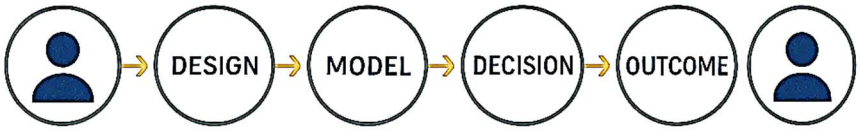
The Responsibility Rule- AI SAFE 4: Why “the algorithm did it” is unacceptable

The Illusion of Blame-Free AI, why “the algorithm did it” is unacceptable. This is why following AI SAFE 1: Safety-First Rule, AI SAFE 2: Economic Balance Rule, AI SAFE 3: Transparency & Audit Rule is key.

By Michal Florek, October 2025

1. Executive Summary

Artificial intelligence cannot bear moral or legal responsibility. Yet in public discourse and corporate governance, we increasingly hear: *“the algorithm did it.”*



Visual 1: Accountability Chain Layout

Artificial Intelligence is often described as a black box that “decides.” The **Responsibility Rule** asserts the opposite: AI is never autonomous in a moral sense.

AI is like a power tool — if it hurts someone, the manufacturer and user are responsible, not “the hammer.”

This white paper dismantles that illusion. It defines the **Responsibility Rule (AI SAFE 4)**—a framework ensuring every AI system remains anchored to identifiable human accountability. It proposes a global **Human Accountability Certification (HAC)** system, integrates responsibility into AI design and deployment lifecycles, and closes the ethical gap between automation and liability.

Key assertions:

- **AI amplifies human choices; it does not replace them.**
- **Lack of clarity creates liability.** Explainability and accountability must coexist.
- **Responsibility must be certified, not implied.**

- **Liability cannot be delegated to code.**
- **Public trust requires visible ownership of outcomes.**

The Rule transforms ethics into infrastructure: a verifiable, continuous chain of accountability from design to oversight.

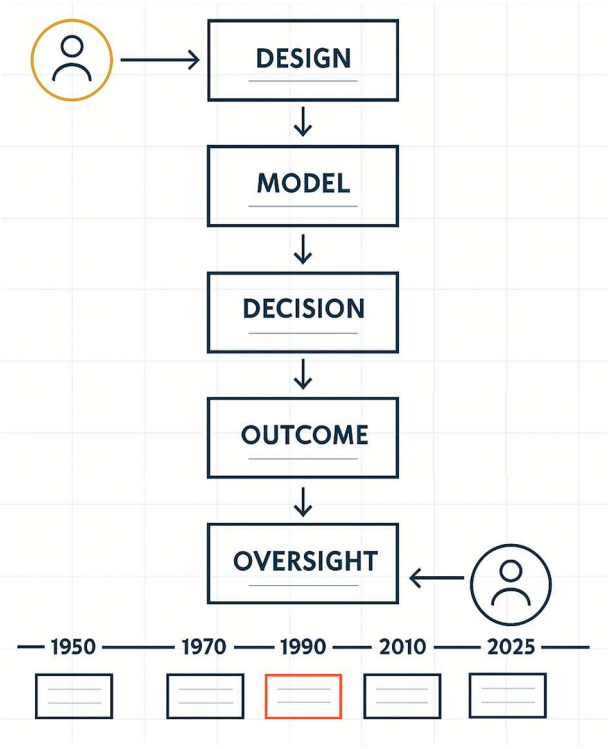
---

2. Context & Problem Definition

Automation has long diffused blame. From factory accidents to autopilot crashes, each era's innovation produced its own *"nobody's fault."* In the age of AI, this **diffusion** becomes institutional, which **destroys trust**.

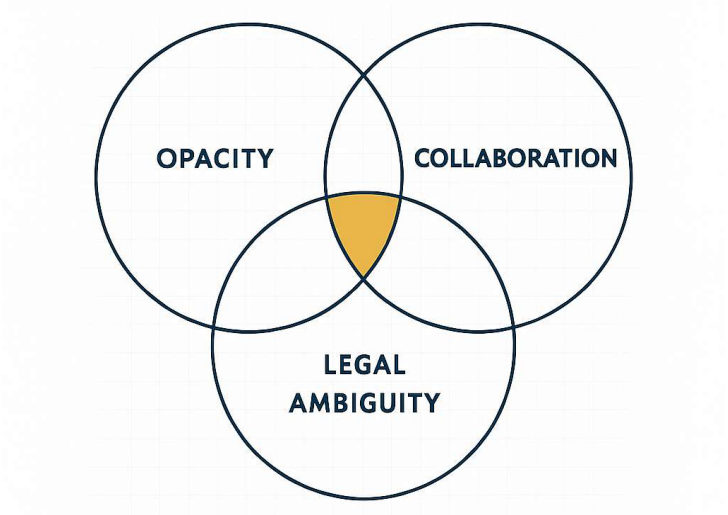
"The algorithm made an error" is not a statement of fact — it is an act of moral disappearance.

When AI automation is impacting many different areas of life at the same time, can we really afford a 'no blame' approach today?



Visual 2: Blame Illusion Timeline (1950–2025)

The Responsibility Gap



Visual 3: *Responsibility Gap Venn Layout*

Emerges from three forces: lack of clarity, complex collaboration, and legal ambiguity. When combined, they create an ethical vacuum where harm is measurable, yet no actor is accountable.

Consequences

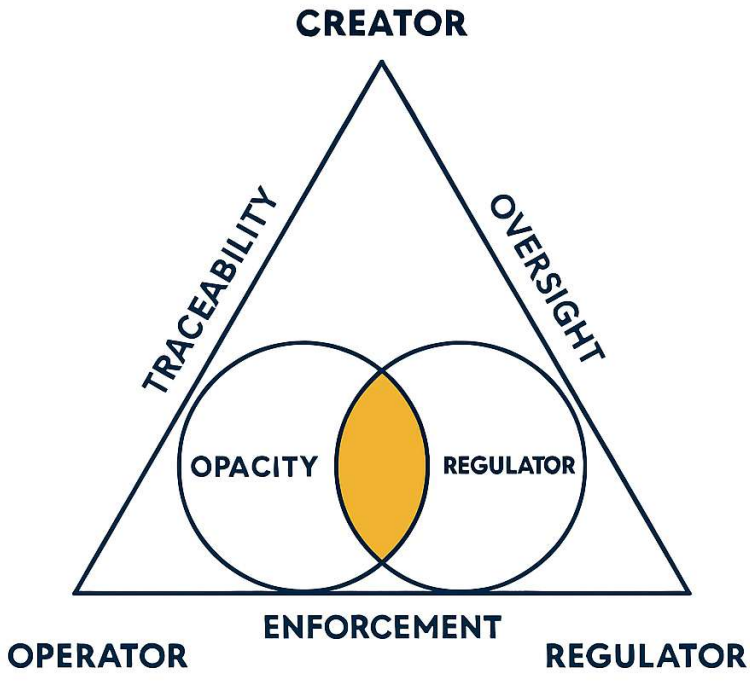
- **Economic:** Moral hazard and unchecked automation.
- **Psychological:** Erosion of trust in systems.
- **Regulatory:** Paralysis due to fragmented ownership.
- **Cultural:** Civic resignation — "machines know better."

Persistence of the "Blame-Free" Mindset

Marketing, legal convenience, and regulatory lag promote the myth of neutral algorithms. The Responsibility Rule challenges this by demanding *traceable human ownership* for every algorithmic act.

---

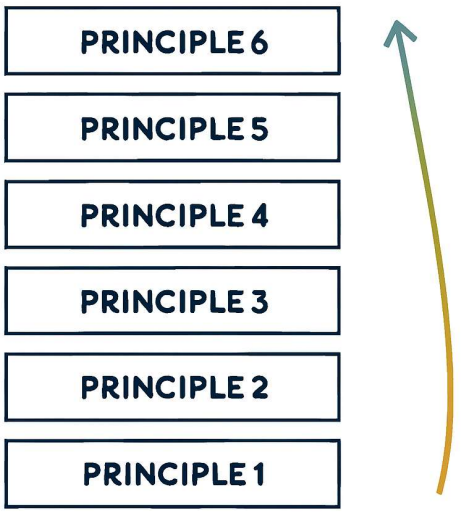
3. The Responsibility Rule – Core Principles



Visual 4: Responsibility Triangle

Every AI system must be traceable to an accountable actor. The Rule is structured around **six enforceable principles:**

#	Principle	Function
1	**Traceability**	Map every output to its human or institutional origin.
2	**Explainability**	Ensure cause–effect clarity for oversight and citizens.
3	**Attestability**	Require human sign-off at lifecycle milestones.
4	**Non-Transferability**	Prohibit delegating blame to machines.
5	**Moral Continuity**	Preserve ethical intent through all updates.
6	**Visibility**	Make accountability records public and verifiable.

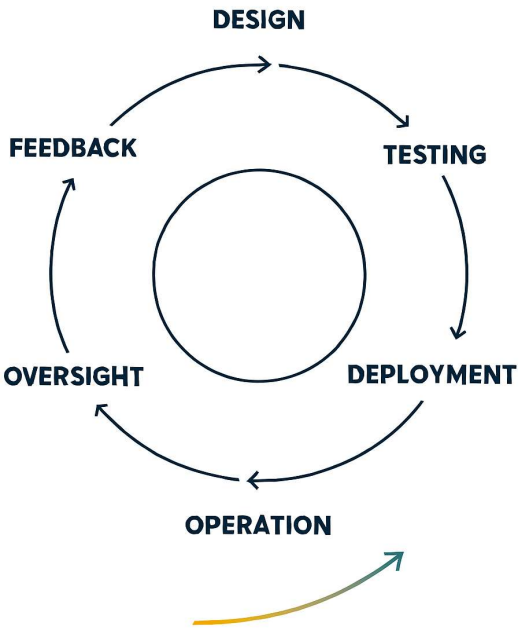


Visual 5: Accountability Stack Skeleton

Accountability without explainability is theatre (an act!), whilst explainability without accountability is noise.

These principles are the architecture of trust. They turn ethics into system design.

4. Governance Model: The Accountability Chain



Visual 6: Accountability Lifecycle Circle

Lifecycle Stages

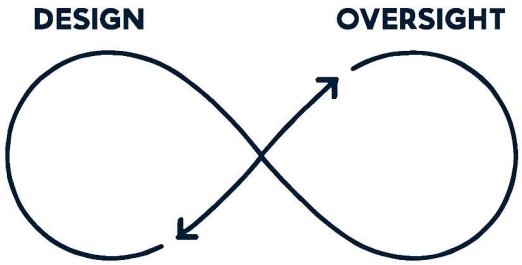
1. **Design Accountability** – Ethical impact, data provenance, bias prevention.
2. **Testing Accountability** – Transparent test logs and independent review.
3. **Deployment Accountability** – Public risk disclosure and human-in-loop guarantees.
4. **Operational Accountability** – Continuous monitoring and incident reporting.
5. **Oversight & Enforcement** – Audits, sanctions, and public registries.

Visual 7: *HAC Chain Layout*

Each stage issues a **Human Accountability Certificate (HAC)**:

- HAC-D (Design)
- HAC-T (Testing)
- HAC-D2 (Deployment)
- HAC-O (Operation)
- HAC-A (Audit)

Together, they form a **closed-loop chain of traceable responsibility.**

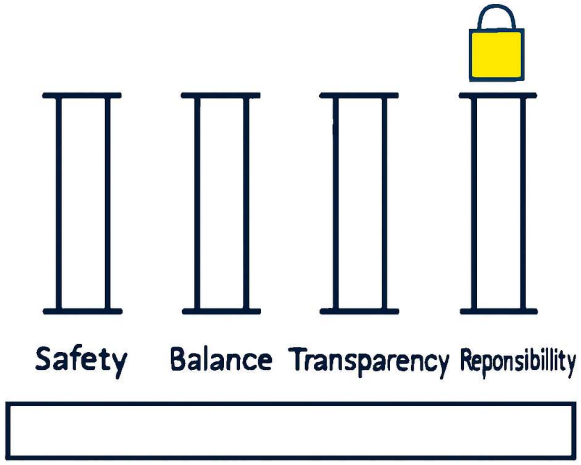


Visual 8: Responsibility Feedback Loop

The loop transforms accountability from paperwork into living infrastructure.

---

5. Framework Integration – The Four Pillars of AI Trust



Visual 9: Four Pillars of AI Trust

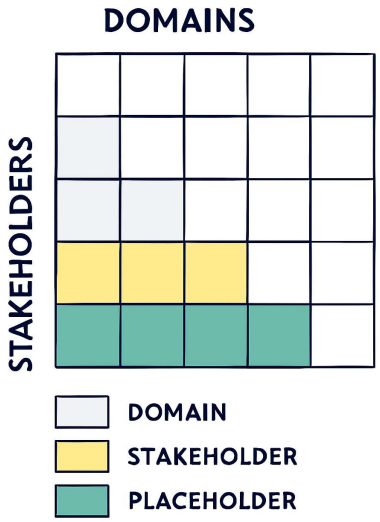
Framework	Question Answered
AI SAFE 1 – Safety-First	*Is it safe?*
AI SAFE 2 – Economic Balance	*Is it fair and stable?*
AI SAFE 3 – Transparency & Audit	*Can we see and verify it?*
AI SAFE 4 – Responsibility	*Who answers for it?*

Each rule reinforces the others. **Responsibility** is the keystone — the lock that secures the entire AI SAFE architecture.

Visual 10: Trust Cycle Diagram

---

6. Policy & Industry Recommendations



Visual 11: *Responsibility Distribution Map*

Human Accountability Certification (HAC)

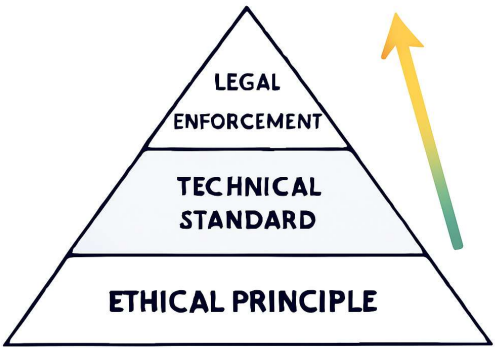
A global regulatory framework verifying that:

- Every AI system has a **named accountable owner**.
- Lifecycle attestations are **digitally signed and auditable**.
- Registries connect systems, actors, and compliance status.

Visual 12: HAC Regulatory Pathway

Legal Clauses

1. Non-Delegation of Liability
2. Mandatory Attestation
3. Liability Continuity
4. Transparency Disclosure
5. Sanctions for Breach

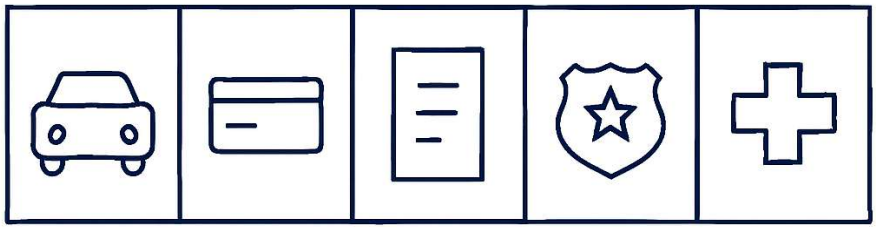


Visual 13: *Policy Flow Pyramid*

Institutional Recommendation

Governments should legislate HAC integration within national AI Acts and require interoperability with ISO 42001 and EU AI Act standards.

7. Case Studies & Future Scenarios



Visual 14: Case Study Carousel

Self-Driving Vehicles

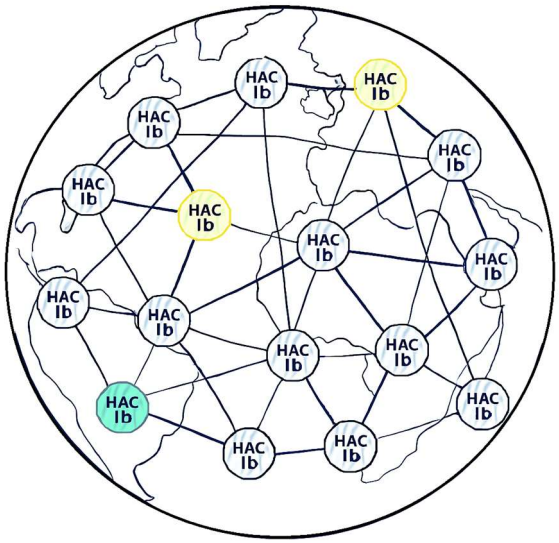
Blame oscillated between software and human monitor. Under HAC, design and deployment teams would bear certified accountability.

Credit Scoring Bias

Non-transparent data led to discrimination. HAC-T bias attestation would have prevented release.

Generative Misinformation

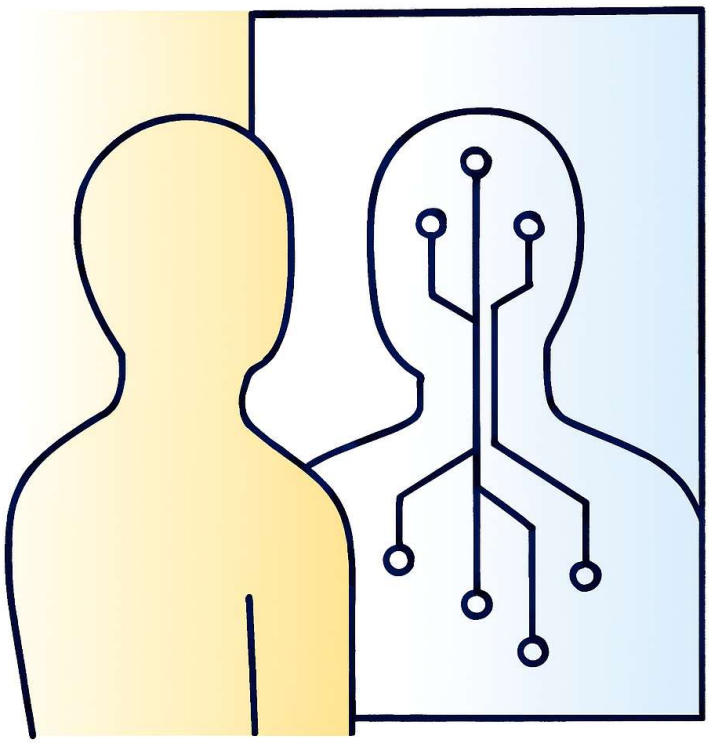
Developers blamed users. Non-transferability would preserve responsibility.



Visual 15: *Accountability Web 2030*

By 2030, global HAC registries enable instant accountability tracing. Every AI decision has a digital signature; every harm has a human answer.

8. Conclusion – The Mirror Principle



Visual 16: *Mirror Metaphor*

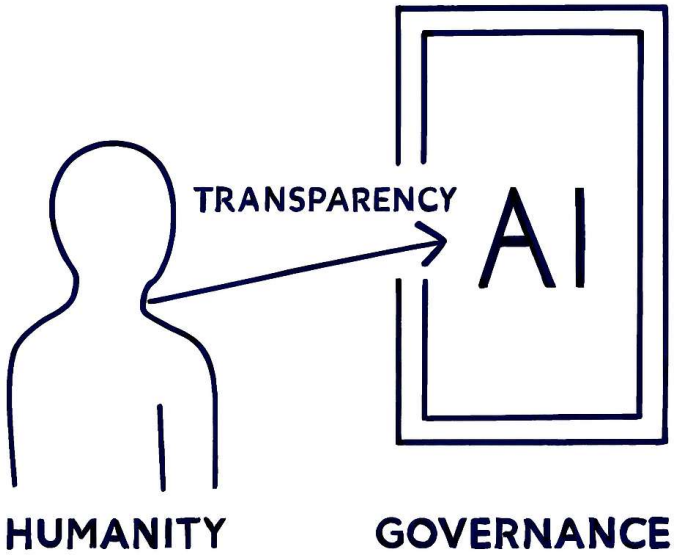
AI is not a moral actor; it is a mirror.
What it reflects depends on who is looking.

Autonomy does not absolve — it obliges.

Responsibility is civilization's boundary.
When machines act, we must ensure humanity remains visible in the reflection, in subconscious.

Final Call

- **Policymakers:** legislate responsibility infrastructure.
- **Industry:** embed HAC in every release pipeline.
- **Academia:** train "AI accountability engineers."
- **Citizens:** demand to know who is responsible.



Visual 17: *Mirror Principle Diagram*

The age of algorithmic innocence ends here.

---

9. Citations / References

1. OECD AI Principles (2019) - https://oecd.ai/en/ai-principles
2. AI - Human Frameworks - https://theailaws.com/ai-human-frameworks-1
3. UNESCO (2021) – Recommendation on the Ethics of AI - https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
4. EU AI Act Full Text (2024) - https://artificialintelligenceact.eu/the-act/
5. ISO/IEC 23894: 2023 – AI Risk Management - https://www.iso.org/standard/77304.html
6. ISO/IEC 42001: 2023 – AI Management Systems - https://www.iso.org/standard/42001
7. IEEE P7009 – Fail-Safe Design Standard - https://ieeexplore.ieee.org/document/10462965
8. RAND (2024) – When AI Gets It Wrong, Will It Be Held Legally Accountable? - https://www.rand.org/pubs/articles/2024/when-ai-gets-it-wrong-will-it-be-held-legally-accountable.html
9. Collina, L. (2023) – Critical Issues about AI Accountability Answered, California Management Review - https://cmr.berkeley.edu/2023/11/critical-issues-about-a-i-accountability-answered/
10. Lima et al. (2023) – Blaming Humans and Machines: What Shapes Reactions to Algorithmic Harm - https://arxiv.org/abs/2304.02176
11. Ryan et al. (2023) – Modelling Responsibility for AI-Based Safety-Critical Systems - https://arxiv.org/abs/2401.09459