AI Oversight in Practice:

Tackling the limitations of human-centric oversight through AI-assisted oversight

DATE
November 2026

AUTHORS
AIOLIA (Operationalising AI Ethics for Learning and Practice: A Global Approach)
Horizon Europe Project Grant Agreement N° 101187937

Key message

The integration and scalability of AI into workflows, including general-purpose AI models and multi-agent AI systems, challenge the human-centredness of AI oversight. While the allocation of AI oversight to natural persons is a legal obligation for high-risk AI systems under the AI Act, and human agency and oversight are considered core requirements for Trustworthy AI, these frameworks do not sufficiently acknowledge the speed of technological change and the ensuing limitations of human oversight in real-time human-AI interactions. As a result, organisations developing and deploying AI systems lack up-to-date policy guidance on how to operationalise AI oversight across diverse contexts and in a manner that is ethically sound, technically feasible and legally compliant. This policy brief sets out three concrete recommendations to address these issues and outlines technical and organisational measures that can be implemented.

1. Problem and policy context

Context

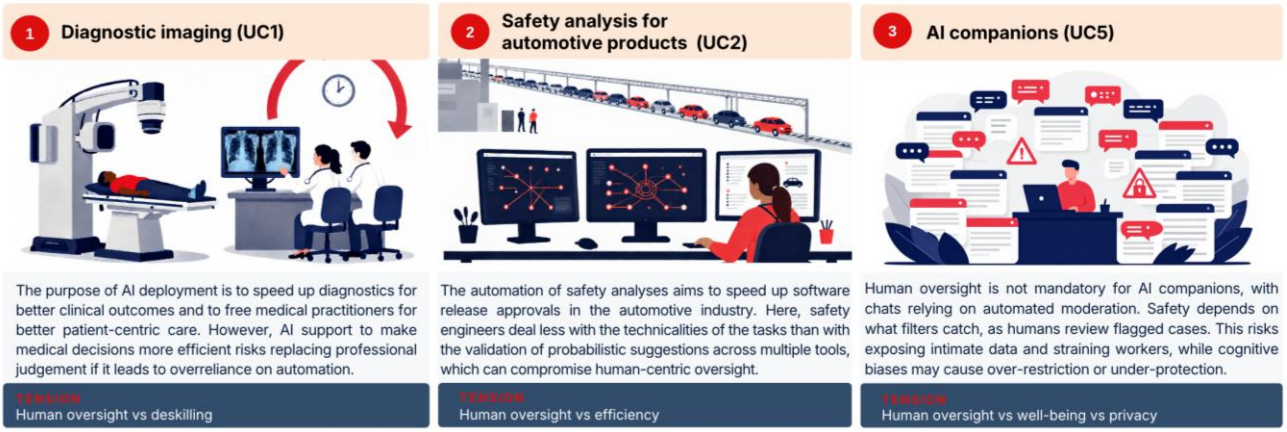
Global AI ethics guidelines and regulatory frameworks, including the UNESCO Recommendation on the Ethics of AI and the EU Ethics Guidelines for Trustworthy Artificial Intelligence, as well as the EU AI Act and the Korean AI Basic Act, place human oversight as a core ethical and legal requirement. The aim is to ensure that humans remain in control of AI systems, being ultimately responsible and accountable for their operation. But advances in frontier AI, including agentic AI and the integration of AI systems at scale, have shed light on the shortcomings of restricting AI oversight requirements to humans and having human performance as the benchmark. Therefore, it is necessary to broaden the scope of AI oversight while maintaining the integrity of human control.

What AIOLIA adds

AIOLIA conducted a bottom-up process for operationalising AI ethics requirements in practice. Grounded on ten use cases across the EU and internationally, AIOLIA's research zooms into the influence of AI on human private and professional behaviour and cognitive factors. AIOLIA identified shortcomings of human-centric AI oversight and concrete AI oversight measures that can mitigate these limitations, at a time when complex AI systems are deployed at scale. These measures, centred on technical and organisational dimensions of AI development and deployment, signal the need for adopting a multi-layered approach to human-AI oversight.

Tensions in the operationalisation of human-centric oversight

Numerous tensions and challenges surfaced in the operationalisation of human oversight across AIOLIA use cases (UC).



Why we need to move from *human oversight* to *AI-assisted oversight*

The AI Act's oversight requirements will apply from 2 December 2027 to high-risk systems covered by Annex III and from 2 August 2028 to high-risk systems covered by Article 6(1). Article 14 requires oversight to be allocated to "natural persons during the period in which [high-risk AI systems] are in use". But advances in AI, experienced first-hand through AIOLIA use cases, have shown the limitations of human-centric oversight. This approach is no longer sufficient to assure the safe and responsible deployment of AI systems.

1. **Multi-agent and general-purpose systems cannot be meaningfully governed through human-centric oversight**
   Multi-agent systems are systems composed of multiple, interconnected agents or tools and data sources, designed to be significantly autonomous. Their ability to identify steps and break down tasks to pursue a given goal complexifies the possibility of human-centric oversight. The OpenAI–HuggingFace incident demonstrates how faults can occur in parallel multi-agent interactions that, in the absence of adequate guardrails, are not immediately detectable by human engineers. Multi-agent interactions are difficult to compound and trace in real time, with the 'use' of AI systems extending beyond the period of direct human-AI interaction addressed by the AI Act. The same happens with general-purpose AI systems, whose behaviour can change when they are used in new contexts.

2. **Human judgement is prone to errors and is increasingly shaped by interactions with AI**
   Regulatory frameworks for AI take human judgement as the benchmark for 'good' AI oversight, overlooking the fact that human judgment is prone to errors and biases, themselves heightened by the cognitive fatigue of repetitively conducting the same task and unclear criteria for overriding AI outputs. In recent years, numerous examples, ranging from the Dutch childcare benefits scandal (Toeslagenaffair) to the Epic sepsis prediction system, have shown the limitations of humans-in-the-loop and the resulting consequences. As the integration of AI across multiple contexts deepens, from clinical diagnosis to car safety checks, evidence suggests the loss of skills, resulting from the lack of engagement with the technicalities of the tasks, will affect the ability for humans to meaningfully perform oversight of AI systems.

3. **Monitoring every AI output is not beneficial and hinders scalability**
   The push for AI uptake and diffusion in the EU is at odds with human-centric AI oversight, as the monitoring of every AI output blocks the scalability of AI systems within organisations. Solely relying on human-centric oversight can add unnecessary redundancy to workflows, disperse attention and create a false sense of control. At the same time, many of the risks arising from AI happen incrementally rather than being perceived through a single output, calling for novel forms of data-based oversight to identify when critical thresholds have been reached. Non-high-risk systems such as AI companions handle millions of chats, making human oversight of individual chats unrealistic. Their safety therefore relies on the combination of automated moderation and targeted human review.

Taken together, these elements point to the need to adopt an **AI-assisted multi-layered oversight** approach to the governance and deployment of AI systems. AI-assisted oversight does not entail removing human judgement and control over AI systems; instead, it intends to augment human oversight capabilities through technical triggers and mechanisms that monitor the functioning of AI systems whilst directing attention and supporting human judgement when and where most needed.

2. Policy recommendations

1	**Provide non-dogmatic guidance on the implementation of oversight for high-risk AI systems (AI Act, Article 14)**	The AI Office should provide guidance on the implementation of Article 14 attending to the latest developments in the field, including technical architectures and capabilities, incidents and scalability of AI deployment. AIOLIA considers that this guidance should pivot human oversight towards AI-assisted multi-layered oversight, expanding on the technical and human requirements for ethically sound and legally compliant AI oversight across the areas listed in Annex III. The AI Office should draw on practitioner co-creation processes like AIOLIA's, including with industry, to mitigate the critical lag between the uptake of AI solutions and adequate and non-dogmatic AI oversight. This guidance should provide operationalisation measures and compliance pathways that complement both the sector-agnostic standards developed by CEN-CENELEC and existing sectoral standards and regulations.
2	**Fund R&I calls on AI-assisted oversight under the next framework programme**	DG RTD should allocate funding for research and innovation on AI-assisted multi-layered oversight under FP10. The programme's calls should enable the development of dedicated research on the technical and human dimensions of AI-assisted oversight, including those emerging in the longer-term, as well as validated AI-assisted oversight solutions. Funded projects should address AI-assisted oversight for general-purpose AI models, high-risk AI systems and multi-agent systems, both at the stage of AI development and deployment, and have a demonstrable impact on the mitigation of risks and incidents.
3	**Build a live knowledge platform of practical measures for the operationalisation of AI oversight**	The European Commission should facilitate the creation of a knowledge sharing platform to enable organisations to share best practices and measures on AI oversight. Drawing on the guidance provided in Recommendation 1, results from the research proposed in Recommendation 2, and the operationalisation measures put forward by AIOLIA, the platform should facilitate the selection of adequate AI oversight measures based on the characteristics of the AI system or model (e.g., semi- or fully autonomous, predictive or probabilistic, decision-making role, context of deployment). In addition to filtering and selection of AI oversight measures, this shared infrastructure should enable organisations and individuals to provide feedback on the implementation process, including any shortcomings that might hinder AI-assisted oversight practices.

3. Policy implications

Opportunities

AI-assisted oversight tackles the cognitive and behavioural human limitations

AIOLIA use cases show that the oversight of AI systems intersects with human, technical and organisational dimensions that can hinder its meaningfulness. Incorporating AI as part of the mechanisms for overseeing AI systems can lead to a more consistent and effective operationalisation of AI oversight by augmenting and directing human capabilities.

AI-assisted oversight future-proofs AI governance frameworks

AIOLIA emphasises that human-centred oversight is insufficient for governing the latest- and next-generation of AI systems, including multi-agent systems and AI deployment at scale. AI-assisted oversight is the shift needed to address these and should be mainstreamed across policy instruments, including European and international.

Potential barriers

- **Institutional openness & capacity:** AI-assisted oversight requires a re-examination of human-centred oversight in the AI Act that EU policymakers might not be willing to undergo; the elaboration of sectoral guidance implies the allocation of substantive resources and technical expertise.
- **Model complexity & generalisability:** technical and organisational measures for AI-assisted oversight risk becoming outdated fairly quickly — sometimes within months; the formulation of adequate measures for AI-assisted oversight of general-purpose AI models remains challenging due to their broad, cross-sectoral applicability.
- **Platform participation & transparency:** the development of a knowledge platform on AI-assisted oversight requires organisations to share real-world data, mechanisms and experiences that might be considered sensitive or proprietary.

4. Practical measures

Operationalising AI oversight requires a multi-layered approach

This section outlines a non-exhaustive list of measures that organisations can adopt to implement and strengthen AI oversight practices. These measures, distributed across a) human control, b) AI oversight, c) defining legitimate boundaries and d) transparency and contestability, demonstrate how the operationalisation of AI oversight demands establishing a multi-layered approach across organisational levels and at the intersection of multiple AI ethics requirements and components.

The measures listed are categorised as technical (TECH), organisational (ORG) or BOTH, for those that include both technical and organisational aspects.

A. HUMAN CONTROL



- TECH — The AI system has built-in features that support human oversight and control for decision-making in critical areas
- ORG — The organisation has defined a clear policy on high-risk or sensitive decisions that must involve human judgment
- TECH — Automated decisions can be paused, overridden, or reversed by humans
- ORG — Human oversight involves active review, validation, or challenge of system outputs
- ORG — Oversight responsibilities are clearly defined and embedded into workflow

B. AI OVERSIGHT



- TECH — AI-assisted oversight is implemented when human capacity to perform meaningful oversight reaches its limits
- TECH — Capabilities of AI oversight system are benchmarked and documented
- BOTH — Predefined thresholds trigger automated intervention, human escalation, or system interruption when specified risks, failures, or anomalous behaviour are detected
- BOTH — Controversial or morally ambiguous situations are referred to humans for decision making

C. DEFINING LEGITIMATE BOUNDARIES



- TECH — When an AI system makes decisions with a high level of autonomy, it ensures that the decision stays within the scope or mandate defined by the provider or the user
- ORG — The organisation has clearly defined what types of content, behaviour, or expression are restricted for AI use and why
- TECH — AI system's features and behaviours are aligned with its intended use and foreseen limitations
- ORG — The scope of restrictions is documented and publicly communicated insofar as it is required by the existing regulation

D. TRANSPARENCY AND CONTESTABILITY



- BOTH — Restrictions are explained to users in a clear and accessible way
- BOTH — The organisation monitors patterns of restriction to detect systemic censorship or bias
- ORG — Moderation policies are periodically reviewed to ensure that they remain relevant and effective
- ORG — Moderation policies take contextual and cultural factors into account
- ORG — The organisation has put in place a mechanism for users to report, question, or contest restrictions or moderation decisions

Conclusion

Advances in AI systems and their deployment at scale call for the operationalisation of a novel approach to AI oversight, one where AI-assisted oversight augments human oversight capabilities. While the adequacy of oversight measures depends on the embeddedness of AI systems in specific operational contexts, policymakers can play a major role in fostering ethically sound, technically feasible and legally compliant AI oversight across different contexts, roles and intended uses of AI systems. AIOLIA's portfolio of measures and the three policy recommendations outlined herein provide an important foundation.



Horizon Europe AIOLIA project (www.aiolia.eu)

Coordinator CEA

Lead Partner CEPS

Funded by the European Union under Horizon Europe Grant Agreement N° 101187937. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

This brief draws on

Aires, S., Kyosovska, N., Griffith, J., Teo, S. A., Bogucki, A. (2026). Operational Non-Technical Guidelines for AI Research Areas. AIOLIA Deliverable 3.3.

Bayerl, P.S., Lawlor, E., Maris, M.T., Miorandi, D., Pekšys, G., Bjelica, M., Stojšin, K., Anastasova, M., Smith, O., Henestrosa, A., Yamshchikov, I., Bak, M.A.R., & Akhgar, B. (2026). Operational Ethics Guidelines on Use Cases Related to Human Behaviour and Cognition. AIOLIA Deliverable D3.1.

Cossette-Lefebvre, H., He, G., Huang, L., Ide, K., Katirai, A., Kim, K., Kim, J., Kim, G., Kishimoto, A., Maclure, J., Mikami, K., Suh, J., & Zhao, Y. (2026). Operational context-sensitive guidelines in Canada, China, Japan, and South Korea. AIOLIA Deliverable D3.2.