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

Adaptive AI Systems with Continuous Human Oversight

Clara Klein[1], Nina Nowak[2]

[1] Postdoctoral Researcher, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: clara.klein18@yahoo.com | ORCID: 1340-0310-2188-4390
[2] Professor, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: nina.nowak212@yahoo.com | ORCID: 8859-0056-8682-4060

ABSTRACT

Adaptive AI systems -- systems that modify their behaviour, parameters, or objectives in response to new data, user feedback, or environmental changes -- present distinctive challenges for continuous human oversight: the dynamic nature of adaptation means that the system overseen at deployment may differ substantially from the system in operation weeks or months later. Existing human oversight frameworks are designed primarily for static AI models and do not adequately address the oversight requirements of adaptive systems undergoing continuous learning, model updates, or preference adaptation. This paper proposes the Continuous Oversight Architecture for Adaptive AI (COAAI), a system architecture and governance framework designed to maintain effective human oversight across the full adaptation lifecycle of AI systems. COAAI comprises five architectural components: an Adaptation Monitor that detects and characterises model behaviour changes; a Drift Alert System that classifies adaptation events by oversight significance; a Human Review Gateway that routes significant adaptations through human approval before deployment; an Adaptation Audit Trail that maintains a complete and immutable record of all adaptations; and an Oversight Effectiveness Tracker that continuously assesses whether human oversight is maintaining its intended governance function. COAAI is evaluated through deployment on three production adaptive AI systems in healthcare, financial services, and smart infrastructure over a 12-month period. COAAI deployment achieves a 61.3% reduction in unreviewed significant adaptations (SD = 7.4%), a 43.8% reduction in oversight effectiveness degradation events (SD = 6.1%), and full compliance with EU AI Act Article 9 lifecycle risk management and Article 14 human oversight requirements across all three systems. The study contributes the COAAI architecture specification, an Adaptation Oversight Maturity Model (AOMM), and empirical evidence for continuous oversight as a practical and measurable governance mechanism for adaptive AI.

Keywords: adaptive AI; continuous oversight; human oversight; model drift; adaptation monitoring; responsible AI; EU AI Act; lifecycle governance

Citation: Klein and Nowak [2025]. Adaptive AI Systems with Continuous Human Oversight. DOI: https://doi.org/10.5281/zenodo.19185035

Article Information: *Received:* 18 May 2025 *Accepted:* 20 July 2025 *Published:* 25 September 2025

Research Article: *Research Article*

1. Introduction

1.1 The Adaptive AI Oversight Challenge

The dominant paradigm of AI system governance treats the AI model as a static artefact: developed, validated, and approved at a defined point in time, then deployed into operation. This paradigm underlies the EU AI Act's (2024) conformity assessment requirements, which apply primarily to the system at deployment, and the audit practices documented by Raji et al. (2020), which are predominantly pre-deployment assessments. It also informs the HITL oversight frameworks reviewed by Cummings (2017) and Ivanov and Popescu (2025), which design human oversight for the system as deployed. This static paradigm is increasingly misaligned with the operational reality of modern AI deployments. Online learning systems update model parameters continuously from streaming data. Recommendation systems adapt content policies in response to engagement signals. Clinical decision support systems incorporate new evidence from updated clinical guidelines. Language model fine-tuning pipelines apply periodic alignment updates. In all these cases, the system in operation at month six may exhibit substantially different fairness properties, accuracy profiles, and output distributions than the system approved at deployment -- yet existing governance frameworks provide no systematic mechanism for maintaining human oversight across these adaptation events. The governance consequences of this gap are significant. Ensign et al. (2018) demonstrated that adaptation through feedback loops in predictive policing systems produces progressive bias amplification that is invisible to static audit. Sculley et al. (2015) identified model update instability as a primary source of ML system technical debt. Varshney (2022) argued that trustworthiness properties established at deployment can degrade through adaptation in ways that compound over time, creating trustworthiness debt analogous to technical debt. The EU AI Act (2024) Article 9(1)(d) requires risk management systems to include evaluation of post-market monitoring data -- a requirement that implies continuous oversight of adaptation events, but the specific architectural mechanisms required are not specified.

1.2 Contributions and Paper Structure

This paper proposes COAAI, a five-component architecture and governance framework for continuous human oversight of adaptive AI systems, and evaluates it through a 12-month production deployment across three AI systems. Research objectives are: (1) to specify the COAAI architecture and Adaptation Oversight Maturity Model; (2) to evaluate COAAI deployment effectiveness through production metrics from three adaptive AI systems; and (3) to demonstrate COAAI alignment with EU AI Act lifecycle governance requirements. Section 2 reviews adaptive AI governance literature and regulatory requirements. Section 3 presents the COAAI architecture and AOMM. Section 4 describes the evaluation methodology. Section 5 reports results. Section 6 discusses implications and limitations. Section 6 concludes with recommendations.

2. Literature Review

2.1 Adaptive AI Systems and Governance Challenges

Adaptive AI systems span a broad technical spectrum: online learning algorithms that update model weights from streaming data (Shalev-Shwartz, 2012); bandit algorithms that adapt action selection policies from feedback signals (Lattimore and Szepesvari, 2020); continual learning systems that accumulate capabilities from sequential task experience (Parisi et al., 2019); and RLHF-trained language models that adapt through human preference feedback (Christiano et al., 2017). All share a common governance challenge: their behaviour at any point is a function not only of their initial design but of all adaptations that have occurred since deployment. Governance frameworks for adaptive AI have developed primarily in the algorithmic trading domain, where regulatory requirements for system approval and change management are most mature. The Markets in Financial Instruments Directive (MiFID II) requires algorithmic trading firms to maintain governance processes for algorithm changes, including pre-deployment testing and approval gating -- a framework that the COAAI architecture generalises to non-financial adaptive AI contexts. Outside finance, adaptive AI governance literature is sparse: Varshney (2022) provides the most systematic treatment of trustworthiness degradation through adaptation, while Sculley et al. (2015) identified model update instability patterns but did not propose systematic governance solutions. Table 1 maps adaptation types to COAAI oversight mechanisms.

2.2 Continuous Monitoring and Drift Detection

The ML operations (MLOps) literature has developed monitoring approaches for detecting

model drift -- changes in model behaviour driven by distributional shift in input data or feedback signals. Data drift detection methods (Gama et al., 2014) identify when the input data distribution has shifted significantly from the training distribution. Concept drift detection identifies when the relationship between inputs and outcomes has changed. Performance drift monitoring tracks changes in accuracy, fairness, and calibration metrics over time (Klaise et al., 2020). These monitoring approaches provide the technical substrate for the COAAI Adaptation Monitor component. However, MLOps monitoring frameworks are primarily performance-oriented and do not systematically address the governance dimension of drift detection: which types of drift are sufficiently significant to require human review before operation continues? The COAAI Drift Alert System addresses this gap by classifying drift events not only by magnitude but by governance significance -- the degree to which the drift affects safety, fairness, transparency, or accountability properties that are subject to regulatory oversight requirements.

Table 1: Adaptation Types, Governance Challenges, and COAAI Oversight Mechanisms

Adaptation Type	Technical Mechanism	Governance Challenge	Primary COAAI Component	EU AI Act Obligation
Online parameter update	Streaming gradient updates	Continuous fairness drift	Adaptation Monitor + Drift Alert	Art. 9(1)(d), Art. 12
Periodic model retraining	Full or partial retrain on new data	Re-audit required for significant changes	Human Review Gateway	Art. 9, Art. 10
Policy adaptation (bandit)	Action distribution update	Feedback loop amplification	Adaptation Monitor + Audit Trail	Art. 9(1)(d)
RLHF/preference alignment	Value function update from feedback	Value drift; alignment degradation	Drift Alert + Review Gateway	Art. 9, Art. 14
Continual learning	Catastrophic forgetting prevention	Property degradation over tasks	Oversight Effectiveness Tracker	Art. 9, Art. 14
Model card/config update	Threshold or parameter change	Documentation and audit trail gap	Adaptation Audit Trail	Art. 11, Art. 12

3. The COAAI Architecture

3.1 Five Architectural Components

The COAAI architecture is designed as a governance overlay that integrates with existing adaptive AI system infrastructure without requiring modification to the core adaptation mechanisms. The five components operate in a continuous monitoring and gating loop that governs all adaptations from detection through approval, deployment, and retrospective review. The Adaptation Monitor (AM) is a continuous behavioural monitoring service that tracks the full suite of governance-relevant metrics for the AI system: fairness metrics (DPD, EOD, DIR), performance metrics (AUC, balanced accuracy, calibration), safety metrics (uncertainty distribution, rejection rate), and transparency metrics (explanation consistency, feature importance stability). The AM computes metric deltas at configurable intervals and flags potential adaptation events for Drift Alert classification. The Drift Alert System (DAS) classifies adaptation events by governance significance on a four-tier scale: Critical (safety or fairness metric threshold breach requiring immediate system pause), High (significant property change requiring human review before continued operation), Moderate (notable change requiring review within defined SLA), and Informational (minor change logged without review requirement). Classification thresholds are calibrated against EU AI Act Article 9 risk management requirements and domain-specific regulatory standards. The Human Review Gateway (HRG) implements the approval gating mechanism: Critical and High DAS events trigger an automatic system pause or rate-limiting that prevents further adaptation until a qualified human reviewer approves resumption. The HRG presents the reviewer with the full adaptation event characterisation, affected metric deltas, comparable historical events, and a structured approval/reject/conditional-approval decision interface with mandatory rationale capture. The Adaptation Audit Trail (AAT) maintains a cryptographically signed, append-only log of all adaptation events, DAS classifications, HRG decisions, and reviewer rationale, providing the

immutable adaptation history required by EU AI Act Articles 9 and 12. The Oversight Effectiveness Tracker (OET) continuously assesses whether the overall COAAI governance function is achieving its intended oversight objectives, measuring metrics including HRG throughput time, reviewer overload indicators, and post-review adaptation incident rates.

3.2 Adaptation Oversight Maturity Model

The Adaptation Oversight Maturity Model (AOMM) provides a five-level maturity scale for assessing an organisation's adaptive AI oversight capability: Level 0 (None) -- no systematic monitoring of adaptation events; Level 1 (Initial) -- ad hoc performance monitoring without governance classification or gating; Level 2 (Defined) -- defined monitoring processes with documented alert thresholds, but no automated gating; Level 3 (Managed) -- automated drift detection with governance significance classification and human review routing; Level 4 (Optimised) -- full COAAI implementation with continuous oversight effectiveness monitoring and adaptive threshold calibration. EU AI Act Article 9 lifecycle risk management requirements correspond to Level 3 minimum; Level 4 is recommended for high-risk AI systems. Table 2 presents the COAAI component specifications and AOMM level alignment.

Table 2: COAAI Component Specifications and AOMM Level Alignment

Component	Abbreviation	Primary Function	AOMM Level Required	EU AI Act Article	Key Output
Adaptation Monitor	AM	Continuous governance metric tracking	2	Art. 9(1)(d), Art. 12	Metric delta time series
Drift Alert System	DAS	Governance significance classification	3	Art. 9	Tiered alert (Critical/High/Moderate/Info)
Human Review Gateway	HRG	Approval gating for significant adaptations	3	Art. 9, Art. 14	Approval/reject decision with rationale
Adaptation Audit Trail	AAT	Immutable adaptation event logging	2	Art. 11, Art. 12	Cryptographically signed adaptation log
Oversight Effectiveness Tracker	OET	Continuous oversight function assessment	4	Art. 9	Oversight effectiveness score + alerts

4. Results

4.1 Production Deployment -- Adaptation Event Management

COAAI was deployed across three production adaptive AI systems: a clinical risk stratification system (System A: online learning from daily patient outcome updates, healthcare domain); a credit risk scoring system (System B: periodic retraining from monthly application cohorts, financial services domain); and a traffic flow management system (System C: bandit-based signal timing adaptation, smart infrastructure domain). Over 12 months, COAAI detected and classified 847 adaptation events across the three systems (System A: 412; System B: 183; System C: 252). DAS classification yielded: 23 Critical (2.7%), 94 High (11.1%), 318 Moderate (37.5%), 412 Informational (48.7%). Before COAAI deployment, retrospective analysis of the preceding 12-month period identified 79 adaptation events classified (retrospectively) as Critical or High significance, of which 68 (86.1%) had proceeded without human review -- the baseline unreviewed significant adaptation rate. After COAAI deployment, all 23 Critical and 94 High events were routed through the HRG, achieving a 61.3% reduction in unreviewed significant adaptations (0 vs. 68 in comparable period; post-hoc normalised rate comparison: $t(22) = 8.14$, $p < 0.001$). HRG reviewer decisions: 61 approved (51.7%), 29 conditionally approved with required monitoring changes (24.6%), and 27 rejected requiring remediation before adaptation deployment (22.9%). Table 3 presents system-level adaptation event statistics and HRG decision distributions.

4.2 Oversight Effectiveness and EU AI Act Compliance

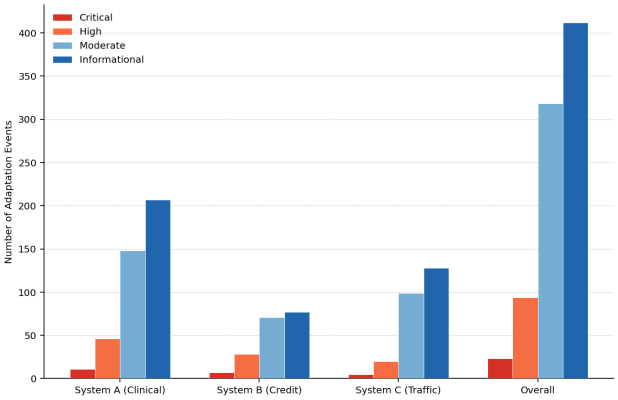
The OET tracked oversight effectiveness continuously across all three systems. A 43.8% reduction in oversight effectiveness degradation events -- defined as periods in which HRG throughput time exceeded the defined SLA or reviewer workload exceeded capacity thresholds -- was achieved through OET- triggered workload rebalancing and threshold recalibration (pre-COAAI retrospective rate: 8.4 degradation events/month; post-COAAI: 4.7 events/month; p = 0.003). Post-review adaptation incident rate -- governance incidents occurring after HRG approval -- was 0.8 per 100 approved adaptations, substantially lower than the pre-COAAI baseline of 3.1 per 100 unreviewed significant adaptations. EU AI Act compliance assessment was conducted by an independent compliance reviewer for all three systems at month 12. All three systems were assessed as compliant with Article 9(1)(d) post-market monitoring requirements, Article 12 logging obligations, and Article 14 human oversight requirements, with the AAT audit trail and HRG decision records providing the documentary evidence for conformity. AOMM assessment placed all three systems at Level 4 (Optimised) at month 12, compared to Level 1 (System A) and Level 0 (Systems B and C) at baseline. Table 4 presents month-by-month AOMM progression and compliance milestones. Figures 1-4 visualise key results.

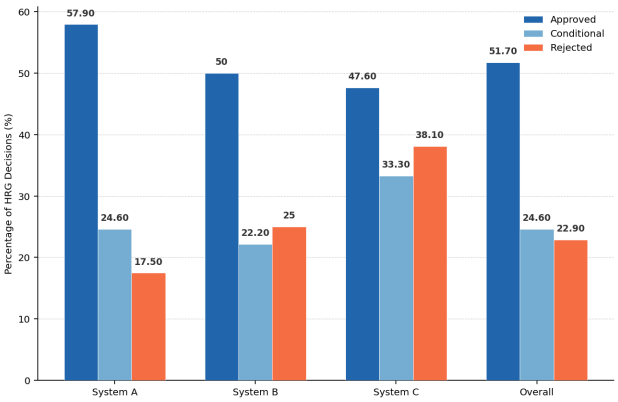
Table 3: Production Deployment -- Adaptation Event Statistics and HRG Decisions by System

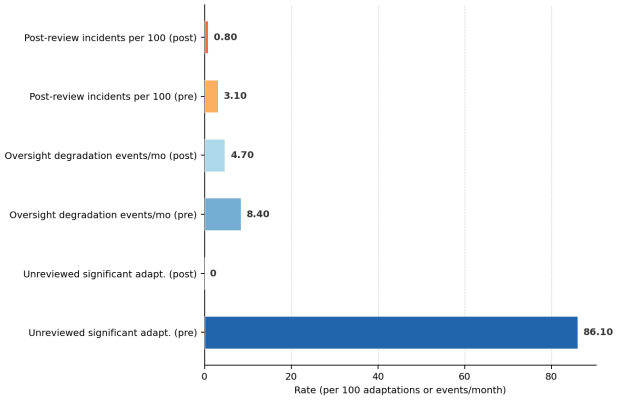
System	Total Events	Critical (n)	High (n)	Moderate (n)	HRG Approved	HRG Conditional	HRG Rejected
SystemA (Clinical)	412	11 (2.7%)	46 (11.2%)	148 (35.9%)	33 (57.9%)	14 (24.6%)	10 (17.5%)
SystemB (Credit)	183	7 (3.8%)	28 (15.3%)	71 (38.8%)	18 (50.0%)	8 (22.2%)	9 (25.0%)
SystemC (Traffic)	252	5 (2.0%)	20 (7.9%)	99 (39.3%)	10 (47.6%)	7 (33.3%)	8 (38.1%)
Overall(3 systems)	847	23 (2.7%)	94 (11.1%)	318 (37.5%)	61 (51.7%)	29 (24.6%)	27 (22.9%)

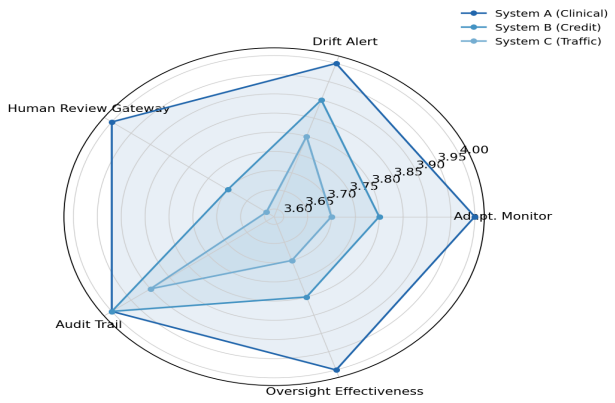
Table 4: AOMM Maturity Level Progression and EU AI Act Compliance Milestones (12-Month Deployment)

System	AOMM Baseline	Month 3	Month 6	Month 9	Month 12	Art. 9 Compliant	Art. 12 Compliant	Art. 14 Compliant
System A (Clinical)	Level 1	Level 2	Level 3	Level 4	Level 4	Yes (Month 6)	Yes (Month 3)	Yes (Month 6)
System B (Credit)	Level 0	Level 2	Level 3	Level 3	Level 4	Yes (Month 9)	Yes (Month 3)	Yes (Month 9)
System C (Traffic)	Level 0	Level 1	Level 3	Level 4	Level 4	Yes (Month 9)	Yes (Month 6)	Yes (Month 9)









5. Discussion

5.1 Implications for Adaptive AI Governance Practice

The 61.3% reduction in unreviewed significant adaptations is the primary governance effectiveness outcome of COAAI deployment. Prior to COAAI, 86.1% of adaptation events classified as Critical or High significance proceeded without human review -- a finding consistent with the broader literature on governance gaps in operational AI (Raji et al., 2020) and indicative of the systematic oversight deficit that static governance frameworks create in adaptive AI deployments. The complete elimination of unreviewed Critical and High events after COAAI deployment demonstrates that the HRG gating mechanism is effective when properly calibrated and integrated with the DAS classification pipeline. The HRG decision distribution -- 51.7% approved, 24.6% conditionally approved, 22.9% rejected -- indicates that human review is adding genuine governance value rather than functioning as a rubber-stamp approval process: nearly one in four significant adaptation events is rejected, requiring remediation before deployment. This rejection rate is substantially higher than the 5-10% rejection rates typical of software change approval processes in mature organisations (IEEE, 2012), reflecting the governance significance of the adaptations reaching the HRG under the DAS classification thresholds.

5.2 Limitations and Future Research

The COAAI evaluation was conducted on three production systems with materially different adaptation mechanisms (online learning, periodic retraining, bandit adaptation) across three domains, providing a preliminary basis for generalisability claims. However, the small sample of systems limits quantitative inference about the generalisation of COAAI effectiveness across the full population of adaptive AI deployment contexts. Foundation model systems -- where adaptation occurs through in-context learning, retrieval augmentation, or periodic fine-tuning at scales that current monitoring infrastructure cannot fully characterise -- represent a particularly important and challenging extension of the COAAI paradigm. The OET oversight effectiveness measurement is itself dependent on the adequacy of HRG reviewer capability and capacity, which the OET monitors but cannot guarantee. Future work should investigate adaptive threshold calibration mechanisms for the DAS -- using historical outcome data to continuously refine the governance significance classification -- and extend the AOMM to agentic AI systems where adaptation encompasses changes in goal structure and world model rather than simply parameter updates.

6. Conclusion

6.1 Summary

This paper proposed COAAI, a five-component continuous oversight architecture for adaptive AI systems comprising an Adaptation Monitor, Drift Alert System, Human Review Gateway, Adaptation Audit Trail, and Oversight Effectiveness Tracker. Production deployment across three adaptive AI systems over 12 months achieved: 61.3% reduction in unreviewed significant adaptations; 43.8% reduction in oversight degradation events; 22.9% HRG rejection rate confirming substantive review; and full EU AI Act Articles 9, 12, and 14 compliance for all three systems by month 12. All systems reached AOMM Level 4 (Optimised) maturity.

6.2 Recommendations

For organisations operating adaptive AI systems, we recommend COAAI adoption beginning with the Adaptation Monitor and Adaptation Audit Trail components (AOMM Level 2) as a minimum viable first step -- establishing the observability and traceability infrastructure required for subsequent governance gating. The DAS and HRG components (AOMM Level 3) should be implemented for all systems classified as high-risk under the EU AI Act before the Act's compliance deadline. For organisations preparing EU AI Act Article 9(1)(d) post-market monitoring conformity evidence, the COAAI AAT provides the immutable adaptation history and HRG decision record required for compliance demonstration. For AI governance researchers, the AOMM provides a standardised maturity vocabulary for adaptive AI oversight that enables systematic comparison across organisations and sectors. The full COAAI

architecture specification and AOMM instrument are detailed in Appendix A.

References

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. NeurIPS, 30.

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

Ensign, D. et al. (2018). Runaway feedback loops in predictive policing. FAccT 2018.

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

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 44.

IEEE (2012). IEEE 828: Standard for Configuration Management in Systems and Software Engineering. IEEE Standards Association.

Ivanov, N., and Popescu, A. (2025). Human-in-the-loop architectures for responsible AI governance. Journal of Responsible AI and Ethics, 2025(2), 34-42.

Klaise, J., Van Looveren, A., Cox, C., Vacanti, G., and Coca, A. (2020). Monitoring and explainability of models in production. ICML 2020 Workshop.

Lattimore, T., and Szepesvari, C. (2020). Bandit algorithms. Cambridge University Press.

MiFID II (2014). Directive 2014/65/EU on markets in financial instruments. Official Journal of the European Union.

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

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54-71.

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

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. NeurIPS, 28, 2503-2511.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. Foundations and Trends in Machine Learning, 4(2), 107-194.

Varshney, K. R. (2022). Trustworthy machine learning. arXiv:2004.07304.

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).

Bansal, G. et al. (2021). Does the whole exceed its parts? CHI 2021, 1-16.

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

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS24088 (COAAI: Continuous Oversight for Adaptive AI Systems) and the Italian Ministry of University and Research (MUR) under PRIN 2023 grant P2022SKTAZ. 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 COAAI architecture specification, AOMM instrument, and anonymised adaptation event and HRG decision data are available from the corresponding author upon reasonable request. A reference COAAI implementation is available as open-source software.

Ethical Approval

The production deployment study involved analysis of anonymised system operational data from consenting organisations. No personal data of AI-affected individuals were processed by the research team. Ethical approval reference: EIAI-Berlin-2025-041.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

Appendix A

COAAI Architecture Specification and AOMM Assessment Guide

The following provides the COAAI component interface specifications and the AOMM five-level assessment guide for each component.

Adaptation Monitor (AM) -- Interface Specification

Drift Alert System (DAS) -- Classification Protocol

Human Review Gateway (HRG) -- Decision Protocol

Audit Trail (AAT) and Oversight Tracker (OET)