Cambridge Forum on AI: Law and Governance (2025), **1**, e29, 1–17
doi:10.1017/cfl.2025.10010

Cambridge Forum

RESEARCH ARTICLE

Better together? Human oversight as means to achieve fairness in the European AI Act governance

Ana Maria Corrêa, Sara Garsia and Abdullah Elbi

Center for IT & IP Law, KU Leuven, Leuven, Belgium
Corresponding author: Ana Maria Corrêa; Email: anamaria.correaharcus@kuleuven.be

(Received 3 November 2024; revised 24 January 2025; accepted 16 April 2025)

Abstract
In this article, we investigate the relationship between human oversight and fairness within the evolving framework of EU AI regulatory governance. We address two core research questions: (1) How are human oversight and fairness related? and (2) To what extent does the AI Act establish a framework for human oversight that effectively supports the implementation of the various dimensions of fairness? Based on a review of interdisciplinary literature, the article identifies three normative claims linking human oversight to fairness: first, that human oversight can help mitigate bias and error in AI systems; second, that it can function as a mechanism of accountability by assigning oversight to natural persons where AI systems lack legal liability; and third, that it can introduce human empathy and contextual sensitivity into decision-making processes, enabling a substantive notion of fairness that takes into account individual circumstances. A critical analysis of the AI Act reveals that while these normative aspirations are acknowledged, the Act only partially operationalises them, leaving several aspects of fairness insufficiently supported.

1. Introduction

By 2035, it is projected that 47% of jobs will be replaced by Artificial Intelligence (AI) (OECD, 2019). Optimists suggest that AI will radically transform work by automating routine tasks. Between predictions, optimistic visions and uncertainties, however, lies the current reality. In 2025, the notion of AI as a self-sufficient engine remains compelling but oversimplified. Either by technical limitations and legal requirements, AI is deeply entangled with human involvement at every level: design, implementation and oversight. Perhaps, in the future, innovation may allow for the complete automation of the design and implementation stages. Yet, the involvement of humans in AI decision-making – particularly through oversight – is a normative and political choice. Even if automation of oversight becomes technically feasible in any circumstance, involving human beings in complex AI ecosystems may possibly remain a deliberate choice for several reasons, including fairness of AI outcomes.

In the legal sphere, the debate surrounding human oversight over technology has long echoed in the circles of autonomous weapons, considering the implications of allocation of lethal artefacts to non-human agents (Verdiesen, Santoni de Sio & Dignium, 2021). A report adopted by the European Parliament stressed that AI-enabled weapons must always allow humans to exert meaningful control (European Parliament, 2021). Having meaningful human control, in this context, results in the proper

allocation of human responsibility and accountability over the loss of lives. In some situations, responsibility and accountability are largely perceived as drivers for fairer outcomes in AI decision-making processes (Nagtegaal, 2021).

Recently, the discussions on human oversight have expanded beyond the military use of AI, driven by efforts in Brussels to make human involvement a mandatory legal requirement for deploying high-risk AI systems, regardless of their potential lethality. Currently, legal narratives involving human oversight and AI encompasses issues of liability (Botero Arcila, 2024), procedural safeguards to ensure human centrism in AI (Enqvist, 2023) and scepticism about the possibility of creating a legal framework that enables for meaningful human oversight in realities that are increasingly pervaded by AI systems (Green, 2022).

Specialized literature frequently suggests that the requirement of human oversight in AI aims to preserve human autonomy as the right of self-determination (Beck & Burri, 2024; Enqvist, 2023; Green, 2022; Koulu, 2020; Wagner, 2019). In this same sense, non-binding EU policy documents such as the Ethics Guidelines for Trustworthy AI also stressed the connection of human oversight and human autonomy (HLEG AI, 2019, p.12; European Commission, 2020, p.21).

Interesting enough, the recently enacted Artificial Intelligence Act (AIA) does not explicitly state the instrumental objective of human oversight for ensuring human autonomy. Instead, the Act specifies that human oversight is primarily intended to prevent or mitigate the risks AI may pose to health, safety and fundamental rights (Art. 14 (2), AIA).

Given these multiple interests the concept of human oversight has inspired among technologists and human sciences, this article aims to explore the relationship between human oversight and fairness within the EU AI regulatory governance. First, this focus addresses a gap in the legal literature (Enqvist, 2023; Koulu, 2020; Sterz et al., 2024), which has overlooked how involving humans in AI decision-making is expected to enhance fairness, according to the provisions for human oversight in Article 14 of the Act. Article 14 tasks human overseers with the ability to understand and respond to anomalies in high-risk AI outputs, potentially including those related to unfairness (Art. 14, (4), AIA). Second, analysing the link between fairness and human oversight is relevant, considering that in past landmark cases, the Court of Justice of the European Union has ruled that meaningful human intervention in decision-making purely based on automated systems contributed to the fairness of those decisions. The link between fairness and human intervention in automated decision-making was established by the European Union Court of Justice (PNR C-817/19; Schufa C-634/21). This link between fairness and human oversight, therefore, requires further reflection.

Fairness is largely understood as a multi-semantic concept in AI governance (Gerards & Xenidis, 2021, p.47; Wachter, Mittelstadt & Russels, 2021; Hacker, 2018, p.1176), serving as a proxy for various normative ideas, such as inclusion, equal treatment, gender equality and the lack of discriminatory impacts and biases. In this article, besides these broadly discussed dimensions of fairness, we also focus on fairness as accountability and empathy.

Considering this backdrop, this article aims to respond to the following research questions: (i) How are human oversight and fairness related? (ii) To what extent does the AIA establish a framework for human oversight that effectively supports the implementation of the various dimensions of fairness?

Drawing on interdisciplinary literature, we explain that human oversight is related to fairness in three non-exclusive normative ways. First, human oversight seeks to address AI bias to achieve fairness as the right to non-discrimination. Second, human oversight aims to place natural persons to oversee AI systems that themselves lack liability, thus enabling fairness as accountability. Third, human oversight attempts to introduce empathy into AI decision-making processes, allowing for the consideration of contextual factors that can lead to fairer outcomes. Although human oversight and fairness may also intersect with other aspects, such as ensuring transparency of automated systems and economic reasons, our analysis focuses on bias, accountability and empathy to allow for a more

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

in-depth exploration of these dimensions. We conclude that these normative aspirations are only partially embraced in the AIA due to the current allocation of oversight obligations between AI providers and deployers, the lack of organizational oversight provisions, the side-effects of explainability and transparency including over-reliance on AI and lack of provisions regarding the duration of human oversight.

By addressing these issues, the article provides two novel contributions to the literature on AI and human oversight. First, it describes three normative goals for human oversight that are related to fairness. Second, it showcases how the AIA has only partially created a structure for human oversight that will contribute to the materialization of the principle of fairness as non-discrimination, accountability and empathy in AI decision-making. This theoretical discussion is useful for the interpretation and operationalization of the requirement of human oversight as regulated by the AIA in the coming years.

2. Understanding human oversight

Beyond the context of AI, oversight measures have been mobilized in several fields as a relevant aspect of governing complex systems. In this regard, the concept of oversight refers to the process of ensuring that organizations, technologies and individuals comply with established rules, standards, indicators and sometimes expectations (OECD, 2021; Restrepo Amariles, 2017; Keay, 2014, p.279). It involves various measures at both macro and micro governance levels to monitor and enforce adherence. In some cases, oversight also aims to delegate responsibility and agency.

In the context of sociotechnical systems, oversight can be carried out either by other technologies or by human operators (Mökander, 2023). Technological oversight is preferable in scenarios where the scale of information to be processed and accuracy requirements go far beyond human capability, such as monitoring electrical distribution to ensure compliance with energy standards. Alternatively, human oversight is often implemented in scenarios such as healthcare diagnostics, and numerous other instances where automated supervision systems alone may not sufficiently achieve the complex objectives of oversight (Kyriakou & Otterbacher, 2023). In this matter, currently, the International Organization for Standardization (ISO) is developing ISO/IEC AWI 42105, a set of guidelines for human oversight of AI systems applicable to all types of organizations and throughout the AI system life cycle.

In recent years, human oversight has gained broader attention from legal scholars across multiple fields of law, given human intervention was placed as a mandatory legal requirement in the deployment of high-risk AI systems. This debate is not confined to the legal sphere, instead is a complex thread of technical, moral and legal arguments.

In the following sections, we address different justifications of human oversight in AI governance that relate to fairness in AI governance.

2.1. Arguments for human oversight as a driver of fairness

In the context of AI, the necessity of human oversight has emerged as a topic of discussion in different areas of knowledge. Considerations relate to the involvement of human judgement as part of certain decision-making processes, especially where significant risks to safety, discrimination or moral considerations are involved. This section explores three arguments for justifying human oversight that relate to fairness.

2.1.1. Mitigating errors

One of the technical arguments for human oversight lies in the need to enhance safety and counterbalance potential errors or biases made by AI systems (Shneiderman, 2016). Given the focus of this article is fairness, the following example focuses on bias.

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

In migration management, AI systems are increasingly used to analyse passenger name records and publicly available data in order to identify potential criminals on flights (Council of the EU, 2024). Human oversight is still required in this process, given AI only generates probabilities. Police officers review AI-generated analyses to determine whether a passenger should be further investigated before entering the country. One of the alleged reasons for this is that human beings would be in measure to counterbalance AI bias (European Commission, 2022).

The argument that human oversight may mitigate errors and ensure safety in various circumstances, including migration management, faces several objections for different reasons explored further in this article (Banks, Plant & Stanton, 2019). Among these objections, there is the fact that human overseers may (i) over rely on the AI system, (ii) lack of sufficient time which compromises human judgement and (iii) lack of sufficient knowledge to challenge the AI recommendation or prediction. Some of these objections were addressed by the AIA, including the requirement for investment in AI literacy. Despite these objections, the extra safeguard of human oversight still seems preferable to improve safety and mitigate errors in these circumstances, at least from the regulatory point of view (Binns, 2020).

2.1.2. Upholding moral responsibility

Several voices argue that one of the main reasons for ensuring human oversight over technology is grounded in moral responsibility (Loh & Loh, 2017; Santoni De Sio & van den Hoven, 2018). Automated systems can have a significant impact on individuals' health, physical integrity and fundamental rights. By maintaining human oversight, decisions with sensitive human and social impacts would remain subject to human judgement. In this case, human beings can be held accountable for actions that with the support of AI may be detrimental to others.

Human beings are held morally accountable for errors, not AI, nor machines in general. In law, this debate was translated into legal liability (De Bruyne, Van Gool & Gils, 2022; Wagner, 2019). In the legal arena, this binomial of human oversight and liability can be better understood if one looks into what justifies liability, in the first place (De Bruyne & Dheu, 2023, p.54–55). Among numerous theoretical efforts to rationally justify liability in legal systems, there are the ability to provide reparation to the victim and the deterrence to act in deviation due to the consequences put in place by liability. These abilities – providing reparation and being deterred by consequences – are inherently human, so far.

To illustrate the argument, consider an AI system used in healthcare to diagnose diseases. If this system produces an incorrect diagnosis, it could lead to improper treatment and harm to the patient. By maintaining human oversight, healthcare professionals, with proper training and experience, review and confirm AI-generated diagnoses, thus holding individuals accountable for decisions that might ultimately affect patients' health.

2.1.3. Human oversight as a mean of maintaining empathy in decision-making

In several circumstances, human oversight is relevant for maintaining empathy, described as an ability to deeply understand and react to others' emotion and circumstances of life, due to emotional and cognitive responses (Cuff et al., 2016, p.144). Obviously, AI by nature lacks emotions and feelings. At first glance, this might seem advantageous in contexts where emotional detachment can lead to more rational choices. But is it desirable to live in a society where all the choices are rationally made in unempathetically manner? What does this actually mean in practice?

Consider the use of AI to assess students. A teacher notices a student consistently underperforming on exams and assignments. Instead of attributing this to lack of effort, the teacher takes the time to speak with the student and learns they are struggling with a learning disability, undiagnosed mental health issues or challenges at home, such as caring for siblings. By empathizing with the student's situation, the teacher can offer accommodations, such as extended deadlines, alternative assessment formats or access to support services like tutoring or counselling. This approach ensures fairness

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

by recognizing and addressing the structural or personal barriers the student faces, giving them an equitable chance to succeed.

In contrast, an approach devoid of empathy, such as rigidly applying the pre-fixed rigid standards to all students, might fail to account for these barriers, penalizing the student unfairly and perpetuating their disadvantage. Empathy, a distinctly human trait, is relevant to determining that decisions are not only efficient but also informed by other human values such as compassion.

According to literature review and the selected arguments of the previous sections, the quest for human oversight in the deployment of AI is driven by both practical and ethical considerations. We enumerated that human oversight, among other reasons, aims to mitigate bias, preserve moral responsibility and to maintain empathy in certain contexts. How do these different quests for human oversight relate to the principle of fairness more specifically?

3. Relationship between human oversight and fairness

Even though the concept of fairness has been extensively debated in AI governance (Bellamy et al., 2019; Lee, Floridi & Singh, 2021), the AIA, across its 144 pages, mentions fairness only five times and defines it zero time in the legally binding part the document. In the preamble, however, recital 27 defines fairness as inclusion, promotion of equal access, gender equality, cultural diversity, lack of discriminatory impacts and biases (Recital 27, AIA). Despite this shy definition of fairness, we argue that the AIA provides a framework that is committed with other dimensions of fairness as defined by the literature.

The concept of fairness is multi-semantic and serves as a proxy for various normative ideas in AI literature. It can be understood as either substantive or procedural: substantive fairness aligns with specific normative objectives, while procedural fairness outlines processes designed to achieve broader goals set by policymakers or society at large (Naudts & Vedder, 2025).

3.1. Non-discrimination

In the context of decision-making, substantive fairness has been very often defined as the "absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" (Mehrabi et al., 2021, p.1). Fairness is the right to not be discriminated against (Hacker, 2018; Wachter, Mittelstadt & Russels, 2021; Xenidis & Senden, 2020; Zarsky, 2014; Zuiderveen Borgesius, 2018). Within this dimension, fairness is the "equal and just distribution of both benefits and costs" or the "equal opportunity in terms of access to education, goods, services and technology" (HLEG AI, 2019, p.12–13). This concept of fairness is the one mentioned in recital 27 of the AIA. This elaboration of fairness is related to the normative idea that human beings are, at their very fundamental level, equal. Therefore, efforts should be put in place to prevent and correct inequality in treatment (Binns, 2018, p.151; Hardt, Price & Srebro, 2016). In technical literature, the concept of fairness as the right to not be discriminated against is approached in different ways. Individual fairness, for instance, is achieved when individuals who have the same attributes receive similar AI prediction. Alternatively, group fairness seeks to ensure that different demographic groups (e.g. defined by race or gender) receive similar outcomes proportionally, preventing disparities in treatment or results across groups (Hacker, 2018, p.1175; Dwork et al., 2012).

The requirement of human oversight is mandatory to high-risk AI exactly to mitigate risks to fundamental rights, such as the right to not be discriminated against based on arbitrary reasons (recital 66, AIA). The non-discrimination principle is at the cornerstone of fundamental rights in the European Union. It is elevated to one of the fundamental values of the European Union in the Treaty on the European Union (Art. 2, TEU). Accordingly, the Treaty on the Functioning of the European Union requires the EU to combat discrimination based on sex, ethnic origin, religion or belief, disability, age and sexual orientation, when creating and enforcing its policies (Art. 10, TFEU). Additionally,

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

the EU Charter of Fundamental Rights stresses the relevance of the principle of equal treatment in the EU legal order, by containing a specific article that prohibits discrimination on various grounds (Art. 21, EU Charter of Fundamental Rights). Given the AIA mandates the requirement of human oversight to mitigate risks to fundamental rights posed by high-risk AI, in our view, it is not an extrapolation to argue that human oversight in the AIA framework also aims to address bias and illegal discrimination. This may be a high expectation, but the regulator opted to place human beings to counteract AI unfairness.

3.2. Accountability

Beyond equitable distribution of outcomes, fairness also relates to accountability (Binns, 2018). The AIA does not define fairness as accountability, but an extensive literature closely relates the two concepts (Veale, Van Kleek & Binns, 2018; Vedder & Naudts, 2017; Johnson, 2005). Accountability refers to the responsibility of actors to justify their decisions and be held answerable for the impact of their systems. Accountability requires that unfair AI outcomes will result in liability. In legal terms, liability aims at providing means for corrective actions, including the reparations of the victims and deterrence to act in deviation to fair standards. Human oversight facilitates the allocation of liability to certain overseers for eventual unfair AI outcomes. This is particularly relevant because AI does not dispose of legal personhood; therefore, AI systems cannot be held liable when in deviation to legal rules.

Accountability is not a focus in the AIA, as it does not directly regulate liability regimes for AI systems.[1] However, because the AIA governs the allocation of obligations along the AI value chain, including human oversight obligations for AI providers and deployers (the latter through appointed overseers), its provisions will inevitably affect the accountability of actors across the chain in relation to their assigned obligations.

3.3. Empathy

Human oversight is argued to maintain a certain level of empathy in AI decision-making processes, especially when those decisions impact human beings. This can help prevent overly cold or purely rational decision-making (Binns, 2020). Although fairness and empathy are related, they are distinct concepts. Fairness involves treating individuals equally according to consistent standards and ensuring accountability, while empathy is the human ability to understand and share others' emotions (Cuff et al., 2016). Nonetheless, empathy can drive humans to better grasp the nuanced context in which a person affected by a decision finds themselves, potentially leading to fairer outcomes. For example, an AI system tasked with making decisions purely based on objective data might not always reach the fairest result without considering additional context.

One such example is seen in workplace accommodation. An AI lacking empathy can undermine fairness in workplace accommodations by rigidly applying rules without considering individual circumstances. For example, if an employee is struggling to meet deadlines due to a personal loss, an empathetic human manager might offer flexibility or additional support to balance fairness with productivity. In contrast, an AI system might evaluate performance solely based on rigid metrics, flagging the employee as underperforming and triggering penalties.

3.4. Provisional conclusion and further investigation

According to this literature review, the relationship between human oversight and fairness lies in three points. (i) Human oversight attempts to mitigate AI bias and, therefore, to achieve fairness as

[1]The adaption of the liability rules to the digital age is the goal of other two legislative initiatives, the AI liability Directive (2022) and the revised Product Liability Directive (2022).

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

the right to not being discriminated against. (ii) Human oversight ensures that AI systems – which cannot be liable for harm – are overseen by individuals or legal entities who can be held legally liable, allowing the materialization of fairness as accountability. (iii) Human oversight introduces empathy to the decision-making process which may allow understanding of other contextual elements and result in fairer outcomes. Having these relationships between human oversight and fairness delineated, one relevant question deserves further investigation: Does the AIA establish a framework for human oversight that enables the implementation of these three dimensions of fairness? We explore this question in the following section.

4. Challenges and regulatory approaches to human oversight in the EU: is it 'better together'? From the GDPR to the AIA

While from technical and ethical perspectives human oversight over AI is argued to preserve numerous interests and values, including equal treatment, accountability and empathy, the real-life implementation of human oversight faces challenges.

One main challenge is the lack of meaningful human supervision, occurring when a human being is in the loop of the automated system, but this participation is irrelevant for the final outcome, even less for a fairer outcome. This happens, for instance, because people to whom human oversight tasks are entrusted (i) do not have enough time to properly assess the automated decision (Wagner, 2019, p.14); (ii) are not properly trained (Wagner, 2019, p.14) or (iii) lack formal authority to overrule the automated decision (Brennan-Marquez, Levy & Susser, 2019, p.757).

An important contribution to the debate on meaningful human supervision came from the Article 29 Data Protection Working Party, in its interpretation of the General Data Protection Regulation (GDPR). Even though, the GDPR does not explicitly mention "human oversight," article 22 and Recital 71 do refer to the "right to obtain human intervention" for data subjects who are subjected to decisions based solely on automated processing, including profiling, that produce legal effects or similarly significant impact on them. The Working Party sets a standard, by stating that when the human has "the authority and competence to change the decision" there is "meaningful oversight" rather than a mere "token gesture" (Article 29 Data Protection Working Party, 2018, July, p.21). Additionally, the guidelines recommend that data controllers consider implementing a hybrid decision-making process – or better together approach (Solove & Matsumi, 2024) – thereby increasing the level of human intervention to ensure that the decision-making process is no longer fully automated.

The authority and competence to change the decision of the human involved in the decision-making is also the underlying rationale of the recent judgement of the European Court of Justice on the Schufa case (Case C-634/21, SCHUFA Holding I), which emphasized that human involvement must be significant and capable of influencing the outcome, ensuring that the process remains fair and transparent. In the previous PNR case (Case C-817/19), the Court examined the compliance of automated systems processing passenger data under the PNR Directive. The decision stated that the agents in charge of the individual review must be provided by Member States with "clear and precise rules capable of providing guidance and support" to respect the fundamental rights of privacy, data protection and non-discrimination.

Academic literature has further explored the mechanisms of human intervention in the context of AI and automated systems, from a GDPR perspective. As noted by scholars like De Hert, human oversight must go beyond mere formalities and it must involve substantive review to ensure fairness and accountability (De Hert & Lazcoz Moratinos, 2021; Gil González & De Hert, 2019). Veale and Edwards (2018) elaborate on the complexities of implementing Article 22, arguing that ensuring effective human oversight requires not only legal compliance but also a deep understanding of the algorithms and their potential biases. Goodman and Flaxman (2017) discuss the broader regulatory challenges posed by Article 22, highlighting that a "right to explanation" must be paired with robust oversight mechanisms to be truly effective.

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

Along with the lack of meaningful human supervision, another, even more complex, challenge for the real-life implementation of human oversight pertains to the feasibility and effectiveness of human oversight as such to counterbalance AI risks (Green, 2022, p.11; Koulu, 2020, p.41).

This is the case when the human overseers are endowed with training on how the AI system works, authority to overrule the system's output and time to conduct their assessment, but they poorly perform in detecting AI errors, essentially not being able to reliably oversee the AI system's functioning. This is explained by a variety of causes, related to the behavioural responses to AI systems. Psychological pitfalls to which, as humans, we are variably subject, span from automation bias leading to over reliance on the system's output, to confidence bias resulting, on the contrary, in under reliance on the system's output and more in general to the development of heuristics about the competence of the system (Sterz et al., 2024, p.2501; Buçinca, Malaya & Gajos, 2021, p.2). Moreover, it has been highlighted how the increased reliance on automation leads to the degradations of the skills of the operators of the systems (Jones, 2020, p.11), such that being relieved from some functions do not allow the operators (i.e. the overseers) to perform better judgements (Elish, 2019, p.50). The root cause of the challenges to the feasibility and effectiveness of human oversight lies in the underlying contradiction between the use, in every sphere of human life, of technologies that are structurally meant to support or perform tasks that are better executed by machines – such as the analysis of enormous quantity of data to identify recurrent patterns – and the need to ensure that the same not-human-friendly tasks are effectively overseen by humans (Crootof et al., 2023, p.469; Green, 2022, p.11; Sterz et al., 2024, p.2497; Zerilli, Knott & Maclaurin et al., 2019, p.560).

However, all is not lost. Behavioural responses to AI can be better understood and improved, ultimately leading to more effective human oversight, as showcased by Langer, Baum and Schlicker (2025). In particular, the authors propose a framework to identify the factors that impact on the reliable detection of AI errors by humans, including both inaccuracies and unfair results.

Overall, what emerges from this account is that human oversight is not a panacea for the inherent flaws posed by AI technologies. Challenges to the implementation of meaningful human supervision, on the one hand, and the feasibility and effectiveness of human oversight, on the other hand, hinder the potential of this requirement to counterbalance AI risks. Nevertheless, the requirement for human oversight represents a crucial component of AI governance.

4.1. Human oversight under the AIA: what is the place of fairness?

Article 14, AIA explicitly requires that high-risk AI systems must be developed and deployed in ways that can be effectively overseen by natural persons during the period they are being used. In practice, who are the overseers, and which qualities should they have? What should be overseen, how and when? In essence, what is human oversight under the AIA?

And, most importantly, to what extent does the AIA establish a framework for human oversight that effectively supports the implementation of fairness as non-discrimination, fairness as accountability and fairness as empathy and why does it matter?

In each of the following sections, we will start by unfolding the elements of human oversight under the AIA in a structured way ("who," "what," "how," "when"). This will allow us to better frame the relationships between human oversight and fairness in the context of the AIA, and assess whether the AIA provides a framework for human oversight that enables fairness, in the three dimensions of non-discrimination, accountability and empathy that we have delineated above.

In particular, we argue that the different quests for human oversight and how they relate to the principle of fairness matter in practical terms. We will show how the three dimensions of fairness are relevant to a functioning human oversight and how they may influence the same interpretation, design and implementation of human oversight in the AIA.

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

4.1.1. Who are the overseers, and which qualities should they have?

The AIA foresees three categories of overseers, intervening in different capacities.

The first category consists of the AI system's provider, namely anyone – person, public authority, company, etc. – who either develops, places on the market or puts into service an AI system under its name, for payment or free of charge (Art. 3 (3) (9) (11), AIA). Providers must develop the high-risk AI systems with operational constraints by design (Art. 14(1) and (3) and Recital 73, AIA). In fact, such measures must be identified by the providers before the system becomes operational, in adherence with the system's features, as autonomy level, envisaged risks and intended use. Then, article 14(3) gives the option to the provider to directly build the oversight measures into the system or to leave the practical implementation to the deployer – any entity using an AI system under its authority, excluding the use for personal non-professional activities (Art. 3(4), AIA). Thus, this first oversight level consists of technical measures, e.g. interfaces, which are the responsibility of the provider, who however, in application of article 9(5), has to take into consideration the expertise to be expected by the deployer and the context of use.

The second category of overseers is the AI system's deployer, to be intended as the organization responsible for using the AI system in the course of a professional activity. The deployer has to (i) select the AI system that best fits its organization (private or public entity, as per Art. 3(4) of the AIA; (ii) if determined by the provider, implement the technical measures identified by the provider, as per Art. 14(3) of the AIA; and (c) assign in practice the human oversight tasks to personnel endowed of competence, training, authority and the necessary support, as per Art. 26(2) of the AIA.

The third category of overseers consists of the human operators, appointed by the deployer, to practically carry out the oversight tasks such as monitoring the AI system or interpreting its output, on behalf of the same deployer, as required by Articles 14(4) and 26(2) of the AIA. We define this category as "the appointed overseers" and we distinguish their role from that of the deployer, as the latter intervenes in a different capacity, by selecting which AI system to use in the first place, implementing the technical oversight measures identified by the provider and enabling the appointed overseers to perform their tasks. Finally, the obligation to ensure AI literacy established by Article 4 of the AIA impacts on the allocation of human oversight tasks. At all levels, human oversight must be entrusted to persons with adequate technical knowledge, experience, education and training, allowing them to make informed choices (Art. 3(56), Art. 4 and Recital 20, AIA).

Fundamentally, the AIA establishes a threefold structure of oversight that is in principle distributed between AI providers (those who develop or place AI systems on the market); deployers (those who use AI under their authority) and appointed overseers (those who practically execute oversight on behalf of the deployer). While the provider must technically ensure that the system can be overseen, the deployer has to execute the oversight through appointed human operators. However, the technical burden to identify the oversight measures is entirely placed on the provider, while the deployer is depicted as a mere executor of these measures.

How does this relate to fairness? In particular, are the fairness dimensions, explored in this article, relevant for the "who" of human oversight and do they influence its interpretation, design or implementation?

We argue that the distribution of oversight roles in the AIA has relevant implication for fairness as the right to not be discriminated against and as accountability.

First, while the roles of providers, deployers and appointed overseers appear to be sequential in time, they are not. It can be argued that providers need to continually reassess the technical oversight measures if the system shows adaptability after deployment, as is the case with machine learning (Laux, 2023, p.3) and deployers need to provide the appointed overseers with the necessary support to perform their tasks, including monitoring the adequacy of the oversight measures implemented.

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

Second, much of the decision-making power about the type of oversight measures to be set up is placed on the provider, who, however, does not have all the information about the actual context of use of the AI system at the disposal of the deployer.

As for fairness as non-discrimination, the adaptiveness of the AI system in the deployment phase and the lack of context in which the oversight measures are identified by the provider may compromise the efficiency of these technical measures to address discrimination. The provider bears the responsibility to minimize risks, including biases that could lead to discrimination. However, as providers may not have visibility into the specific contexts of AI deployment and the AIA does not prescribe a cooperation obligation between providers and deployers, there is potential for gaps in addressing context-specific biases. This impacts the right to non-discrimination, as general technical design cannot cover all nuances of fairness across diverse deployment contexts, possibly changing in the course of the deployment.

The uneven distribution of oversight roles described above, without a proper cooperation obligation between providers and deployers, impacts fairness as accountability, as the competences assigned respectively to providers, deployers and appointed overseers may not reflect their actual capacity to exercise such competences, with a consequent misplacement of accountability.

In light of these considerations, setting up human oversight measures in line with the fairness dimensions of non-discrimination and accountability implies a closer cooperation between providers and deployers, to remedy the lack of information of the former. It is relevant to consider that Art. 72 of the AIA establishes post-marketing monitoring obligations for providers, such that providers can acquire from deployers information on the effectiveness of the human oversight measures originally identified and possibly adapt them. However, it can be argued that even if not explicitly required by the AIA, dialogic mechanisms between providers and deployers should be established before the first use of the AI system by the deployer, to design and implement appropriate oversight measures from the very beginning.

4.1.2. What should be overseen and how?

What specifically needs to be overseen is not defined by the AIA. This is in line with the fact that the Regulation is founded on the proportionality principle according to which "one size does not fil all," so the actual human oversight – including the "what" – must be proportionate to the system's features, risks and context of use (Art. 14(3), AIA). While the AIA does not specify the "what" of the oversight, it contains many details about the "how" of the oversight.

First, the type of oversight required from the provider is prominently of a technical nature such as the development of interface tools to facilitate human supervision, while organizational oversight measures are not explicitly mentioned (Art. 14, AIA; Lazcoz & De Hert, 2023, p.10; Smuha & Ahmed-Rengers, 2021, p.35). Second, article 14(4) of the AIA contains a list of objectives that the deployer – more precisely the appointed overseers – must be enabled to realize in execution of the oversight measures identified by the provider. These objectives can be divided into two main groups: the appointed overseers must have the capability to (i) understand and (ii) to act.

Regarding the first group, the appointed overseer must be in the position to understand the capacities and the limitations of the system, be aware of the automation bias related to the overreliance on the system's output, correctly interpret the system's output (Art. 14(4) (a) (b) (c), AIA).

Understanding the systems, then, should allow the appointed overseer to act, by monitoring the system's operation, deciding when it is better not to use the system or disregard the system's output and intervening during the operation or halting the system (Art. 14(4) (a) (d) (e), AIA).

The understandability and the monitoring of the system is closely related to the requirements of transparency and explicability, enshrined by Articles 12 and 13 of the AIA (Enqvist, 2023, pp.517–518; Green, 2022, p.6). The recording of log mandated by Article 12 allows to trace the functioning of the system and thus facilitate the monitoring. Meanwhile, Article 13 prescribes that the system must be designed so that the deployer can interpret the output and use it appropriately (Art. 13(1), AIA) and

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

that the instructions for use made available to the deployer must include the specifications of human oversight measures, in particular of the technical measures facilitating the interpretation of the outputs (Art. 13(3)(d), AIA). Therefore, transparency and explicability, on the one hand, and human oversight, on the other, are functionally related: the overseer capability to interpret and monitor the system can result in better transparency and explicability of the system functioning towards its users, while only a system with good built-in transparency features can enable "effective" human oversight (Enqvist, 2023, pp.517–518).

How does this relate to fairness? In particular, are the fairness dimensions relevant for the "what" and "how" of human oversight and do they influence their interpretation, design or implementation?

First, we argue that fairness as non-discrimination should guide the execution of human oversight as per Art. 14(4) of the AIA, given that the understanding and the action upon the AI system is meant to prevent discrimination to happen. This means in practice that the appointed overseers must be endowed by the deployer (Art. 26(2), AIA) with contextual knowledge about the specific situation of use of the AI system, for instance in terms of targeted people/groups of people and geographical settings. Only with a thorough awareness of the context, the appointed overseer will be able to perform human oversight to prevent non-discrimination.

Second, fairness as accountability is undermined by the fact that the "what" and "how" of human oversight do not require a proper organizational oversight, risking a disproportionate burden on the appointed overseers – the last and weakest link in the oversight chain. However, Art. 26(3) of the AIA states the deployer's organizational freedom to implement the human oversight measures identified by the provider. We argue that this freedom comes along with the obligation of the deployer to identify and test in the first place the oversight tasks prescribed by Art. 14(4) of the AIA, then practically assigned to the appointed overseers. In fact, only the deployer is in the position to specify the tasks of the overseer and to be accountable, given that it is the deployer that primarily identifies within its organization the role assigned to the human–AI interaction.

Third, fairness as empathy is extremely relevant for the functioning of the "how" of human oversight and may significantly influence its design and implementation.

Empathy has been described above as a vector for the human overseer to better understand and investigate contextual elements, resulting in fairer outputs when there are peculiar circumstances that the AI system cannot factor in its decision-making process.

While empirical research focuses on how explainability may enhance human trust in AI decision-making (Leichtmann et al., 2023), the empathetic abilities of the overseers can be hampered by the assumption that a more transparent or explainable system allows for (better) human oversight, enriching what we defined the capability to understand. Research has showcased that humans tend to look at explanations as generic evidence of the system's competence, without engaging with them, because of the cognitive effort required that as humans we tend to avoid (Buçinca, Malaya & Gajos, 2021, p.2). Moreover, explanations about the system's functioning and output can be counterproductive by increasing the overseer's trust, when the system is actually malfunctioning (Green, 2022, p.7). Notably, the AIA does not consider these possible side-effects of the requirements of transparency and explainability, but only the positive effects.

Similarly, as regards what we defined the capability to act, it has been studied that humans do not usually perform well when they have to decide whether to disregard an automated output (Green, 2022, p.7) or to take over the control of the system during difficult or unexpected situations, while the system's functioning is regularly automated (so called hand-off problem) (Elish, 2019, pp.50, 53). And again, the AIA does not consider the problematic aspects related to hand-off scenarios.

Ultimately, the "how" of human oversight and its procedural elements, such as counting on the system explanation to accept or reject an AI decision, may negatively impact the human overseer quest for empathy to better engage and investigate the contextual elements at hand.

Awareness of the possible side-effects of transparency and explainability and, more broadly, of the mental processes set in motion by the specific human–AI interaction adopted in a given organization

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

is crucial to design and implement effective human oversight measures, as explored by (Buçinca, Malaya & Gajos, 2021; Langer, Baum & Schlicker 2025; Sterz et al., 2024; Zerilli, Knott & Maclaurin et al., 2019). In this sense, fairness as empathy can usefully guide the interpretation of the "how" of human oversight, by requiring providers and deployers – each with their own competences – to set up human oversight measures that take into account human reactions to automation. For instance, deployers could provide appointed overseers with training dedicated to the broader psychological biases of human–AI interaction, besides automation bias.

4.1.3. When should the oversight take place?

Article 14(1) requires that human oversight must be enabled during the period in which the high-risk AI systems "are in use." This indication per se is not much useful. However, the wording of the whole Article 14 and the combined reading of Art. 9, which demands the establishment of a "continuous iterative" process for risk management, suggest a continuative cycle of human oversight, from first use until shutdown (Laux, 2023, p.3). In fact, functions as monitoring, halting the system and even interpreting the output can be performed only if human oversight is enabled continuously.

What remains unclear under the AIA is the duration of the human intervention and the possible triggers that should activate or increase human oversight (and how) (Enqvist, 2023, p.525 and ff.).

How does this relate to fairness and does fairness influence the interpretation, design or implementation of the "when" of human oversight?

We argue that the three dimensions of fairness come into play.

The (inadequate) duration of the human intervention significantly affects fairness as non-discrimination and empathy. While it is understandable that the AIA does not prescribe a specific timeframe, given its broad application, it could have required that, in line with the specific features of the high-risk AI system in question, the appointed overseer must have sufficient time to conduct human oversight. The time factor, in fact, has been emphasized as one of the conditions that can render human oversight merely symbolic (Sterz et al., 2024, p.2502; Wagner, 2019, p.115). In essence, when the overseer lacks adequate time, the human contribution – particularly in carefully considering non-discriminatory impacts and understanding contextual elements – is diminished or even compromised. While sectoral legislation can compensate for this shortcoming, a horizontal safeguard in the AIA would have created a level playing field.

The choice of when human oversight takes place affects fairness as accountability. Deployers, through appointed overseers, must be enabled to exercise oversight in any situation and should decide when specific oversight tasks are performed. For instance, to avoid overreliance on AI system's output, deployers may demand the appointed overseers to consult the AI's output only after having conducted an independent check. However, the technical design choices made by providers may constrain when oversight is required, for instance by displaying the output before the appointed overseers are able to conduct their independent checks. The recurring issues of providers' lack of contextual knowledge and the potential adaptiveness of the AI system in the course of the deployment jeopardize the correct allocation of accountability for identifying the appropriate moments and specific triggers for human oversight tasks.

Therefore, interpreting the "when" of human oversight in line with the three fairness dimensions entails that (i) appointed overseers should be given sufficient time – for instance, based on the state-of-the-art in the relevant domain – to perform tasks which are not continuous, like interpreting the output and disregarding it; (ii) providers should design flexible human oversight interfaces allowing the deployers to identify the appropriate moments and specific triggers for human oversight tasks; (iii) deployers should instruct the appointed overseers about the duration and moments in which human oversight tasks must be performed, through organizational measures as protocols.

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

5. Final remarks

The requirement for human oversight is prominent in the AI governance debate and has been the subject of a specific political decision by the EU, as the AIA has established human oversight obligations for high-risk AI systems, although its implementation in practice faces challenges of meaningfulness, feasibility and effectiveness. Against this background, in this paper, we aimed to investigate (i) whether there are relationships between human oversight and fairness, a focus that is less explored in the legal literature to date, and (ii) whether and to what extent the AIA establishes a framework for human oversight that effectively supports the implementation of different dimensions of fairness.

Building on interdisciplinary literature review on human oversight, we could establish three normative ways in which human oversight relates to fairness. First, human oversight aims to mitigate AI bias, promoting fairness as the right to non-discrimination. Second, since AI systems themselves cannot be held liable for harm, human oversight ensures accountability by placing responsibility on individuals or legal entities who can be held accountable, thus enabling fairness as accountability. Third, by integrating empathy into the decision-making process, human oversight enables a deeper understanding of contextual factors, potentially leading to fairer outcomes in certain situations.

By drawing conclusions from our assessment of the human oversight infrastructure of the AIA vis-à-vis the three fairness dimensions, we argue that the AIA only partially supports these normative aspirations. The analysis conducted highlights (i) the uneven and strict distribution of oversight roles among providers, deployers and appointed overseers, which overlooks the importance of contextual information and the possible adaptiveness of AI systems during the deployment; (ii) the prominent technical nature of the human oversight prescribed and the lack of explicit reference to organizational measures; (iii) the lack of consideration of the possible side-effects of transparent and explainable AI systems and of the other mental processes set in motion by the specific human–AI interaction; and (iv) the lack of safeguards on the duration of the human intervention.

These four factors put at risk the achievement of fairness, in its three dimensions. Non-discrimination is affected because of the strict distribution of oversight roles leading to general technical oversight and because of the absence of safeguards on the duration of human intervention. Accountability is undermined due to the distribution of oversight roles which does not reflect the actual capacities and knowledge of each category of overseers and due to the lack of organizational oversight. Empathy may be compromised because mental processes – besides automation bias – triggered by human–AI interaction are not considered and the adequate duration of human intervention is not safeguarded.

Nevertheless, we consider the relationship between human oversight and fairness a valid guidance for interpreting the human oversight measures prescribed by the AIA and implementing them in view of the defined goals of non-discrimination, accountability and empathy. In this way, the weaknesses of the AIA can be addressed, and the three fairness dimensions acquire practical relevance for a functioning human oversight.

To remedy the strict distribution of competences made by the AIA, cooperation mechanisms between providers and deployers should be established, even before the first use of the AI system, and providers should design flexible interfaces for human oversight allowing deployers to make all the necessary adjustments. Deployers should identify and test in the first place the oversight tasks and instruct the appointed overseers on the duration and triggers of the human intervention, remedying the lack of organizational oversight and safeguards on duration. Finally, the mental processes set in motion by human–AI interaction should be carefully considered by providers and deployers implementing human oversight measures and *ad hoc*-training should be provided to appointed overseers.

Ultimately, implementing the human oversight measures prescribed by the AIA in view of the defined goals of fairness as non-discrimination, accountability and empathy positively addresses the challenges to meaningfulness, feasibility and effectiveness of human oversight.

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

Those we have proposed are only examples of actions that even if not mandated by the AIA, are nonetheless possible – and we argue necessary – in view of its implementation through the expected EU Commission Guidelines (Art. 96(1)(a), AIA), harmonized standards and common specifications (Art. 40 and 41, AIA) and even Codes of conduct for non-high risk AI (Art. 95, AIA).

Funding statement. This research was supported by the European Union trough the projects Themis 5.0 (Grant agreement 101121042) and PopEye (Grant agreement 101168317). The funding did not compromise the independence of this research.

Competing interests. The authors declare that they have no competing interests, financial or otherwise, that could have influenced the content or conclusions of this research.

References

Article 29 Data Protection Working Party. (2018, July). *Guidelines on automated decision-making and Profiling for the purposes of Regulation 2016/679*. https://ec.europa.eu/newsroom/article29/items/612053 Accessed November 1, 2024.

Banks, V. A., Plant, K. L., & Stanton, N. A. (2019). Driving aviation forward: Contrasting driving automation and aviation automation. *Theoretical Issues Ergonomics Science*, *20*(3), 250–264. https://doi.org/10.1080/1463922X.2018.1432716

Beck, J., & Burri, T. (2024). From 'human control' in international law to 'human oversight' in the new EU act on artificial intelligence. In D. Amoroso, and F. S. D. Sio (Eds.), *Research handbook on meaningful human control of artificial intelligence systems* (104–130). Cheltenham: Elgar. https://doi.org/10.4337/9781802204131.00014

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., … Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy and Technology*, *31*, 543–556. https://doi.org/10.1007/s13347-017-0263-5

Binns, R. (2020). Human judgement in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*, *16*(1), 197–211. https://doi.org/10.1111/rego.12358

Botero Arcila, B. (2024). AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight? *Computer Law & Security Review*, *54*, Article 106012. https://doi.org/10.1016/j.clsr.2024.106012

Brennan-Marquez, K., Levy, K., & Susser, D. (2019). Strange loops: Apparent versus actual human involvement in automated decision making. *Berkeley Technology Law Journal*, *34*(3), 745–771. https://doi.org/10.15779/Z385X25D2W

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1)*, Article 188, New York: Association for Computing Machinery. https://doi.org/10.1145/3449287

Council of the EU. (2024). Air Passenger Data: Council and European Parliament Reach Provisional Agreement to Increase Security and Enhance Border Management. https://www.consilium.europa.eu/en/press/press-releases/2024/03/01/air-passenger-data-council-and-european-parliament-reach-provisional-agreement-to-increase-security-and-enhance-border-management/

Crootof, R., Kaminski, M. E., Price, W., & Nicholson, I. I. (2023). Humans in the loop. *Vanderbilt Law Review*, *76*(2), 429–510. https://scholarship.law.vanderbilt.edu/vlr/vol76/iss2/2

Cuff, B. M., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review*, *8*(2), 144–153. https://doi.org/10.1177/1754073914558466

De Bruyne, J., & Dheu, O. (2023). Liability for damage caused by artificial intelligence: Some food for thought and current proposals. In P. Morgan (Ed.), *Tort liability and autonomous systems accidents common and civil law perspectives* (27–62). Cheltenham: Edward Elgar Publishing. https://doi.org/10.4337/9781802203844

De Bruyne, J., Van Gool, E., & Gils, T. (2022). Tort law and damage caused by AI systems. In J. D. Bruyne, and C. Van Leenhove (Eds.), *Artificial intelligence and the law* (2nd ed., pp.395–445). Intersentia.

De Hert, P., & Lazcoz Moratinos, G. (2021, October 13). *Radical rewriting of Article 22 GDPR on machine decisions in the AI era*. European Law Blog. https://www.europeanlawblog.eu/pub/radical-rewriting-of-article-22-gdpr-on-machine-decisions-in-the-ai-era/release/1

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference (ITCS '12). Association for Computing Machinery, USA*, 214–226. https://doi.org/10.1145/2090236.2090255

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Enqvist, L. (2023). 'Human oversight' in the EU artificial intelligence act: What, when and by whom? *Law, Innovation and Technology*, *15*(2), 508–535. https://doi.org/10.1080/17579961.2023.2245683

European Commission. (2020). *White Paper on Artificial Intelligence. A European approach to excellence and trust*. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en Accessed August 1, 2024.

European Commission. (2022). *The use of digitalisation and artificial intelligence in migration management*. https://home-affairs.ec.europa.eu/system/files/2022-02/00_EU_EMN_Digitalisation%20inform%20February%202022_EN_0.pdf Accessed August 1, 2024.

European Parliament. (2021). Guidelines for military and non-military use of Artificial Intelligence. https://www.europarl.europa.eu/news/en/press-room/20210114IPR95627/guidelines-for-military-and-non-military-use-of-artificial-intelligence Accessed August 1, 2024.

Gerards, J., & Xenidis, R. (2021). *Algorithmic discrimination in Europe – Challenges and opportunities for gender equality and non-discrimination law*. European Commission: Directorate-General for Justice and Consumers, Publications Office. https://data.europa.eu/doi/10.2838/544956 Accessed August 1, 2024.

Gil González, E., & De Hert, P. (2019). Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles. *ERA Forum*, *19*, 597–621. https://doi.org/10.1007/s12027-018-0546-z

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, *45*, Article 105681. https://doi.org/10.1016/j.clsr.2022.105681

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review*, *55*(4), 1143–1185. https://doi.org/10.54648/cola2018095

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), USA*, 3323–3331. https://doi.org/10.5555/3157382.3157469

High-Level Expert Group on Artificial Intelligence (HLEG AI). (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Accessed August 1, 2024.

Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology*, *7*(2), 99–107. https://doi.org/10.1007/s10676-005-4585-0

Jones, M. L. (2020). The ironies of automation law: Tying policy knots with fair automation practices principles. *Vanderbilt Journal of Entertainment and Technology Law*, *18*(1), 77134. https://scholarship.law.vanderbilt.edu/jetlaw/vol18/iss1/3

Keay, A. (2014). Comply or explain in corporate governance codes: In need of greater regulatory oversight? *Legal Studies*, *34*(2), 279–304. https://doi.org/10.1111/lest.12014

Koulu, R. (2020). Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law*, *27*(6), 720–735. 10.1177/1023263X20978649

Kyriakou, K., & Otterbacher, J. (2023). In humans, we trust: Multidisciplinary perspectives on the requirements for human oversight in algorithmic processes. *Discover Artificial Intelligence*, *3*(44). https://doi.org/10.1007/s44163-023-00092-2

Langer, M., Baum, K., & Schlicker, N. (2025). Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. *Minds and Machines*, *35*(1), 1–30. https://doi.org/10.1007/s11023-024-09701-0

Laux, J. (2023). Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. *AI and Society*, *39*, 2853–2866. https://doi.org/10.1007/s00146-023-01777-z

Lazcoz, G., and De Hert, P. (2023). Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities. *Computer Law & Security Review*, *50*, 1.

Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI and Ethics*, *1*, 529–544. https://doi.org/10.1007/s43681-021-00067-y

Leichtmann, B.,Humer, C.,Hinterreiter, A.,Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computer in Human Behavior*, *139*. https://doi.org/10.1016/j.chb.2022.107539

Loh, W., & Loh, J. (2017). Autonomy and responsibility in hybrid systems. In P. Lin, R. Jenkins & K. Abney (Eds.), *Robot ethics 2.0.: From Autonomous Cars to Artificial Intelligence* (35–50). New York: Oxford University Press. https://doi.org/10.1093/oso/9780190652951.003.0003

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), Article 115. https://doi.org/10.1145/3457607

Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches. *Digital Society*, *49*(2), 1–32. https://doi.org/10.1007/s44206-023-00074-y

Nagtegaal, R.(2021). The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly*, *38*(1), Article 101536. https://doi.org/10.1016/j.giq.2020.101536

Naudts, L., & Vedder, A. (2025). Fairness and artificial intelligence. In N. A. Smuha (Ed.), *The Cambridge handbook on the law, ethics and policy of artificial intelligence* (pp. 79–100). Cambridge: Cambridge University Press.

OECD. (2019). *Artificial Intelligence in Society*. Paris: OECD Publishing. https://doi.org/10.1787/eedfee77-en

OECD. (2021). *OECD Regulatory Policy Outlook 2021*, *Chapter 3*. Paris: OECD Publishing. https://doi.org/10.1787/38b0fdb1-en

Restrepo Amariles, D. (2017). Supping with the Devil? Indicators and the rise of managerial rationality in law. *International Journal of Law in Context*, *13*(4), 465–484. https://doi.org/10.1017/S1744552317000398

Santoni De Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, Article15. https://doi.org/10.3389/frobt.2018.00015

Shneiderman, B. (2016). Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(48), 13538–13540. https://www.jstor.org/stable/26472631

Smuha, N. A., and Ahmed-Rengers, E.(2021). How the EU can achieve legally trustworthy AI: a response to the European Commisson's proposal for an Artificial Intelligence Act. https://strathprints.strath.ac.uk/85567/1/Smuha_etal_SSRN_2021_How_the_EU_can_achieve_legally_trustworthy_AI.pdf Accessed August 1, 2024.

Solove, D. J., & Matsumi, H. (2024). AI, algorithms, and awful humans. *Fordham L. Rev*, *92*(5), 1923–1940. https://fordhamlawreview.org/issues/ai-algorithms-and-awful-humans/

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, USA*, 2495–2507. https://doi.org/10.1145/3630106.3659051

Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, *34*(2), 398–404. https://doi.org/10.1016/j.clsr.2017.12.002

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. 10.1145/3173574.3174014. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, USA*, Paper 440, 1–14. https://doi.org/10.1145/3173574.3174014

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology*, *31*(2), 206–224. https://doi.org/10.1080/13600869.2017.1298547

Verdiesen, I., Santoni de Sio, F., & Dignium, V. (2021). Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. *Minds and Machines*, *31*, 137–163. https://doi.org/10.1007/s11023-020-09532-9

Wachter, S., Mittelstadt, B., & Russels, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, Article 105567. https://doi.org/10.1016/j.clsr.2021.105567

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet*, *11*(1), 1–19. https://doi.org/10.1002/poi3.198

Xenidis, R., Senden, L. (2020). EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination. In U. Bernitz, *et al.* (Ed.). *General principles of EU law and the EU digital order* (151–182). lphen aan den Rijn: Kluwer Law International.

Zarsky, T. Z. (2014). Understanding Discrimination in the Scored Society. *Washington Law Review*, *89*(4), 1375–1412. https://digitalcommons.law.uw.edu/wlr/vol89/iss4/10

Zerilli, J., Knott, A., Maclaurin, J. et al. (2019). Algorithmic decision-making and the control problem. *Minds and Machines*, *29*, 555–578. https://doi.org/10.1007/s11023-019-09513-7

Zuiderveen Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making Council of Europe. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic%20decisionmaking/1680925d73 Accessed August 1, 2024.

Dr. Ana Maria Corrêa is a senior researcher at the KU Leuven Centre for IT & IP Law (CiTiP). Her research explores the intersection of law, digital technologies, and human rights. She leads multiple EU-funded research projects on AI, data spaces, and robotics, collaborating with universities, policymakers, and technology firms across Europe and the United States.

Sara Garsia is a Doctoral Researcher at the Centre for IT and IP Law (CiTiP) of the KU Leuven Faculty of Law and Criminology. Since October 2023 she has worked as a legal research associate at CiTiP, with a research focus on human-centered AI Governance, in the context of the EU funded project THEMIS 5.0. She has recently started her PhD, where along with the research of the FLEXIQ project, she questions how energy-related data are leveraged for the clean energy transition in the EU and which implications arise from a fairness and ultimately a social justice perspective.

https://doi.org/10.1017/cfl.2025.10010 Published online by Cambridge University Press

Abdullah Elbi is a PhD researcher at the Centre for IT & IP Law (CiTiP) at KU Leuven Faculty of Law and Criminology and a member of the Biometric Law Lab. He holds an LL.B. from Bilkent University's Faculty of Law (June 2019) and is a qualified lawyer in Türkiye. As a Jean Monnet scholar, he earned his LL.M. in Law and Digital Technologies from Leiden University (September 2021). His research explores the fundamental-rights implications of emerging technologies, with particular emphasis on data protection, trustworthy artificial intelligence, AI literacy, meaningful human oversight mechanisms, and biometrics. He has contributed to several EU- and nationally funded research projects, including iMARS, FAITH, PopEye, CORTEX2, and SALT.

Cite this article: Corrêa A.M., Garsia S. and Elbi A. (2025). Better together? Human oversight as means to achieve fairness in the European AI Act governance. *Cambridge Forum on AI: Law and Governance* **1**, e29, 1–17. https://doi.org/10.1017/cfl.2025.10010