Accountability

2026-05-06 · cheerful mango Haubentaucher

The Responsibility Gap

A visiting traveller witnesses a penal machine execute a condemned man for a sentence no living person can read or explain, and is invited to endorse the system.

Algorithmic governance and black-box AI in high-stakes domains highlight opacity and accountability gaps: audits, explainability requirements, and governance mechanisms all attempt to close the distance between automated decision and human answerability.

Required reading	*In der Strafkolonie*[1] by Franz Kafka (1919; Reclam UB 9900)
Preparation	Complete the Guided Reading Sheet individually before the session.

In der Strafkolonie
Franz Kafka (1919)
Full text via Wikisource:

Minority Report
Steven Spielberg (2002)
A pre-crime unit arrests people before they offend, based on algorithmic prediction. When the system targets its own chief, the question of who holds the algorithm accountable becomes impossible to avoid.

Case.

1. 2014 to 2018: Amazon develops ML-based recruiting tool to automate hiring for technical and managerial roles.
2. System trained on historical hiring data; penalised resumes containing the word "women's" or female-coded language.
3. Tool systematically downranked female candidates; hired only male managers in cohorts where bias was strongest.
4. October 2018: tool scrapped after internal audit. No individual held accountable. The algorithm had no name.

Guided Reading Sheet

Read *In der Strafkolonie* before the session. Work through the questions below individually, then bring your notes to the discussion.

1. The condemned man never learns what he is accused of or what the machine will inscribe on his body. What does this tell us about the relationship between legibility and legitimacy? Can a system be just if it cannot explain itself to those it acts upon?
2. Three figures surround the machine: the officer who understands and maintains it, the condemned man who cannot, and the visiting traveller who is invited to endorse it. How does responsibility

shift across this triangle? Who, if anyone, is morally culpable for what happens?

3. The story illustrates the idea of the *responsibility gap*: when a complex system causes harm, no single human actor is clearly at fault. Can you name a contemporary case (in law, medicine, finance, or infrastructure) where this gap has appeared?

4. The EU AI Act (2024) requires human oversight for high-risk AI systems. What does "oversight" mean in practice if the system's decisions are too fast, too complex, or too numerous for humans to review individually?

5. Compare *causal responsibility* (who caused the harm) with *moral responsibility* (who is blameworthy) and *legal liability* (who must compensate). Do these three always converge? Give an example where they do not.

Opinion Landscape and Debate

Format: Surrounded-style debate. See Appendix A for the full protocol, speaker's list rules, and moderator scripts.

Opening question: If a system causes injustice at this scale, and every individual who used it acted within their authorised role, is anyone guilty of wrongdoing; and if not, what does that mean for the concept of accountability? [1]

Opinion A, Accountability. Named human accountability is achievable and required: when an autonomous system causes harm, the architecture of governance must ensure that specific individuals remain legally and morally answerable for its outputs.

The Defence Lawyer argues the tool violated the defendant's right to a fair trial and demands a named responsible party.

- *My client cannot cross-examine a dataset. She cannot see the features the model weighted. She cannot confront the humans whose historical decisions trained it. That is not a fair trial; it is a verdict delivered by an oracle.*

- *The right to a reasoned decision is foundational to due process. The algorithm said so is not a reason. It is an abdication.*

- *If a biased human expert testified in court, we could challenge their credentials, their methods, their prior record. We have no equivalent procedure for an algorithm. That asymmetry is a constitutional problem.*

The Regulator proposes a new accountability framework and must defend it against both sides.

- *The existing categories (product liability, professional negligence, criminal intent) were designed for individual actors and single causal chains. None of them fit distributed, automated decision systems. We need a new category: algorithmic liability.*
- *Auditability must be a pre-condition for deployment in high-stakes domains, not a post-incident investigation. If you cannot explain why your system made a decision before harm occurs, you should not be permitted to deploy it.*
- *The question is not who is liable for this specific incident. It is what governance regime prevents the next thousand incidents at scale.*

RESPONSIBILITY GAP. The structural absence of a single culpable actor in distributed sociotechnical systems. Closing it requires deliberate governance design, not post-hoc blame.

OPINION B, COMPLIANCE. The existing system already provides accountable humans; the motion is either already met or unworkable: demanding named individual accountability for each algorithmic output ignores how distributed decision-making necessarily operates.

The Judge defends reliance on the tool as consistent, evidence-based sentencing practice.

- *Documented bias in human sentencing (based on time of day, defendant ethnicity, and courtroom demeanour) is well-established in the research literature. The tool reduces that variance. Less variance is more equal treatment.*
- *I did not delegate my judgement to the tool. I consulted it as one input among many, as I would a psychiatric assessment or a victim impact statement. The tool is not the verdict.*
- *If we ban algorithmic tools from courtrooms, we return to a system where the most important variable in a sentencing outcome is which judge you happened to draw. Is that preferable?*

The Software Vendor insists the tool performs within documented parameters and liability rests with the deployer.

- *Our documentation states in the executive summary that human review is mandatory before any recommendation is acted upon. The court chose to rely on the output without applying that review. We cannot be held liable for misuse.*

EXPLAINABILITY. A system is explainable when it can account for its outputs in terms a human decision-maker can evaluate. A necessary but not sufficient condition for accountability.

- *The training data was provided by the court system itself. If the historical sentencing record reflects bias, that is a problem with a model that learned from it.*
- *Product liability law requires that a product fails to perform as documented. Ours did not. Extending vendor liability to downstream misuse would make the development of any decision-support tool legally untenable.*

KEY LEARNINGS The responsibility gap is not caused by negligence but by the architecture of distributed sociotechnical systems: standard categories of fault (criminal intent, professional negligence, product liability) apply only awkwardly when causation is diffuse. Explainability is a necessary but not sufficient condition for accountability; a system can be interpretable and still have no human willing to own its outputs. Meaningful human control is contested: it may mean the ability to intervene, to understand, to reverse, or merely to authorise, and these are different requirements with different institutional costs. The lessons of aviation safety culture, medical liability law, and financial auditing all offer partial models for algorithmic governance, but none transfers cleanly.

MEANINGFUL HUMAN CONTROL. A regulatory concept requiring that humans retain the ability to understand, intervene in, and reverse AI decisions. What this means in practice is contested.