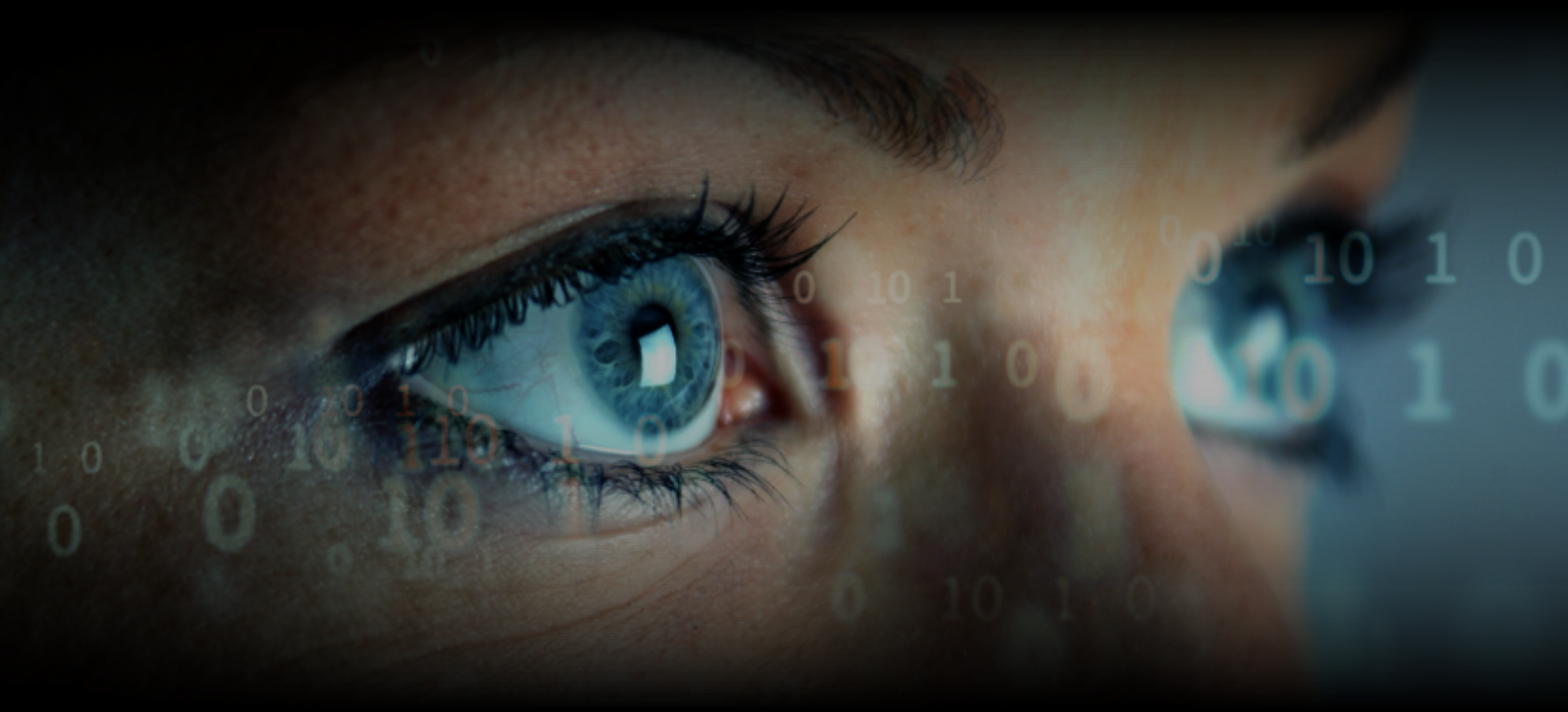
TechDispatch

HUMAN OVERSIGHT

OF AUTOMATED DECISION-MAKING



EDPS

European Data Protection Supervisor

Luxembourg: Publications Office of the European Union, 2025

PDF ISBN 978-92-9242-930-0 ISSN 2599-932X doi:10.2804/8957972 QT-01-25-003-EN-N

CONTENTS

1. Introduction

Two centuries of technological progress have redefined society, as industrial and digital automation have modified how people live, work, and how we interact with each other. What began with mechanised production lines to increase efficiency has evolved into autonomous and semi-autonomous systems capable of analysing data and making decisions with minimal human involvement. Propelled by breakthroughs in Artificial Intelligence (AI) - such as deep learning and transformer architectures - this automation now spans a wide range of sectors, including healthcare, finance, military defence, transportation, and public administration.[1]

A significant development in this evolution has been the automation of decision-making processes with the development of systems that not only execute tasks but also make decisions that can affect individuals' lives and rights. **Automated decision-making** (ADM) is increasingly relied upon to process personal data and make consequential determinations, from eligibility assessments for social benefits to medical diagnoses, recruitment procedures, and credit approvals.

While technological advances in ADM offer significant potential, they also introduce risks of opacity, bias, and discrimination in decision outcomes. Such risks can undermine trust in technology and lead to violations of individual rights, as well as broader harm to democratic processes and societal cohesion.[2] To mitigate these concerns, ADM systems that significantly impact people's lives must be subject to rigorous monitoring to safeguard privacy, data protection, non-discrimination, and other fundamental rights.

Individuals might not always be aware that they are subjected to ADM. This can create an imbalance of power between those affected by these systems and those who design, deploy, or control them. As ADM becomes increasingly integrated into processes, tools, and services, it is essential to ensure that these systems are not left to make autonomous, uncontrolled decisions that affect individuals' fundamental rights.

Therefore, the involvement of humans as a safeguard against the risks associated with ADM systems (e.g., algorithmic bias and misclassification[3]) is increasingly perceived as necessary. Integrating human judgment at various stages - during the design, real-time monitoring, or post-decision audits - can help ensure that ADM systems align with ethical standards, societal values, and regulations.

However, simply adding a human within the decision-making process does not inherently ensure better outcomes, nor should it serve as a means to deflect accountability for the system's decisions.[4] In fact, just including a human is unlikely to prevent systems from producing wrongful

---

1 In many cases, decision-making systems are implemented to support robotic operations across various industrial domains, including automotive manufacturing, electronics assembly, and logistics. Although automation and robotics are closely interrelated, this document focuses exclusively on automated decision-making processes and does not address robotics as a distinct subject. For the purposes of this document, automation refers specifically to the automation of the decision-making process itself, regardless of whether the resulting decision is operationalized through robotic systems.

2 Future of Privacy Forum. (2017). Unfairness by algorithm: Distilling the harms of automated decision-making. https://fpf.org/wp-content/uploads/2017/12/FPF-Automated-Decision-Making-Harms-and-Mitigation-Charts.pdf

3 ADM misclassification occurs when an ADM system assigns an incorrect label or category to input data - such as flagging a legitimate transaction as fraud.

4 In her work "*Moral Crumple Zones Cautionary Tales in Human-Robot Interaction*", Madeleine Elish introduces the idea that human operators sometimes act as *moral crumple zones* for the systems. "*While the crumple zone in a car is meant to protect the human driver, the moral crumple zone protects the integrity of the technological system, at the expense of the nearest human operator*" (Elish, 2025).

or harmful outcomes for individuals. This is frequently due to inadequate implementations or lack of control over the system - issues that will be further examined in this document.

In any case, for human oversight to be **meaningful** and **effective**,[5] it must be carefully structured, taking into account both the limitations of ADM systems and the complex dynamics between human operators and machine-generated outputs.

The objective of this TechDispatch is twofold. First, it examines common assumptions about how humans interact with and monitor decision-making systems, highlighting the **overly optimistic nature of many of these assumptions**. Accepting these assumptions uncritically can lead to inadequate or flawed implementations, posing significant risks - including harm to individuals and potential violations of fundamental rights. Second, it explores practical measures that providers[6] and deployers[7] of ADM systems can take to ensure that human oversight supports democratic values and safeguards human rights.

It is important to note that this Tech Dispatch does not aim to offer any legal interpretation. Instead, it focuses on exploring how the implementation of human oversight can impact its overall effectiveness.

This TechDispatch builds upon the knowledge gathered during the Internet Privacy Engineering Network (IPEN) event, organised by the EDPS and the Karlstad University in September 2024.[8]

2. Definition of human oversight of ADM systems

Before analysing the potential role of human oversight for safeguarding against the risks posed by ADM systems to fundamental human rights, it is important to first clarify the meaning of 'human oversight' as defined in this document.

For the purposes of this document, *human oversight* refers to the active involvement of at least one human *operator*[9] in monitoring ADM system operations, evaluating the system's decisions, and having the ability to intervene if necessary (Sterz, S., et al., 2024).

While human oversight can occur at different stages of a system's lifecycle, including before deployment (ex-ante),[10] *real-time* oversight on system operations (i.e. the supervision of systems deployed in production environments) is considered the one that can be most consequential.

5 Chapter 2 (*Definition of human oversight of ADM systems*) provides a detailed description of what is meant by 'meaningful' and 'effective' human oversight in the context of this document.

6 This document uses the term 'providers' to refer to organisations that develop and make ADM technology available.

7 This document uses the term 'deployers' to refer to any natural or legal person, public authority, agency or other body using an ADM system under its authority except where the ADM system is used in the course of a personal non-professional activity.

8 Information and recordings of the event can be found in the EDPS webpage in: https://www.edps.europa.eu/data-protection/technology-monitoring/ipen/ipen-event-human-oversight-automated-making_en

9 The human 'operator' is a person who is properly trained to be able to supervise the system and has enough authority to override the system's decisions.

10 Human oversight of AI systems - legal aspects of new technologies. (Krzysztof Wojdylo). - Legal Aspects of New Technologies. https://newtech.law/en/articles/human-oversight-of-ai-systems [Accessed in 01/07/2025]

This stage represents the critical window in which a human operator can still review the system's behaviour and intervene before its output takes effect, helping to prevent potential harm to human lives or infringements on individuals' fundamental rights.

Therefore, this document will focus on ***real-time*** human oversight.[11]

For the purposes of this document, '**meaningful**' human oversight is defined as active involvement that improves the quality of the decisions taken by the system, rather than as a merely procedural formality, or symbolic gesture. A Human is empowered and positioned to monitor the system in a substantive way.

In turn, '**effective**' human oversight, as interpreted in this document, refers to human involvement that has a tangible, positive impact on outcomes - specifically by preventing or mitigating harm, promoting fairness, and enhancing the overall quality, reliability, and accountability of decisions.

This document presents examples of scenarios where human operators monitor systems designed to automate decision-making processes. The most prominent types of ADM systems are those powered by AI technologies - such as machine learning and deep learning - due to their advanced capabilities and widespread use across multiple sectors.

However, while many examples focus on AI-based ADM systems, others do not. It is important to emphasize that **ADM systems can also be implemented without AI**, using predefined rules, deterministic algorithms, or programmed logic.

Therefore, for the purposes of this document, we consider ADM systems in a broad sense - whether AI-based or not.

3. Flawed assumptions about human oversight of ADM

When an ADM system is deployed with human oversight, there is a significant risk of creating unfounded expectations regarding its effectiveness. Both system deployers and affected individuals may erroneously assume that the system will invariably perform reliably, believing that any anomalies will inevitably be detected and addressed by human operators. Such assumptions can lead to unrealistic expectations and foster overreliance on the system.

At the same time, if regulatory frameworks are based on overly simplistic assumptions about how human oversight should be implemented in practice, the resulting obligations for providers and deployers may be unclear or insufficiently detailed. This can undermine the effectiveness of safeguards and negatively impact all stakeholders - including system developers, operators, and the individuals affected by automated decisions.

The following sections categorises these assumptions, distinguishing whether they are associated to the system itself, the behaviour of the human operator, or the nature of the human interaction with the system.

[11] Nevertheless, Chapter 4 (*Promoting more effective and meaningful human oversight*) will detail particular steps that are to be taken following the decision-making process. One example of this is regular auditing.

3.1 Wrong assumptions about the technology used in ADM systems

Assumption: ADM systems will operate within specific and predetermined conditions

It is important to acknowledge that the design and implementation of technology typically involve a set of assumptions about the environments in which they will operate. For example, when designing a car, manufacturers generally assume it will be used in climates with temperatures suitable for human life - even though some regions on Earth may fall well outside that range. Such design limitations are common and often necessary; designing vehicles to withstand extreme arctic conditions, for instance, would be prohibitively expensive and unnecessary for the vast majority of users.

However, assuming that ADM systems will consistently operate under expected or controlled conditions can be misleading, and potentially dangerous, particularly in dynamic or unpredictable environments. The proper functioning of a self-driving car, for instance, may depend on the presence and good condition of road markings or signs, which might be missing, damaged, or intentionally manipulated.[12] In these situations, human oversight becomes essential, enabling human judgment to address exceptional or unforeseen scenarios - but only if the operator has sufficient time to react and is supported by interfaces that facilitate timely and effective intervention.

> **Example:** *Semi-autonomous driving systems rely heavily on cameras, sensors, and radars to detect road signs, lane markings, and other vehicles. However, there have been multiple reports of incidents where the system failed to detect obstacles or road markings accurately, leading to crashes, including fatal ones. For instance, in 2016, a Tesla vehicle in autopilot mode failed to detect a white truck crossing a highway against the bright sky, leading to a fatal crash.*[13]
>
> In this case, the system operated beyond its capacity to accurately interpret real-world complexities, such as unusual lighting conditions and atypical road obstacles that were underrepresented in its training data.

Importantly, it is not sufficient for human operators to be aware of the automated system's limitations - **systems should also incorporate safeguards that prevent them from operating beyond their intended function.**

This leads to another concerning assumption.

---

[12] A 2018 article warned of the risk of tricking self-driving cars into considering stop road signs as speed limit signs by placing stickers in specific areas of the road signs (called 'perturbations'). Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634). https://arxiv.org/pdf/1707.08945

[13] Yadron, D., & Tynan, D. (2016, June 30). Tesla driver dies in first fatal crash while using autopilot mode. The Guardian. Retrieved from https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk [Accessed in 01/07/2025]

Assumption: ADM systems will transfer control to humans when handling atypical or outlier situations

As mentioned previously, ADM systems are often not equipped to reliably assess the limits of their own capabilities or to account for how the complexity of how real-world complexity affects task performance. These systems typically operate in complex and dynamic environments, where unexpected or novel situations can arise - many of which may not have been foreseen during programming, training, testing or fine-tuning.

Even when a system detects high levels of uncertainty, it may not interpret this as a signal to relinquish control, particularly if it lacks an explicit mechanism to defer decision-making or escalate to a human. The assumption that such handovers will eventually occur can therefore be misleading.

To ensure responsible use, ADM systems should be designed not only to perform tasks but also with guardrails halting or modifying their actions when operating under conditions for which they are ill-equipped, and to communicate this clearly to the human users. Without appropriate safeguards, the assumption that systems will defer to human control in outlier situations—where they lack reliable or safe courses of action—remains unfounded.

Ideally, systems should be designed to evaluate whether the conditions needed to reliably complete a task are met - without going beyond their intended capabilities. For example, an automated welfare eligibility system that reviews documentation related to benefit allocation should be able to identify cases with complex or unclear circumstances that don't fit predefined criteria - such as income records that conflict with declared income or instances of undocumented hardship. In these situations, the system should proactively alert a human operator and, when necessary, transfer control to them to allow for timely and informed decision-making.

Finally, ADM systems and decision-support systems (DSS)[14] should provide confidence scores that help operators assess the reliability of their outputs. In addition, ADM system should offer a clear rationale[15] for how a decision was reached. This can support operators in evaluating whether the underlying logic is sound.

[14] Decision support systems (DSS) are computer-based tools designed to assist individuals or organisations in making informed decisions. Rather than making decisions on their own, DSS help users analyse data, evaluate options, and anticipate outcomes to support better judgment - especially in complex or uncertain situations. Decision support systems are considered, within the scope of this document, to be **partially automated systems**.

[15] See Chapter 4, sub-section "*The system decisions should be explainable*"

Example: *The Babylon Health AI-powered symptom checker was a tool integrated in the UK's National Health Service (NHS) General Practitioner at Hand service, aimed to provide patients with 24/7 access to health assessments and advice. However, it faced significant challenges in accurately diagnosing conditions and appropriately triaging patients. In a 2018 report, the Care Quality Commission (CQC), the UK's healthcare regulator, raised concerns about the safety and effectiveness of Babylon's AI chatbot. The report highlighted instances where the AI failed to detect serious conditions, such as heart attacks, and provided inappropriate advice, like suggesting that a woman with a breast lump might be experiencing "hysteria" rather than a medical issue.*[16] *Furthermore, a 2020 investigation by BBC Newsnight revealed that Babylon's AI chatbot could not consistently identify critical symptoms, leading to potential misdiagnoses and delayed treatments.*

3.2 Wrong assumptions about how humans and technologies interact

Assumption: Automation does not influence human judgment

In many decision-making contexts, humans rely on automated inferences, recommendations, or data-driven insights generated by algorithms before making a final choice. This is the case of decision-support systems that analyse vast amounts of data, identify patterns, and present suggested outcomes or filtered options that effectively shape the decision landscape for the human user. As a result, the system's outputs can frame the available choices, subtly nudging users toward specific conclusions or reinforcing existing biases - a phenomenon known as ***automation bias***.[17]

Operators' decisions are often shaped by a strong trust in the system's perceived expertise, especially when the system performs challenging calculations or inferences, or operates on complex or highly specialized data for which the human may feel unqualified to intervene.

In fields like medical diagnosis, financial investment, or recruitment, the recommendations or risk assessments generated by DSS tools can significantly influence a person's perception of the available options. Although the final decision may rest with a human, the process may already have been steered or constrained by the system's suggestions in ways that are not always apparent or easily overridden. This integration of automated reasoning into human decision-making blurs the boundary between the roles of the human and the machine, revealing that the decision-making process is not wholly human-driven but rather co-dependent on machine processes.

[16] Browne, G. (2023, September 19). The fall of Babylon is a warning for AI unicorns. WIRED. https://www.wired.com/story/babylon-health-warning-ai-unicorns/ [Accessed in 01/07/2025]

[17] The tendency for humans to favour suggestions from automated systems has been well documented over the years. Research has shown that individuals often follow system recommendations, even when those recommendations conflict with their training or other available information (Skitka, 1999).

Example: *A 2023 study published in Radiology*[18] *revealed that radiologists, regardless of experience level, are susceptible to automation bias when interpreting mammograms with AI assistance. The study involved 27 radiologists who evaluated 50 mammograms, with AI-generated Breast Imaging Reporting and Data System (BI-RADS) suggestions provided. In cases where the AI's suggestions were incorrect, radiologists' accuracy significantly declined.*

The findings underscore the need for implementing appropriate safeguards when incorporating AI into radiological processes to mitigate the negative consequences of automation bias. Such measures may include presenting users with the confidence levels of AI predictions and providing training to help radiologists understand the strengths and limitations of AI tools.

Assumption: There is no automated decision-making if there is a human supervising the system

The phenomenon of automation bias is closely linked to the erroneous assumption that **there are no automated decisions if there is a human supervising the process** - an assumption that can foster unrealistic expectations and lead to inadequate implementation of human oversight.

Deployers of decision-support systems - that is, systems that provide recommendations, without making the final decision - may be led to believe that, when they deploy their systems, the decision-making is not automated as long as it is the human who makes the final decision.

However, even if the final action or decision requires human intervention, it is important to recognize that the choice is often made based on a course of action suggested by the system.

To provide suggestions, decision-support systems typically evaluate what data is relevant to their analysis and perform inferences on that data[19] throughout the process that leads to their outputs, whether in the form of predictions or recommended decisions.[20]

In doing so, the outputs presented to the user often reflect a series of intermediate analytical steps and decisions already made by the system.[21]

While it is understandable that these intermediate steps are fundamental to the value of automated systems, it is crucial for both deployers and operators to recognize that the system's outputs may result from a sequence of automated processes.

If a human operator accepts the system's output as valid primarily because it was made by an automated system, a behaviour commonly known as *automation bias*, then the decision-making process, in effect, **becomes equivalent to an automated one**.

[18] Yala, A., Lehman, C. D., Schapiro, M., & Barzilay, R. (2023). The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. Radiology, 307(4), e222176. https://doi.org/10.1148/radiol.222176

[19] In AI, an inference refers to the process by which a system uses a model (such as a machine learning model) to draw conclusions, make predictions, or produce outputs based on input data.

[20] **Feature selection** and **attention mechanisms** are two aspects of modern AI systems in which relevance evaluation is particularly evident. In **feature selection**, the system identifies and utilizes only the most pertinent input variables for analysis, while **attention mechanisms**, particularly in deep learning models, dynamically prioritize the most important parts of the input data during prediction or output generation.

[21] See Reuben Binns, Michael Veale, Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR, *International Data Privacy Law*, Volume 11, Issue 4, November 2021, Pages 319–332, https://doi.org/10.1093/idpl/ipab020

Some authors have referred to this configuration, in which human decisions are made while influenced by system recommendations, as *"quasi-automation"* (Wagner, 2019).

> **Example:** *In 2017, the Durham Constabulary*[22] *deployed the Harm Assessment Risk Tool (HART), an algorithmic system designed to support officers in assessing the risk of reoffending - categorizing individuals as low, moderate, or high risk. Although the final decision on custody or rehabilitation was left to officers,* ***in practice they frequently followed the algorithm's recommendation****. Observers noted*[23] *that HART's predictive scores "guided decisions as to whether a suspect should be charged or released onto the Checkpoint rehabilitation programme", illustrating how the system's suggestion heavily influenced the application of human discretion.*
>
> This phenomenon reflects a classic case of automation bias - where humans defer to machine outputs even when they hold ultimate authority. In this case, concerns emerged that officers might rely excessively on HART, effectively turning a decision-support tool into a *de facto* automated decision-maker.

The purpose of this section is not to assert that recommendation systems (such as DSS) constitute ADM systems. Rather, it aims to highlight that such systems, while not making final decisions themselves, **can exert unforeseen influence on human decision-making**.

For this reason, in contexts where human operators rely heavily on system recommendations, there should be a *presumption of automation by default* -meaning that the deployer should treat the system as if it was operating autonomously and apply effective human oversight. For instance, this is one reason for the design of the system's interface to stimulate the critical thinking of the operator.

Assumption: That human operators possess the authority to decide

Another common misconception is that human operators will always retain meaningful authority to override the suggestions provided by automated systems - some authors refer to this as *agency*.[24]

This assumption fails to account for the practical limitations on independence and expertise that might constrain an operator's ability to deviate from system recommendations.

Operators may choose not to override a system's output due to fear of potential consequences, particularly if the system's recommendation is later found to be correct. **This hesitation reflects a lack of perceived agency.**

Moreover, when systems operate on complex or highly specialized data, operators may delegate to the system's apparent expertise, feeling insufficiently qualified to challenge its

[22] The territorial police force responsible for policing the council areas of County Durham and Darlington in North East England - UK.

[23] Waddell, K. (2019, October 19). In AI we trust - too much. AXIOS. https://www.axios.com/2019/10/19/ai-automation-bias-trust [Accessed in 01/07/2025]

[24] See (Wagner, 2019), (HLEG, 2019), and (Taddeo, 2021)

outputs.[25] This perceived imbalance in knowledge can discourage intervention and reinforce reliance on the system, even in situations where human judgment is necessary.

> **Example:** *In 2020, the UK exams regulator, Ofqual, deployed an algorithm to standardize GCSE[26] and A-level grades based on historical school performance. This approach led to approximately 40% of students receiving lower grades than their teachers had predicted, disproportionately affecting students from disadvantaged backgrounds.*[27]
>
> *Despite the algorithm's significant impact, human oversight was insufficient and lacked clear authority. The Royal Statistical Society offered to assist in reviewing the algorithm but was met with non-disclosure agreements and limited engagement. Furthermore, the External Advisory Group set up by Ofqual had no power to alter the algorithm's application.*
>
> This lack of human agency rendered the oversight ineffective, highlighting the risks of symbolic rather than substantive human involvement in ADM systems.

Assumption: Systems combining human and machine work better

The assumption that hybrid systems - those that integrate human judgment with machine outputs - are inherently superior is misleading. It overlooks the complex challenges involved in coordinating human and ADM, often referred to as the "MABA-MABA trap" ("Men Are Better At, Machines Are Better At").[28] This concept reflects the overly optimistic assumption that machines and humans will seamlessly complement each other - machines managing structured, rule-based tasks, while humans handle ambiguity, context, and nuance.

In practice, this synergy is difficult to achieve. **Human operators may struggle to interpret or trust machine-generated outputs, especially when systems are opaque or complex**. This can result in overreliance on automation or, conversely, hesitation to override flawed recommendations - both of which increase the likelihood of error. Meanwhile, machines often lack the contextual awareness needed to adapt their recommendations to human judgment, which can lead to misalignment between system suggestions and human intent. In the work (Crootof, R., et al., 2023) the authors warn that *"hybrid system can all too easily foster the worst of both worlds, where human slowness roadblocks algorithmic speed, human bias undermines algorithmic consistency, or algorithmic speed and inflexibility impair humans' ability to make informed, contextual decisions"*

Without deliberate design, clear role allocation, and appropriate training, hybrid systems may end up compounding the weaknesses of both components rather than harnessing their strengths. Instead of enhancing decision-making, such systems can introduce confusion, reduce accountability, and ultimately degrade performance.

---

[25] See 'automation bias' in subsection "*Assumption: Automation does not influence human judgment*"

[26] GCSE is the qualification taken by 15 and 16 year olds to mark their graduation from the Key Stage 4 phase of secondary education in England, Northern Ireland and Wales.

[27] Burgess, M. (2020, August 20). The lessons we all must learn from the A-levels algorithm debacle. WIRED. https://www.wired.com/story/gcse-results-alevels-algorithm-explained/ [Accessed in 01/07/2025]

[28] The MABA-MABA list was introduced by Paul Fitts in the 1950s and was meant to help designers decide which parts of a task should be handled to people or automated systems, aiming to leverage the complementary strengths of humans and machines.

Example: *The Boeing 737 MAX airplane models operate with the Manoeuvring Characteristics Augmentation System (MCAS). Designed to automate stabilizer adjustments, Boeing assumed that MCAS could handle structured tasks while pilots would manage nuanced situations. However, this division failed in practice: pilots became distrustful of MCAS as it sometimes pushed the airplane nose down unexpectedly, and due to limited transparency and inadequate training, they hesitated or struggled to override the system during emergencies.*[29]

This overreliance on automation, combined with reduced situational awareness and misaligned roles, contributed to two fatal crashes in 2018 and 2019, demonstrating how the assumption that humans and machines will seamlessly complement each other can lead to critical failures when the coordination between human judgment and automated systems breaks down.

Assumption: Human operators have appropriate mechanisms to override the system's decisions

This assumption is problematic because it overlooks frequent deficiencies that exist in the interfaces designed for operators to interact with automated systems. Without accurate, interpretable, and timely information, operators may lack sufficient situational awareness to detect system errors or anomalies in real time.

Even when an issue is identified, the tools available to intervene could be inadequate. Interfaces may lack responsiveness, clarity, or intuitive controls, particularly in time-sensitive or high-stakes environments.[30] Furthermore, many systems are configured in ways that prioritize automation over manual input, limiting the operator's ability to act swiftly and effectively when intervention is needed.

As a result, **human operators may be formally involved but functionally unable to monitor the systems in a meaningful way**. This disconnect renders the assumption of straightforward or reliable human override unrealistic in many real-world applications, particularly when system design fails to support timely human awareness and control.

[29] The New York Times (Sept. 26, 2019), Boeing Underestimated Cockpit Chaos on 737 Max, N.T.S.B. Says. Retrieved from https://www.nytimes.com/2019/09/26/business/boeing-737-max-ntsb-mcas.html [Accessed in 01/07/2025]

[30] *"In high-risk situations, the assumption that the human will be capable of taking over control immediately without disruption can result in severe miscoordination and ultimate system failure."* (Methnani, L., et al., 2021)

Example: *Between 2014 and 2019, Poland's Public Employment Services (PES) deployed a profiling algorithm that sorted job seekers into one of three categories - based on data such as age, education, and work history - to allocate varying levels of support.*[31] *Client advisors were designated as human supervisors, with the authority to override the system's classifications.* ***However, the system's design*** *and organisational constraints limited their ability to intervene effectively. Advisors faced challenges such as excessive workloads,*[32] *insufficient training, and a lack of clear guidelines on when and how to override the system's decisions. In some instances, local managers discouraged or prohibited overrides, fearing scrutiny from higher authorities. Additionally,* ***the system's outputs were presented in ways that made it difficult*** *for advisors to assess their accuracy or relevance to individual cases. Consequently, despite their formal role, human operators were functionally unable to monitor and intervene in the ADM system in a meaningful way.*

Assumption: Transparency and explainable AI improve human oversight

While transparency and explainability are often promoted as tools to enhance human oversight, this assumption can sometimes be misleading. Although explainable AI (XAI) can help users understand a system's general functioning, it does not necessarily enable them to identify when the system is ill-equipped to handle exceptional or unforeseen situations - precisely when human oversight is most critical.

XAI tools typically focus on clarifying routine decision-making processes, but may not expose the model's limitations, especially in edge cases or outlier scenarios. As a result, users may develop a false sense of confidence in the system's reliability, accepting its outputs without fully questioning underlying assumptions, biases, or blind spots. In some cases, **the presence of seemingly coherent explanations can even legitimize flawed or biased models**, masking deeper issues under the appearance of transparency.

Therefore, while explainability can in general support understanding, it does little to address the system's vulnerabilities in atypical contexts - ultimately limiting its value for meaningful human oversight.

Example: *A 2021 study*[33] *investigated the impact of machine learning (ML) recommendations and their explanations on clinicians' decision-making in selecting antidepressant treatments. In a controlled experiment involving 220 clinicians, participants were presented with patient vignettes, some accompanied by ML-generated treatment recommendations and various forms of explanations. The findings revealed that, when clinicians were provided with incorrect ML recommendations accompanied by explanations that were limited yet easily interpretable, there was a significant decrease in treatment selection accuracy.*

[31] Poland: Government to scrap controversial unemployment scoring system - AlgorithmWatch. (n.d.). AlgorithmWatch. https://algorithmwatch.org/en/poland-government-to-scrap-controversial-unemployment-scoring-system/ [Accessed in 01/07/2025]

[32] See also assumption "*That human operators are able to do the task*" in this section

[33] Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry, 11(1). https://doi.org/10.1038/s41398-021-01224-x

These results suggest that explanations, especially those that are simplistic and easily digestible, may inadvertently foster overreliance on ML recommendations, even when those recommendations are incorrect.

The deployers of ADM systems must also provide the data subjects with clear explanations of how the decision affecting them was made. These explanations should be sufficient to enable the data subject to understand, review, and challenge the automated decision effectively.

Assumption: Human oversight can help address the system's shortcomings

This assumption is problematic because it treats human oversight as a safety net to justify the deployment of defective or underperforming algorithms. While human oversight can play a critical role in mitigating risks, it is not a substitute for the need to build systems that are accurate, reliable, and ethically sound from the outset.

Relying on human operators to detect or correct errors assumes that system defects will be obvious and that operators will have the time, tools, and authority to intervene effectively. In reality, flawed algorithms often produce subtle, complex issues that may go unnoticed - even by well-trained users. Expecting humans to act as fail-safes places an unrealistic burden on them and can lead to user fatigue, hesitation, or misplaced trust in the system.

Moreover, overreliance on human can discourage proper system validation, delaying necessary improvements or updates. Organisations may deprioritize thorough testing, assuming that any deficiencies can be detected and addressed by human checks. Responsible deployment of automated decision-making systems requires rigorous pre-deployment evaluation and continuous refinement - human oversight should complement, not compensate for, systemic weaknesses.

Example: *IBM Watson for Oncology was designed to help oncologists by providing evidence-based treatment recommendations for cancer patients. The system was trained on vast amounts of medical literature and data to suggest the most appropriate courses of treatment based on patients' medical records. However, the Watson for Oncology's deployment in real-world settings highlighted significant limitations. The system was found to recommend unsafe or incorrect treatments in a number of cases. One notable issue was that the AI suggested chemotherapy treatments that were not suitable for the patients' specific conditions. These errors were particularly concerning in cases involving rare cancers or atypical patient conditions that Watson had not been adequately trained on.*

The system's providers relied on the assumption that oncologists would act to catch any errors or inconsistencies in Watson's recommendations. However, in practice, many doctors did not have the time, tools, or sufficient understanding of the system's underlying processes to confidently challenge Watson's recommendations. In some cases, this led to delays in treatment or, worse, the administration of potentially harmful treatments.[34]

[34] Adlluru, S. K. (2024, November 18). The Rise and Fall of IBM Watson: Lessons from AI's Journey in Healthcare. Medium. https://medium.com/@14asaikiran06/the-rise-and-fall-of-ibm-watson-lessons-from-ais-journey-in-healthcare-8d43bb60cc85 [Accessed in 01/06/2025]

It should also be considered that an ADM system, such as one used to identify individuals for arrest and search, can generate harmful outcomes (for instance, discriminatory effects) **before** real-time human intervention prevents further adverse consequences (e.g. suspects identified by the system who are subsequently released after human assessment).

3.3 Wrong assumptions about how humans behave

Assumption: That human operators know what to do

The belief that human operators supervising ADM systems inherently know how to intervene might be unfounded. In practice, operators often face significant uncertainty about their role and responsibilities, particularly in high-stakes or ambiguous scenarios.

Many organisations deploy automated systems without providing clear protocols or adequate guidance on when human intervention is necessary. This lack of structure can leave operators uncertain about whether they are expected - or even permitted - to challenge a system's output. The problem is further exacerbated when organisations fail to equip operators with the training, or technical competence needed to evaluate system performance and recognise potential errors. Bainbridge underscores the paradox of entrusting operators with a seemingly arbitrary assortment of tasks - specifically those that system designers were unable to automate initially (Bainbridge, 1983).

Without clearly defined responsibilities and sufficient support, human oversight risks becoming symbolic rather than substantive, weakening its intended role as a safeguard against system failure.

> **Example:** *Since 2010, the Basque Country has integrated the EPV-R (Escala de Predicción de Violencia Grave contra la Pareja) risk assessment tool into its judicial system to evaluate the likelihood of severe intimate partner violence. Initially developed to assist police officers in assessing gender-based violence cases, the tool was later incorporated to aid judges in their decision-making processes. However, the implementation of EPV-R revealed significant challenges in its effective monitoring and application. Reportedly, judges and police officers often* ***lacked sufficient training to interpret the tool's outputs effectively, leading to uncertainty about how to incorporate the risk scores into their decision-making processes.***[35] *In some instances, the algorithmic recommendations were followed without critical evaluation, potentially reinforcing existing biases and leading to decisions that might not align with the best interests of the individuals involved.*
>
> The lack of clarity regarding the role of human oversight in these cases underscores the importance of providing adequate training and establishing clear guidelines for operators to ensure that ADM systems are used responsibly and ethically.

[35] Valdivia, A., Hyde-Vaamonde, C. & García Marcos, J. (2025) Judging the algorithm. *AI & Society,* **40**, 2633–2650. https://doi.org/10.1007/s00146-024-02016-9

Assumption: That human operators are able to do the task

The belief that human operators of automated systems are inherently capable of intervening effectively is often flawed. In practice, operators may not be in the appropriate psychological or cognitive state to assume control, especially during high-stakes or unexpected situations.

In her seminal 1983 paper, *Ironies of Automation* (Bainbridge, 1983), Lisanne Bainbridge highlighted a paradox: automation often assigns humans the most challenging tasks - those that are difficult to automate. She observed that:

"A more serious irony is that the automatic control system has been put in because it can do the job better than the operator, but yet the operator is being asked to monitor that it is working effectively.

There is therefore no way in which the human operator can check in real-time that the computer is following its rules correctly. One can therefore only expect the operator to monitor the computer's decisions at some meta level, to decide whether the computer's decisions are 'acceptable'. (…) The human monitor has been given an impossible task."

Building on this, Methnani *et al.* argued that assuming that humans can immediately take over control without disruption in high-risk scenarios can lead to severe miscoordination and potential system failure (Methnani, L., et al., 2021). The authors emphasize the importance of designing systems that account for the possibility that human operators **may not always be ready or able to respond** promptly to requests for input.

MC Elish posited, *"Unfortunately, creating this kind of role for humans, who must jump into an emergency situation at the last minute, is something humans do not do well as was the case in the Air France crash. Human factors research has proven this "handoff" scenario detracts from, rather than enhances, human performance"* (Elish, 2025).

> **Example:** *In March 2018, an Uber self-driving car struck and killed a pedestrian in Tempe, Arizona, marking the first fatality involving an autonomous vehicle.*[36] *The vehicle's safety driver was reportedly distracted during the incident. Investigations revealed that the driver was watching a television show on her phone at the time of the crash, despite being responsible for overseeing the vehicle's operation.*
>
> *The National Transportation Safety Board (NTSB) found that the driver's distraction impaired her ability to take timely corrective action. Additionally, the NTSB noted that Uber's system design did not adequately ensure the safety driver's oversight, relying on a system that was insufficient to detect and mitigate such distractions.*

To ensure effective human oversight of automated systems, it is crucial to design interfaces that present information clearly and accessibly, enabling operators to be promptly alerted, comprehend the system's status and make informed decisions. Operators must be afforded sufficient time to assess the context and evaluate the decisions proposed by the system, particularly in high-stakes or complex scenarios. Human operator responsibilities should be

[36] Smiley, L. (2022, March 8). 'I'm the Operator': The Aftermath of a Self-Driving Tragedy. WIRED. https://www.wired.com/story/uber-self-driving-car-fatal-crash [Accessed in 01/06/2025]

structured to prevent cognitive overload, ensuring that operators can maintain situational awareness without being overwhelmed.

Assumption: That human operators will have 'fitting intentions'

Effective human oversight of automated systems hinges not only on the technical capabilities and situational awareness of operators but also critically on their intentions. As highlighted in the interdisciplinary study *On the Quest for Effectiveness in Human Oversight,*[37] an operator's intentions must align with the responsibilities of their role to mitigate risks effectively.

The concept of 'fitting intentions' is defined as the operator's proactive commitment to address risks while thoughtfully considering the interests of all relevant stakeholders - including the affected individuals both individually and collectively, where applicable. This commitment is essential for ensuring effective human oversight. Conversely, operators with ill intentions can exacerbate risks. Indeed, both the ADM system and the human reviewer may **converge** in producing unfair outcomes - such as excessively high insurance, mortgage, or loan costs for consumers, or wage allocations that prioritize the provider's profits at the expense of workers.

In such cases, human oversight might prove to be ineffective and could be unable to address the inherent and structural flaws embedded within the ADM systems.

Cultivating the right mind-set is essential, and this begins with clear role definition, robust ethical training, and a workplace culture that emphasizes accountability and ethical responsibility.

Equally vital is the meaningful involvement of those affected by AI systems in their design and deployment. This means bringing diverse stakeholders - including workers' representatives,[38] and impacted communities - into decision-making processes ("a seat at the table") in contexts like work-management or hiring systems.[39] Such inclusive co-creation helps ensure that human operators are not only capable but also genuinely motivated and accountable, balancing technical functionality with ethical legitimacy and stakeholder trust.

> **Example:** *Robodebt was an automated debt recovery system implemented between 2016 and 2019, designed to identify and recover overpayments made to welfare recipients. The system matched income data from the Australian Taxation Office with self-reported income data provided to Centrelink, Australia's social security agency. Discrepancies often resulted in debt notices being issued without adequate human oversight.*[40]

[37] (Sterz, S., et al., 2024),

[38] De Stefano, V., & Taes, S. (2022). Algorithmic management and collective bargaining. *Transfer: European Review of Labour and Research*, *29*(1), 21-36. https://doi.org/10.1177/10242589221141055 (Original work published 2023)

[39] On multistakeholders' participation fostering effective and meaningful human review, see Chapter 4, Sub-section "Integrating feedback from affected individuals" of this TechDispatch.

[40] Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2024). Managing unintended consequences of algorithmic decision-making: The case of Robodebt. Journal of Information Technology Teaching Cases, 14(1), 165-171. https://journals.sagepub.com/doi/full/10.1177/20438869231165538 [Accessed in 01/07/2025]

While human operators were formally responsible for reviewing and validating these automated debt notices, investigations revealed that many employees processed them without proper review. This was influenced by internal policies and performance targets that prioritized the volume of debt notices issued over its accuracy. Such an environment discouraged critical evaluation and fostered a culture in which the intentions of human supervisors were misaligned with the ethical responsibility to ensure fairness and legality.

Although human operators had the theoretical power and understanding to override Robodebt's automated outputs, management's directive to prioritize throughput over accuracy meant that their intentions were not aligned with protecting claimants from harm. This misalignment - operators motivated by performance goals rather than risk mitigation - rendered their oversight function ineffective and ultimately led to widespread injustice.

4. Promoting more effective and meaningful human oversight

The discussion thus far shows that implementing human oversight within ADM systems, or partially automated systems such as DSS, is not straightforward. In many cases, human oversight is implemented in a superficial manner, serving more as a symbolic gesture than a functional safeguard. Simply assigning a person to verify a system's output is not sufficient if the conditions under which that verification occurs are not properly established. This section will focus on the organisational and technical measures that could be put in place to ensure that human oversight is not only present, but genuinely effective and accountable.

Building on the work of (Sterz, S., et al., 2024) the primary objective of human oversight should be risk mitigation. In this sense, human oversight is not the ultimate solution for ensuring the safety of partially or fully ADM systems. Rather, human oversight is one of several complementary measures aimed at safeguarding fundamental human rights in the face of these systems' growing prevalence and societal impact.

(Sterz, S., et al., 2024) identify four necessary conditions for human oversight to be effective:

- The system should provide means for operators to intervene and, when necessary, override its decisions.
- The operator should have access to the relevant information needed to understand and evaluate the system's decisions.
- The operator should have agency and, as such, they must be able to override the decision of the system when appropriate (Green, 2022).
- The operator should have fitting intentions in line with their oversight role. For instance, operators should actively strive and mandate for the systems' decisions to be fair, unbiased and to respect fundamental rights.

The aim is not to assign the sole responsibility for meeting the above conditions to individual operators.

Rather, this section emphasizes the organisational and technical measures to be implemented to support effective human oversight. It examines the responsibilities of organisations, providers, and operators to increase the effectiveness of human oversight.

Finally, this section provides a non-exhaustive survey of prevailing approaches and proposals aimed at the systematic integration of the aforementioned human oversight requirements into such systems.

4.1 Organisational measures for effective and meaningful human oversight

Organisations should provide stable, healthy, and fair work conditions for adequate human oversight

A respectful work environment facilitates effective human oversight. Stable contracts, fair wages, and healthy working hours enable human operators to perform their work in humane conditions. Undoubtedly, stable, healthy and fair work conditions are the basis for oversight to be effective and meaningful, as they are the basis for any kind of meaningful work.

> Example: *In the realm of social media content moderation, major technology companies often outsource content moderation tasks to subcontractors that train and review their automated tools for the moderation of harmful and triggering content.*[41] *These subcontractor companies often provide workers with sub-optimal work conditions. After an automated system flags media as potentially harmful, an operator must review the content and assess its harm within seconds (Gillespie, 2020). These content moderators review distressing content, including graphic violence and child abuse material, often without mental health support and counselling. The absence of stable and supportive work environments not only jeopardises the well-being of these workers but also undermines the effectiveness of human moderation of harmful content.*[42]

Organisations should provide appropriate and sufficient time for review

Even when operators have stable working conditions, the environment in which human oversight occurs might still not be optimal. One critical factor is the amount of time available to perform supervision tasks, which can vary significantly depending on the nature of the decision. For instance, an operator may need less time to verify that an image categorised as a person driving by an ADM system actually shows a person driving (a purely classification

[41] UNI Global Union, "Content moderators launch first-ever global alliance, demand safe working conditions and accountability from tech giants," 30 April 2025. [Online]. Available: https://uniglobalunion.org/news/moderation-alliance/ [Accessed in 01/07/2025]

[42] NeuroLaunch.com. (2025, March 25). Content Moderator Mental Health: Addressing the psychological toll of digital sanitation. https://neurolaunch.com/content-moderator-mental-health/ [Accessed in 01/07/2025]

task) than to use their judgement to assess whether an image of a person driving depicts an infraction.

Given that ADM systems are typically deployed to enhance decision speed, it is crucial to recognize that human operators require sufficient time to respond - particularly when intervention is needed to modify or halt a system operating beyond its intended scope or lacking appropriate safeguards.

> **Example:** *A 2017 study* (Cummings, 2017) *explores how time constraints influence human oversight of automated decisions. The paper discusses the influence of automation bias in decision support systems, particularly in aviation domains. It notes that automation bias occurs in decision-making because humans have a tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct, and this bias can be exacerbated in time-critical domains.*
>
> The study emphasizes that in high-pressure environments, such as aviation, the reliance on automated systems can lead to errors if operators do not adequately monitor or question the system's outputs.

Human oversight should be considered an important safeguard by the organisations

Organisations must acknowledge human oversight of ADM systems as an essential safety mechanism. The effectiveness of any safeguard depends not only on its design, but also on the organisation's genuine commitment to its implementation. Acknowledging the serious risks posed by inadequate oversight is, therefore, essential. Operators can only exercise agency on the functioning of the system if they do not fear implicit or explicit consequences by their company or organisation if they do override the system.

Prioritising the protection of fundamental rights over profit-driven motives enables organizations to create conditions where human oversight is truly effective. Compliance with fundamental rights - and the sector-specific laws that uphold them - is not optional but a legal requirement for both providers and deployers of ADM systems. This compliance must be secured from the earliest stages of policy development and system design, well before human review mechanisms are implemented as safeguards. [43]

[43] "given the vast scale of the harm that can arise from the problematic use of algorithmic regulation and the potential irreversibility of the damage, mere reliance on *ex post* remedies is insufficient. This does not mean that *ex post* remedies have no role to play in countering the identified threat. To the contrary, it is equally important to reflect on how they can be strengthened to ensure that not only individual but also systemic review of the executive's action through algorithmic regulation can be carried out. However, *ex ante* protection mechanisms are also needed, for instance in the form of certain requirements that should be fulfilled before algorithmic regulation can be used. In light of the importance of the decisions made during the design and implementation phase of algorithmic systems, oversight also needs to occur at the upstream level, rather than only at the level of the outcomes proposed or adopted by the system. The translation of legal rules and policies from law to code is not a techno-scientific matter, but an exercise that entails normative and political choices." Smuha NA. Algorithmic Rule By Law. In: *Algorithmic Rule By Law: How Algorithmic Regulation in the Public Sector Erodes the Rule of Law*. Cambridge University Press; 2024:i-ii., Chapter IV, online version last accessed on 17 June 2025.

Example: *A recent study by the JRC investigated how effectively human oversight can override biased decisions from decision-support system powered with AI in the areas of human resources and credit lending (Gaudeul, A. et al., 2025). The investigation encompassed field experiments with 1400 professional and semi-structured interviews and workshops. In the field experiments, some professionals were provided with the system's recommendations, while others did not. Furthermore, two AI systems were deployed: one designed to mitigate bias, and another that did not. The investigators found no tendency for operators to follow the fair recommendations over the unfair ones. Additionally, employees tended to prioritise the interests of the company over fairness, whether these interests were implicit or explicit.*

The **reluctance for operators to defy the goals of their organisation**, even when these goals are biased and discriminatory, shows that the organisation's culture or strategy can have a detrimental effect on human oversight.

Ensuring human oversight is not misused as a fall back responsibility in system failures

Human oversight is essential in complex, safety-critical systems. Yet when failures occur, blame is often placed on the individual at the controls, rather than on the broader systemic issues that enabled the incident. This phenomenon is not unique to AI. In capital-intensive industries like aviation, pilots are the most visible decision-makers, and their mistakes are frequently labelled as 'pilot error' - a term that often obscures deeper contributing factors such as inadequate training, flawed procedures, poorly designed human-machine interfaces, or organisational pressures. By focusing solely on the individual, investigators and the public risk overlooking latent hazards in areas like maintenance practices, air traffic control, aircraft design, and corporate culture.

A realistic approach to safety is one that recognizes that errors emerge from the interplay of technical, human, and organisational factors. Assigning blame to a single person not only hinders accountability but also impedes meaningful learning and system-wide improvement.

Likewise, in ADM systems where human operators monitor or intervene in automated processes - whether in industrial control rooms, autonomous vehicles, or algorithmic decision-making-supervision personnel should not serve as convenient scapegoats when errors occur. Instead, such incidents should trigger a thorough root-cause analysis that examines the design of human oversight tools, the appropriateness of alert thresholds, and the performance of communication channels under stress. For example, if a ground-based controller misses a critical alert in an automated safety system, the question should not be "Why didn't the controller stop it?" but rather "Why did the system allow an unmanaged alert burden? Why were the procedures insufficient to guide timely intervention?"

By shifting the focus from individual blame to systemic accountability, organisations can strengthen resilience, foster open reporting cultures, and drive continuous improvement rather than perpetuating a cycle of blame that ultimately compromises safety.

Organisations should provide adequate training and support to the operators

Introducing a machine in a workflow that used to be performed solely by humans can change the skill requirements for human operators. The organisation should, therefore, provide training to prepare workers for this change. The specifics of the training will depend on the system and the task that the system is aiming to solve. In general, the training provided to the operator should at least include:

- The **functioning of the ADM system.** The operator should have sufficient understanding of the capabilities and limitations of the ADM system in place, as well as a general understanding of how the system works.
- A sample of **cases of system failures**, e.g., cases in which the recommendations provided by the ADM system were incorrect. The operators should be instructed on how to respond to - at least the most common - system failures. Research has shown that direct exposure to failures during training can reduce automation bias.[44]
- The **operational procedures (actions)** that the operator is expected to perform within the system, such as approving or rejecting the output of the system. Operators might also need to provide feedback or an explanation if an output is rejected. Alternatively, the operators may be asked to provide a prior decision on the task before the system's automated decision is given, to better detect the presence of automation bias in tasks that are not time-sensitive.
- The **tools** that the operator can use to effectively perform its task. For example, the means that the operator has to regain control of the decision-making process or to correct the behaviour of the system.
- The **objective** of the human task. This is likely to be risk mitigation, but it might also be accuracy improvement, or an increase of the system's trustworthiness. (Sterz, S., et al., 2024) Depending on the context, the operator might be given instructions to be particularly attentive to false positives or false negatives given their impact on individuals. For example, in banking, wrongly flagging a legitimate transaction as fraud (false positive) can lead to customer dissatisfaction and disruption. In contrast, in healthcare, failing to detect breast cancer (false negative) can have serious, potentially life-threatening consequences.
- If the task at hand requires human discretion, i.e., it does not possess an *objective* and a clear ground truth, then the operator must have some **guidance criteria** on how to assess that the decision is good or correct. For instance, whether a music recommendation is good, or whether some language is toxic will vary depending on the individual or the community that is affected by the decision (Lai, V et al., 2023).
- The training should be **tailored** to people that have different levels of expertise, experience and prior knowledge (Crootof, R., et al., 2023).

[44] Misuse of Diagnostic Aids in Process Control: The Effects of Automation Misses on Complacency and Automation Bias https://journals.sagepub.com/doi/epdf/10.1177/154193120805201906

The catastrophic crash of Air France Flight 447 in 2009[45] serves as a stark reminder of the consequences of insufficient training. The pilots' inability to effectively respond to system failures were partly due to inadequate understanding of the automated systems combined with their reliance on them, which provided them with insufficient tools to manually operate the plane when the system failed.

4.2 Technical measures for effective and meaningful human oversight

The system decisions should be explainable

ADM systems deployed in real-world settings are growing increasingly complex, especially as advancements in technology lead to larger-scale models. For example, state-of-the-art large language models have billions of parameters, making their internal workings difficult to interpret, even for experts. These 'black box' systems deliver impressive performance but pose significant challenges for meaningful human oversight due to their opacity.[46]

In many high-stakes or less intuitive scenarios - such as credit scoring or job performance evaluations - human operators often lack a natural baseline to independently assess decisions. In these cases, interpretable-by-design models like rule-based systems or decision trees[47] provide clear advantages. Because their decision-making processes are transparent and understandable, these models naturally facilitate effective human oversight. In light of this, many scholars recommend prioritising interpretable models over complex black-box systems for high-stakes decisions. This ensures greater scrutiny of the decision-making process and, consequently, greater accountability. [48]

However, many modern ADM systems are too complex to be inherently interpretable. These black-box models, including advanced machine learning algorithms, operate with internal logics that are difficult to understand even for experts. This opacity challenges human reviewers' ability to evaluate, question, or contest system outputs effectively.

To bridge this gap, Explainable AI (XAI) techniques such as SHAP, LIME, and counterfactual explanations offer practical solutions. These methods generate user-friendly, local explanations

[45] On May 31, 2009, Air France Flight 447, an Airbus A330 traveling from Rio de Janeiro to Paris, encountered a brief loss of airspeed data due to ice on the pitot tubes (small, forward-facing devices mounted on the exterior of the aircraft, typically on the nose or wing, which are used to measure airspeed). This led to the autopilot disengaging, requiring the pilots to manually control the aircraft. Despite being highly trained professionals operating a modern aircraft, the pilots became overwhelmed and confused by the situation. They failed to recognize and correct a high-altitude stall, ultimately resulting in the aircraft crashing into the Atlantic Ocean and the loss of all 228 people on board. Investigations revealed that the pilots lacked adequate training to handle such scenarios without automated assistance.

[46] There is often an inverse relationship between the performance of AI models and their explainability: as the complexity and predictive accuracy of AI systems increase, their transparency and interpretability tend to diminish.

[47] Decision trees in AI are a type of model that makes decisions by following a tree-like structure of rules, where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a final decision or prediction.

[48] Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Cynthia Rudin, available at: https://arxiv.org/abs/1811.10154

that clarify why specific decisions were made (e.g., why a loan application was denied) or global explanations describing how the system generally behaves (e.g., how income and credit history are weighted). While these tools improve transparency and support human oversight, they have inherent limitations and cannot fully replace the benefits of interpretable models.[49]

Research by (Vasconcelos, H. et al., 2023) supports the value of explainability, finding that AI explanations are most appreciated when tasks are difficult, explanations are simple, and there are incentives like monetary value. Their study suggests that well-designed explanations can reduce automation bias by promoting more critical engagement and meaningful human oversight of AI recommendations.

However, it is important to note that despite their benefits, explainability techniques can sometimes increase overreliance on the system, as highlighted in previous research (Bansal, G. et al., 2021). [50] This paradox underscores the need for balanced approaches that enhance understanding without fostering blind trust.

The system interface should have a clear, intuitive, and accessible design

Careful consideration should be given to the design of the system interface through which operators interact with the ADM system. These design choices are likely to influence the operator's review of the system. Even seemingly minor elements - such as the colours in the display of an ADM decision - can prone the operator towards accepting a decision hastily.

To design the interface for effective human oversight, it must be considered that humans and machines work differently. Interfaces must be designed to account for these differences and stimulate the critical thinking of the operator. One key consideration is **the difference in reaction time** between humans and machines. The interface should be designed in such a way that the human receives control of the system with sufficient time to react and for the operator to be critically engaged with the task.

Another consideration for building the interface is the cognitive load[51] of the operator. The interface should avoid providing excessive information to the operator and it is advisable to use **simple, clear and intuitive** language and visualisations. Note that the language used must not be over-simplified either. Therefore, the clarity of language must be weighted with sufficiently detailed explanations on the functionality of the ADM system.[52]

[49] See the Tech Dispatch on Explainable AI from 2023 for more details on the limitations of XAI techniques.

[50] See the Tech Dispatch on Explainable AI from 2023 for an in-depth investigation of the advantages and limitations of the topic. The document can be found in the EDPS webpage in: https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en

[51] In the context of human oversight of ADM, cognitive load refers to the mental effort required by a human operator to understand, monitor, and make decisions about an AI system's outputs. It's a key factor in determining whether human oversight is effective or just a token review, non-meaningful.

[52] Wannes Ooms, Lotte Cools, Thomas Gils and Frederic Heymans (Knowledge Centre Data & Society), "From Policy To Practice: Prototyping The EU AI Act's Human Oversight Requirements", March 2025. https://data-en-maatschappij.ai/uploads/Publications-items/Policy-Prototyping-Human-Oversight-KCDS_2025-03-27-085643_hapv_2025-04-04-145830_xwgv.pdf [Accessed in 02/07/2025].

4.3 Operator's profile for effective and meaningful human oversight

The operator should have fitting intentions

The operator should have fitting intentions that align with the fundamental purpose of human oversight. This means striving, to the best of their ability, to ensure that the ADM system's predictions are fair, non-discriminatory, and do not cause undue harm.

It is important to recognize that an operator's intentions are not solely rooted in personal values but are often shaped by broader organizational influences - such as pressure to meet time constraints, cost limitations, or profit targets. These external pressures and directives can undermine the quality and integrity of human oversight.

Human oversight is not merely a technical function - it also involves ethical judgment, especially in contexts where automated decisions significantly affect individuals' rights and well-being.

The operator should have the necessary expertise

In some cases, ADM systems are deployed to make decisions that are difficult or impossible to understand and challenge without the appropriate expert knowledge (Sterz, S., et al., 2024). For example, in situations where the operator monitors a system designed to detect pathologies in Magnetic Resonance Images (MRIs) or a system designed to provide a credit score to a person applying for a mortgage. In cases like these, effective human oversight might only be possible if the human operator possesses the necessary level of expertise in the relevant field (e.g. radiology or finance).

4.4 Practical approaches for effective human oversight

The design of human oversight processes must be driven by empirical evidence. This section introduces a number of proposals for implementing effective human oversight in practice.[53]

First, this section will cover the importance of **auditing** and **sampling** as mechanisms for verifying that effective human oversight is in place.

It will then discuss the **four-eyes principle** and **awareness checkpoints**, which are additional layers of trust in the human oversight mechanism that stimulate the operator's critical awareness.

[53] The complexity of implementing effective human oversight in practice is a matter of especial attention and on-going investigation. This section presents a list of current approaches that will undoubtedly increase in the future from theoretical research to empirical investigations.

Institutionalised distrust, a holistic approach to embed distrust in the endeavour of human oversight, will also be covered as a reaction to the unrealistic expectations it fosters (see chapter 3). Lastly, this section will explore ways **to incorporate feedback from affected individuals**, ensuring their perspectives inform both system improvement and the trust placed in these systems.

The implementation of human oversight may involve some, all, or a combination of the approaches discussed throughout this section. It is essential to recognise that there is no one-size-fits-all model for human oversight. Instead, human oversight mechanisms should be adapted to the specific context, task, and risks associated with the use of the ADM system.

Auditing

For a human oversight role to be effective, it must be subject to regular external auditing. These audits should assess not only the technical performance of the system, but also the human factors involved in its operation. Key areas to evaluate include (but are not limited to):

- **Operator dependence on automated decisions**: Auditing should examine the extent to which operators rely uncritically on the system's outputs. For example, an external audit of the *Viogén* system - a Spanish algorithm designed to assess the risk of domestic violence - found that police officers followed the system's recommendations in 95% of cases.[54] While high concordance may reflect trust in the tool, it also raises questions about the **genuineness of autonomous human judgment** in the oversight process.
- **Timing and presentation of system outputs**: Auditors should also assess whether it is more effective for the operator to make an initial judgment autonomously, with the system's evaluation/assessment revealed only upon request or after the operator's decision is recorded. For example, in some ADM workflows, such as AI-assisted radiology, clinicians are encouraged to make an initial judgment **before receiving the system's recommendation**. This delayed disclosure can help reduce automation bias and promote deeper cognitive monitoring with the task (Gaube, S. et al., 2021).
- **System usability and cognitive ergonomics**: Audits should also evaluate whether interfaces support effective oversight. This includes the clarity of system feedback, alert fatigue risk, and whether operators have sufficient time and information to intervene meaningfully.
- **Training and organisational culture**: Monitoring quality depends on well-trained operators who understand the system's strengths and limitations. Audits should examine whether training programs support critical thinking and whether the organisational culture encourages challenge, rather than blind compliance with automated outputs.

Regular auditing of these elements is essential to ensure that human oversight remains critical, autonomous, and genuinely capable of correcting or overriding system errors when necessary.

[54] Eticas Foundation (2022), Can AI solve gender violence? Auditing the use of AI to assess risk. The Case of Viogén. https://eticasfoundation.org/wp-content/uploads/2024/12/Eticas_Audit_of_VioGen.pdf [Accessed in 01/07/2025]

Sampling

In high-volume decision-making environments - such as welfare distribution, law enforcement triage, or healthcare triage - it is often **impractical or resource-intensive** to subject *every* automated decision to real-time human oversight. In such cases, a **sampling approach** may be employed, whereby human operators review a representative subset of the system's outputs to monitor performance, detect errors, and assess fairness.

While this strategy offers scalability and operational feasibility, it also presents **significant limitations**. First, sampling is inherently retrospective: decisions are reviewed **after** they have been implemented, which means harmful outcomes may have already occurred by the time they are identified - if they are identified at all. This delay undermines the capacity of human oversight to function as a timely corrective mechanism, particularly in high-risk contexts where early intervention is critical.

Second, sampling raises **equity and accountability concerns**. Individuals whose cases are not selected for review may be subject to unnoticed errors or biases, especially if the system underperforms for specific subpopulations. This creates the risk of unfair or discriminatory treatment without any immediate pathway for redress. Moreover, errors affecting non-sampled cases may not be captured in the oversight process, obscuring broader systemic failures and delaying necessary interventions at design level.

To mitigate these issues, sampling must be strategically designed to ensure:

- **Diverse and representative** sampling criteria to detect differential impacts across groups.
- **Dynamic sampling rates** that increase scrutiny in high-risk or low-confidence scenarios.
- **Feedback loops** that link audit findings to system updates and operator training.
- **Transparency around sampling methods** to maintain public trust.

In summary, while sampling can be a pragmatic tool in large-scale ADM oversight, its use must be carefully structured to avoid turning human review into a superficial safeguard. Effective sampling requires thoughtful design, regular re-evaluation, and strong governance to ensure it contributes meaningfully to both accountability and system improvement.

Enhancing human oversight effectiveness: redundancy and oversight mechanisms

Operators responsible for monitoring ADM systems might be tasked with oversight processes that rarely exhibit obvious failures. While this may sound reassuring, the rarity of errors can actually lead to a decline in vigilance over time - a well-documented phenomenon in supervisory control settings. To counteract this and maintain high levels of attentiveness and decision quality, it is crucial to design workflows that actively support and challenge human oversight.

One such approach is the implementation of the '**four-eyes principle**', where critical decisions made by one operator are reviewed or validated by a second individual. This redundancy, as discussed by (Schikora, 2010), serves multiple purposes: it helps detect occasional mistakes, reduces the impact of individual cognitive biases, and bolsters accountability. In safety-critical or ethically sensitive domains - such as aviation, healthcare, or weapon systems - this second layer of verification can significantly enhance the reliability and transparency of the oversight.

The four-eyes principle can be implemented in two primary ways: **sequentially** or in **parallel**. In a sequential implementation, one operator makes the initial decision, which is then reviewed and either confirmed or challenged by a second operator. This approach promotes accountability, allows for reflective analysis, and is particularly well-suited to situations where time permits deliberate review - such as in regulatory or high-stakes administrative settings. In contrast, a parallel implementation involves two operators making autonomous assessments simultaneously, without knowledge of each other's decisions. This method minimizes potential bias introduced by the first judgment, encourages independent thinking, and can lead to higher reliability in fast-paced, safety-critical environments like air traffic control or medical diagnostics.

In addition to decision redundancy, **awareness checkpoints** can be embedded into the operator's workflow. These are structured prompts or brief tasks designed to periodically assess and recalibrate an operator's situational awareness. Examples include quick scenario-based queries, simulated anomalies, or interactive diagnostics that require the operator to re-engage with system state information.

Such checkpoints not only refresh the operator's cognitive ability to monitor the system but also provide valuable data on operator readiness and attention over time.

> **Example:** *Technologies like Threat Image Projection (TIP)*[55] *can play a vital role in keeping human operators actively engaged in the human oversight of ADM systems by transforming routine oversight tasks into dynamic, feedback-driven learning experiences. TIP maintains operator vigilance and attention by introducing realistic but fictional threat items at unpredictable intervals during X-ray screening, prompting human action in real time. This unpredictability prevents complacency and encourages constant attentiveness. Moreover, TIP provides instant, personalized feedback - highlighting both successful identifications and missed threats - which reinforces learning and promotes continuous improvement.*
>
> By integrating performance analytics and tailored training insights, TIP ensures that human oversight remains not only functional, but cognitively meaningful.

Together, these mechanisms - peer validation and cognitive refresh - transform human oversight from a passive supervisory role into a more active, interactive, and resilient process. They help ensure that human operators remain both effective and meaningfully engaged, even in environments characterized by rare but high-consequence failures.

Institutionalised distrust: holistic proposal

Drawing on democratic theory, the concept of *institutionalised distrust* emphasizes that trust in governance is not achieved by assuming good intentions or infallibility, but by designing systems that anticipate failure, bias, and self-interest. As Braithwaite argues, embedding **mechanisms of distrust** within institutions can ultimately cultivate **well-placed public trust** (Braithwaite, 1998). Similarly, Sztompka describes *"the institutionalisation of distrust in the*

[55] Cutler, V., & Paddock, S. (2009, October). Use of threat image projection (TIP) to enhance security performance. In 43rd Annual 2009 International Carnahan Conference on Security Technology (pp. 46-51). IEEE. See also the "Threat Image Projection (TIP) Guidance", QinetiQ and CPNI – Centre for Protection of National Infrastructure, 2016, https://www.npsa.gov.uk/system/files/documents/cc/6c/Threat-image-projection-TIP-guidance.pdf [Accessed on 01/07/2025]

architecture of democracy" as a cornerstone of legitimate governance, where transparency, accountability, and contestability enable trust to be earned rather than assumed (Sztompka, 2000).

When applied to ADM governance, this principle implies that distrust in the abilities and motivations of human operators should not be dismissed, but rather leveraged as a foundation for institutional design. Monitoring systems should be built with structural safeguards that assume the possibility of human error or bias, thereby encouraging more resilient and trustworthy governance.

Laux (Laux, 2023) proposes an approach with six principles based on Sztompka's institutionalised distrust concept:

1. Justification. Organisations should be required to produce detailed reports justifying the deployment of ADM systems, including their necessity, goals, and expected outcomes.

2. Periodic mandates. Supervision roles should be subject to term limits or rotation (e.g., for auditors or review panels) to enhance impartiality and reduce the risk of alignment with provider or institutional interests.

3. Collective decision-making. Involving multiple individuals or bodies in supervision of decisions can reduce individual biases and disincentives. While it does not guarantee higher competence, it supports a more balanced and transparent evaluation process.

4. Limited institutional competence. Organisations should have narrowly defined human oversight mandates to prevent overreach or monopolisation of decision-making power.

5. Justiciability and accountability. Human oversight mechanisms must be subject to legal accountability. In liberal democracies, individuals must be able to challenge both public and private institutions in court to defend their rights, ensuring recourse in cases of algorithmic harm or injustice.

6. Transparency. Institutions must ensure meaningful algorithmic transparency, including:

    - Clarifying whether human oversight is genuine or merely symbolic;
    - Disclosing whether human oversight relies on AI and, if so, the data and methods used;
    - Sharing the results of empirical testing on human oversight performance (as per the principle of justification);
    - Describing the design and operational details of human oversight practices.

For institutionalised distrust to have a chance at being successful, it is particularly important that organisations consider human oversight as a critical, important and nuanced safety task to tackle discrimination and bias. This is due to a study indicating that institutionalised distrust might not be effective if the norms of the organisation, i.e., the implicit expectations guiding decisions within an organisation, are prone to biases.

Lastly, it is important to consider the limitations highlighted by Laux himself: *"Of course, institutionalising distrust in human oversight would not guarantee risk-free AI systems. Instead, it would provide a general scaffolding along which trustworthy local implementations of human oversight may be built."*

Integrating feedback from affected individuals

Human oversight of ADM systems must go beyond periodic audits or technical performance evaluations. It should include **structured and continuous mechanisms** for capturing feedback directly from the individuals most affected by the system's decisions. Whether the context involves a credit-scoring algorithm, an automated hiring system, or a medical triage system, affected individuals offer unique, ground-level insights into how these systems perform in real-world conditions - insights that are often invisible to developers or auditors.

To operationalize this, organisations should embed accessible and user-centred feedback channels within the systems themselves. These could include:

- **In-platform feedback** forms or surveys;
- **Clear appeal** mechanisms;
- Integrated 'voice of the user' tools;
- Anonymous reporting options for perceived errors or biases.

These tools enable the collection of both **quantitative and qualitative data** - not only about false positives and false negatives, but also about user experience, perceived fairness, and trust. This feedback becomes an essential input for human operators, helping to:

- Identify systemic blind spots;
- Detect patterns of bias or harm across demographic groups;
- Prioritize areas for model recalibration or workflow redesign.

This feedback-informed oversight loop allows organisations to shift from a reactive stance (correcting mistakes after harm is done) to a proactive approach focused on continuous improvement.

Moreover, closing the loop with affected individuals - by transparently communicating how their input led to system changes - is critical for building and maintaining public trust. It signals that human oversight is not a symbolic gesture, but a living process grounded in accountability and responsiveness.

In sum, human oversight grounded in stakeholder feedback—particularly from those directly impacted by the ADM system, such as loan applicants - serves to bridge the gap between technical optimization and social legitimacy. It ensures that ADM systems evolve not only based on performance metrics, but also in alignment with the **values, needs, and rights** of the people they impact.

5. Conclusion

This TechDispatch explains what ADM is and critically examines common assumptions underlying human oversight of ADM systems, revealing why many such beliefs are unsupported and in practice do not improve the quality of the decisions taken by the system. These misconceptions risk fostering a false sense of security, potentially allowing flawed algorithms to operate unchecked under the veneer of human control.

Drawing on these insights, the report outlined a comprehensive set of organizational, technical, and operator-focused measures to make human oversight meaningful and effective. Organisational commitment from both providers and deployers is essential to prioritize human oversight at all levels. Thoughtful system design should embed capabilities that enhance human understanding and control, while operator training must equip individuals with the skills and situational awareness needed to intervene appropriately.

It is vital to emphasise that inherent flaws in ADM systems - whether due to scientific unreliability or conflicts with fundamental rights - cannot be fully addressed by human oversight alone. Relying solely on human operators as a final safeguard risks legitimizing defective and controversial systems that may cause harm without tackling root causes. To prevent this, human oversight should be complemented by institutional review processes, such as public evaluation and approval before deployment. [56] Systems that violate laws or fundamental rights must be prohibited outright rather than remedied through after-the-fact human intervention. For example, automated scoring is already a form of decision-making subject to data protection and other legal frameworks.

Moreover, the effectiveness of human oversight varies significantly depending on the decision-making context. Monitoring purely technical systems differs fundamentally from monitoring socio-technical systems - such as welfare allocation, workforce management, credit or insurance pricing, or law enforcement applications. Socio-technical systems must be designed to uphold fundamental rights by design, rather than depending on human reviewers to correct violations after deployment. Systemic biases - such as discrimination based on race or socio-economic status - often reflect inherent flaws in the ADM itself. Problematic systems, like emotion recognition in recruitment, should therefore be rejected outright, rather than relying on human oversight to mitigate risks.

To ensure that human oversight fulfils its intended purpose, there is an urgent need to develop and adopt clear, standardised frameworks and metrics for assessing human oversight performance. Such standards will enable providers, deployers, and regulators to systematically evaluate whether human involvement effectively detects, mitigates, and corrects errors or harmful outcomes. Without this rigor, human oversight risks becoming a mere procedural formality rather than a genuine safeguard.

By embracing the organisational, technical, and operational measures outlined here - and committing to robust standards for human oversight - stakeholders can better ensure that automation enhances, rather than undermines, human dignity, ethical responsibility, and societal trust.

[56] Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 45, 105681.

6. Bibliography

AI, H. H.-l. (2019). Ethics guidelines for trustworthy AI.

Anderson, J. &. (2023). The future of human agency. *Pew Research Center.*

Bainbridge, L. (1983). Ironies of automation. *Analysis, design and evaluation of man–machine systems*, 129-135.

Bansal, G. et al. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *In Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-16).

Biewer, S. et al. (2024). Software doping analysis for human oversight. *Formal Methods in System Design*, 1-50.

Braithwaite, J. (1998). Institutionalizing distrust, enculturating trust. *Trust and governance*, 343, 356.

Crootof, R., et al. (2023). Humans in the Loop. *Vand. L. Rev.*, 76.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. *In Decision making in aviation* (pp. 289-294).

Elish, M. C. (2025). Moral crumple zones: cautionary tales in human–robot interaction. *Robot Law: Volume II*, 83-105.

Gaube, S. et al. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1), 31.

Gaudeul, A. et al. (2025). The Impact of Human-AI Interaction on Discrimination.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 2053951720943234.

Gilpin, L. H. (2018). Explaining explanations: An overview of interpretability of machine learning. *In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681.

HLEG, A. (2019). A definition of AI: Main capabilities and disciplines. *High-Level Expert Group on Artificial Intelligence*.

Jacobs, M. P.-V. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*, 11.

Jones, M. L. (2015). The ironies of automation law: Tying policy knots with fair automation practices principles. *Vand. J. Ent. & Tech. L.*, 18, 77.

Lai, V et al. (2023). Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. *In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1369-1385).

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 & Society*, 39(6), 2853–2866.

Methnani, L., et al. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence*, 4.

Miller, T. (2023). Explainable AI is dead, long live explainable ai! hypothesis-driven decision support using evaluative AI. *In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 333-342).

Scantamburlo, T. et al. (2023). Moral Exercises for Human Oversight of Algorithmic Decision-Making. *In CEUR WORKSHOP PROCEEDINGS* (Vol. 3537, pp. 57-66). CEUR-WS.

Schikora, J. (2010). Bringing the four-eyes-principle to the lab.

Skitka, L. J. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991-1006.

Sterz, S., et al. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. *arXiv (Cornell University)*.

Sztompka, P. (2000). Trust, distrust and the paradox of democracy. *Polish Pol. Sci. YB*, 29, 5.

Taddeo, M. &. (2021). How AI can be a force for good–an ethical framework to harness the potential of AI while keeping humans in control. *Ethics, governance, and policies in artificial intelligence*, 91-96.

UNI Global Union. (2025, April 30). *Content moderators launch first-ever global alliance, demand safe working conditions and accountability from tech giants*. Retrieved May 5, 2025, from https://uniglobalunion.org/news/moderation-alliance/

Vasconcelos, H. et al. (2023). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1-38.

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet*, 11(1), 104-122.

This publication is a brief report produced by the Technology and Privacy Unit of the European Data Protection Supervisor (EDPS).

It aims to provide a factual description of an emerging technology and discuss its possible impacts on privacy and the protection of personal data. The contents of this publication do not imply a policy position of the EDPS.

Issue Authors: Vítor Bernardo, Laura Hernández

Editors: Luis Velasco, and Xabier Lareo.

Contact: techmonitoring@edps.europa.eu

To subscribe or unsubscribe to TechDispatch publications, please send a mail to techmonitoring@edps.europa.eu.

The data protection notice is online on the EDPS website.


EDPS
edps.europa.eu