Institute for Youth in Policy

5214F Diamond Heights Blvd #3055
San Francisco, CA 94109
+1 (715) 469-6884
fellowship@yipinstitute.org
www.yipinstitute.org/fellowship/

Fellowship Capstone | Policy Brief

Black Box Models and the Governance of Mechanistic Science

Bhavana Rupakula

I. Executive summary

As black box models increasingly drive decision-making in science and policy, there are rising concerns in their reliability and ethics in high stakes environments, particularly science research. This brief suggests that governance of mechanistic science in the age of AI requires rethinking the role of transparency, suggesting a hybrid governance model.

II. Overview

In 1956, two years after the death of computer science pioneer Alan Turing, John McCarthy, a professor at Dartmouth College [1], organized a summer workshop about thinking machines - 'artificial intelligence'. That convening launched decades of innovation centered on replicating, and even surpassing, human cognition using machines.

Today, the most powerful AI models, particularly 'black box' systems, are re-shaping scientific discovery but also governance and policy making. These models, while phenomenal in their abilities, lack the ability to explain their decisions. This contradiction affects the most basic principle of mechanistic science [2].

This issue leads to an interesting roadblock in policy making, where-in leaders have to decide between innovation and transparency.

A. Relevance

Black box models are widespread in critical sectors like finance, healthcare, and autonomous vehicles, showcasing high accuracy in complex tasks. It is essential to understand how these models operate and the rationale behind their decisions, yet their opaque nature makes this challenging [3].

In high-stakes areas such as healthcare and criminal justice, these models can exacerbate biases, leading to serious risks. This lack of transparency creates a reproducibility crisis, making it difficult for researchers to audit their logic or verify claims. While the General Data Protection Regulation (GDPR) [4]. provides a right to explanation for automated decisions, governance of these technologies remains insufficient and needs urgent attention.

III. History

A. Current Stances

Today, black box models are increasingly used in various scientific fields, such as genomics, neuroscience, climate modeling, and drug discovery, due to their ability to handle large datasets and make highly accurate predictions. However, this dependence on opaque algorithms has sparked a growing debate among scientists about their fit with the principles of mechanistic science.

One major issue is the lack of interpretability and transparency. Cynthia Rudin, a professor at Duke

University and a supporter of interpretable AI, warns that [5] "There is a huge cost to using black box models in high-stakes decisions when interpretable models exist and perform equally well." Many researchers believe that black box models should not be implemented when human health, legal accountability, or policy formation is at stake unless their inner workings can be understood and explained.

The reproducibility crisis is worsened by reliance on black box models. A 2019 report from the National Academies of Sciences [6] pointed out that "computational reproducibility is increasingly difficult in modern science due to algorithmic opacity, data unavailability, and lack of standardization in modeling workflows." In these situations, the failure to replicate findings damages public trust in science and slows down progress.

Some people defend black box models by highlighting their practical benefits. Jeff Dean, Google's Chief Scientist for AI [7] has said, "We care about performance, and if the best-performing model is a black box, we'll use it—but we're actively investing in ways to make them more interpretable." This practical view is shared by many industry researchers, but it remains controversial in academic science, where explanation and understanding of cause-and-effect relationships are seen as essential. The conflict lies not in whether black box models are useful—they are—but in how science can manage their use responsibly without losing transparency, reliability, and accountability.

IV. Policy Problem

A. Stakeholders

There are a range of stakeholders involved in the black box model dilemma, with very differing goals [8]. Scientists and regulators are accustomed to casual clarity through mechanistic frameworks, and they expect similar explanatory standards when science intersects with technology. Indiana's flagship institutions like Indiana University, Purdue University, and the University of Notre Dame are increasingly deploying AI to accelerate biomedical, agricultural, and engineering discoveries.

Other important stakeholders include school districts: AI in education and work can affect performance; local governments - cities across the state like Fishers and Zionsville are implementing smart-city initiatives; and the general public - AI can monitor every aspect of their lives.

B. Risks of Indifference

Lack of supervision or regulation of black box model application in science and civic contexts could reinforce structural inequalities and erode institutional trust. In health science, for instance, IU and Purdue researchers now rely more and more on deep learning models to analyze patient data—but with no explainability, the derived information may infuse and pass along racial or socioeconomic bias.

C. Nonpartisan Reasoning

Mechanistic governance shows this through its strict data handling and verification protocols. AI risk assessment models also demonstrate the importance of adaptive control. By combining these methods, hybrid governance can create common ground that crosses political divides. Regulatory strategies that change based on system risk, instead of ideological beliefs, encourage innovation while ensuring safety:

1) **Risk-Based Governance Fits Both Technocratic and Libertarian Values**

Across different political beliefs, many support evidence-based, risk-calibrated policy frameworks. Centrists and technocrats prefer structured regulations that safeguard public welfare without hindering innovation. At the same time, libertarian-leaning stakeholders often oppose heavy government intervention but agree to minimal, focused oversight where real risks exist. A governance strategy that adjusts requirement and oversight intensity based on system risk (like critical healthcare versus customer support bots) meets both sides' needs: it protects citizens while reducing unnecessary bureaucratic burden. By applying risk-based thresholds instead of a one-size-fits-all approach, policy avoids ideological conflicts over government size and reach.

2) **Public Trust Is a Shared Concern Across the Political Spectrum**
Rebuilding and maintaining public trust in science and technology is a bipartisan issue, especially in a time of growing skepticism toward institutions. Governance that demands hybrid AI-mechanistic systems to be auditable, explainable, and accountable earns legitimacy from both sides. Transparent systems also support democratic oversight and media attention, making it easier for everyone—regardless of their beliefs—to hold developers and regulators responsible.
3) **Science and Innovation Should Remain Neutral Amid Political Conflicts**
Keeping science neutral and credible is vital for social stability and long-term success. When unclear AI systems lead to mistakes or unfairness, they risk not just causing harm but also eroding trust in the scientific process, which can become a political tool. A well-designed hybrid governance framework protects science from polarization by maintaining consistent standards of transparency and rigor, regardless of which party is in power. This way, both conservative and progressive governments can rely on the same system to assess new technologies. Standardized governance also promotes global regulatory cooperation, boosting international competitiveness—a goal shared by most political views.

V. Tried Policy

In 2024, the Indiana Senate introduced Senate Bill 150 [9], which established the Indiana Artificial Intelligence Task Force. The force is responsible for evaluating the risks and benefits of AI for Hoosiers as well as recommending guidelines for state bodies' use of AI. Similarly, the Management Performance Hub (MPH) [10] shared a webpage which contained a comprehensive overview of the enterprise-level policy governing the use of AI within the state government. This policy was issued and is being monitored by the Office of the Chief Data Officer (OCDO), Chief Privacy Officers (CPO), and MPH.

Indiana University and Purdue University have established AI ethics committees, such as IU's Center for Bioethics and Purdue's Institute for Physical Artificial Intelligence (IPAI) [11]. However, these efforts are mostly advisory and lack enforcement.

VI. Policy Options

1. Epistemic-Layered Explanation

This option suggests a tiered approach. Lower-stakes systems can rely on statistical validation, while high-stakes systems, like medical diagnostics or prescribing tools, must meet stricter standards for interpretation. This may include requiring causal models along with predictive algorithms, ensuring uncertainty quantification, and integrating counterfactual query abilities. This way, regulators can make sure stakeholders receive explanations that match the potential societal impact.

2. Tiered Institutional Oversight

To put this approach into practice, we propose creating Epistemic Review Boards. These would work like Institutional Review Boards (IRBs) but focus on hybrid mechanistic-AI systems. This would involve interdisciplinary panels that include AI scientists, domain experts, ethicists, and public representatives who assess whether a system's interpretability and performance meet specific criteria. In this model, getting system approval balances scientific integrity with the need for technological advancement.

3. Formal Documentation & Provenance

Taking cues from the IPCC and open science movements, this option calls for required documentation of model cards, data lineage, drift tracking, and interpretability techniques, along with compliance checks. Instead of vague statements about transparency, developers would need to clearly explain how models justify decisions, how they track performance over time, and how they address biases. These reports would be available to the public, promoting accountability and independent review.

4. Standards & Certification

Building on existing frameworks like IEEE-USA and NIST RMF, we recommend expanding the standards to include interpretability markers tailored for biomedical applications. Organizations looking to implement biomedical AI would need to get certification similar to medical device approval, confirming they meet interpretability benchmarks, uncertainty requirements, and fairness reviews. Such a framework could match the FDA's pathways for investigational devices, providing clear regulations.

5. Adaptive & Hybrid Governance

Finally, the governance of hybrid scientific systems should be developed through adaptive, hybrid models that combine state regulation, professional standards, and civil society oversight. Similar to medical regulatory "regulatory sandboxes," these frameworks allow for limited, supervised deployment to collect real-world data and support ongoing policy development. Meanwhile, public discussions, such as citizen panels or expert-stakeholder committees, can help establish guidelines for responsible deployment and implicit social norms.

VII. Conclusions

In hybrid domains, a one-size-fits-all approach does not work. Interpreting AI requires clear causation, but strict mechanical standards can stifle innovation. Governance needs to be more inclusive, matching explanation expectations with system risks and the specific context of the domain. To make this change, we support hybrid strategies that combine levels of interpretability, oversight at the institutional level, documentation standards, and certification processes. Starting in biomedical fields can create case studies that help

develop governance on a larger scale. This layered approach seeks to improve transparency, build public trust, and encourage innovation at the intersection of mechanical science and AI systems.

Acknowledgment

The Institute for Youth in Policy wishes to acknowledge Mason Carlisle, Lilly Kurtz, Asher Cohen, Paul Kramer. and other contributors for developing and maintaining the Fellowship Program within the Institute.

References

[1] Lawrence Livermore National Laboratory. (n.d.). *The birth of Artificial Intelligence (AI) research | Science and Technology*. Science and Technology. Retrieved July 15, 2025, from https://st.llnl.gov/news/look-back/birth-artificial-intelligence-ai-research

[2] Craver, C., Tabery, J., & Illari, P. (2015, November 18). *Mechanisms in Science (Stanford Encyclopedia of Philosophy)*. Stanford Encyclopedia of Philosophy. Retrieved July 16, 2025, from https://plato.stanford.edu/entries/science-mechanisms/

[3] Forbes Technology Council - Expert Panel. (2025, May 14). *Improving Transparency In Black Box AI: Expert Strategies That Work*. Forbes. https://www.forbes.com/councils/forbestechcouncil/2025/05/14/improving-transparency-in-black-box-ai-expert-strategies-that-work/

[4] General Data Protection Regulation. (n.d.). *Art. 22 GDPR – Automated individual decision-making, including profiling - General Data Protection Regulation*. GDPR. Retrieved June 26, 2025, from https://gdpr-info.eu/art-22-gdpr/

[5] Rudin, C. (2019, May 13). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *National Library of Medicine*, *1*(5), 206-215. https://pmc.ncbi.nlm.nih.gov/articles/PMC9122117/

[6] Fineberg, H., Stodden, V., & Meng, X.-L. (2020). Highlights of the US National Academies Report on "Reproducibility and Replicability in Science." *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.cb310198

[7] Perkel, S. (2025, May 19). *Google chief scientist predicts AI could perform at the level of a junior coder in a year*. Yahoo! Tech. Retrieved June 26, 2025, from https://tech.yahoo.com/ai/articles/google-chief-scientist-predicts-ai-084702595.html

[8] Adler, P., Falk, C., Friedler, S., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2016). Auditing black-box models for indirect influence. Knowledge and Information Systems, 54, 95 - 122. https://doi.org/10.1007/s10115-017-1116-3.

[9] Artificial Intelligence and Cybersecurity, SB0150, 123th General Assembly. https://legiscan.com/IN/text/SB0150/id/2953970/Indiana-2024-SB0150-Enrolled.pdf

[10] Management Performance Hub. (2025). *State of Indiana Artificial Intelligence Policy and Guidance*. Indiana State Government. https://www.in.gov/mph/AI/

[11] IPAI Steering Committee. (n.d.). *IPAI Steering Committee - Purdue Computes*. Purdue University. Retrieved July 16, 2025, from

https://www.purdue.edu/computes/institute-for-physical-artificial-intelligence/steering-committee/