Adaptive Governance through AI Prediction Errors: Bridging World Model Adaptation and Policy Innovation

Naoyasu Yoshimura
National Graduate Institute for Policy Studies (GRIPS)
Japan
DOC24056@grips.ac.jp,naoysmr1029@outlook.jp

ABSTRACT

As AI systems increasingly mediate high-stakes decisions, adaptive governance is vital for balancing innovation and risk. Current frameworks like the EU AI Act rely on static risk models that struggle with AI's evolving nature. This paper proposes Adaptive Transparency—a novel approach using fluctuations in AI prediction errors as real-time regulatory signals. Instead of revealing full algorithms, it discloses error trends, enhancing adaptability while protecting proprietary information. From both AI governance and HCI perspectives, Adaptive Transparency enables dynamic risk monitoring, proactive intervention, and user-centered transparency tools. It promotes public trust and innovation. Future research should empirically assess the impact of prediction-error-driven transparency on trust calibration, regulatory compliance, and AI accountability.

CCS CONCEPTS

• **Social and professional topics** → **Technology governance**; • **Computing methodologies** → *Artificial intelligence*; • **Human-centered computing** → *User interface design.*

KEYWORDS

Prediction Error, Adaptive Governance, AI Transparency, Human-AI Interaction, Risk Monitoring, Sociotechnical Systems

ACM Reference Format:
Naoyasu Yoshimura. 2025. Adaptive Governance through AI Prediction Errors: Bridging World Model Adaptation and Policy Innovation. In *Proceedings of First Workshop on Sociotechnical AI Governance (STAIG@CHI'25).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

AI models, particularly those utilizing world models, continuously update their internal representations of external environments. However, unexpected real-world dynamics—environmental shifts, adversarial input, or emerging behaviors—can trigger prediction error surges. Such surges are valuable signals of risk, misalignment, or drift of the model. While frameworks such as the EU AI Act emphasize static, risk-based governance, these structures struggle to accommodate AI's evolutionary and context-sensitive nature. This paper proposes an adaptive governance framework that leverages prediction error fluctuations as socio-technical regulatory signals. These signals serve as input to dynamic human-AI regulatory processes that balance accountability, transparency, and innovation. This paper explores the policy and HCI implications of prediction-error-driven governance, emphasizing adaptive transparency, the human-AI interaction for adaptive regulation, and the adaptive role of the world model as a governance tool.

STAIG@CHI'25, April 27, 2025, Yokohama, Japan

ACM ISBN 978-1-4503-XXXX-X/25/04
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORKS

Traditional AI governance frameworks, such as the EU AI Act, rely on static compliance models, but adaptive regulation is essential for AI systems that continuously evolve (Reuel & Undheim, 2024; Janssen, 2025)[5, 8]. A promising approach to enabling adaptive governance involves focusing on AI prediction errors. According to the Free Energy Principle (Friston, 2010)[4], AI systems minimize prediction errors to optimize decision making, and monitoring these fluctuations can provide insight into system uncertainty and emerging risks (Bereska & Gavves, 2024; Zeng et al., 2024)[1, 10]. Research in human-computer interaction (HCI) has explored how the representation of uncertainty and adaptive trust calibration can support decision-making in AI-assisted systems (Okamura & Yamada, 2020)[6], providing a foundation for designing regulatory transparency tools. Recent HCI studies also demonstrate that user-facing explanations can improve trust and interpretability (Ribeiro et al., 2016; Doshi-Velez & Kim, 2017)[3, 9], reinforcing the viability of prediction error–based transparency mechanisms.

3 FRAMEWORK: PREDICTION ERROR–DRIVEN GOVERNANCE

This paper proposes an adaptive AI governance framework in which fluctuations in prediction errors serve as regulatory signals to dynamically adjust oversight mechanisms in response to evolving system behavior. For example, in the context of autonomous driving, prediction errors can increase due to weather changes, unfamiliar road conditions, or adversarial input. Such anomalies often indicate model drift, environmental changes, or emerging ethical concerns that warrant timely regulatory responses, such as adjusting risk thresholds or requiring human oversight. Rather than disclosing all the details of the algorithms, the proposed framework emphasizes *adaptive transparency*, where AI systems communicate prediction error trends in real time. This enables risk-based governance while preserving corporate confidentiality. Regulatory bodies can establish flexible compliance thresholds based on observed error patterns, while visualization techniques from human-computer interaction (HCI) research help stakeholders intuitively interpret AI uncertainty. Dashboards informed by HCI principles can categorize prediction

errors by risk level and activate human-in-the-loop mechanisms when uncertainty exceeds acceptable limits.

4 IMPLEMENTATION STRATEGY

To operationalize the proposed framework, a multi-tiered governance model should integrate policy mechanisms with principles from human-computer interaction (HCI). In this model, AI systems self-report fluctuations in prediction errors, triggering automated risk assessments aligned with cross-sector thresholds for high-stakes applications. Adaptive dashboards offer user-centric visualizations of uncertainty, enabling stakeholders to intuitively interpret system behavior. When significant anomalies arise, audit mechanisms initiate regulatory intervention. Developers must then dynamically adjust AI models in line with updated risk conditions. On the interface side, predictive design elements—guided by HCI research—communicate AI instability and embed trust calibration mechanisms based on empirical studies of uncertainty representation in decision-support contexts. Transparency tools, co-designed with policymakers and industry professionals, aim to balance usability with accountability. This framework also aligns with recent efforts to ground AI systems in socially constructed human judgments. For example, Chen & Zhang (2023) [2]propose "case law grounding" to align AI behavior with normative expectations in content moderation and legal reasoning. Similarly, this approach treats prediction errors as normative signals indicating misalignments between AI world models and societal expectations. Even without real-time ground truth, simulations can construct error profiles across varied conditions, revealing early warning signs like confidence drops, distributional shifts, or variance spikes. Simulations thus offer actionable insights. In deployment, disclosing prediction errors to regulators or human operators enables context-sensitive responses, surfaces policy-relevant risks, and supports adaptive oversight. Through this feedback loop, prediction errors help monitor AI systems while enhancing institutional trust, accountability, and adaptability.

5 EMPIRICAL EVALUATION AND EXPERIMENTAL DESIGN

To empirically validate the proposed governance framework while ensuring safety, controllability, and inclusive stakeholder participation, I propose to implement and test the full system—comprising prediction error generation, human-in-the-loop intervention, adaptive HCI components, and policy feedback mechanisms—within a simulation environment. This approach enables experimentation in realistic yet risk-free domains such as autonomous driving. In addition, the simulation setting supports the experimentation of the Society-in-the-Loop[7], where users, developers, and policymakers can collaboratively engage in regulatory decisions. Simulation-based governance prototyping is essential not only for technical validation but also for exploring the institutional and behavioral feasibility of adaptive oversight mechanisms. The empirical component of this study follows a mixed-methods research design: (1) Conceptualization involves the development of a theoretical model using loop diagrams and systems modeling; (2) Prototyping applies the framework to autonomous driving scenarios through UI mockups and dashboard design; (3) User Studies investigate how humans respond to AI uncertainty through A/B testing and behavioral logging; (4) meta-learning design adapts the interface based on user behavior using contextual bandits and reinforcement learning simulations; and (5) Policy Integration translates experimental findings into governance tools such as compliance dashboards and institutional guidelines via stakeholder interviews. The experimental evaluation focuses on three aspects: the validity of regulatory signals, the effectiveness of uncertainty displays, and the adaptiveness of HCI components—each assessed through relevant behavioral and system performance metrics. Together, these support the refinement and scaling of prediction-error-driven governance systems.

6 CONCLUSION

This study presents prediction error not merely as a technical metric but also as a dynamic interface between AI systems and governance. By framing prediction error as a shared language among developers, policymakers, and users, the framework promotes interdisciplinary alignment and adaptive policy responses. Participatory mechanisms—such as Society-in-the-Loop feedback, transparency tools, and fairness monitoring—enable more inclusive and accountable oversight. Real-time evaluation using threshold triggers and scenario-based simulations helps governance evolve with AI capabilities. Rather than viewing errors as failures, this approach interprets them as signals of misalignment with societal norms—offering a practical pathway for ethical, adaptive regulation. The framework contributes to AI governance by embedding prediction error monitoring as a form of adaptive transparency, balancing confidentiality with accountability. It also highlights the importance of human-centered design, ensuring that transparency tools are both technically effective and socially relevant. Future research should validate this approach through empirical studies and real-world integration, further establishing prediction error as a socio-technical bridge for resilient and inclusive AI governance.

REFERENCES

[1] L. Bereska and E. Gavves. 2024. Mechanistic Interpretability for AI Safety – A Review. (2024). arXiv:2404.14082 http://arxiv.org/abs/2404.14082

[2] Q. Z. Chen and A. X. Zhang. 2023. Case Law Grounding: Using Precedents to Align Decision-Making for Humans and AI. (2023). arXiv:2310.07019 http://arxiv.org/abs/2310.07019

[3] F. Doshi-Velez and B. Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (2017). arXiv:1702.08608 http://arxiv.org/abs/1702.08608

[4] K. Friston. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11 (2010), 127–138.

[5] M. Janssen. 2025. Responsible governance of generative AI: conceptualizing GenAI as complex adaptive systems. *Policy and Society* (2025). https://doi.org/10.1093/polsoc/puae040

[6] K. Okamura and S. Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLOS ONE* 15, 2 (2020). https://doi.org/10.1371/journal.pone.0229132

[7] Iyad Rahwan. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.

[8] A. Reuel and T. A. Undheim. 2024. Generative AI Needs Adaptive Governance. (2024). arXiv:2406.04554 http://arxiv.org/abs/2406.04554

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144. https://doi.org/10.1145/2939672.2939778

[10] Z. Zeng, C. Zhang, F. Liu, J. Sifakis, Q. Zhang, S. Liu, and P. Wang. 2024. World Models: The Safety Perspective. (2024).