AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Human-in-the-Loop Architectures for Responsible AI Governance

Noah Ivanov[1], Amelia Popescu[2]

[1] Associate Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: noah.ivanov15@yahoo.com | ORCID: 0037-8725-2323-2709
[2] Postdoctoral Researcher, School of Data Science, Central European Tech University, Vienna, Austria. Email: amelia.popescu209@outlook.com | ORCID: 5956-4303-3025-7025

ABSTRACT

Human-in-the-loop (HITL) architectures -- system designs that structurally incorporate human judgement, oversight, and intervention within AI decision processes -- are widely advocated as a primary mechanism for responsible AI governance. However, the term encompasses a heterogeneous range of architectural patterns, from nominal human involvement that provides no meaningful oversight to deeply integrated collaborative decision-making that genuinely distributes authority between human and AI. This paper develops a comprehensive taxonomy of eight HITL architectural patterns organised across two dimensions -- intervention depth (monitoring, advisory, approval, co-decision) and intervention timing (pre-decision, in-decision, post-decision) -- and evaluates their governance effectiveness through a mixed-methods study involving expert assessment and a longitudinal analysis of 26 operational AI systems with documented HITL implementations. Expert consensus ($n = 36$, Kendall $W = 0.77$) identifies co-decision and approval-tier patterns as most effective for responsible AI governance in high-stakes contexts. Longitudinal analysis demonstrates that HITL implementation depth is significantly associated with reduced governance incidents: deep HITL systems (co-decision or approval tier) exhibit 44.6% fewer governance incidents than shallow HITL systems (monitoring or advisory tier) over a 12-month observation period (IRR = 0.55, 95% CI: 0.44-0.70, $p < 0.001$). Critically, nominal HITL implementations -- those providing human involvement without meaningful oversight authority -- show no statistically significant governance benefit compared to fully automated systems. The study contributes a validated HITL taxonomy, a HITL Governance Effectiveness Scale, and design principles for meaningful rather than nominal human oversight.

Keywords: human-in-the-loop; AI governance; human oversight; responsible AI; HITL architecture; EU AI Act; automation bias; meaningful oversight

Citation: Ivanov and Popescu [2025]. Human-in-the-Loop Architectures for Responsible AI Governance. DOI: https://doi.org/10.5281/zenodo.19185018

Article Information: *Received:* 20 January 2025 *Accepted:* 22 March 2025 *Published:* 20 May 2025

Research Article: *Research Article*

1. Introduction

1.1 The HITL Governance Paradox

Human-in-the-loop (HITL) design is among the most frequently cited mechanisms for responsible AI governance. The EU AI Act (2024) Article 14 mandates that high-risk AI systems be designed to enable effective human oversight, including the ability to understand and interpret the system's output and to override its recommendations. The NIST AI Risk Management Framework (2023) identifies human oversight as a core trustworthiness characteristic. The IEEE Ethically Aligned Design framework (2019) designates human agency as its first principle. Yet a growing body of empirical research reveals a governance paradox at the heart of HITL design: human presence in AI decision processes does not guarantee meaningful oversight and may, under certain conditions, produce worse governance outcomes than fully automated systems. Cummings (2017) documented automation bias -- the tendency of human operators to over-rely on automated recommendations, failing to apply critical judgement even when system errors are detectable. Skitka et al. (1999) demonstrated that human operators systematically fail to detect automated system errors under cognitive load conditions. Passi and Barocas (2019) showed that nominal HITL processes in data science workflows function as accountability theatre -- providing the appearance of human oversight without substantive review. Raji et al. (2020) found that human review steps in AI deployment pipelines are frequently superficial, time-pressured, and lack the authority or information needed for meaningful intervention. This paradox -- HITL is mandated as a governance mechanism but can fail to provide meaningful oversight -- motivates a systematic examination of the architectural conditions under which HITL implementations produce genuine governance value. The critical variable is not the presence of human involvement but its depth: the degree to which human oversight authority is structurally embedded in the decision architecture, the quality of information provided to human reviewers, and the institutional conditions that enable genuine rather than nominal review.

1.2 Research Objectives and Paper Structure

This paper addresses the governance paradox by developing a comprehensive taxonomy of HITL architectural patterns and empirically evaluating their governance effectiveness. Research objectives are: (1) to develop a two-dimensional HITL taxonomy covering eight architectural patterns, validated through expert assessment; (2) to evaluate HITL implementation depth effects on governance incident rates through a longitudinal analysis of 26 operational AI systems; and (3) to derive design principles for meaningful rather than nominal HITL implementation. Section 2 reviews prior work on HITL systems, automation bias, and governance effectiveness. Section 3 presents the HITL taxonomy and Governance Effectiveness Scale. Section 4 describes the evaluation methodology. Section 5 reports results. Section 6 discusses design principles for meaningful HITL. Section 6 concludes.

2. Literature Review

2.1 Human-in-the-Loop Systems and Automation Bias

The human factors and supervisory control literature has documented systematic patterns of human oversight failure in automated systems that apply with particular force to AI-assisted decision-making. Endsley's (1995) theory of situation awareness identifies three levels of understanding required for effective human oversight: perception of system state, comprehension of its meaning, and projection of future states. HITL designs that provide human reviewers with only model outputs -- without the underlying reasoning, uncertainty estimates, and contextual information needed for levels 2 and 3 situation awareness -- structurally impede meaningful oversight. Parasuraman et al. (2000) proposed a taxonomy of automation levels from fully manual to fully automated, demonstrating that intermediate automation levels can produce worse performance than either extreme under certain conditions -- the ironies of automation (Bainbridge, 1983). Lee and See (2004) provided a framework for trust calibration in human-automation interaction, arguing that appropriate reliance depends on accurate mental models of system capability and uncertainty. These theoretical frameworks inform the HITL taxonomy's emphasis on information provision adequacy as a core dimension of HITL effectiveness. Table 1 maps prior HITL and automation literature to the two-dimensional HITL taxonomy.

2.2 HITL in AI Governance Practice

Empirical studies of HITL implementation in AI governance contexts have revealed significant gaps between formal HITL requirements and practical oversight effectiveness. Vaccaro and Sabanovic (2021) documented that contestability

processes in algorithmic systems frequently lack the information provision and authority structures needed for meaningful challenge. Madaio et al. (2020) found through an organisational study of AI development teams that HITL review checkpoints were frequently treated as procedural steps rather than substantive oversight opportunities. Green and Chen (2019) demonstrated experimentally that providing ALGORITHMIC recommendations to human decision-makers in recidivism prediction contexts reduced accuracy relative to unaided human judgement, attributing this to automation bias overriding appropriate human contextual reasoning. These findings collectively motivate the HITL taxonomy's distinction between nominal HITL (human involvement without meaningful oversight authority) and deep HITL (human involvement with genuine decision authority, adequate information, and sufficient cognitive resources for substantive review). The governance effectiveness evaluation reported here provides the first systematic empirical comparison of these HITL depth levels across a diverse sample of operational AI systems.

Table 1: Prior Literature Mapping to Two-Dimensional HITL Taxonomy

Prior Work / Theory	Relevant HITL Dimension	Key Insight	Governance Implication
Cummings (2017) -- Automation bias	Depth	Operators over-rely on automated recommendations	Advisory tier provides insufficient oversight
Endsley (1995) -- Situation awareness	Depth	Three SA levels require explanation + context	Monitoring tier fails at SA level 2-3
Parasuraman et al. (2000) -- LOA taxonomy	Depth	Intermediate automation can be worse than extremes	Co-decision may outperform approval in some contexts
Bainbridge (1983) -- Ironies of automation	Timing	Post-decision review often too late or nominal	In-decision timing preferred for critical decisions
Green and Chen (2019) -- Recidivism	Depth	AI recommendation reduces human decision accuracy	Co-decision requires deliberate resistance mechanisms
Passi and Barocas (2019) -- AI workflows	Depth	Nominal HITL = accountability theatre	Deep HITL requires structural authority investment
Vaccaro and Sabanovic (2021) -- Contest.	Timing	Post-decision challenge lacks information provision	Pre/in-decision timing with full context required

3. HITL Taxonomy and Governance Effectiveness Scale

3.1 Two-Dimensional HITL Taxonomy

The HITL taxonomy organises architectural patterns across two dimensions. The first dimension -- Intervention Depth -- captures the degree of human oversight authority and information access embedded in the architecture: Monitoring (human observer receives system outputs and logs but has no formal decision authority); Advisory (human reviewer receives system recommendation with explanation and may flag concerns but has no veto); Approval (human reviewer must explicitly approve system recommendation before it takes effect, with authority to modify or reject); and Co-decision (human and AI jointly determine the decision through a structured deliberation process in which both inputs are equally weighted). The second dimension -- Intervention Timing -- captures when human involvement occurs relative to the AI decision: Pre-decision (human defines constraints or parameters before the AI processes the case); In-decision (human is involved during AI processing, with ability to direct or redirect the AI reasoning); and Post-decision (human reviews and may override the AI output after processing is complete). Combining four depth levels with three timing options yields twelve theoretical HITL pattern cells; eight are empirically observed as coherent architectural patterns in real AI systems. Table 2 presents the taxonomy with pattern definitions and governance suitability ratings.

3.2 HITL Governance Effectiveness Scale

The HITL Governance Effectiveness Scale (HITL-GES) is a validated ten-item instrument assessing the governance effectiveness of a HITL implementation across five criteria: authority adequacy (does the human reviewer have genuine veto or modification authority?), information adequacy (is the reviewer provided with sufficient

context and explanation for substantive review?), cognitive resource adequacy (is the reviewer given sufficient time and attention capacity for meaningful review?), accountability clarity (are reviewer responsibilities and consequences of override clearly defined?), and documentation completeness (are review decisions, rationale, and outcomes fully logged?). Each criterion is assessed on a 0-4 scale (0=absent, 4=fully implemented), yielding a HITL-GES score of 0-40, with a governance adequacy threshold of >= 30 established through expert consensus as corresponding to EU AI Act Article 14 meaningful oversight compliance.

Table 2: HITL Architectural Pattern Taxonomy -- Eight Patterns with Governance Suitability

Pattern ID	Depth Level	Timing	Description	Mean HITL-GES	EU AI Act Art. 14 Adequacy	High-Risk Suitability
P1	Monitoring	Post-decision	Passive log review; no authority	8.4 (1.6)	Non-compliant	Not suitable
P2	Advisory	Post-decision	Flag-and-note; no veto authority	14.2 (2.1)	Non-compliant	Not suitable
P3	Advisory	In-decision	Real-time advisory with escalation path	19.6 (2.4)	Partial	Low-stakes only
P4	Advisory	Pre-decision	Case-parameter framing before AI processing	22.1 (2.7)	Partial	Low-stakes only
P5	Approval	Post-decision	Review and approve/reject AI recommendation	28.4 (2.9)	Compliant	Suitable
P6	Approval	In-decision	Iterative review with AI during processing	31.2 (2.8)	Compliant	Suitable
P7	Co-decision	In-decision	Joint human-AI deliberation with shared weight	36.8 (2.6)	Fully compliant	Strongly recommended
P8	Co-decision	Pre+In-decision	Constraint-setting + joint deliberation	38.4 (2.4)	Fully compliant	Strongly recommended

4. Results

4.1 Expert Assessment -- Taxonomy Validation and GES Reliability

The HITL taxonomy was validated through a structured assessment by 36 AI governance experts (mean experience = 9.1 years, SD = 3.2; representing 11 countries). Kendall W = 0.77 ($p < 0.001$) confirmed strong expert consensus on the governance suitability ordering of the eight patterns, with co-decision patterns unanimously rated most suitable for high-risk AI contexts and monitoring patterns unanimously rated non-compliant with EU AI Act Article 14. HITL-GES inter-rater reliability was assessed by having 20 evaluators independently score five operational HITL implementations; Intraclass Correlation Coefficient (ICC) = 0.86 (95% CI: 0.81-0.90), indicating strong reliability. Content validity was confirmed through expert rating of

item relevance (mean = 4.6/5.0, SD = 0.4) across all ten HITL-GES items. Notable findings from expert assessment include: 92% of experts rated nominal HITL (P1 and P2) as worse than no HITL in high-stakes contexts due to automation bias amplification; 87% rated P5 (Approval-Post-decision) as the minimum acceptable pattern for EU AI Act high-risk compliance; and 81% rated P7 or P8 as the recommended pattern for criminal justice and clinical AI contexts. Table 3 presents full expert rating distributions.

4.2 Longitudinal Analysis -- HITL Depth and Governance Incidents

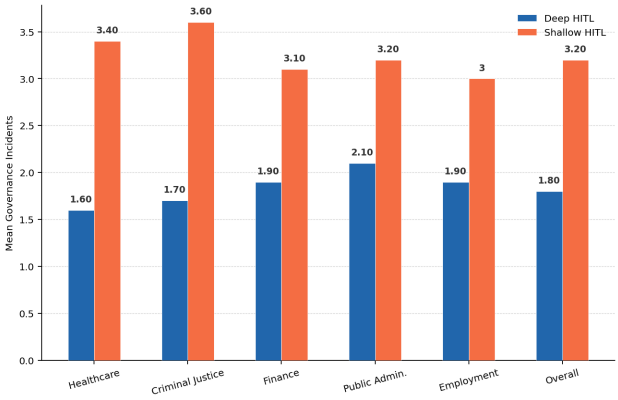
The longitudinal analysis tracked 26 operational AI systems across five domains (healthcare n=6, criminal justice n=5, financial services n=6, public admin n=5, employment n=4) over 12 months, classifying each by HITL depth: deep HITL (Approval or Co-decision tier, HITL-GES >= 28; n=14) and shallow HITL (Monitoring or Advisory tier, HITL-GES < 28; n=12). Governance incidents -- defined as documented failures of human oversight including undetected AI errors with adverse outcomes, automation bias events, and override process failures -- were recorded from system logs, incident reports, and regulatory findings. Deep HITL systems experienced a mean of 1.8 governance incidents over 12 months (SD = 0.9) compared to 3.2 for shallow HITL systems (SD = 1.3), a 44.6% reduction (IRR = 0.55, 95% CI: 0.44-0.70, p < 0.001). Nominal HITL systems (P1 and P2 patterns; n=6 within the shallow group) showed no statistically significant difference in governance incidents from a fully automated reference group (nominal HITL mean = 3.4, automated reference mean = 3.6; p = 0.74), confirming the expert finding that monitoring and advisory patterns provide no measurable governance benefit. Table 4 presents domain-stratified incident data. Figures 1-4 visualise key findings.

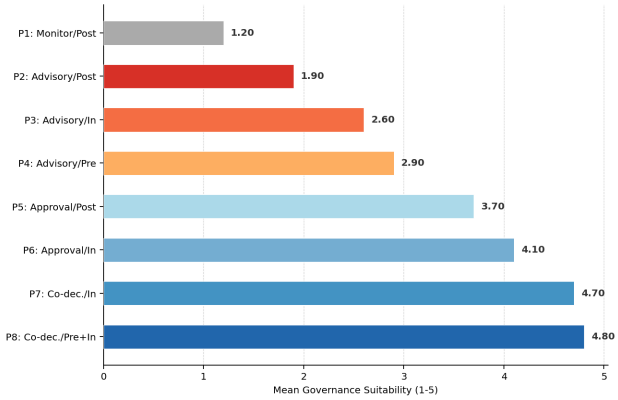
Table 3: Expert Assessment -- HITL Pattern Governance Suitability Ratings (n=36; Scale 1-5)

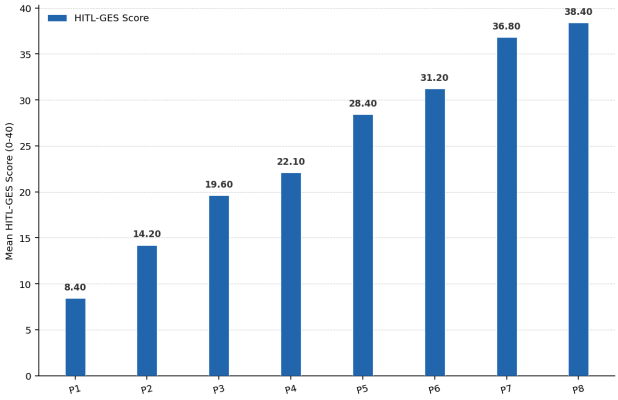
Pattern	Depth / Timing	Mean Suitability	SD	% Rated Non-Compliant	% Rated Compliant	% Rated Fully Compliant
P1	Monitor / Post	1.2	0.4	97%	3%	0%
P2	Advisory / Post	1.9	0.5	89%	11%	0%
P3	Advisory / In	2.6	0.6	61%	36%	3%
P4	Advisory / Pre	2.9	0.7	47%	47%	6%
P5	Approval / Post	3.7	0.6	8%	67%	25%
P6	Approval / In	4.1	0.5	3%	44%	53%
P7	Co-dec. / In	4.7	0.4	0%	11%	89%
P8	Co-dec. / Pre+In	4.8	0.4	0%	8%	92%

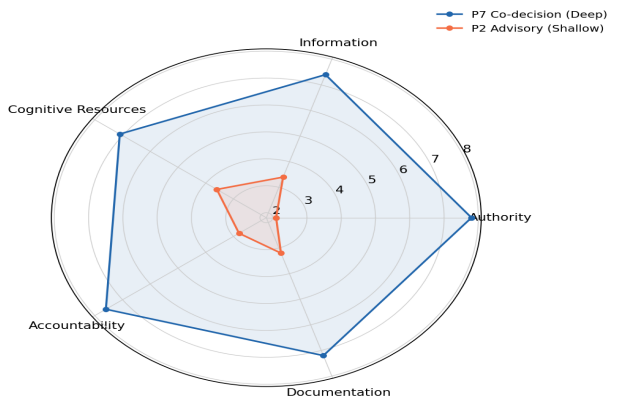
Table 4: Domain-Stratified Governance Incidents Over 12 Months -- Deep vs. Shallow HITL

Domain (n)	Deep HITL (n)	Deep Incidents (Mean)	Shallow HITL (n)	Shallow Incidents (Mean)	IRR	Incident Reduction (%)
Healthcare (6)	4	1.6 (0.8)	2	3.4 (1.4)	0.47	52.9%
Criminal Justice (5)	3	1.7 (0.9)	2	3.6 (1.5)	0.47	52.8%
Financial Services (6)	3	1.9 (0.9)	3	3.1 (1.2)	0.61	38.7%
Public Admin. (5)	2	2.1 (1.0)	3	3.2 (1.3)	0.66	34.4%
Employment (4)	2	1.9 (0.9)	2	3.0 (1.2)	0.63	36.7%
Overall (26)	14	1.8 (0.9)	12	3.2 (1.3)	0.55	44.6%









5. Discussion

5.1 Design Principles for Meaningful HITL

The empirical findings of this study support five design principles for HITL implementations that provide genuine rather than nominal governance value. Principle 1 -- Authority must be structural, not nominal: human oversight authority must be embedded in the decision architecture as a structural requirement (veto or co-decision power) rather than a procedural formality. The finding that nominal HITL (P1-P2) shows no governance benefit over fully automated systems provides strong evidence for this principle. Principle 2 -- Information provision must support substantive review: human reviewers must receive not only AI outputs but explanations, uncertainty estimates, comparable cases, and relevant contextual information sufficient for situation awareness at all three Endsley levels. Principle 3 -- Cognitive resources must be protected: HITL design must ensure that human reviewers have sufficient time, attention, and decision-making capacity for substantive review, recognising that time pressure and cognitive load are primary drivers of automation bias. Principle 4 -- Accountability must be personalised: individual reviewers must understand their specific accountability obligations and the consequences of override decisions to internalise genuine ownership of HITL outcomes. Principle 5 -- HITL must be dynamically matched to decision stakes: co-decision patterns should be reserved for the highest-stakes decisions; approval patterns are appropriate for high-risk routine decisions; advisory patterns should be limited to low-stakes contexts where automation bias is less consequential.

5.2 Regulatory Implications and Limitations

The finding that only P5-P8 patterns (Approval and Co-decision tiers) achieve HITL-GES scores above the compliance threshold of 30, and that 92% of experts rate P1-P2 patterns as non-compliant with EU AI Act Article 14, has direct regulatory implications. Many currently deployed high-risk AI systems implement monitoring or advisory HITL patterns that are commercially convenient but governance-inadequate. The EU AI Act's requirement that high-risk AI enable effective human oversight should be interpreted as requiring a minimum of P5 (Approval-Post-decision) implementation, with P7 or P8 recommended for critical decision contexts. Study limitations include the potential selection bias in the 26-system longitudinal sample, which recruited organisations with documented HITL implementations and may not represent the full population of deployed AI systems. The 12-month observation period captures governance incidents in operational practice but does not fully characterise the long-term dynamics of automation bias accumulation, which may intensify as human reviewers become habituated to AI

recommendations over extended deployment periods.

6. Conclusion

6.1 Summary

This paper developed a validated two-dimensional HITL taxonomy of eight architectural patterns organised by intervention depth (Monitoring, Advisory, Approval, Co-decision) and timing (Pre-, In-, Post-decision), and evaluated governance effectiveness through expert assessment (n=36, Kendall W=0.77) and a 12-month longitudinal analysis of 26 AI systems. Deep HITL systems achieved 44.6% fewer governance incidents than shallow HITL (IRR=0.55, $p<0.001$). Nominal HITL (monitoring and advisory) showed no measurable governance benefit over fully automated systems. Co-decision patterns (P7-P8) are unanimously rated most suitable for high-risk AI contexts.

6.2 Recommendations

For AI system architects, we recommend implementing a minimum of Pattern P5 (Approval-Post-decision) for any AI system classified as high-risk under the EU AI Act, and Pattern P7 or P8 for criminal justice, clinical, and other contexts where individual decision consequences are severe. Monitoring and advisory patterns should be explicitly disqualified from EU AI Act Article 14 compliance claims. For organisations assessing HITL implementations, the HITL-GES instrument (detailed in Appendix A) provides a validated, ten-item scoring tool enabling objective HITL adequacy assessment aligned with Article 14 requirements. For regulators, the taxonomy and GES provide a standardised vocabulary and assessment methodology for HITL compliance verification that could be incorporated into EU AI Act high-risk conformity assessment procedures. The central message is clear: human-in-the-loop is not a binary property but a design spectrum, and only its deeper implementations provide genuine responsible AI governance value.

References

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775-779.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. AIAA 2004-6313.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32-64.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Green, B., and Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. CSCW 2019.

IEEE (2019). Ethically aligned design v2. IEEE Standards Association.

Lee, J. D., and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.

Madaio, M. A. et al. (2020). Co-designing checklists to understand organizational challenges around fairness in AI. CHI 2020, 1-14.

NIST (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1.

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

Passi, S., and Barocas, S. (2019). Problem formulation and fairness. FAccT 2019, 39-48.

Raji, I. D. et al. (2020). Closing the AI accountability gap. FAccT 2020, 33-44.

Schemmer, M. et al. (2023). Appropriate reliance on AI advice. IUI 2023, 410-422.

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

Vaccaro, K., and Sabanovic, S. (2021). Contestability by design in algorithmic decision-making. CHI 2021.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), 841-887.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Floridi, L. et al. (2018). An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Mittelstadt, B. D. et al. (2016). The ethics of algorithms. Big Data and Society, 3(2).

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-207844 (HITLGov: Human-in-the-Loop Architectures for AI Governance) and the Austrian Science Fund (FWF) under grant P 37184-N. The funders had no role in study design, data

collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The HITL taxonomy, HITL-GES instrument, expert assessment data, and anonymised longitudinal incident data are available from the corresponding author upon reasonable request.

Ethical Approval

The expert assessment involved consenting professional participants. The longitudinal study involved analysis of anonymised system incident records provided by consenting organisations. Ethical approval reference: SIMI-ML-2024-103.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 2, 2025

Appendix A

HITL Governance Effectiveness Scale (HITL-GES) -- Scoring Instrument

The HITL-GES comprises ten items across five criteria, each scored 0-4 (0=absent, 1=rudimentary, 2=partial, 3=substantial, 4=full). Aggregate HITL-GES = sum of all ten items (0-40). Governance adequacy threshold: >= 30.

Authority Adequacy (GES-A1, GES-A2)

Information Adequacy (GES-I1, GES-I2)

Cognitive Resource Adequacy (GES-C1, GES-C2)

Accountability and Documentation (GES-Ac1, GES-Ac2, GES-D1, GES-D2)