Human Oversight under Article 14 of the EU AI Act

Melanie Fink

APART-GSK Fellow of the Austrian Academy of Sciences, Central European University

Assistant Professor, Europa Institute, Leiden University

Forthcoming in:

Gianclaudio Malgieri, Gloria González Fuster, Alessandro Mantelero, and Gabriela Zanfir-Fortuna (eds), *AI Act Commentary: A Thematic Analysis* (Hart-Bloomsbury, forthcoming 2026)

Contents

1 Introduction

Born out of the goal to ensure human-centric AI, Article 14 AI Act requires that high-risk AI systems can be 'effectively overseen by natural persons'. It sets out detailed obligations for providers of AI systems to create the conditions for effective human oversight during the period in which they are in use. This is complemented by Article 26(2) AI Act, which requires deployers of high-risk AI systems to assign human oversight.

The following first defines what Article 14 means with 'human oversight' (Section 2). Section 3 then provides the context for Article 14's applicability, elaborating on its role within the AI Act's network of risk management obligations (Section 3.1), and its relationship with human oversight requirements outside the AI Act (Section 3.2). Section 4 discusses the purpose human oversight is designed to fulfil in AI governance in general (Section 4.1) and in the AI Act in particular (Section 4.2). Section 5 then elaborates on the concrete human oversight obligations for providers and deployers respectively, distributing them across the three key markers for effectiveness in human oversight: authority (Section 5.1), comprehension (Section 5.2), and environment (Section 5.3). Section 6 provides a look ahead, focusing on how to apply Article 14 against the background of the limits of human oversight mechanisms. Section 7 concludes.

2 Definition: What is Human Oversight under Article 14?

Article 14 AI Act requires that high-risk AI systems can be 'effectively overseen by natural persons'.[1] In its broadest sense, 'human oversight' refers to the involvement of a natural person in an algorithmic work process. It can be achieved through different mechanisms, with oversight exercised at different stages and in varying degrees of intensity. These variants of human oversight are often labelled differently, depending on when and how humans intervene.

In the governance context, two commonly distinguished mechanisms are 'human-in-the-loop' and 'human-on-the-loop'.[2] Both tend to refer to real-time human intervention during the operation of an AI system. In a 'human-in-the-loop' mechanism, a human needs to guide or validate every output before it can take effect, whereas in a 'human-on-the-loop' mechanism, a human only monitors the decision-cycle and has a possibility to intervene at any moment, but does not need to do so.[3] Outside real-time human intervention, two further types of mechanisms can be distinguished. The first is 'human review', where human intervention is ensured (or may be obtained) after an AI system's output has taken effect, for instance to overturn that output.[4] The second is 'human design', where humans are involved at the design, training, and testing

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence [2024] OJ L2024/1689, in the following 'AI Act'.

[2] On the origins of these terms see Riikka Koulu, 'Human Control over Automation: EU Policy and AI Ethics' (2020) 12 European Journal of Legal Studies 9.

[3] Guillermo Lazcoz and Paul de Hert, 'Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities' (2023) 50 Computer Law & Security Review 105833, 7; Rebecca Crootof, 'A Meaningful Floor for "Meaningful Human Control"' (2016) 30 Temple International & Comparative Law Journal 53, 54.

[4] European Commission, 'White Paper on Artificial Intelligence – A European Approach to Excellence and Trust', 19 February 2020, COM(2020) 65 final, 21; for Lazcoz and Hert (n 3) this is a 'human out of the loop'.

stages, but not at all during the operation of the AI system.[5] None of these labels is used consistently across different fields and contexts or have a fixed legal meaning.[6] 'Human-in-the-loop', for instance, may be used more broadly, encompassing mechanisms that require human involvement during the operation of an AI system more generally, thus also including 'human-on-the-loop' mechanisms.[7] Similarly, also 'human-on-the-loop' may be understood more broadly, to include mechanisms that operate during the design stage of an AI system.[8]

Article 14 does not explicitly prescribe a specific human oversight mechanism. The key requirement is that whatever oversight mechanism is chosen, needs to be 'effective' and 'commensurate with the risks, level of autonomy and context of use of the high-risk AI system'.[9] 'Effective' seems to correspond, in essence, to 'meaningful',[10] which has emerged as a benchmark to assess human intervention required under the EU's General Data Protection Regulation (GDPR), but also in the discussion surrounding autonomous weapon systems.[11] Clearly, a *pro forma* human that 'rubber-stamps' the AI system's output does not meet this criterion. Beyond this, the line between effective and not effective human oversight can only be drawn on a case-by-case basis and with due regard to the list of concrete abilities the human overseer should have as set out in Article 14(4). This is discussed in detail in Section 5.

While Article 14 focuses on the effectiveness of human oversight without specifying how to achieve that, it does exclude some oversight mechanisms by implication. Human oversight of AI systems as envisaged by Article 14 is to be exercised 'during the period in which they are in use'.[12] It thus begins with deployment and can never be satisfied by human involvement at the design or training stage of an AI system *alone*. Strictly speaking, this also means that Article 14 itself, structured around provider obligations, sets out an obligation to *create the conditions for* human oversight, whereas the actual human oversight obligation is found in Article 26(2), which requires deployers to 'assign human oversight'.

Whether Article 14 demands a real-time intervention possibility in all cases is less clear. This may be deduced from Article 14(4)(e) which requires that the person to whom human oversight is assigned is able to 'intervene in the operation [...] or interrupt the system through a "stop" button

[5] For Lazcoz and de Hert (n 3), this is 'human back in control'; for the High-Level Expert Group on Artificial Intelligence, this is a form of 'human-on-the-loop', see AI HLEG, 'Ethics Guidelines for Trustworthy AI', 8 April 2019, 16.

[6] Therese Enarsson, Lena Enqvist and Markus Naarttijärvi, 'Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts' (2022) 31 Information & Communications Technology Law 123, 126.

[7] Rebecca Crootof, Margot E Kaminski and W. Nicholson Price II, 'Humans in the Loop' (2023) 76 Vanderbilt Law Review 429, specifically note 37; Meg L Jones, 'The right to a human in the loop: Political constructions of computer automation and personhood' (2017) 47 Social studies of science 216; Kiel Brennan-Marquez, Karen Levy and Daniel Susser, 'Strange Loops: Apparent versus Actual Human Involvement in Automated Decision Making' (2019) 34 Berkeley Technology Law Journal 745, 749.

[8] See AI HLEG (n 5) 16. However, they further distinguish a 'human in command', which 'refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation'.

[9] AI Act, art 14(1, 3).

[10] Sarah Sterz and others, 'On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives' (2024), fn1

[11] For AWS: Giulio Mecacci and others (eds), *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems* (Edward Elgar Publishing 2024); Crootof (n 3); on the GDPR see below Section 3.2.

[12] AI Act, art 14(1).

or a similar procedure'. However, as Section 5 explains in more detail, it is unclear whether the list of requirements in Article 14(4) has to be understood cumulatively or alternatively. If understood cumulatively, real-time intervention possibility in the form of a 'human-in-the-loop' or a 'human-on-the-loop' mechanism would *always* be required. If understood alternatively, real-time intervention possibilities would only be necessary if that is 'appropriate and proportionate' in a specific case.

The only explicit requirement for a specific human oversight mechanism appears in Article 14(5). It sets out that an output of a remote identification system, such as video surveillance systems that make use of facial recognition technology, cannot produce effects unless it 'has been separately verified and confirmed by at least two natural persons with the necessary competence, training and authority'. Thus, for these types of systems, the AI Act demands a 'human-in-the-loop' mechanism that involves two individuals. This requirement does not apply to high-risk AI systems used 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.

3 Context: The Landscape Around Article 14

3.1 Article 14 within the AI Act

Within the AI Act, Article 14 is part of the requirements for high-risk AI systems that are to be met by the providers of AI systems.[13] To a large extent, these requirements stem from a process that was set in motion in October 2017 with an invitation by the European Council to the Commission to 'put forward a European approach to artificial intelligence'.[14] In the policy documents that were adopted in the following years,[15] the human oversight requirement takes a central place in the overall aim to safeguard 'human-centric' and trustworthy AI.[16] Human oversight emerged as the panacea to limit or even eliminate AI-related risks, as evidenced by both its broad formulation and applicability, as well as its interplay with other safeguards.[17]

Article 14 is exceptionally broad in at least two ways. First, human oversight policies are often specifically concerned with the use of algorithms in decision-making and the role of humans to affect the outcome of a specific individual decision.[18] In contrast, Article 14 applies no matter the sector, context, or role an AI system may fulfil within a work-flow. Second, human oversight requirements usually focus on the deployment stage of AI systems. Article 14, instead, centres on human oversight requirements at the development stage of AI systems, with the legislators

[13] AI Act, art 16(a).

[14] European Council, Conclusions, Brussels, 19 October 2017, 7.

[15] For a detailed account of the process that followed see Nathalie A Smuha, 'The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence' (2019) 20 Computer Law Review International 97; Nathalie A Smuha and Karen Yeung, 'The European Union's AI Act: beyond motherhood and apple pie?' in Nathalie A Smuha (ed), *The Cambridge Handbook on the Law, Ethics and Policy of Artificial Intelligence* (Cambridge University Press forthcoming 2025).

[16] Lazcoz and Hert (n 3) 8.

[17] Lena Enqvist, ''Human oversight' in the EU artificial intelligence act: what, when and by whom?' (2023) 15 Law, Innovation and Technology 508, 514–515.

[18] Crootof, Kaminski and Nicholson Price II (n 7) 440–442. Similarly, see Ben Green, 'The flaws of policies requiring human oversight of government algorithms' (2022) 45 Computer Law & Security Review 105681.

primary preoccupation being the creation of the *conditions for* effective human oversight.[19] Article 14 has only one significant limitation: It is applicable only to those AI systems that qualify as high-risk under the AI Act.

Within the ecosystem of AI safeguards, human oversight takes the position of a *primus inter pares*. This is most clearly expressed in the Ethics Guidelines by the High-Level Expert Group on Artificial Intelligence (AI HLEG) that were later endorsed by the European Commission in its Communication on 'Building Trust in Human-Centric Artificial Intelligence'.[20] The AI HLEG not only lists 'human agency and oversight' as the first among seven key requirements for 'Trustworthy AI'.[21] It also notes that 'the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required'. As Enqvist points out, this suggests an inverse relationship between human oversight and other safeguards, where a decrease in the former triggers an increase in the latter.[22] This idea is not a novelty in EU regulation, but is the approach taken in Article 22 GDPR, which requires additional safeguards to be put in place when automated decisions are made without meaningful human involvement.[23]

This logic was not fully carried forward into the AI Act. Article 14(2) AI Act conceptualises the human overseer as a 'fail safe' that comes into play especially when the other safeguards fail to sufficiently address the risks of an AI system.[24] As such, the AI Act does not generally establish an inverse relationship between the safeguards applicable to high-risk AI systems. In other words, the quality of human oversight – whether extremely good or terribly bad – does not lower or increase the threshold that other safeguards have to meet. Somewhat strangely, though, Article 6(3)(c) excludes AI systems intended to detect decision-making patterns or deviations from the qualification as high-risk when they are not meant to 'replace or influence' a prior human assessment, without 'proper human review'. This seems to suggest that in this context, 'proper human review' justifies the 'downgrading' of a high-risk AI system to a lower risk category, making the safeguards specific to high-risk systems inapplicable.[25]

3.2 Article 14 and its 'Siblings'

The human oversight requirement is by no means unique to the AI Act. In a 2022 survey, Green identified a total of 41 policy documents that contain a mandate or guidance for human oversight of public sector algorithms.[26] In EU legislation, beyond Article 14 AI Act, two sets of legally binding human oversight requirements currently exist.

[19] In this vein see also Lazcoz and Hert (n 3) 8.

[20] AI HLEG (n 5) 16; European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 'Building Trust in Human-Centric Artificial Intelligence', 8 April 2019, COM(2019) 168 final, 4.

[21] AI HLEG (n 5) 12, 14.

[22] Enqvist (n 17) 512.

[23] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data [2016] OJ L2016/119, in the following 'GDPR', art 22(2a, 3).

[24] AI Act, art 14(2). See also David Abbink and others, 'Introduction to meaningful human control of artificially intelligent systems' in Giulio Mecacci and others (eds), *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems* (Edward Elgar Publishing 2024) 7.

[25] I would like to thank Isabella Banks for sharing this insight with me.

[26] Green (n 18) 4; tracing rights to a human decision in European and American law see Aziz Huq, 'A Right to a Human Decision' (2020) 105 Virginia Law Review 611.

Prior to the adoption of the AI Act, legal requirements for human involvement in automated processes stemmed primarily from the GDPR. Article 22 GDPR restricts the use of decision making systems based 'solely on automated processing', when the decision has legal or otherwise significant effects on the person.[27] The dominant interpretation of that provision is that to avoid processing being qualified as 'solely automated', a human must be involved in a 'meaningful' manner, even though what exactly qualifies as 'meaningful' is debated.[28] If no human is (meaningfully) involved, additional measures are required to 'safeguard the data subject's rights and freedoms and legitimate interests'.[29] As opposed to Article 14 AI Act, Article 22 GDPR is only concerned with automated *decision-making*. It is also less demanding in that it does not prohibit automation without human involvement, but only requires additional safeguards to be put in place for those instances. With this in mind, it may be expected that Article 22 GDPR will remain relevant primarily in the context of automated decision-making that does not involve high-risk AI applications, because Article 14 is inapplicable in those cases.

Also the EU's Digital Services Act (DSA) that was adopted in 2022 and governs large online platforms includes a human oversight provision.[30] Article 20 DSA requires platforms to set up internal complaint-handling systems that allow users to challenge the often fully automated content moderation decisions, ranging from deleting a comment to terminating a whole account. According to Article 20(6) DSA, decisions on complaints have to be 'taken under the supervision of appropriately qualified staff, and not solely on the basis of automated means'. With a view to the scale and complexity of content moderation governance systems,[31] it has been argued that Article 20(6) DSA requires humans only to perform high-level supervision, rather than examining every individual complaint.[32] This sets a significantly lower threshold than Article 14 AI Act. Nonetheless, it is an important addition to the AI Act's human oversight safeguards because,

[27] GDPR, art 22; equivalent rights can also be found in Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data [2018] OJ L295/39, art 24; Directive (EU) 2016/680 of the European Parliament and of the Council f 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data [2016] OJ L119/89, art 11; for an in depth analysis of this provision in relation to 'multi stage profiling systems' see Reuben Binns and Michael Veale, 'Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR' (2021) 11 International Data Privacy Law 319.

[28] See, for instance, Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (2018); see also Marco Almada, 'Human intervention in automated decision-making' [2019] ICAIL '19: Seventeenth International Conference on Artificial Intelligence and Law 2; Maja Brkan, 'Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond' (2019) 27 International Journal of Law and Information Technology 91.

[29] GDPR, art 22(2a, 3).

[30] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services [2022] OJ L277/1.

[31] See, for instance, Douek, who argues that from '[a]ccidental beginnings, content moderation governance systems are becoming some of the most elaborate and extensive bureaucracies in history', see Evelyn Douek, 'The Siren Call of Content Moderation Formalism' in Lee C Bollinger and Geoffrey R Stone (eds), *Social Media, Freedom of Speech, and the Future of our Democracy* (Oxford University Press 2022).

[32] Rachel Griffin and Erik Stallman, 'A Systemic Approach to Implementing the DSA's Human-in-the-Loop Requirement', 22 February 2024, Verfassungsblog, https://verfassungsblog.de/a-systemic-approach-to-implementing-the-dsas-human-in-the-loop-requirement/.

even when AI systems are used, neither the content moderation decisions themselves, nor the decisions on complaints qualify as high-risk applications under the AI Act, making Article 14 inapplicable.

Outside the EU context, also the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, the first legally binding international instrument on artificial intelligence, includes human oversight provisions.[33] Article 8 requires 'adequate transparency and oversight'. This is understood broadly, but, as the Explanatory Report clarifies, it also encourages parties 'to implement measures ensuring that these systems are designed, developed and used in such a way that there are effective and reliable oversight mechanisms, including human oversight'.[34] In addition, Article 15(1) requires that where an AI system 'significantly impacts upon the enjoyment of human rights', parties to the Convention have to ensure that effective procedural safeguards are available to persons affected. According to the Explanatory report, in cases where an AI system 'substantially informs or takes decisions impacting on human rights', the guarantees should entail 'human oversight, including ex ante or ex post review of the decision by humans'.[35]

As a treaty under public international law, the AI Framework Convention will be implemented by the contracting parties through the adoption of domestic laws and policies. In the EU, the Union itself will become a party to the Convention and implement it by means of the AI Act.[36] The AI Framework Convention's safeguards are broadly formulated and, in particular, not limited to high-risk AI systems. It thus may set important complementary requirements to the AI Act. To the extent the AI Framework Convention's provisions have direct effect under EU law, they also benefit from the stronger enforcement mechanisms available under EU law.

4 Purpose: Why Human Oversight?

4.1 The Purpose of Human Oversight in AI Governance

Human oversight is not an end in itself, but serves a broad range of purposes that vary between disciplines and contexts.[37] Without the ambition to being exhaustive, the variety of goals pursued by human oversight policies can be clustered into three categories, even though these partially overlap: (1) output-oriented goals, (2) process-oriented goals, and (3) accountability-oriented goals.

Output-oriented goals put humans in charge of identifying and correcting inaccurate or undesirable outcomes. Machine errors may occur when a system malfunctions or is simply prone to mistakes. The idea is that humans can spot these through knowledge, experience, or even intuition. A dramatic example is the 'Petrov incident'. On 26 September 1983, Stanislav Petrov, a

[33] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Vilnius, 5 September 2024.

[34] Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Vilnius, 5 September 2024, para 65.

[35] Ibid, para 103.

[36] Council Decision (EU) 2024/2218 of 28 August 2024 on the signing, on behalf of the European Union, of the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law [2024] OJ L2024/2218.

[37] For a nuanced overview see Crootof, Kaminski and Nicholson Price II (n 7) 473–487.

lieutenant colonel of the Soviet Air Defence Forces, decided to override an automated system that mistakenly detected a nuclear attack on the basis of 'gut feeling', a decision which is considered to have avoided a disastrous nuclear war.[38] The output-correction role of a human comes in especially where humans outperform algorithms. An algorithm's lack of contextual or social knowledge, for instance, or its inability to make moral or value judgments can lead to inaccurate or biased outcomes that a human may be able to correct.[39] In an example described by Kaminski, an algorithm flagged Somali-owned grocery stores in Seattle for food stamp fraud due to 'suspicious transactions' that could, however, be explained by the particular shopping patterns of the East African immigrant community in the area.[40] A human may be able to detect and correct the inaccurate and unfair outcome by filling in the necessary contextual and social knowledge.

Process-oriented goals focus on achieving a 'better' process, regardless of whether the human actually affects the outcome. This is the cluster with most variation, also because the question whether a process is indeed 'better' depends on the perspective taken. From the perspective of individuals affected by AI systems, the addition of a human can safeguard procedural rights, including the right to a reasoned decision, the right to be heard, or the right to an effective remedy.[41] Affected individuals may also simply trust a process more or only feel 'seen' and treated with dignity when other humans are involved.[42] However, many process-oriented goals focus on the perspective of the individual who interacts with the AI system, such as a decision-maker or someone using a chatbot. Human oversight requirements create space for the exercise of human judgment, protecting human agency and autonomy. This has been highlighted as particularly important in the public sector where discretion creates room for the decision-maker to take into account context and respond to novel, marginal, and individual circumstances, and thus needs to be preserved.[43] Finally, process-oriented goals may also have a more systemic or societal focus, with human involvement as a safeguard for the principles of democratic equality and legitimacy.[44]

Accountability-oriented goals, as the name suggests, introduce a human to achieve accountability, which cannot be borne by machines.[45] In the case of self-driving cars, for example, one of the dominant reasons why the law still requires a human driver present in the car

[38] This incident is described by Kiel Brennan-Marquez and Stephen Henderson, 'Artificial Intelligence and Role-Reversible Judgment' (2019) 109 Journal of Criminal Law and Criminology 137, 146.

[39] On the respective strengths of humans and algorithms in this respect see Crootof, Kaminski and Nicholson Price II (n 7) 461–467; Daniel Solove and Hideyuki Matsumi, 'AI, Algorithms, and Awful Humans' (2024) 92 Fordham Law Review 1923; with reference to examples in law enforcement, medicine, finance, and administration, see Brennan-Marquez and Henderson (n 38) 146–148.

[40] Margot E Kaminski, 'Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability' (2019) 92 Colorado Law Review 1529, 1542–1543.

[41] Crootof, Kaminski and Nicholson Price II (n 7) 478–479

[42] Ibid 480–482; Mireille Hildebrandt, 'Privacy As Protection of the Incomputable Self: Agonistic Machine Learning' (2019) 20 Theoretical Inquiries in Law 83; Reuben Binns, 'Human Judgment in algorithmic loops: Individual justice and automated decision-making' (2022) 16 Regulation & Governance 197.

[43] Green (n 18) 3–4. With further references see Enarsson, Enqvist and Naarttijärvi (n 6) 127–129.

[44] Brennan-Marquez and Henderson (n 38); Koulu (n 2) 16.

[45] Crootof, Kaminski and Nicholson Price II (n 7) 482–483; Juliane Beck and Thomas Burri, 'From "human control" in international law to "human oversight" in the new EU act on artificial intelligence' in Giulio Mecacci and others (eds), *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems* (Edward Elgar Publishing 2024) 107–109.

is to ensure that in the event of an accident, someone will be liable.[46] As Elish demonstrates, the human overseer often ends up bearing the brunt of the moral and legal responsibility, even when the undesired outcome was caused by system malfunction.[47] The human in these cases thus also has an 'exculpatory function',[48] 'absorbing' responsibility beyond their actual control over the outcome and obscuring the role of the AI system and humans involved in its development.[49] This 'liability sponge'-role of the human overseer can be actively abused, as was suggested in the case of the self-driving cars company Tesla, which designed the software to transfer control to the human driver less than a second before impact.[50]

4.2 The Purpose of Human Oversight in the AI Act

What does Article 14 envisage the role of the human overseer to be? Article 14(2) addresses this explicitly, stating that human oversight 'shall aim to prevent or minimise the risks to health, safety or fundamental rights'. The formulation suggests a primarily output-oriented focus. This is also reflected in the Commission's Explanatory Memorandum to its proposal for the AI Act, where it notes that the 'obligations for ex ante testing, risk management and human oversight will also facilitate the respect of other fundamental rights *by minimising the risk of erroneous or biased AI-assisted decisions'*.[51]

However, especially the reference to fundamental rights in Article 14(2) also hints at process-oriented goals. EU policy documents leading up to the AI Act's adoption support this interpretation, emphasising process-oriented goals that focus on the perspective of the individual interacting with the AI system.[52] This is clearly expressed in the Ethics Guidelines of the AI HLEG, noting that 'Humans interacting with AI systems must be able to keep full and effective self-determination over themselves'. To make that work, 'The allocation of functions between human and AI systems should [...] leave meaningful opportunity for human choice' by 'securing human oversight over work processes in AI systems'.[53] This was also echoed by the Commission, stating that human oversight 'helps ensuring that an AI system does not undermine human autonomy'.[54] With the idea that human oversight increases trust in an AI system, the human oversight requirement also serves broader systemic and societal goals.[55]

Thus, output-oriented goals dominate Article 14. But it also encompasses process-oriented goals – particularly those focusing on human agency, autonomy, and trust in AI – when viewed

[46] Ben Wagner, 'Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems' (2019) 11 Policy & Internet 104, 109.

[47] Madeleine C Elish, 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction' (2019) 5 Engaging STS 40, 41.

[48] Wagner (n 46) 117.

[49] Elish (n 47) 50.

[50] Crootof, Kaminski and Nicholson Price II (n 7) 438, 483.

[51] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 21 April 2021, COM(2021) 206 final, 11 [emphasis added].

[52] Koulu (n 2) 31

[53] AI HLEG (n 5) 12, see also 16: 'Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects.'

[54] European Commission, 'Building Trust in Human-Centric Artificial Intelligence' (n 20) 4; European Commission, 'White Paper on Artificial Intelligence' (4) 21.

[55] Johann Laux, 'Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act' [2023] AI & Society 1.

within the broader context of the AI Act's adoption. Notably absent from Article 14's aims is the goal of ensuring accountability. However, this absence in the legal text does not preclude the human overseer from absorbing lability when things go wrong, since accountability structures in law are fundamentally human-centric. The concentration of liability on human overseers may therefore emerge as a side effect of Article 14's oversight requirements.

5 Obligations: The Responsibilities of Providers and Deployers

Human oversight is conceptualised in the AI Act as a shared responsibility between providers and deployers.[56] Article 14 is addressed to providers and sets out detailed obligations to *create the conditions* for effective human oversight during the period in which the AI system is in use. The provider can meet these obligations by building human oversight measures into the AI system when it is developed, by identifying measures to be implemented by the deployer, or by doing both.[57] Article 14 is complemented by Article 26(2) AI Act, which, though in much less detail, requires deployers of high-risk AI systems to assign human oversight.

The key requirement under Article 14 is that human oversight mechanisms need to be 'effective' and 'commensurate with the risks, level of autonomy and context of use of the high-risk AI system'.[58] While this is not further specified, Article 14(4)(a)-(e) sets out a list of five more concrete abilities the human overseer should have in order to fulfil their role. The provider has to implement these 'as appropriate and proportionate'. Clearly, this gives providers some flexibility in choosing and developing the specific technical configuration.[59] Whether this flexibility concerns only the question of *how* to implement each requirement in the list or also *which ones* to implement in a given context is difficult to determine. Literal interpretation of Article 14(4) is inconclusive in this respect, even more so when consulting different language versions. Teleological interpretation that focuses on the effective achievement of the objectives pursued by Article 14 suggests that the provision should be interpreted restrictively. That would mean either reducing the providers' flexibility to choosing *how* to implement each requirement, or – in any case – allowing disregard of one or more requirements of the list only when implementing it would be clearly disproportionate.[60]

The five requirements in Article 14(4) do not follow an obvious structure, deeper rationale, or overall consistent approach and have been described as a 'loose collection of items that were deemed useful'.[61] The following aims to bring structure into Article 14(4) by grouping the requirements into three categories, based on what scholars have identified as key markers for effectiveness in human oversight: authority, comprehension, and environment.[62] The resulting taxonomy is shown in Table 5.1.

[56] Ibid 3.
[57] AI Act, art 14(3).
[58] AI Act, art 14(1, 3).
[59] Enqvist (n 17) 519–520.
[60] For a different view see Sterz and others (n 10) 9.
[61] Ibid 9.
[62] Along these lines, see ibid 3–6; Wagner (n 46) 115; Green (n 18) 6–7.

Table 5.1: Taxonomy of Human Oversight Obligations in the EU AI Act

	Authority (provider and deployer)	**Comprehension** (mainly provider)	**Environment** (mainly deployer)
Provider Sphere ***Create conditions for human oversight***	14(4)(d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system; 14(4)(e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.	14(4)(a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance; 14(4)(c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;	14(4)(b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
Deployer Sphere ***Assign human oversight***	26(2) assign human oversight to natural persons who have the necessary authority	26(2) assign human oversight to natural persons who have the necessary competence and training	26(2) assign human oversight to natural persons who have the necessary support

5.1 Authority

The first category concerns the authority of the human overseer in relation to the AI system. In this respect, Article 14(4) sets out two specific requirements. According to Article 14(4)(d), the natural person human oversight is assigned to must be able to decide 'not to use [...] or to otherwise disregard, override or reverse the output of the high-risk AI system'. This requirement focuses on output- and process-oriented goals by ensuring space for the exercise of human judgment that should in turn make it possible for the human overseer to change inaccurate or unfair outcomes. In addition, Article 14(4)(e) requires the human overseer to be able to intervene in or interrupt an AI system 'through a "stop" button or a similar procedure that allows the system to come to a halt in a safe state'. Stopping a system's operation would appear to be particularly relevant in case of malfunction or a threat to safety.

While the provider has to create the technical conditions for the human overseer to be able to exercise authority over the AI system, it is the deployer who has to ensure that ability in practice. This is recognised in Article 26(2) which requires deployers specifically to 'assign human oversight to natural persons who have the necessary [...] authority'.

5.2 Comprehension

The second category concerns the level of comprehension the human overseer has of how the AI system works. Clearly, to make any type of human oversight measure useful in practice, it is important that the human overseer has some understanding of the AI system. However, what degree is necessary or even feasible is debated.[63] Article 14(4) sets out two requirements in this respect. According to Article 14(4)(a), the human overseer must be able 'to properly understand the relevant capacities and limitations of the high-risk AI system' and 'to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance'. This aims at a broad and general understanding of the AI system to know what it can and cannot do or be used for and to be able to notice when it does not perform as expected.

Article 14(4)(c), in addition, requires the human overseer to be enabled 'to correctly interpret the high-risk AI system's output'. This echoes the provider's obligation under Article 13(1) of the AI Act to ensure that high-risk AI systems are designed and developed so 'that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately'. Absent a general definition of 'interpretability',[64] the concrete requirement providers need to fulfil in this respect will depend the goal interpretability serves in a specific context. One of these purposes is that of ensuring the deployer of an AI system can meet their obligations to provide explanation or justification to those affected by the outcome. Public authorities, in particular, typically have reason-giving duties under constitutional or administrative law.[65] In addition, instruments that regulate digital technologies have increasingly also included more or less specific explanation duties, with corresponding rights of those affected.[66] In the AI Act itself, Article 86 grants persons subject to decisions based on a high-risk AI system's output that produce legal or similarly significant effects a right 'to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken'. Interpreted within this context, Article 14(4)(c) requires, as a minimum standard, the level of interpretability necessary for the deployer to comply with Article 86.

The provider's obligations under Article 14(4)(a) and (c) to ensure adequate levels of understandability and interpretability are complemented by Article 26(2) which requires deployers to assign human oversight to natural persons who have the necessary competence and training. At the very least, this entails that the person designated as the overseer has a sufficient level of AI literacy to follow the provider's instructions of use that, according to Article 13(3)(d), have to detail the human oversight measures taken 'including the technical measures put in place to facilitate the interpretation of the outputs of the high-risk AI systems by the deployers'.

[63] For a contextualsation of different types and degrees of transparency requirements in the context of the AI Act see Madalina Busuioc, Deirdre Curtin and Marco Almada, 'Reclaiming transparency: contesting the logics of secrecy within the AI Act' (2023) 2 European Law Open 79.

[64] Adrien Bibal and others, 'Legal requirements on explainability in machine learning' (2021) 29 Artificial Intelligence and Law 149.

[65] In the EU, see in particular Charter of Fundamental Rights of the European Union [2016] OJ C202/389, art 41(2)(c); see Melanie Fink and Michèle Finck, 'Reasoned A(I)dministration: Explanation Requirements in EU Law and the Automation of Public Administration' [2022] European Law Review 376.

[66] See in particular the discussion surrounding the existence of a 'right to an explanation' in the GDPR, which is analysed here: Andrew D Selbst and Julia Powles, 'Meaningful information and the right to explanation' (2017) 7 International Data Privacy Law 233

5.3 Environment

The third category concerns the environment within which the human oversight mechanism is set up. It entails two critical aspects: the biases that influence human interaction with AI systems, and the practical circumstances under which human overseers operate.

First, it is well documented that humans may over-rely on an AI system's outcome either automatically, because they place trust in machines that leads them to neglect counter-evidence ('automation bias'), or selectively, because it confirms their pre-existing stereotypes ('selective adherence').[67] At the same time, humans also under-rely on an AI system's output when they are predisposed to prefer human predictions over algorithmic ones, even when the AI system has been shown to be more accurate ('algorithmic aversion').[68] Article 14(4)(b) addresses the problem of automation bias explicitly, requiring oversight measures that enable the person to whom oversight is assigned 'to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons'. Neither 'selective adherence' nor 'algorithmic aversion' are addressed explicitly. However, it may be assumed that the requirement in Article 26(2) that human overseers be 'appropriately trained' includes training specific to avoiding these known problems. Importantly, though, while such training may help, empirical evidence suggests that training alone is insufficient to solve the problem.[69]

The second aspect concerns the incentive structures around the human overseer.[70] Article 26(2) specifically requires deployers that human oversight is assigned to persons 'who have [...] the necessary support'. This should at least include an environment in which the person using the AI system does not entirely depend on the algorithmic output and can, as Green points out, 'thoroughly consider all of the information relevant to a given decision'.[71] This includes, for instance, having enough time to make informed choices.[72]

6 Looking Ahead: The Limitations of Human Oversight

It is safe to say that high expectations are placed on human oversight requirements in general, and on Article 14 in particular. However, research has shown that human involvement does not always meet these expectations. Empirical evidence suggests that human overseers often do not spot wrong or undesirable suggestions made by an AI system. Human overseers may even correct perfectly good outputs, thus not only failing to improve decision-making quality, but actively

[67] Saar Alon-Barkat and Madalina Busuioc, 'Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' (2023) 33 Journal of Public Administration Research and Theory 153, 155.

[68] Laux (n 55) 6.

[69] Green (n 18) 7–8.

[70] Laux (n 55) 6–7.

[71] Green (n 18) 6; similarly see Sterz and others (n 10) 7–8.

[72] Wagner (n 46) 114–115; Sterz and others (n 10) 8.

undermining it.[73] A number of reasons have been identified that account for this problem, including the biases that influence human interaction with AI systems, and the practical circumstances under which human overseers operate.[74] As discussed in more detail in Section 5, Article 14(4) recognises and attempts to address these issues by placing additional obligations on providers and deployers. However, not all challenges can be resolved through additional obligations. Some limitations, particularly those inherent in human cognitive capacity, cannot be overcome through regulatory requirements alone and need to be taken into account when designing oversight mechanisms.

It is beyond the scope of this contribution to delve into suggestions for reform of human oversight policies and alternative safeguards.[75] However, a number of lessons can be learned that are important to take into account when interpreting and applying Article 14. First, it is vital to recognise that human oversight is not a panacea for the risks AI systems pose. Overreliance on human oversight as a safeguard becomes particularly problematic when it affects the implementation of other safeguards.[76] Thus, within the AI Act's system of safeguards for high-risk systems, the presence of human oversight measures should not be used to justify lowering standards for other protective measures. This is also relevant in relation to Article 6(3)(c), which explicitly allows – albeit only in a specific context – for the lowering of other safeguards when 'proper human review' is exercised (see above Section 3.1). This provision should be interpreted and applied with caution.

Second, the effectiveness of human oversight can only be assessed and improved in relation to the goals it is to achieve. At the very least, this means it is essential to clearly specify these goals in advance, ideally with a view to the different contexts in which AI systems are deployed. While Article 14 could have provided more clarity in this respect, providers and deployers should articulate specific oversight objectives when designing and implementing oversight mechanisms for their AI systems. What might be useful to take into account when doing so, is that the limitations of human oversight particularly affect output-oriented goals. Instead, many of the process-oriented goals seem to be easier to achieve through the human oversight requirement. For instance, while a human will not be able to ensure the accuracy of all of the AI system's outputs, they can still provide 'the ear' necessary to safeguard the right to be heard. Knowing this is useful because it allows defining goals human overseers can actually fulfil and to design alternative safeguards for goals that cannot reasonably be achieved through human oversight mechanisms.

Third, the relationship between human oversight and accountability is complex. Human overseers with tasks that are impossible to fulfil risk becoming 'liability sponges',[77] taking 'the fall' when things go wrong without having the corresponding agency that justifies responsibility. This effect has also been observed by Green, who noted that as a consequence of human oversight policies, accountability often shifts to street-level bureaucrats.[78] At the same time, being able to hold the humans in charge accountable is essential for the rule of law and individual justice. This conundrum is not explicitly addressed by Article 14, but it would be wise to specifically take into

[73] Johannes Walter, 'Human oversight done right: The AI Act should use humans to monitor AI only when effective' (2023), ZEW policy brief 02/2023.

[74] For detail see Green (n 18); Laux (n 55); Sterz and others (n 10); Walter (n 73).

[75] For suggestions on how to go ahead see Green (n 18) 11–16; Laux (n 55).

[76] With examples, see Green (n 18) 9–11.

[77] Crootof, Kaminski and Nicholson Price II (n 7) 483.

[78] Green (n 18) 9.

account this tension when designing both oversight mechanisms and accountability frameworks. The latter should distribute responsibility more broadly across the AI system's lifecycle, rather than concentrating accountability solely on the human overseer at the point of use.

7 Conclusion

Article 14 of the EU AI Act represents a significant milestone in the regulation of artificial intelligence, establishing human oversight as a central safeguard for high-risk AI systems. The provision's detailed requirements reflect an ambitious attempt to ensure that AI systems remain under effective human control, to prevent risks to health, safety, and fundamental rights. Article 14 requires providers to create the conditions for effective oversight as appropriate and proportionate to the circumstances. This includes implementing measures that enable the human overseer to exercise authority over the system, properly understand its capacities and limitations, correctly interpret its output, and remain aware of automation bias. The provision's primary strength lies in its exceptionally broad scope, applying to all high-risk AI systems regardless of sector, context, or the system's role in the workflow.

EU policy places the human oversight requirement at the heart of its regulatory framework for digital technologies. This safeguard, more than any other, may thus be expected to be scrutinised against demanding benchmarks. This is a particular challenge, given the empirical evidence of the challenges for humans to effectively oversee complex AI systems, including human cognitive constraints, biases, and practical operational circumstances. While Article 14 attempts to address these through specific requirements, some limitations cannot be overcome through regulatory requirements alone. Against this background, the design of human oversight requirements should be guided by the capabilities of humans vis-à-vis AI systems and ensure additional safeguards for risks that cannot be addressed through human oversight.

Importantly, the effectiveness of oversight mechanisms can only be properly assessed and improved when specific goals are clearly articulated in advance. Article 14 lacks precision in this respect. It also over-relies on output-oriented goals that put humans in charge of correcting wrong or undesirable outputs, the most difficult to achieve for humans. As the AI Act moves from adoption to implementation, this suggests that successful human oversight will require not just careful attention to context, but also clear goal-setting, and realistic expectations about what human oversight can achieve.