Multilevel oversight of AI systems in line with the AI Act

Diana M. Popa
Delft University of Technology
Email: d.m.popa@tudelft.nl

Abstract

The AI "oversight – control" nexus is a matter of debate in both scientific literature and policy papers, given the complex and disruptive nature of the technology and the intricate legislative systems regulating the deployment of AI systems. Most approaches give precedence to the need for a multilayered governance model, while in the same time taking into account the lifecycle stages of the AI product. Oversight over the ecosystem in which the AI is deployed (the governance layer) should not be confused with the process level human oversight. At process level, human oversight is needed in the case of "high-risk systems", and AI literacy is a precondition for this oversight to be relative. Effective oversight requires many resources at both ecosystem and organisational levels. Additionally, oversight of AI systems is also highly regulated in democratic societies and therefore can be seen as a reflection of the importance that democratic values have in the way disruptive technologies are deployed within society.

Keywords: oversight, control, governance, AI Act.

The need for control and human oversight of AI systems is acknowledged and set down in both legislation texts and research and policy papers, but operationalisation of both concepts is either broad or diverse and in practice address different levels of applicability in relation to the given AI system. Control and oversight *of* or *over* the functioning of the AI system itself should not be equalled to *how* the legislative framework address measures of control and oversight *of the ecosystem* in which the AI systems are deployed. The multitude of initiatives aiming to establish generally accepted definitions and determine the practical limitations of the "meaningful human oversight" – control nexus (Verdiesen, et al. 2021) underline the difficulty of determining the control capacity over the AI system. Efforts for identifying the correct balance of human oversight and control are meant to counter the "black box" effect that the disruptive nature of the technology brings with itself, especially when used as a decision support tool with a high impact on the individuals.

The Artificial Intelligence Act (AIA) itself mentions both control and human oversight as risk mitigation measures, using terms such as "relevant" / "appropriate"/ "meaningful" human oversight. What the specific implementation measures are remain at a general level (Article 14 primarily) while in the same time, the scientific literature addresses the fact that there is ambiguity in the way human oversight is operationalized and implemented, even more so when it comes to "meaningful human oversight". This is not necessarily a drawback of the legislation, given the fact that the AI Act was just approved by the European Parliament and also considering that, like in the case of the GDPR, the AI Act wants to be an overarching legislation, leaving room for personalized application at Member State level and remaining broad enough so that the AIA itself needs not reviewing or updating every year, taking into consideration also the alert rhythm of technological development in the field.

From the policy and research approaches, different initiatives have addressed the operationalization of human oversight and the "control – oversight" relationship, either as a nexus or as opposed measures, with approaches depending on the applications field of the concepts. Both control and oversight are defined and embedded within the given broader ecosystem in which the AI model is implemented, and most approaches adhere to a three layer ecosystem: the governance layer (the supra-national and/or national ecosystem in which the AI applications are developed), the socio-technical or organisational layer (based on internal regulations and sector rules) and the technical or process layer (also addressing the product safety regulations) (Verdiesen et al., 2021; Adams et al., 2024; Novelli et al.,2024). Oversight takes place at European level, within dedicated structures (still in development), at national level, with data protection authorities, at organizational level and at process level. Multilayer oversight overlaps (even if one might argue not perfectly) with a multistakeholder governance system of AI systems implementation within a certain regional or national context. Oversight is implemented both vertically and horizontally: the former within national supervisory authorities, expanding their structures to take on formal oversight and control roles with sanctioning capabilities and the latter within the organisations deploying the AI system themselves.

From a component perspective, building blocks for an effective oversight system include good functioning of the democratic rule of law principles, sound legislative frameworks accompanied by binding powers (both for the pre and post deployment phase), appropriate financial and human resources in all three layers and technical capabilities (technological maturity). In line with democratic principles, an oversight system should address the principles of transparency, accountability and responsibility regarding the way the AI systems are used in public context and incorporate human rights standards. Binding involvement of oversight bodies (in the sense of consultation before the implementation of a tool) and binding powers post factum are necessary for oversight to be effective (Wetzling, 2024), the latter with supervisory/ investigative and sanctioning powers.

Not addressing here prohibited practices that are equally regulated by the legislation, in the case of high-risk or impactful AI systems, the risk approach managed through measures and degrees of oversight is also influenced by the type of deployer, the sector it activates in and the process in question. While in the case of internal processes, it is a reflection of the risk appetite of an organisation, in the case of processes that have an impact on individuals outside the organisation, such as is the case with AI used by public authorities in the execution of the public administration act, oversight is tighter regulated from the governance level, since it should also be in line with the social values of the system it operates in. Although actual alignment with public values is not required by the legislation, it is a practice in line with democratic values that increases transparency of the governing act and public decision making and is a reflection of the governing style of a certain nation state, such as is the case of the Netherlands.

The ecosystem division is also often overlapped with the stages of the product's lifecycle, given that AI is a product deployed on the unique market and also has to comply with product safety regulations, approach which identifies key points during the life of the AI system when human control or oversight is needed, also in relation to the inherent risks of that certain stage. Therefore, within each layer, different control and oversight measures are put in place, at different moments in the AI product's lifecycle: before deployment, during deployment and after deployment.

A well-structured oversight model therefore includes both ex-ante and ex-post elements. Expert assessment in the form of expert bodies should be included in both stages, with pre-deployment expert advice going in the design phase of the system and in the high level regulating frameworks and with ex-post oversight in the form of expert assessment and democratic scrutiny (Oetheimer, 2024). These recommendations are also valid in the case of regulatory AI frameworks.

From this temporal perspective, control is implemented at three different points in time: ex-ante or pro-active control, on-going or simultaneous control, and ex-post control. Ex-ante or pro-active controls are (or should be) implemented by way of the AIA or by the means of the AI national strategy or organisational policies and practices through:

- bans on prohibited (high risks) systems (AIA);
- enforcement through market surveillance and control (AI Act);
- (quick) scans/ risk analysis for selection of trustworthy suppliers and safe acquisitions (NCTV, 2024) and identification of supply chain risks (Bluebird & Hawk BV, et. al, 2024);
- DPIAs and FRIAs, (AI Act, National AI strategy, GDPR, Organisational policies);
- attribution of supervisory roles and compartmentalisation;
- establishment of internal ethics committees addressing issues such as data ethics compliance or data "pedigree".

Ongoing or simultaneous control:

- supervisory authorities at national and international levels with sanctioning powers;
- Meaningful human oversight within the process (organisational policies);
- Process- related activity of ethics committees for deviating cases.

Equally relevant, for oversight to be effective, practical and technical capabilities giving external parties (either supervisory authorities of the broader public) relevant insight into log frames to evaluate how the data processing took place, details on the way the training data was processed and altered at every step should also be made available. These are forms of **ex-post** control, giving external parties the possibility to get insight into the system, either through information made public by the deployer beforehand or through standard access requests. A different form of ex-post control would be the requirement to make the AI models FAIR (Findable, Accessible, Interoperable and Reusable) and have metadata or a more extensive documentation package made available, either openly or upon request. This is partially addressed by the requirement to have high-risk AI systems registered in the European data base for High Risk systems that is under development (Article 71 of the AIA).

Article 14 of the AIA specifically addresses human oversight measures at the process level, placing responsibilities on both developer and deployer: on the developer during the design phase (therefore an ex-ante measure) for the identified risks, and for the deployer for the attribution of the actual human oversight during the process to human resources with the appropriate AI literacy. For high-risk systems, technical provisions include measures similar to kill switches (Article 14, point 4, e.) and oversight measures on supervision of two separate natural persons with appropriate training, competence and authority. Thus, for human oversight to be meaningful, AI literacy is a precondition (Recital 20, Articles 14, 26, 91). Adequate human and financial resources at eco-system and socio-technical levels are also paramount for a functioning oversight system. This is where fragmentation at European level is foreseen to remain high, For example, the Dutch Data Protection Authority (AP), as the dedicated supervisory authority for the use of AI systems that have an impact on personal data, has had an increased budget of 3.6 million euro for 2026 and 2027 (AP, 2024) to address the new tasks under its coordination streaming from the upcoming of the AIA. Yet, despite the budget increase, the Dutch AP is still challenged by finding the necessary numbers of trained specialists for the exercise of the given tasks. While the oversight and governance systems involve multiple entities, data protection authorities, as the responsible oversight entities of the AI also have the role of curbing the data appetite of public authorities. At the implementation level, human oversight remains both challenging and fragmented, influenced by level of digital maturity within a certain society, financial capabilities of each deployer (hyper-scalers having the higher ground), the availability of a trained workforce or lack thereof, the difficulty to keep up with technological and legislative developments, efforts by central government to harmonize implementation of EU legislation in a top bottom approach within national context and uniformise/change already established daily practices.

The fact that AI is a disruptive technology is also reflected by the intense debates and analyses in legislative and policy-making circles in order to set it into a unified, comprehensive and EU level actionable text. Oversight of AI systems is highly regulated in democratic societies and while democratic states spend years and a great deal of financial resources to harmonize these legislative frameworks, less democratic states or ones that rather focus on market competition principles seem more inclined to invest in and focus on the development and implementation of the technology self, focusing on capitalizing on the benefits without the restrictions of ethical and robust legislative frameworks, and as such gaining larger market shares in the world competition stage.

Addressing the “why” of the need for human oversight includes the fact that AI is a disruptive technology, the level of trust in technology in general (at social level for example – for example the World Values Survey) and AI technology in particular, and the approach within a certain society towards rules and regulations. As Adams (2024) suggests, AI governance does not translate into responsible AI. Oversight structures are a form of foresight and a practice of a healthy ecosystem, in which decision makers, weather at supra-national, national or sectorial and even company levels implement checks and balances in a proactive manner. Foresight is limited though to the overview of the known variables of the given environment at a certain moment and unforeseen uses remain a reality and very much in the control and responsibility of the deployer. Taking the discussion on when human oversight is needed further, it comes down to the cases and moments of high risk situations. Despite the fact that automated instruments for decision making have been hailed for the benefits they bring regarding time saving with repetitive tasks, it is in "life or death" situations where they failed and when ultimate responsibility was passed to the human. Identifying such "life or death" situations should be addressed in the design phase of the AI system, despite the limitations or impossibility of mapping out all possible real life situations in a design/ test phase or lab setting. And while the AIA does address ex-ante or pre-deployment controls, lists prohibited AI practices, and identify highly regulated sectors and exceptions, in practice, specific sector level knowledge and also case by case analysis are needed in order to identify what qualifies as such and to identify the correct moment and forms of human oversight.

Empirical case studies involving the use of responsible AI in line with the AIA within different contexts are much awaited. Paradoxically, despite extensive regulatory systems, what gives rise to the need for case by case approaches is the uniqueness stemming from the intersection of legislations (for an overview see Annex 1 of the AI Act: list of Union Harmonization legislation), technological regulations, ethics, organizational settings and target groups affected by the AI. Harmonizing legislation attempts such as those in the case of products for the single market are underway. However, if the harmonization process will not work, it is foreseeable that the EU commission will come with its own standards, with the consideration that technical safety is easier to implement and standardize, whereas human oversight is more difficult both to regulate and ethical implications even more difficult to evaluate, reflecting also political choices. As practice evolves, it is expected that in the next years the scientific literature will abound with these case studies. It will be interesting to see how early adaptors, independently of their size, will set trends, how fast good practices will be absorbed and, as was the case with the GDPR, how case law will evolve. These will also serve as measures for measuring the effectiveness of the human oversight and control nexus, currently a challenge that will take the form of consistent practices in the coming years.

References

Adams, R., Adeleke, F., Florido, A., de Magalhaes Santos, L.G., Grossman, N., Junck, L., Stone, K. (2024) Global Index on Responsible AI 2024 (1st Edition). South Africa: Global Center on AI Governance.

Autoriteit Persoonsgegevens. (2024). Directie Coordinatie Algoritmes Werkagenda coördinerend algoritmetoezicht in 2024. Accessed Mey 2024 at: https://www.autoriteitpersoonsgegevens.nl/documenten/werkagenda-coordinerend-algoritmetoezicht-2024

Bluebird & Hawk BV., De Nederlandse Vereniging van Banken, ICT Group, Nederlandse Spoorwegen en Technolution; Nationaal Coördinator Terrorismebestrijding en Veiligheid (NCTV), Nationaal Cyber Security Centrum (NCSC); Algemene Inlichtingen en Veiligheidsdienst (AIVD); CIO Rijk (2024). Cybercheck: ook jij hebt supply chain risico's. Available at: https://www.ncsc.nl/documenten/publicaties/2024/april/18/cybercheck-ook-jij-hebt-supply-chain-risicos

Oetheimer, M. (2024) Presentation during the European Data Protection Summit. Rethinking data in a democratic society. Session 3: Zooming out onto democracy and the rule of law. How to build a functioning democratic oversight. Available at: https://20years.edps.europa.eu/en/summit/media

Nationaal Coordinator Terrorismebestrijding en Veiligheid (2024). Quick Scan nationale veiligheid bij inkoop en aanbesteding. Accessed June 2024 at: https://www.nctv.nl/onderwerpen/economische-veiligheid/documenten/publicaties/2024/02/01/quick-scan-nationale-veiligheid-bij-inkoop-en-aanbesteding

Novelli, Claudio and Hacker, Philipp and Morley, Jessica and Trondal, Jarle and Floridi, Luciano, A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities (May 5, 2024). Available at SSRN: https://ssrn.com/abstract=4817755 or http://dx.doi.org/10.2139/ssrn.4817755

Verdiesen, I.; Santoni de Sio, F., Dignum, V. (2021). Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight. Minds and Machines (2021) 31:137–163. https://doi.org/10.1007/s11023-020-09532-9

Wetzling, T. (2024) Presentation during the European Data Protection Summit. Rethinking data in a democratic society. Session 3: Zooming out onto democracy and the rule of law. How to build a functioning democratic oversight. Available at: https://20years.edps.europa.eu/en/summit/media