EDPS

EUROPEAN DATA PROTECTION SUPERVISOR

The EU's independent data protection authority

Checklist on human intervention on automated decision-making (ADM)

Purpose and scope

The processing of personal data through systems that automate decision-making (ADM), with limited or no human intervention, can significantly impact individuals' fundamental rights and freedoms. This is particularly the case where these decisions have legal effects or can significantly affect individuals, including when the presence of bias or system errors could result in discrimination, unequal treatment, or other adverse outcomes. Accordingly, such systems should be deployed with appropriate governance measures, including robust procedures and control mechanisms, to ensure their lawful, fair, and accountable use.

One important control is **human oversight**. Human oversight refers to the active involvement of humans in meaningfully supervising the operation of ADM systems, assessing the system's outputs and, where necessary, intervening - by overriding, suspending, or disregarding an automated output - particularly in cases where the system produces an erroneous, biased, or inappropriate result.

In this document, "human oversight" refers to a set of measures aimed at minimising potential negative effects on individuals during and after the operation of ADM systems.

Human oversight is often proposed as a key measure for mitigating risks associated with the deployment of ADM systems. However, while it remains an important safeguard, it should not be regarded as a 'silver bullet' capable of fully eliminating the risks inherent in such systems, including those stemming from the biases that influence human interaction with ADM systems as well as the circumstances under which the oversight system operates.

In fact, the effectiveness of human oversight depends significantly on how it is designed and implemented in practice.[1]

Oversight mechanisms should be tailored to the specific context in which an ADM system is used and be proportionate to the level of risk it poses to individuals and society. Processing that is likely to result in a high risk to the rights and freedoms of natural persons will require more formalised, robust, and fail-safe oversight processes, whereas lower-risk processing activities uses may justify lighter approaches. In all cases, oversight should be clearly defined, properly documented, and supported by appropriate organisational practices, training, and technical safeguards.

---

[1] The **EDPS TechDispatch on Human Oversight of Automated Decision-Making** (2025) presents an analysis of common misconceptions and provides some measures on how to improve it. The attached checklist uses the TechDispatch as a starting point.

This document provides a checklist that EUIs can use as a self-assessment tool to **measure the maturity** of the human oversight measures applied to their ADM systems.

Disclaimer

This checklist is intended to support reflection, gap identification, and improvement of oversight processes. While the checklist can support controllers' accountability, it is not designed to be used as a formal compliance assessment tool. Its objective is to list the features that human oversight mechanisms could have to be as effective as possible.

Although both the EUDPR and the AI Act provide for human intervention in automated decision-making, they do so in different contexts. This checklist does not seek to determine where applicable legislation mandates human oversight. Rather, it offers structured guidance on critical factors to consider when implementing human oversight measures.

It is important to note that the requirements of the EUDPR apply to any system that processes personal data, regardless of whether it qualifies as an ADM.

Lastly, it should be borne in mind that human oversight is not a substitute for sound system design.

Oversight mechanisms play a critical role in mitigating risks but should not serve as a substitute for addressing fundamentally flawed, unsafe, or unreliable automated systems. If an ADM system requires frequent human intervention to function correctly or to avoid harmful outcomes, this indicates deficiencies in its design and suitability for deployment, and such a system should not be placed into production until those risks have been properly mitigated.

Self-assessment checklist

Section 1

The controls in this section apply broadly to any system whose decisions, or recommendations, may impact individuals.

Governance: Documented human oversight and accountability

- [ ] There is a clearly defined and documented governance framework for human oversight of the ADM system, including roles, responsibilities, and escalation procedures

User training: Users are prepared to interact with decision-support systems

- [ ] Users receive a mandatory on boarding training on the system capabilities/ limits (including on possible failure scenarios)
- [ ] User training includes scenario-based exercises that incorporate known failure cases and lessons learned
- [ ] User training includes components on ethical decision-making
- [ ] User training addresses cognitive and automation bias to prevent "excessive trust on the system"

Accountability: Failures trigger systemic reviews, not scapegoating

- [ ] There are mechanisms to ensure accountability for decisions influenced by ADM (e.g., audit trails, decision logs), particularly where automation bias may occur
- [ ] Lessons learned are documented, reinforced through targeted technical and organisational measures, and incorporated into the users' training
- [ ] Implementation of root-cause analysis protocol (RCA)[2] after failures (i.e. "what went wrong, and why?")
- [ ] When the ADM system is updated, the human oversight protocol is also reevaluated (a change in the system's logic can render previous human training obsolete)

[2] A systematic, step-by-step process for identifying the fundamental, underlying causes of a problem or failure, rather than just the symptoms

Engagement: Oversight is treated as a safeguard, not a formality

- [ ] There are KPIs defined for oversight (e.g., error detection rate, override frequency)
- [ ] Oversight impact in the quality of the decision-making process is measured and incorporated into organizations' KPIs
- [ ] Metrics are periodically reviewed and reported to management
- [ ] There is a training/refresher on ADM oversight that is provided to top management

Explainability: Decisions are interpretable enough for users to evaluate

- [ ] System users are provided with plain-language summaries of the ADM logic, or the reasoning for each decision under review
- [ ] There is clear evidence demonstrating that all decisions are traceable to the underlying grounds and justification used
- [ ] ADM system outputs are displayed with confidence scores [3]

4 eyes principle: High-stakes decisions are checked by more than one user

- [ ] More than one user oversees independently decisions with legal effects or having a high impact on the rights and freedom of individuals (e.g., disciplinary proceedings, withdrawal of funds)
- [ ] A formalised escalation mechanism is in place to resolve disagreements between human overseers, or between a human and a high-confidence ADM system, including a documented tie-breaker process and, where necessary, escalation to senior management
- [ ] Agreement/disagreement rates between system operators are documented

Feedback loops: Affected individuals can provide feedback, which is then integrated back into the system

- [ ] Accessible and user-friendly appeal mechanisms are available to individuals affected
- [ ] User feedback is integrated into AI model re-training, or system updates

[3] A confidence score is a statistical measure that quantifies the certainty or reliability of a prediction, or decision made by AI models. It is often expressed as a percentage, indicating how confident the model is in its output.

Expertise: Users are qualified in the relevant domain

- [ ] There is an updated competency matrix for each user function dealing with critical systems (i.e. which features are sought when assessing candidates)

Critical awareness: Training and protocols prevent blind trust in automation

- [ ] Periodically test operator decision-making without ADM support
- [ ] Operators are rotated across tasks to avoid habituation

Section 2

The controls in this section apply when there are operators responsible for supervising ADM systems who have the authority to intervene in their action (e.g. suspend system or disregard decisions).

Authorisation: Operators have the proper authority to carry out their duties

- [ ] The supervisory tasks assigned to operators have been clearly defined, communicated, and are documented
- [ ] The possibility of overriding the system's decisions is explicitly stated in the task description
- [ ] There is a formal mechanism that ensures operators can exercise "stop-the-line" authority to pause or halt the entire system in response to suspected systemic issues, without risk of reprisal or adverse consequences

Time for review: Oversight tasks are assigned with realistic timeframes

- [ ] Operators are assigned a maximum workload quota for each shift
- [ ] When dealing with oversight tasks, operators are not involved in other tasks
- [ ] There is a formalised requirement that mandates "cool-down" periods for operators responsible for high-stakes or emotionally demanding decisions, ensuring sustained judgment quality and operator well-being

Interface design: Clear, intuitive, avoids overload and allows quick intervention

- [ ] Usability and accessibility testing conducted with representative operators (unexpected reaction from operators' remarks are duly noted)
- [ ] The average reaction time from operators is measured
- [ ] Override buttons/commands on interface are easily accessible
- [ ] User interface is designed to mask sensitive data that is not strictly necessary for decision-making review, thereby minimising data protection risks

Auditing: Regular internal & external audits include technical + human factors

- [ ] All overrides of system decisions are recorded and subject to analysis Independent third-party audit to ADM (including operator roles)
- [ ] There are periodic internal audits of decision logs

Vigilance aids: Simulated anomalies/checkpoints maintain operator attention

- [ ] "Red team" anomaly injections are introduced regularly whenever testing can be performed without adversely affecting operator performance or impacting individuals. Operator response accuracy and speed are documented and measured for each exercise

Sampling: Selection of a subset of decision logs for auditing

- [ ] Logs are sampled and analysed, with a defined sample rate (e.g., 10% of the total)
- [ ] Samples contain different types of ADM outcomes (e.g. positive and negative)