The Empty Chair in AI Governance: Why Human Oversight may Fail When we Need it Most

Cosiamo Bavuso
UMANIAMO®, USA

cosiamo@umaniamo.org

Abstract: AI governance frameworks increasingly rely on the assurance that a "human in the loop" will provide accountability, judgment, and control when artificial intelligence systems are used in consequential decisions. This paper argues that this assurance is incomplete and potentially misleading unless organizations examine whether the human system surrounding the AI is actually capable of exercising judgment when it matters. The paper defines artificial intelligence as machine inference: computational systems that detect patterns, generate outputs, recommend actions, and simulate features of reasoning without conscience, responsibility, embodiment, or wisdom. It introduces the concept of the Empty Chair: the absence of a clearly designated owner for the formation, sustainment, and exercise of human judgment in AI-enabled decision systems. The paper proposes that meaningful oversight requires at least seven conditions: cognitive capability, behavioral authority, relational safety, decision discipline, cognitive sustainment, survivability, and relational sustainment. Drawing on applied organizational work with governance bodies, enterprise leaders, public-facing institutional systems, and cross-organizational cohorts, the paper identifies a recurring capability-authority split in which those most likely to detect certain categories of risk are not always positioned or authorized to interrupt decision momentum. The analysis further examines how automation bias and ethical fading can weaken human responsibility by making AI outputs appear more complete, objective, and procedurally sufficient than they are. The paper reframes cognitive diversity as governance infrastructure rather than a team-building concept and presents the D.E.E.P.E.R.™ practice as a structured method for recurring reflective diversification across decision cycles. It also offers practical interventions, including cognitive diversity audits, cognitive role charters, decision pause points, Empty Chair protocols, authority alignment, and recurring reflective practice. The paper concludes that AI governance cannot stop at governing machines. It must also govern the human systems responsible for supervising them. The central question is not whether a human is present in the loop, but whether that human is formed, positioned, protected, sustained, and authorized to exercise judgment under pressure.

Keywords: AI governance, Cognitive diversity, Human judgment, Human architecture, AI risk management, Human-in-the-loop

1. Executive Summary

Every serious artificial intelligence governance framework now has some version of the same thing: keep a human in the loop. The phrase is reassuring. It suggests accountability, judgment, and control. It suggests that when an AI system produces a recommendation, a real person will be there to review it, challenge it, and stop it if necessary.

In this paper, artificial intelligence is understood as machine inference: computational systems that detect patterns, generate outputs, recommend actions, and simulate certain features of reasoning without conscience, responsibility, embodiment, or wisdom. Machine inference does not mimic or refer to intelligence in the human sense. However, there is a problem. Most organizations have not tested whether the human in the loop is actually able to exercise judgment when it matters. Their leadership assigned responsibility, created committees, implemented policy, and documented approval processes. They may have even trained staff on AI risk, ethics, privacy, bias, and compliance. They have not asked a deeper question: When the system is wrong, who in the room is cognitively likely to notice and behaviorally able to stop the decision?

That is what I call the Empty Chair.

The Empty Chair is the absence of a clearly designated owner for the formation, sustainment, and exercise of human judgment in AI-enabled decision systems. It is the gap between the person formally assigned to oversight and the actual conditions required for oversight to work. This paper argues that the central risk in AI governance may not be the algorithms alone. It may be the architecture of the human system surrounding it. A government agency can comply with AI governance requirements and still leave the chair empty. The danger is not only that the machine may be wrong. The danger is that the machine may make the human feel less responsible for finding out. That should make every senior official very uncomfortable.

2. The Comforting Fiction of "Human in the Loop"

Imagine the following scenario. An AI-enabled system produces a recommendation. The output is fluent, confident, well-structured, and consistent with available data. The recommendation is reviewed by a designated

official and then discussed in committee; all of the required documentation is completed and the process satisfies existing policy. The decision is made, and then, weeks or months later, the decision is found to be wrong.

The error did not occur because no human was present at the time of the decision, humans were present. The failure occurred because the humans present were not positioned to exercise judgment. Some lacked the cognitive disposition to detect the issue. Others sensed something was wrong but lacked the behavioral authority to interrupt momentum. Still others assumed someone else had already checked.

The governance process did not fail visibly. It failed without notice, until the consequences became undeniable, yet hopefully in time to avert catastrophe. This is the uncomfortable reality behind many governance systems: they are designed to demonstrate that a human occupied the chair, but not that the judgment required by the chair was actively exercised. The distinction is subtle, but it is the difference between governance as documentation and governance as practice. The *RMS Titanic* did not sink to the bottom of the Atlantic, because the bridge was empty. It suffered from an empty moment, a moment in which warnings, assumptions, authority, and decision-making failed to converge quickly enough to change the outcome. AI governance presents a similar challenge. Regulations may require human oversight, but oversight is not the same as human judgment. Presence can be documented. However, human judgment must be cultivated, sustained, and exercised.

The phrase “human in the loop” hides this distinction. It allows organizations to satisfy the literal requirement of oversight while leaving the functional requirement unexamined. The question is not whether a human is in the loop. The real question is whether the human system surrounding the Empty Chair has the capacity, permission, and discipline to intervene when intervention is required.

3. The Empty Chair Defined

The Empty Chair is a structural condition in which no person, role, or process clearly owns the conditions necessary for human judgment to function. It appears when organizations assume that because a person has been assigned to review a decision, meaningful oversight has occurred. But meaningful oversight requires more than assignment. It requires at least seven conditions:

1. Cognitive Capability: someone must be able to notice a flaw, anomaly, inconsistency, or unexamined assumption.
2. Behavioral Authority: that person must have the standing, confidence, or formal permission to interrupt the process.
3. Relational Safety: the organizational governance systems must tolerate challenge, dissent, and hesitation without punishing the person who raises it.
4. Decision Discipline: the process must require unanswered questions to be resolved before momentum carries the decision forward.
5. Cognitive Sustainment:[1] the deliberate, ongoing maintenance of an individual's or group's capacity for independent judgment: the practiced ability to generate original assessment, detect anomaly, and resist convergence toward a dominant or externally-supplied conclusion, treated as a capability requiring continuous exercise rather than a trait assumed to persist once established.
6. Survivability:[2] the capacity of an individual's or group's judgment to remain intact and exercisable under the specific conditions most likely to degrade it: time pressure, confident contradictory input, social cost for dissent, and prolonged low-stakes deference that erodes the habit of intervention before a high-stakes moment requires it.
7. Relational Sustainment: the deliberate, ongoing maintenance of the trust, mutual legibility, and psychological safety required for judgment, once exercised, to actually be transmitted, heard, and acted upon across a group or hierarchy, treated as its own capability subject to its own atrophy, independent of whether individual cognitive capacity remains intact.

Most AI governance frameworks remain heavily focused on system-side controls: model validation, bias testing, audit documentation, risk classification, data governance, and compliance procedures. These controls are necessary. However, they do not answer the more critical question: whether the human system responsible for oversight is actually capable of exercising judgement when it matters.

[1]Cognitive Sustainment is a term framed by the author in reference to the Systems Engineering and Human Systems Integration disciple of sustainment.

[2]Survivability is a term framed by the author in reference to the Human Systems Integration disciple of Survivability and Habitability.

The human side of oversight remains dangerously underdeveloped. We have built governance structures around the comforting assumption that judgment will simply appear when needed and that the right person will see the issue, identify it as such, and act with enough authority to interrupt momentum. That assumption is increasingly unsafe.

As AI systems become more fluent, persuasive, and embedded in decision workflows, the gap between unbiased discernment, sound judgment, and aligned action becomes more unstable. The risk is not that only the machine may be wrong. The risk is that the human system around the machine may no longer be formed, positioned, or authorized to notice, question, and act.

4. Why This Matters More in Government

Government decisions carry public consequence. A flawed decision in a private firm may affect revenue, reputation, or operational performance. However, a flawed decision in government may affect eligibility, benefits, enforcement, surveillance, public safety, immigration, healthcare access, education, procurement, employment, policing, or national security. When AI systems enter these environments, the human oversight problem becomes more serious, not less.

Government agencies are especially vulnerable because they often operate within: complex reporting chains, formal hierarchies, risk-averse cultures, compliance-heavy procedures, fragmented ownership, contractor-supported systems, and politically sensitive decision environments.

In such systems, many people may touch a decision, but no one may fully own the judgment architecture around it. A senior official may believe oversight exists because the organizational chart or project charter says that it does. But the chart may be wrong, and the human system may contain a gap that cannot be seen from the organizational chart, governance or RACI chart, policy manual, or dashboard. That gap or seam is the Empty Chair.

5. The Fourth Question AI Governance Is Missing

Most AI readiness and governance efforts ask three familiar questions: Is the system technically reliable? Is the system legally and procedurally compliant? Is there a human in the loop? This paper proposes a fourth question: Is the human in the loop actually positioned to exercise judgment?

This question radically changes the conversation, not in terms of unfairly judging the person, but by moving AI governance from procedural assurance to human capability within the context of the organization itself. It asks whether the people charged with oversight have the cognitive diversity, behavioral authority, relational trust, and structured process required to challenge machine output.

Any system can be technically sound and still be misused. A process can be compliant and still be hollow of value. A human can be present and still be functionally absent. That is why the fourth question matters.

6. Cognitive Diversity as Governance Infrastructure

Cognitive diversity is often treated as a team-building or communication topic. That framing is too small for the AI era. In AI governance, cognitive diversity is infrastructure. Different people notice different risks. Some are more likely to ask for evidence. Some are more likely to see second-order consequences. Others are more likely to detect procedural gaps. Some are more likely to sense relational harm, public trust implications, or political risk. Some are more likely to imagine future misuse, mission drift, or future benefit. Others are more likely to ask whether a decision can actually be implemented. No single leader naturally sees all of this. No committee automatically covers all of it. And there is no AI system that can substitute or proxy for a properly formed human deliberative group.

Using the Emergenetics® attributes as one applied lens, the relevant attributes can be described across four thinking patterns and three behavioral patterns:

- Analytical thinking: What do the facts show? Where is the logic? How much will this cost? Show me the value.
- Structural thinking: What are the steps, risks, details, and controls? Is there a project plan, and what are the gates?
- Social thinking: Who is affected, and how? Who else needs to be in the room? Who else has missing information?

- Conceptual thinking: What future are we creating? Is this part of our mission? What will the future look like if we don't do this?
- Expressiveness shapes whether insight is voiced or held in reflection.
- Assertiveness shapes whether challenge interrupts momentum, speeds it up, or waits for everyone to agree.
- Flexibility shapes whether the group can adapt when new evidence appears, or commit to a plan and stick with it.

The oversight risk emerges when the attributes most needed to detect a problem are underrepresented, unheard, or behaviorally unable to act. For example, a group may have someone who can see the flaw but cannot stop the room from acting. Another group may have decisive authority but lack the cognitive disposition to notice what needs to be challenged or adapted. Both conditions represent the Empty Chair.

Compare those to an accessible cultural analogy is the character of Captain James T. Kirk in Gene Roddenberry's Star Trek. The person in the chair is not a passive monitor or a checkbox reviewer. The archetype is a human integrator: someone who listens to the people surrounding him, attends to the human system architecture of the bridge, questions the machine, trusts capable specialists, weighs incomplete information, and still accepts responsibility for the ultimate decision. The Chair requires judgment, courage, and consequence-bearing authority. Without those conditions, the Empty Chair remains empty even when someone is physically present.

7. A Note on "Smooth" Meetings

Many facilitators are trained to read a group's health through a familiar progression: a team begins by forming, moves through a period of friction often called storming, settles into what is usually called norming, and finally reaches performing. The vocabulary is common enough that most leaders have heard it, even if they could not name where it came from.

That third stage deserves closer scrutiny than it usually receives. "Norming" is generally read as convergence: a group settling its differences, finding common ground, smoothing out disagreement. A quiet, orderly, single-voice-at-a-time meeting is typically read as the sign of a team that has finally matured.

This reading should concern any official thinking seriously about AI governance, for a simple reason: AI output is itself a catalyst of convergence. A model returns the most statistically probable answer, which is the smoothed center of the bell curve of its training data. If a human group has also been culturally trained to treat the disappearance of dissent as the mark of a well-functioning team, then the group and its AI tool are now optimizing for the same outcome, and reinforcing each other while doing it. The model converges. The room, watching its own disagreement quiet down, reads that as success. Nobody in the room has a reason to be suspicious of the quiet, because the quiet is what they were taught a healthy meeting sounds like.

A genuinely cognitively diverse group does not necessarily sound quiet when it is functioning well. It sounds like every distinct way of thinking in the room finding its function: the evidence-seeker still asking for evidence, the relationship-reader still asking who is affected, the systems-thinker still asking what comes next, simply without those differences competing for the same airtime they once did. From the outside this can look calm. It is something closer to difference that has found its place, not difference that has disappeared.

The practical caution for any AI oversight body is this: a quiet, agreeable review meeting is not, by itself, evidence that judgment was exercised. It may be evidence that judgment was exercised well. It may equally be evidence that dissent has already been priced out of the room and that the AI's confident, convergent output will meet no resistance at all, not because it is correct, but because resistance was never going to be voiced in that room regardless of what the output said.

8. What the Case Patterns Reveal

Across applied work with governance bodies, enterprise HR leaders, public-facing institutional systems, and cross-organizational cohorts, a repeated pattern emerges: the people most likely to detect certain categories of risk are not always the people most likely to be heard. This is not a problem of intelligence or goodwill. It is a design problem.

In several examined groups, leadership bodies were heavily concentrated in Social and Conceptual thinking. These groups were often strong in relationship-building, vision, communication, and stakeholder sense-making. Those are real strengths. But the same groups were comparatively weaker in Analytical and Structural scrutiny:

the capacities most associated with evidence challenge, procedural discipline, detail verification, and systematic follow-through.

That creates a predictable vulnerability when AI enters the room. AI output often arrives in a form that appears highly Analytical and Structural: it is ordered, confident, formatted, and it sounds complete. For a group that does not naturally live in those thinking patterns, the output may feel more authoritative than it deserves to be. The system appears to supply the very capability the group lacks. That is where automation bias becomes not merely a technical concern, but a human architecture concern.

9. Automation Bias and Ethical Fading

Two additional risks make the Empty Chair more dangerous in AI-enabled organizations: automation bias and ethical fading. Automation bias is the tendency to over-trust automated system output, especially when that output appears consistent, confident, data-rich, or technically sophisticated. In practice, this means a recommendation produced by an AI system may be granted more authority than it has earned simply because it looks ordered, precise, and complete. This is especially dangerous in bureaucratic environments.

A well-formatted AI recommendation can appear to carry the weight of analysis even when its assumptions are incomplete, its data appears biased, or its reasoning cannot be independently verified. The output arrives with the visual and linguistic signals of competence: categories, rankings, probabilities, summaries, risk scores, and recommended actions. The human in the loop may not think, "the machine is [morally] right." They may simply assume, "this has already been analyzed" and that is enough to make the moral question and, more importantly, the moral authority vanish from the decision.

Once that assumption enters the room, genuine oversight begins to weaken: questions become checklist items; there is less dissent or rigorous interrogation; review feels slower to start and faster to finish and then the final review becomes confirmation. Afterwards, the human in the loop becomes a human near the loop.

Ethical fading compounds the problem. Ethical fading occurs when the moral dimensions of a decision disappear from view because the situation is framed in technical, procedural, financial, or administrative terms. This occurs when a decision about a person becomes a case; a case becomes a data point; a data point becomes an output; an output becomes a recommendation; and a recommendation becomes a workflow step. In this case, logic of a linear process is unexamined and while at no point does anyone explicitly decide to ignore human dignity, the moral and ethical dimensions simply fade into a workflow.

This is one of the most serious dangers of AI-enabled bureaucracy: not open cruelty or inhumane, but vague abstraction. The system does not need officials to become unethical, only to follow the procedures. Some examples:

- Benefits determination becomes "eligibility optimization."
- Child welfare alerts become "risk stratification."
- Hiring, promotion, or termination decisions become "candidate ranking."
- Public safety interventions become "predictive resource allocation."
- Healthcare denials become "utilization management."

The language changes, and with it the felt responsibility of the decision-maker. Automation bias tells the brain: the system probably got this right. Ethical fading tells the brain: this is not really a moral choice. Together, they create the perfect conditions for the Empty Chair.

The person who might ask, "Is this right?" is replaced by a process question: "Was this reviewed?" The person who might ask, "Who is affected?" is replaced by a dashboard showing "impact." The person who might ask, "Should we stop?" is replaced by a workflow with a pending approval task.

This is why human oversight cannot be reduced to presence, review, or sign-off. The problem is not merely whether a human sees the AI output. The problem is whether the decision environment preserves the human capacity to recognize that a moral judgment is being made. A government agency can satisfy procedural oversight requirements while allowing both doubt and conscience to fade from the room. That is not governance. It is ethical risk disguised as administrative efficiency.

10. The Capability | Authority Split

The most dangerous Empty Chair is not simple absence, rather misalignment. One person may have the capability to detect the issue, while another may have the authority to act. If those are not connected by process, trust, or role design, oversight fails. This is especially important in hierarchical organizations.

Frontline staff often see operational reality first. They notice procedural inconsistencies, implementation risks, and user-facing consequences. But they may not have the authority or confidence to interrupt a decision moving through senior channels. Senior leaders may have authority, but not the ground-level pattern recognition required to see the problem early.

Each group holds part of the oversight function, yet neither holds the whole.

That is a bidirectional Empty Chair, and it looks like this: a staff member sees but cannot stop, and the executive can stop but does not see, and so the system proceeds. If AI governance does not explicitly design for this split, it will reproduce it over and over again.

11. The Positive Deviant in the Chair

The Empty Chair is common, but it is not inevitable. One observed organization provides a useful contrast. Rather than treating cognitive diversity as a one-time workshop, it embedded cognitive-diversity-informed facilitation into the way teams were composed, meetings were structured, and decisions were reviewed. This did not mean every group had perfect balance because that is not realistic. It meant the organization developed recurring practices to notice what and who was missing, and to deliberately invite those into the process.

In some cases, this was as simple as designating a literal empty chair in a meeting and labeling it with the missing perspective. If Analytical thinking was absent, the group had to ask: What evidence have we not examined? If Structural thinking was absent: What steps, risks, and controls are unclear? If Social thinking was absent: Who is affected? If Conceptual thinking was absent: What future are we creating?

This practice is simple, which is why it works. It turns an invisible governance vulnerability into a visible discipline. It makes the people in the room answer for what they do not naturally see.

12. The D.E.E.P.E.R.™ Practice for AI Oversight

A practical response to the Empty Chair is not more training alone. It is recurring reflective diversification. The D.E.E.P.E.R.™ method provides one structured way to do this by requiring a deliberative body to move through all seven Emergenetics® attributes before a decision cycle is considered complete:

- Details: What actually happened? What is the process? What are the steps?
- Evidence: What do the facts show, not only the "facts" produced by the machine? What assumptions are unsupported?
- Empathy: Who is affected by this decision?
- Prospective: What future does this decision create?
- Execution: What action will be taken, by whom, when, and under what conditions?
- Reflection: How did the group process the decision? Who spoke, who stayed silent, and what was avoided?

This practice does not replace technical AI governance, it strengthens it. The D.E.E.P.E.R. practice ensures that human oversight is not reduced to a signature, a review field, or a committee agenda item. It makes judgment a repeated organizational discipline.

13. Implications for Senior Officials

For senior government officials responsible for digital transformation, AI readiness, or enterprise governance, the practical implications are direct. Before deploying or approving AI-enabled decision systems, leaders should ask:

- Who is responsible for human judgment in this process?
- Which cognitive perspectives are missing from the decision body?
- Who is most likely to notice error, bias, or procedural weakness?
- Does that person have the authority to stop or slow the decision?
- Are dissent and hesitation treated as governance assets or bureaucratic irritants?

- Is the "human in the loop" a real decision-maker or a procedural placeholder?
- Have we tested the oversight process under conditions of pressure, ambiguity, and confident machine output?

If these questions have not been answered, documented, and embedded into organizational processes, the agency may have governance documentation without governance capability. That is not governance readiness. It is exposure.

14. What Needs to Change

We are not suggesting that the answer is to find better people. It is to design better human oversight architecture. This architecture includes:

- *Cognitive Diversity Audits*: Map the cognitive and behavioral composition of boards, committees, review teams, and AI oversight bodies.
- *Cognitive Role Charters*: Assign explicit oversight roles: evidence challenger, procedural verifier, stakeholder impact reviewer, future-risk examiner, implementation owner.
- *Decision Pause Points*: Require structured challenge before high-impact AI-enabled decisions proceed.
- *Empty Chair Protocols*: When a perspective is missing, make the absence visible and require the group to answer from that viewpoint before deciding.
- *Authority Alignment*: Ensure the person most likely to detect a class of risk has a formal path to interrupt, escalate, or require reconsideration.
- *Recurring Reflective Practice*: Use a structured method such as D.E.E.P.E.R.™ to prevent dominant cognitive habits from becoming governance blind spots.

None of these interventions are especially expensive. They are, however, uncomfortable. That discomfort is precisely why they are needed. AI-enabled systems often make organizations feel more confident, more informed, more efficient, and more in control. The outputs are clean, the summaries are fluent, the dashboards are orderly and always up to date, the workflow moves, and the water feels warm. That is the danger.

The old image of the frog in slowly boiling water is useful here, not as biology, but as bureaucracy. Most organizations will not lose human judgment all at once. They will acclimate to its absence. They will become comfortable with machine-generated confidence, comfortable with faster approvals, comfortable with fewer objections, comfortable with oversight that appears complete because the form was completed. By the time the risk becomes obvious, the organization may already have normalized the conditions that made real oversight unlikely.

The purpose of human oversight architecture is to interrupt that warming process before it feels dangerous. Cognitive diversity audits, role charters, pause points, Empty Chair protocols, authority alignment, and recurring reflection all introduce deliberate friction into systems that would otherwise reward speed, fluency, and compliance over judgment. That friction is not inefficiency. It is the sound of human responsibility re-entering the room.

15. Conclusion: The Chair Is Either Filled or it is not

AI governance cannot stop at governing machines. It must also govern the human systems responsible for supervising them. The most important question may not be: Is there a human in the loop? The more important question is: What conditions must exist for human judgment to function when it matters most? The difference is not semantic. It is the difference between procedural reassurance and genuine oversight. Every AI governance framework ultimately rests on an assumption: that when a critical decision arrives, the right human being will recognize the issue and possess the authority to intervene. Few organizations have tested that assumption, and almost none have designed for it.

The Empty Chair is not a metaphorical problem. It is a governance vulnerability hiding in plain sight. Once seen, it becomes difficult to ignore, and once ignored, it becomes a choice. Machine inference is not human intelligence. Fluency is not wisdom and output is not judgment.

Ethics Declaration: This paper is a conceptual and theoretical work that did not include human subject research or human participants, identifiable personal data, animal subjects, or experimental interventions requiring IRB or institutional ethical review. Accordingly, formal ethical clearance was not required. Any observations drawn

from the author's professional practice or publicly available information are presented in aggregate and without disclosure of confidential or personally identifiable information.

AI Declaration: Generative artificial intelligence (AI) tools were used during the preparation of this manuscript to assist with editorial refinement, language enhancement, and structural organization. All substantive concepts, theoretical frameworks, arguments, interpretations, examples, and conclusions were independently conceived, evaluated, and authored by the author, most of which were conceived by the author in his professional practice over the past two decades. The author accepts full responsibility for the content of this manuscript.

References

Baverso, P. (2015). Human performance architecture framework: A strengths-based lean approach to human systems performance. AI Practitioner: The International Journal of Appreciative Inquiry, 17(4), 44–55.

Browning, G. (2006). Emergenetics: Tap into the new science of success. HarperCollins.

Cooperrider, D. L., & Whitney, D. (2005). Appreciative inquiry: A positive revolution in change. Berrett-Koehler.

International Organization for Standardization & International Electrotechnical Commission. (2023). ISO/IEC 42001:2023 - Artificial intelligence management system requirements. ISO.

National Institute of Standards and Technology. (2023). AI risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans, 30(3), 286–297.

Roddenberry, G. (Creator). (1966–1969). Star Trek: The original series [TV series]. Desilu Productions; Paramount Television.