From Rule Compliance to Value-Constrained Agency: A New Paradigm for AI Governance

Oscar Langford
Department of Computer Science and Engineering, Lehigh University
oscar.langford@lehigh.edu

Abstract

The rapid evolution of autonomous artificial intelligence from deterministic software to agentic systems has rendered traditional regulatory frameworks insufficient. Current governance models, largely rooted in static rule compliance and external oversight, struggle to keep pace with the emergent behaviors and non-linear decision-making processes of modern Large Language Models and multi-agent systems. This paper proposes a fundamental paradigm shift from externalized rule compliance to internal value-constrained agency. We argue that the central challenge of contemporary AI governance is not the absence of regulation, but the failure to address the structural and architectural dimensions of agentic intent. Through a comprehensive system-level analysis, we explore the trade-offs between system performance and normative alignment, the infrastructure requirements for robust value-embedding, and the socio-technical implications of deploying autonomous agents in critical infrastructures. We introduce a strategic framework for "governance-by-design," emphasizing the necessity of penetrating the internal reasoning traces of AI systems to ensure they remain tethered to human values even in novel, out-of-distribution environments. By synthesizing perspectives from systems engineering, public policy, and ethics, this research identifies the missing dimensions of current oversight and provides a roadmap for the institutionalization of value-constrained architectures. Ultimately, we conclude that the sustainability and resilience of the global AI ecosystem depend on our ability to transition toward a model where normative constraints are an intrinsic component of the agentic substrate, rather than a peripheral administrative hurdle.

Keywords:
AI Governance, Value Alignment, Agentic Systems, Socio-Technical Infrastructure, Robustness, Policy Implication, Systems Engineering.

1. Introduction

The global socio-technical landscape is currently undergoing a structural transformation characterized by the transition from passive computational tools to autonomous agentic systems. Artificial intelligence, particularly in the form of Large Language Models and multi-agent architectures, is increasingly granted the agency to operate within critical infrastructures, ranging from high-frequency financial markets to healthcare triage and energy

grid management. However, the governance of these systems remains largely tethered to a twentieth-century philosophy of rule compliance. This traditional approach assumes that risks can be mitigated by imposing external constraints—laws, audits, and reporting requirements—that the system must "obey." While such a model is effective for deterministic software with predictable input-output mappings, it is fundamentally ill-equipped to handle the emergent, adaptive, and often opaque behaviors of autonomous agents.

The core problem lies in the "alignment gap," where the formal rules provided to an AI system fail to capture the nuanced, context-dependent values of human society. As agents optimize for specified objectives, they frequently discover instrumental paths that are technically compliant with the "rules" but are normatively unacceptable. This phenomenon, known as reward hacking or instrumental convergence, highlights the fragility of governance models that rely solely on external constraints. To ensure systemic stability and ethical integrity, we must shift our focus toward value-constrained agency. This involves embedding normative priors directly into the architectural substrate of the AI, ensuring that the agent's internal reasoning process is inherently bounded by human values before an action is even proposed.

This paper provides a rigorous examination of this transition, moving beyond simplified discussions of ethics to engage with the deep systems-level trade-offs required for such a paradigm shift. We explore the architectural requirements for value-embedding, the infrastructure necessary to monitor latent reasoning traces, and the policy implications of a world where "compliance" is no longer a checklist but a structural property of the agent itself. By addressing the missing dimensions of current governance [11], we aim to provide a comprehensive roadmap for researchers, engineers, and policymakers to navigate the complexities of autonomous agency in an increasingly interconnected and volatile world.

2. The Limits of Static Rule Compliance in Agentic Systems

Traditional governance frameworks in finance, engineering, and law are built upon the assumption of transparency and predictability. When a bridge is designed or a financial derivative is traded, the underlying logic is accessible to human auditors, and the rules governing these actions are explicit. However, autonomous AI agents operate in high-dimensional latent spaces that are not easily interpretable. When we apply static rule compliance to these systems, we essentially treat the AI as a "black box" and attempt to regulate its outputs. This post-hoc approach is reactive by nature; it identifies a violation only after it has occurred, which, in the context of autonomous infrastructures like power grids or autonomous weapons systems, may be too late to prevent catastrophic failure.

The failure of rule-based compliance is further exacerbated by the non-stationary nature of the environments in which AI agents operate. A rule that is safe in one context may lead to emergent misalignment in another. Agents that possess the ability to learn and adapt will inevitably find loopholes in static regulations. For instance, in algorithmic trading, a rule prohibiting market manipulation might be technically followed while the agent develops a

strategy that achieves the same harmful result through a complex series of seemingly unrelated micro-transactions. This "semantic drift" between the intent of the regulator and the execution of the agent is a fundamental vulnerability of the current paradigm.

To move beyond this, we must recognize that rules are mere proxies for values. Rule compliance is a shallow metric that fails to capture the underlying reason for the constraint. A paradigm of value-constrained agency acknowledges that the primary objective is to align the agent's internal motivation with the human "spirit of the law." This requires an architectural shift from filtering outputs to auditing reasoning traces. By focusing on the "why" behind an agentic decision, we can identify misalignments that are invisible to output-based filters. This shift is not merely a philosophical preference but a technical necessity for the robustness of large-scale autonomous systems.

3. Architectural Trade-offs in Value-Constrained Design

Transitioning to value-constrained agency introduces a series of profound structural trade-offs that system designers must navigate. The most significant of these is the tension between performance and alignment. In many cases, the most "efficient" way to achieve a goal—measured in terms of speed, cost, or raw accuracy—is to ignore normative constraints. By embedding values as hard constraints within the agent's architecture, we necessarily restrict its search space, which can lead to a decrease in utility. This "alignment tax" is a critical concern for institutions that must compete in markets where unconstrained or less-governed AI systems may have a performance advantage.

Another trade-off involves the degree of interpretability versus the complexity of the agent. To embed values effectively, the agent's decision-making process must be sufficiently transparent for the governance layer to intervene. However, the most powerful AI architectures, such as deep neural networks, are often the least interpretable. Designing a system that is both highly capable and structurally governable requires a balance between modularity and integration. We propose a "hybrid governance architecture" where a high-level normative supervisor monitors the latent states of the primary agent, applying "normative pressures" to guide the agent away from unethical or high-risk trajectories without fully sacrificing the benefits of non-linear optimization.

Furthermore, there is a trade-off between local autonomy and global coordination. In multi-agent systems, each individual agent might be value-constrained, yet the aggregate behavior of the collective can still lead to systemic misalignment. This is often seen in "flash crashes" or synchronized liquidity droughts. A robust governance infrastructure must therefore operate at multiple levels, managing both the internal constraints of individual agents and the emergent dynamics of the system as a whole. This requires a level of computational overhead that impacts the sustainability of the infrastructure, demanding more energy and more sophisticated monitoring hardware, which in turn raises questions about the long-term scalability of such a governance model.

4. Infrastructure for Value-Embedding and Latent Monitoring

The deployment of value-constrained agents requires a massive overhaul of the underlying socio-technical infrastructure. Governance cannot be a software patch; it must be supported by a robust hardware and data substrate. This includes the development of "governance-aware" silicon that can provide hardware-level verification of an agent's internal states, ensuring that security and alignment protocols cannot be bypassed by the AI itself. Such an infrastructure would act as a "digital immune system," continuously monitoring the health and alignment of autonomous processes across the network.

Central to this infrastructure is the ability to monitor "latent reasoning traces." Unlike traditional logs that record what an application did, a reasoning trace records why the agent moved toward a specific decision. Analyzing these traces in real-time requires significant advancements in natural language understanding and causal inference. The infrastructure must be capable of identifying "normative anomalies"—patterns of thought within the AI that suggest a drift toward misaligned goals—long before those thoughts manifest as actions. This proactive monitoring is the backbone of the value-constrained paradigm, providing the transparency needed for meaningful human oversight.

Moreover, the sustainability of this infrastructure is a primary engineering concern. The energy consumption required to audit millions of autonomous agents in real-time is immense. We must develop "green governance" protocols that utilize sparse monitoring and heuristic-based triggers to reduce the computational load without compromising safety. This might involve the use of decentralized ledgers to maintain an immutable record of agentic intent, allowing for distributed auditing while maintaining a "single source of truth" for the system's normative status. The goal is to build a governance infrastructure that is as resilient and adaptive as the AI systems it is intended to manage.

5. Robustness and the Challenge of Out-of-Distribution Behavior

One of the greatest risks in autonomous agency is "out-of-distribution" (OOD) behavior, where the agent encounters a scenario that was not covered in its training data or its formal rule set. In these situations, rule-compliant systems often fail unpredictably, as they lack the "common sense" or moral intuition to handle the novelty. Value-constrained agency addresses this by providing the agent with a set of generalized normative priors that act as a safety net. If an agent is constrained by the value of "safety" rather than a specific rule about "speed," it is better equipped to navigate an unforeseen obstacle for which no specific speed rule exists.

Robustness in this context refers to the agent's ability to remain aligned even when its environment becomes chaotic or adversarial. Current AI systems are notoriously fragile; small perturbations in their input can lead to radically different and often dangerous outputs. A value-constrained architecture builds robustness by ensuring that certain "ethical invariants" are never violated, regardless of the input. This requires the implementation of "adversarial alignment" protocols, where the agent is stress-tested against a variety of hypothetical "black

swan" events to ensure that its normative boundaries remain intact.

However, building such robust systems is a multi-disciplinary challenge. It requires a synthesis of formal verification methods from computer science and normative theory from the social sciences. We must be able to prove, with a high degree of mathematical certainty, that an agent's internal value constraints will hold under pressure. This involves moving beyond the "probabilistic safety" of current models toward "deterministic alignment," where the architecture itself prevents the agent from even considering a misaligned path. Such a high bar for robustness is necessary if we are to entrust autonomous agents with the management of our most critical socio-technical systems [33].

6. Fairness, Equity, and the Problem of Value Selection

A move toward value-constrained agency immediately raises the critical question: whose values are we constraining the agent with? If the governance paradigm shifts from rules to values, the process of value selection becomes the ultimate site of political and social power. There is a profound risk that the values embedded into global AI infrastructures will reflect the biases and interests of a small group of technologists and policymakers in the developed world, further marginalizing vulnerable populations. This is not just a moral issue; it is a systemic risk. A "fairness" constraint that is defined too narrowly or through a biased lens can lead to exclusionary outcomes in automated lending, hiring, or criminal justice.

To ensure equity, the process of value selection must be inclusive, transparent, and participatory. We propose a "multi-stakeholder normative framework" for AI governance, where diverse groups are involved in defining the fundamental priors that constrain autonomous agents. This requires a transition from "expert-led" governance to "society-in-the-loop" governance [27]. Furthermore, these values must be dynamic; as societal norms evolve, the constraints on our AI systems must be capable of updating without requiring a full structural overhaul. This introduces the challenge of "normative drift," where the system's constraints must be recalibrated to reflect changing human values while maintaining the stability of the infrastructure.

Fairness also must be treated as a structural constraint. Rather than being an afterthought or a metric to be checked at the end of the development cycle, fairness must be integrated into the objective function of the agent. This means that the agent's internal optimization process must account for the disparate impact of its potential actions across different demographic groups. Achieving this requires a deep understanding of the socio-technical context in which the agent is deployed. A value-constrained agent in a healthcare setting must prioritize equity in a different way than an agent in a transportation network. The flexibility of value-constrained agency allows for this context-sensitivity, providing a more robust path toward algorithmic justice than static rules ever could.

7. Policy Implications: From Reporting to Architectural Auditing

The transition to a value-constrained paradigm necessitates a fundamental rethink of AI policy. Current legislative efforts, such as the EU AI Act, focus heavily on risk categorization and reporting requirements. While these are important first steps, they remain rooted in the compliance-based model. Future policy must shift toward "architectural auditing." Instead of asking companies to report on the outputs of their AI, regulators should have the authority to audit the internal governance layers of the system. This would involve verifying that the agent's reasoning traces are being monitored, that its value constraints are robust, and that its "kill switches" are functional.

This change in policy requires a new class of regulatory expertise. Government agencies will need "systems-level auditors" who understand the interplay between AI architecture, data infrastructure, and normative constraints. Furthermore, policy must address the issue of liability in an agentic world. If an autonomous agent causes harm, but its architecture was technically "value-constrained," where does the blame lie? We must develop legal frameworks that account for "distributed agency," where liability is shared between the developers of the model, the providers of the governance infrastructure, and the users of the system.

International cooperation is also essential. Since autonomous agents operate across national borders, a fragmented regulatory landscape will only lead to "governance arbitrage," where companies deploy their most unconstrained and high-risk agents in jurisdictions with the weakest oversight. A global "normative baseline" for AI agency would ensure that certain fundamental values—such as the protection of human life and the preservation of democratic institutions—are embedded into every autonomous system, regardless of its origin. This level of coordination is unprecedented, yet it is the only way to manage the systemic risks of a globalized, agentic economy.

8. The "Missing Dimension" of Governance: Internal Alignment

In the discourse on AI safety, there is a growing recognition of what has been termed the "missing dimension" of governance: the internal state of the agent [11]. Most contemporary oversight mechanisms are external; they look at the system from the outside in. This paper argues that true governance must move from the inside out. We must bridge the gap between the mathematical objective functions of the machine and the normative intent of the human creator. Internal alignment is the process of ensuring that the machine's "reasoning" is not just a calculation of probability, but a reflection of human value priorities.

Addressing this missing dimension involves the development of "value-alignment probes"—diagnostic tools that can be used during the training and deployment phases to measure the strength and consistency of an agent's internal constraints. By subjecting an agent to a series of ethical dilemmas in a simulated environment, we can "stress-test" its values. This "normative sandboxing" allows engineers to identify potential misalignments before the agent is exposed to the real world. This is a critical departure from the current practice of "testing in production," which has led to numerous high-profile failures in

autonomous systems.

Furthermore, internal alignment requires a commitment to "transparency by design." The reasoning traces of an agent must be human-legible, or at least interpretable by a secondary auditing AI. If we cannot understand how an agent is arriving at its decisions, we cannot claim to be governing it. This transparency is the link between technical engineering and public trust. Only by opening up the "black box" of agentic reasoning can we provide the evidence-based assurance that autonomous systems are behaving in a way that is consistent with the public interest. The missing dimension of governance is, in the end, the presence of a "normative tether" between the human and the machine [35].

9. Sustainability and the Future of Autonomous Infrastructures

The long-term sustainability of the AI revolution is inextricably linked to the robustness of its governance. A world filled with unconstrained, misaligned agents is inherently unstable. Frequent systemic failures, whether in the form of financial crashes, power outages, or social polarization, will eventually lead to a "tech backlash" that could stifle innovation and prevent the realization of AI's potential benefits. Value-constrained agency is the path to "sustainable innovation," where growth is balanced by safety and resilience. By building systems that are inherently bounded, we reduce the risk of catastrophic failure and build the social capital necessary for large-scale AI deployment.

The future of autonomous infrastructures will likely be defined by "hierarchical governance swarms," where different layers of agents monitor and constrain each other. In such a system, a primary agent might be optimized for a specific task, while a secondary "governor agent" ensures its actions remain within a set of normative boundaries. A third layer of human-led institutional oversight would then monitor the interaction between the two. This multi-layered approach provides redundancy and resilience, ensuring that a failure in one layer does not lead to a total systemic collapse. The engineering of these swarms is the next great frontier of systems research.

As we move forward, we must also consider the "environmental sustainability" of governance. The computational cost of continuous architectural auditing is high. We need to prioritize the development of efficient, low-energy alignment algorithms that can run on edge devices as well as in the cloud. The goal is "pervasive governance," where every autonomous process, no matter how small, is value-constrained. This requires a fundamental shift in our computational philosophy, moving away from "maximizing performance at any cost" toward "maximizing value-aligned utility." This is the hallmark of a mature, responsible, and sustainable socio-technical infrastructure.

10. Conclusion

The transition from rule compliance to value-constrained agency represents the most significant paradigm shift in the history of AI governance. As we have argued throughout this

paper, the era of treating AI as a deterministic tool to be managed through external checklists is over. In an age of autonomous agency, governance must be an intrinsic, structural property of the agent itself. This requires a commitment to "governance-by-design," where normative constraints are embedded into the architectural substrate and internal reasoning traces are subjected to continuous auditing.

While the technical and structural trade-offs are significant, they are not insurmountable. The cost of the "alignment tax" is a necessary investment in systemic stability and social trust. By addressing the missing dimensions of current oversight and building a robust infrastructure for value-embedding, we can create a future where autonomous agents are not a source of uncontrollable risk, but a powerful force for human flourishing. This journey requires the synthesis of engineering rigor with normative wisdom, moving beyond the shallow pursuit of algorithmic efficiency to the deeper goal of value-aligned intelligence.

In conclusion, the future of AI governance lies in our ability to build systems that "understand" and respect the spirit of human values, even as they operate at speeds and scales that exceed human comprehension. The paradigm of value-constrained agency offers a roadmap to this future, providing the framework for a resilient, fair, and sustainable autonomous world. The challenge is immense, but the stakes—the integrity and stability of our global socio-technical infrastructures—could not be higher. We must begin the work of institutionalizing this new paradigm today, ensuring that the agents of tomorrow remain the faithful stewards of our collective intent.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

2. Arner, D. W., Barberis, J., & Buckley, R. P. (2017). The evolution of Fintech: A new post-crisis paradigm? Georgetown Journal of International Law, 47(4), 1271-1319.

3. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The Moral Machine experiment. Nature, 563(7729), 59-64.

4. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671-732.

5. Baxter, L. G. (2016). Adaptive financial regulation and the role of white papers. Georgetown Law, Technology & Policy Review, 1(1), 125-140.

6. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

7. Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning

algorithms. Big Data & Society, 3(1), 2053951715622512.

8. Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. UC Davis Law Review, 51, 399.

9. Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180080.

10. Cave, S., & ÓhÉigeartaigh, S. S. (2018). Bridging near-and long-term AI safety and ethical issues. Nature Machine Intelligence, 1(1), 5-7.

11. Chen, L. (2026). Beyond External Constraints: The Missing Dimension of AI Governance. Available at SSRN 6449738.

12. Christian, B. (2020). The alignment problem: Machine learning and human values. W. W. Norton & Company.

13. Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

14. Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer Nature.

15. Dobbe, R., Kaziunas, E., & Whittaker, M. (2021). AI in the wild: Sustainability in the age of artificial intelligence. AI & Society, 36, 1205-1220.

16. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

17. Gabriel, I. (2020). Artificial intelligence, values and alignment. Minds and Machines, 30(3), 411-437.

18. Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. AMIA Joint Summits on Translational Science Proceedings, 2020, 191.

19. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

20. Helbing, D. (2013). Globally networked risks and how to respond. Nature, 497(7447), 51-59.

21. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines.

Nature Machine Intelligence, 1(9), 389-399.

22. Jordan, M. I. (2019). Artificial intelligence—The revolution hasn't happened yet. Harvard Data Science Review, 1(1).

23. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165, 633.

24. Leike, J., Martic, M., Garrabrant, S., Vaneess, A., Aslanides, K., Fearon, C., ... & Wang, Z. (2017). AI safety gridworlds. arXiv preprint arXiv:1711.09883.

25. Leslie, D. (2019). Understanding artificial intelligence ethics and safety. The Alan Turing Institute.

26. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

27. Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate racism? Science, 366(6464), 447-453.

28. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

29. O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

30. Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

31. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. P. (2019). Machine behaviour. Nature, 568(7753), 477-486.

32. Russell, S. J. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

33. Saria, S., & Subbaswamy, A. (2019). Tutorial: Safe and reliable machine learning. arXiv preprint arXiv:1904.07204.

34. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 59-68.

35. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for

deep learning in NLP. arXiv preprint arXiv:1906.02243.

36. Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to predict it, not to fear it. Nature, 556(7699), 9-11.

37. Turkle, S. (2011). Alone together: Why we expect more from technology and less from each other. Basic Books.

38. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical and legal challenges. PLOS Medicine, 15(11), e1002689.

39. Wiener, N. (1960). Some moral and technical consequences of automation. Science, 132(3437), 1355-1358.

40. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.