

[31] See e.g. Horsley (2025).

[32] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Article 14, para. 2.

[33] Ambrus (2017).

[34] Shelton (2022).

foreseeably endangered, regulators must take the most efficient general measures to contain those risks.

In this light, the deployment of a system that substantially informs – or directly takes – decisions capable of negatively affecting the enjoyment of human rights may well qualify as a source of risk that States (and the European Union) are under a duty to govern and mitigate. After all, this idea finds a foothold in the Framework Convention itself, whose Article 16 obliges Parties to establish a "risk and impact management framework" tailored to the potential dangers that AI systems pose to human rights. Once it is accepted that human involvement enhances decision-making processes by reducing the likelihood of rights violations, the introduction of a human oversight requirement emerges as a regulatory measure envisioned under Article 16, para. 2, which requires that "adverse impacts of artificial intelligence systems to human rights [...] are adequately addressed".

More importantly, given that the distinctive features of human judgement cannot be replicated by automated processes (and are unlikely to be in the foreseeable future), the role of human oversight should be regarded, at present, as *irreplaceable*. Its mitigating function, in other words, cannot be effectively substituted by technical safeguards alone (e.g. self-checking AI). If this is the case, human oversight ceases to be merely a recommended procedural safeguard. Instead, it becomes an indispensable bulwark for the protection of human rights and, as such, constitutes the object of a positive obligation under human rights law.

5. How to Ensure that Human Oversight Effectively Serves as a Safeguard?

The challenge of designing effective human-machine interactions in sensitive domains long predates the AI Framework Convention, with a prominent precedent being the diplomatic and scholarly debate around ensuring Meaningful Human Control (MHC) over Autonomous Weapons Systems.[35] While no consensus has yet been reached on the normative and operational content of this concept, the fundamental assumption conveyed by the qualifier "meaningful" is uncontested: to fulfill the normative expectations associated with human involvement, it is not enough to simply keep a human in or on the loop; the quality of such involvement must be assured, so that so that it does not provide solely symbolical contribution[36].

[35] For a recent overview, see Sehrawat (2025). The notion of Meaningful Human Control has recently been extended to other domains as well, including autonomous driving and robotic surgery. See Mecacci *et al*. (2024).

[36] Roff, Moyes (2016), p. 2.

To be fair, this is an area where legal expertise, while still relevant, takes a back seat to other disciplines such as ergonomics, systems engineering, human-computer interaction, and cognitive psychology. Accordingly, the following observations aim only to outline briefly the key conditions under which human oversight can meaningfully serve its safeguard function, drawing in particular on the valuable cues provided by the Explanatory Report. At the same time, it is useful to consider how a closely related notion has been articulated in a more detailed regulatory instrument, namely Article 14 of the EU AI Act on "human oversight" for high-risk AI systems, which does provide a concrete, structured template that may serve as a source of inspiration when giving effect, under the Framework Convention, to the requirement of "effective procedural guarantees" in rights-sensitive contexts[37].

First of all, human oversight cannot be confused with the fact that AI systems are programmed or trained by human beings. While human involvement at the development stage is undeniably important – and rightly acknowledged in the Explanatory Report, which highlights the need for "built-in operational constraints that cannot be overridden by the system itself"[38] – this alone may be a necessary, but not sufficient, condition for fulfilling the safeguard function. For human oversight to serve as a fail-safe, indeed, operators must retain control privileges during the system's actual operation. These include the power to approve a system's recommendation before it takes effect ("*ex ante* review") or to override a decision after it has been made ("*ex post* review")[39]. In this respect, the EU AI Act is instructive in two ways. On the one hand, it frames human oversight as something that must be enabled by design, including through "appropriate human-machine interface tools", so that the system can be "effectively overseen by natural persons during the period in which [it is] in use" (Art. 14, para. 1). On the other hand, it specifies that human overseers should be enabled to disregard, override, or reverse the system's output, as well as to "intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button

[37] In this connection, one cannot help noting the near-verbatim convergence between the Explanatory Report and Recital 73 of the AI Act, which explains the human-oversight requirement laid down in Article 14. Both texts, in particular, envisage that human oversight measures should ensure that the AI system is subject to "built-in operational constraints that cannot be overridden by the system itself" and is "responsive to the human operator", and that the "natural persons" assigned to oversight have the "necessary competence, training and authority" to carry out that role. The convergence is plausibly explained by the temporal and institutional overlap between the two regulatory processes, which unfolded largely in parallel and in which the European Union participated through the European Commission, as well as with the Commission's broader negotiating posture to keep the Convention closely aligned with the AI Act's architecture. See Almada, Radu (2024).

[38] Explanatory Report, para. 103.

[39] *Ibid*.

or a similar procedure that allows the system to come to a halt in a safe state" (Art. 14, para. 4, lit. d) and e)).

Second, individuals entrusted with oversight, which the Report quite significantly refers to as "*natural* persons", must possess "the necessary competence, training and authority" to perform this task effectively[40]. Domain expertise (e.g. in public administration or health care) will not be sufficient. Human operators should also be required to have a functional understanding of the AI system itself. Training programs, however, should extend beyond technical operation to include awareness of the system's known failure modes, its susceptibility to manipulation, and the inherent unpredictability of its outputs. Just as importantly, such training must help build a culture of critical engagement, in which human operators are encouraged to rely on AI tools without abdicating independent judgment or critical scrutiny. Here too, Article 14 of the AI Act offers a useful concretization of what "competence" and "training" are supposed to achieve. It requires that overseers be enabled to understand the system's "relevant capacities and limitations", to monitor its operation in view of "detecting and addressing anomalies, dysfunctions and unexpected performance", and crucially to remain aware of the tendency to automatically rely or over-rely on AI outputs, by expressly referring to the notion of "automation bias" (Art. 14, para. 4, lit. a and b).

Third, human-machine interfaces should be designed in a way that supports active intervention, rather than nudging users toward default acceptance. As the Explanatory Report emphasizes in its commentary on Article 8, two design features are key here: interpretability and explainability[41]. If operators are expected not to blindly trust the machine, they should be put in a position to get a sufficient amount of humanly understandable information about machine data processing ("interpretability"); and to additionally obtain an account of the reasons why the machine is suggesting or embarking on a certain course of action ("explainability")[42]. The EU AI Act confirms this perspective by insisting not only on interface tools (Art. 14, para. 1), but also on enabling overseers to "correctly interpret" the system's output, taking into account "interpretation tools and methods available" (Art. 14, para. 4, lit. c)). Quite remarkably, then, Article 14 treats interpretability not as a purely "transparency" concern, but as a functional precondition for intervention: without intelligible outputs and usable interface affordances, the right to override, disregard or interrupt risks remaining theoretical.

[40] *Ibid*.

[41] Explanatory Report, paras. 60-62. See also Amoroso, Tamburrini (2021), pp. 263-264.

[42] It should be observed, however, that the achievement of interpretable and explainable AI systems poses daunting challenges to AI experts. See Roy, Mondal, Roy, Roy (2025).

Fourth and finally, effective human oversight also depends on a dimension not explicitly addressed by the Explanatory Report, nor by Article 14 of the AI Act: the organizational and institutional context in which it is embedded. Empirical studies have shown that organizational pressures – such as time constraints, risk aversion, hierarchical deference, or punitive tracking of human error – can severely inhibit the willingness of individuals to challenge or override AI-generated outputs[43]. In such environments, oversight may devolve into a rubber-stamping exercise. By contrast, institutions that encourage careful scrutiny of AI decisions, by rewarding critical engagement and creating a psychologically safe space for intervention, are far more likely to foster meaningful oversight.

Three further features of Article 14 AI Act deserve brief mention, insofar as they may inform the "calibration" of human oversight under the Framework Convention. First, Article 14 explicitly adopts a proportionality logic, in that oversight measures must be "commensurate with the risks, level of autonomy and context of use" of the system (Art. 14, para. 3). Second, these measures may be either built into the system by the "provider" (i.e. the developer) and/or implemented by the deployer (i.e. the "user") to implement (Art. 14, para. 3, lit. a and b)[44]. This provider/deployer bifurcation is conceptually important for Convention implementation as well, because it clarifies that meaningful human oversight is seldom achieved at a single touchpoint in human-machine interactions. Rather, it is typically the result of layered measures (technical affordances, training, procedural steps, institutional safeguards), distributed across different actors. Third, Article 14, para. 5, contains an "enhanced oversight" model for certain remote biometric identification systems, requiring that no action be taken on the basis of identification unless it is verified and confirmed by at least two competent natural persons[45]. This illustrates a more general principle: where the error costs are particularly acute (or where the technology is particularly prone to false positives), oversight may need to be institutionalized as redundancy rather than left to a single reviewer.

The foregoing remarks offer only a preliminary sketch. Much more work is needed to translate the abstract notion of human oversight into concrete design features and organizational routines. In this respect, the Council of Europe's Committee on Artificial Intelligence could play a pivotal role. As a multistakeholder body, the CAI is well placed to

[43] See e.g. Haduong, Smith (2025).

[44] For more accurate definitions of "provider" and "deployer" under the AI Act, see Article 3, para. 1, nn. (3) and (4).

[45] Quite disappointingly, such "enhanced oversight" does not apply to the most human rights-sensitive uses of biometric identification systems, namely those employed "for the purposes of law enforcement, migration, border control or asylum, where Union or national law considers the application of this requirement to be disproportionate" (Art. 14, para. 5).

identify best practices, elaborate interpretative guidance, and promote shared benchmarks on what human oversight should entail in different contexts of use. This could involve, for instance, articulating indicators to assess the quality of human-machine interaction, issuing recommendations on interface and organizational design, mapping sector-specific risks where heightened human oversight is needed, and drafting model protocols. In this way, the CAI would provide crucial support to Parties and businesses in meeting their obligations under the Convention and human rights law in general.

References

Almada, M., Radu, A. (2024). The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy. *German Law Journal*, *25*, pp. 646-663.

Ambrus, M. (2017). The European Court of Human Rights as Governor of Risk. In M. Ambrus, R. Rayfuse, W. Werner (Eds.), *Risk and the Regulation of Uncertainty in International Law*. New York: Oxford University Press, pp. 99-116.

Amoroso, D., Tamburrini, G. (2021). Toward a Normative Model of Meaningful Human Control over Weapons Systems. *Ethics & International Affairs*, *35*, pp. 245-272.

Bertuzzi, L. (2024). Tug of war continues on international AI treaty as text gets softened further, https://www.euractiv.com/news/tug-of-war-continues-on-international-ai-treaty-as-text-gets-softened-further/ (Accessed 30 October 2025).

Dörr, O. (2018). General rule of interpretation. In O. Dörr, K. Schmalenbach (Eds.), *Vienna Convention on the Law of Treaties. A Commentary*, 2nd edn. Berlin: Springer, pp. 559-616.

Dubrocard, M. (2024). The Council of Europe Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law: perhaps a global reach, but an absence of harmonisation for sure, https://free-group.eu/2024/05/06/the-council-of-europe-convention-on-artificial-intelligence-human-rights-democracy-and-the-rule-of-law-perhaps-a-global-reach-but-an-absence-of-harmonisation-for-sure/ (Accessed 30 October 2025).

Hickock, M., Rotenberg, M., Caunes, K. (2023). The Council of Europe Creates a Black Box for AI Policy. In *Verfassungsblog*, https://verfassungsblog.de/coe-black-box-ai/ (Accessed 30 October 2025).

Horsley, J. (2025). Mind the semantic gap: semantic efficiency in human computer interfaces. *Frontiers*, *8*, pp. 1-11.

Jha, P.K. (2025). Adversarial Machine Learning: Attacks, Defenses, and Open Challenges. *arXiv*, https://arxiv.org/html/2502.05637v1 (Accessed 30 October 2025).

Lepori, M.A., Mozer, M.C., Ghandeharioun, A. (2025). Racing Thoughts: Explaining Contextualization Errors in Large Language Models. In L. Chiruzzo, A. Ritter, L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics:*

Human Language Technologies (Volume 1: Long Papers). Albuquerque, New Mexico: Association for Computational Linguistics, pp. 3020-3036.
Mecacci, G. *et al*. (Eds.) (2024). *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*. Cheltenham: Edward Elgar Publishing.
Polakiewicz, J. (1999). *Treaty-making in the Council of Europe*. Bruxelles: Council of Europe Publishing.
Rathkopf, C. (2025). Hallucination, reliability, and the role of generative AI in science. *arXiv*, https://arxiv.org/html/2504.08526v1 (Accessed 30 October 2025).
Roff, H.M., Moyes, R. (2016). Meaningful Human Control, Artificial Intelligence and Autonomous Weapons, Article 36 Briefing paper prepared for the CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems, https://article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf (Accessed 30 October 2025).
Roy, S., Mondal, S., Roy, B., Roy, C. (2025). From Questions to Insights: Exploring XAI Challenges Reported on Stack Overflow Questions. *arXiv*, https://arxiv.org/abs/2504.03085 (Accessed 30 October 2025)
Sehrawat, V. (2025). Autonomous weapon systems and meaningful human control. In *International Journal of Law and Information Technology*, *33*.
Shelton, D. (2022). Negative and Positive Obligations. In C. Binder, M. Nowak, J.A. Hofbauer, P. Janig (Eds.), *Elgar Encyclopedia of Human Rights*. Cheltenham: Edward Elgar Publishing, pp. 545-550.
Sinclair, I. (1984). *The Vienna Convention on the Law of Treaties*, 2nd edn. Manchester: Manchester University Press.
Villiger, M.E. (2009). *Commentary on the 1969 Vienna Convention on the Law of Treaties*. Leiden-Boston: Martinus Nijhoff Publishers.

Human oversight as a procedural safeguard under the Council of Europe AI Framework Convention

The contribution explores the possibility of identifying an obligation of human oversight of AI systems within the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Although the final text of the Convention does not explicitly impose such an obligation, the need (or at least the desirability) to ensure human oversight played a central role throughout the entire negotiation process and is still mentioned in the Explanatory Report. After retracing the main stages of how the issue was addressed during the negotiations, the contribution focuses on the interpretative value of the reference to human oversight contained in the Explanatory Report in the light of Articles 31 and 32 of the Vienna Convention on the Law of Treaties, concluding that, while it does not qualify as an "authentic" (and thus legally binding) interpretation, the reference to human oversight in the Report provides valuable interpretative guidance. It will also be argued that an obligation of human oversight may nevertheless be derived from the broader human rights framework that the Convention is intended to implement in the AI domain, having particular regard to States' positive obligation to protect individuals against human rights violations. Finally, the article identifies a

number of practical conditions for human oversight to operate effectively as a safeguard mechanism, with particular attention to system design, operator training, and the organisational context.
Keywords: Human oversight, Council of Europe AI Framework Convention, Positive obligations under human rights law.

Daniele Amoroso, Department of Law, University of Cagliari, Viale Sant'Ignazio 17, 09123 Cagliari
daniele.amoroso@unica.it
https://orcid.org/0000-0002-2583-5229