Human-in-the-Loop Reinforcement Learning for AI Governance: A Fast–Slow Decision Paradigm for Responsible LLM Deployment

Tianyi Shao
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
tianyi.shao@uc.edu

Landon R. Martin
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
lmartin@unr.edu

Stefano Phillips
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
sphillips@uab.edu

Enzo Castro
School of Computing, Clemson University, Clemson, SC, USA.
castro801@clemson.edu

Abstract

The rapid deployment of large language models (LLMs) in high-stakes domains such as healthcare, finance, and legal reasoning has intensified concerns regarding their alignment with human values, fairness, and long-term safety. Traditional reinforcement learning (RL) approaches for LLM alignment, including reinforcement learning from human feedback (RLHF), rely on a static reward model and a single loop of human annotation, which fail to adapt to evolving societal norms and context-sensitive ethical dilemmas. This paper proposes a novel governance framework that integrates human-in-the-loop reinforcement learning with a fast–slow decision paradigm inspired by dual-process cognitive theory. The framework distinguishes between fast, automatic LLM responses that are optimized for efficiency and slow, deliberative interventions that involve human oversight and metacognitive reasoning. We introduce a human-in-the-loop RL architecture where a supervisory human agent dynamically adjusts the balance between fast and slow pathways based on risk estimation, uncertainty quantification, and policy compliance. This architecture is implemented through a hierarchical reward structure that couples immediate performance rewards with long-term governance penalties. We analyze structural trade-offs between system responsiveness and regulatory robustness, and discuss deployment considerations including scalability, auditability, and resilience to adversarial manipulation. Cross-domain comparisons with autonomous driving and algorithmic trading illustrate the generality of the paradigm. We conclude by outlining policy implications for responsible LLM deployment and proposing a governance lifecycle that integrates continuous human oversight with adaptive RL mechanisms.

Keywords

human-in-the-loop reinforcement learning, AI governance, fast–slow decision paradigm, large language models, responsible deployment, dual-process theory, reward design, safety alignment.

1. Introduction

Large language models have achieved remarkable fluency and broad competence across numerous natural language tasks, yet their deployment in real-world settings continues to raise serious governance challenges [1,2]. These challenges include the propagation of biased or harmful content, susceptibility to adversarial prompts, and the difficulty of encoding complex moral and legal constraints into a single reward function [3,4]. Existing alignment techniques, such as RLHF, rely on a fixed reward model trained on human preference data, which is then used to fine-tune the LLM via RL [5]. However, the static nature of this process is ill-suited for contexts where human values evolve, where the cost of an error is high, or where the definition of appropriate behavior depends on nuanced situational factors [6].

A central limitation of current approaches is the absence of an explicit mechanism for incorporating deliberative human reasoning into the ongoing decision loop of the model. While humans provide initial preference labels, the model subsequently operates in a fully automated manner, leaving no room for real-time intervention when the environment or societal expectations shift [7]. This gap has motivated the exploration of human-in-the-loop (HITL) frameworks that keep a human decision-maker actively engaged in the model's learning and inference cycles [8]. Yet most HITL proposals treat the human as an external oracle or a simple labeler, without modeling the internal cognitive dynamics that govern when a human should intervene and how that intervention should be integrated with the model's own learning processes.

Drawing on the dual-process theory of cognition, which distinguishes between fast, automatic System 1 processes and slow, deliberate System 2 processes [9], we propose a governance paradigm that explicitly maps these two modes onto the LLM's decision pipeline. In our framework, the LLM operates in a fast mode for routine, low-risk queries, producing outputs based on its learned policy. A separate slow pathway is triggered for queries that exceed a risk or uncertainty threshold, at which point a human-in-the-loop agent engages in deliberative analysis and provides corrective feedback. This feedback is then used to update the LLM's policy through an online reinforcement learning mechanism, creating a continuous cycle of fast action and slow reflection. The human agent is not merely a supervisor but an integral component of a hierarchical control system that balances efficiency and governance.

This paper makes three principal contributions. First, we formalize a fast–slow decision framework for LLM deployment that integrates human cognitive oversight with RL-based adaptation. Second, we design a human-in-the-loop RL architecture that implements this framework, including a dual reward structure that penalizes harmful fast-mode outputs while incentivizing accurate slow-mode interventions. Third, we analyze the governance implications of this architecture, discussing trade-offs in latency, cost, transparency, and accountability. We illustrate the paradigm through case studies in clinical decision support and content moderation, and compare with analogous systems in autonomous driving and algorithmic trading.

2. Background and Related Work

The alignment of LLMs with human intentions has become a central research agenda in AI safety. The most widely adopted method, RLHF, proceeds in three stages: first, human

annotators rank model outputs to create a preference dataset; second, a reward model is trained to predict these preferences; third, the LLM is fine-tuned using a policy gradient algorithm such as PPO to maximize the learned reward [5,10]. While RLHF has achieved notable success in reducing harmful outputs and increasing perceived helpfulness, it suffers from several structural deficiencies. The reward model is static after training, meaning that changes in human values or deployment context cannot be captured without retraining the entire pipeline [11]. Moreover, the reward model itself may exhibit biases present in the annotation process, and these biases can be amplified during RL fine-tuning [12].

Recent work has attempted to address these limitations by incorporating online human feedback, where annotators evaluate model outputs during deployment and the reward model is updated incrementally [13]. Yet these approaches still treat the human as an external evaluator rather than an active decision-maker that can choose when and how to intervene. Human-in-the-loop RL generalizes the idea by allowing a human to directly modify the agent's actions or reward signals during learning [8]. In robotics and autonomous driving, HITL RL has been applied to safety-critical tasks where the human can override the agent to prevent catastrophic failures [14]. However, transferring these ideas to LLM deployment requires careful consideration of the cognitive load on the human, the latency of interventions, and the need to scale oversight to millions of daily queries.

The dual-process theory of cognition, popularized by Kahneman, provides a useful metaphor for designing AI decision systems [9,15]. Fast, intuitive processes (System 1) are efficient but prone to systematic errors, while slow, analytical processes (System 2) are more accurate but resource-intensive. In AI, hybrid architectures that combine fast and slow pathways have been explored for planning, reasoning, and decision-making [9,16]. For example, Dou et al. proposed a decision framework that alternates between a fast neural policy and a slow deliberative component that uses search or symbolic reasoning [9]. We extend this concept to the governance of LLMs, where the fast pathway corresponds to the base LLM response and the slow pathway corresponds to a human-in-the-loop review bolstered by interpretability tools and policy checking.

3. The Fast-Slow Decision Paradigm for LLM Governance

We propose a governance architecture in which every incoming user query is first processed by a risk assessment module that estimates the probability of harmful or non-compliant outputs. If the estimated risk is below a threshold, the query is handled by the fast pathway, where the LLM generates a response directly from its learned policy. If the risk is above the threshold, the query is routed to the slow pathway, where it is presented to a human decision-maker who conducts a deliberative analysis before authorizing or modifying the response. The human's action is then used as a training signal to update the LLM's policy through an online RL algorithm, thereby gradually reducing the need for future slow-pathway interventions on similar queries.

This design reflects a fundamental trade-off between efficiency and governance. The fast pathway minimizes latency and cost, which is essential for high-throughput applications such as customer service or code generation. The slow pathway introduces delays and requires human resources but enables nuanced ethical reasoning, context-sensitive judgment, and adaptation to new norms. The threshold itself is not fixed; it is dynamically adjusted based on the current system state, historical intervention data, and policy updates. For example, after a series of successful fast-mode responses in a particular domain, the threshold may be lowered to allow more autonomy, while repeated failures trigger a tightening.

The role of the human in the slow pathway is multifaceted. The human is not simply a labeler but an active controller who can inspect the LLM's reasoning, query additional context, consult external databases or ethical guidelines, and craft a response that adheres to governance policies. The human can also provide a scalar reward or a comparative preference signal that is fed back to the RL system. This feedback is used to update both the policy network and the risk assessment module, creating a closed-loop system that continuously improves over time. A key insight is that the human's deliberative process itself can be partially automated using interpretability tools, such as saliency maps or counterfactual explanations, that help the human make faster and more accurate decisions.

4. Human-in-the-Loop RL Architecture

The architecture consists of four main components: a base LLM (policy), a risk estimator, a human interface, and an online RL training module. The base LLM is initialized from a pre-trained model and subsequently fine-tuned via RL. The risk estimator is a separate neural network or ensemble that takes as input the user query, the LLM's internal representations, and contextual metadata, and outputs a risk score. The human interface presents the query, the LLM's fast response, and any supporting information to the human operator, who then decides whether to approve, reject, or modify the response. The human's decision is encoded as a supervision signal.

The RL algorithm operates in two loops. In the fast loop, the LLM interacts with users and receives rewards from a learned reward function that incorporates both immediate user satisfaction and long-term governance metrics. In the slow loop, each time a query is routed to the human, the human's corrective action is used to compute an additional reward that overrides the fast-loop reward for that query. The policy is updated using a combination of both reward signals. To ensure stability, we employ a trust region method that constrains policy updates to avoid catastrophic forgetting of previously aligned behaviors.

A critical feature is the dual reward structure. The fast reward includes metrics such as relevance, coherence, and user satisfaction, while the governance reward penalizes outputs that violate predefined policies or that the human judge deems harmful. The governance reward is typically sparse and delayed, since its effect may only be observed after human intervention. To address this, we augment the reward with a credit assignment mechanism that propagates the governance signal back to the risk estimator and the base policy using importance sampling.

Scalability is a primary concern. In a production environment with millions of daily queries, routing every risky query to a human is infeasible. Therefore, we employ a multi-tier system where low-risk queries are handled entirely automatically, moderate-risk queries are flagged for batch auditing, and only high-risk queries trigger real-time human intervention. The thresholds are learned via a cost-sensitive optimization that balances the cost of human labor against the cost of potential harm. Additionally, the human interface can be augmented with AI-assisted decision tools that suggest likely safe responses based on past interventions, reducing the cognitive burden on the operator.

5. Governance Implications

The adoption of a fast–slow human-in-the-loop RL paradigm carries profound implications for AI governance. First, it introduces a mechanism for continuous adaptation to evolving societal norms. Unlike static RLHF, where the reward model is fixed, the proposed framework allows human values to be injected on an ongoing basis through the slow pathway.

This is particularly important in domains such as content moderation, where definitions of acceptable speech shift over time and across cultures. The governance system can be tuned by adjusting the risk threshold and the human's decision criteria, without requiring retraining the entire model.

Second, the architecture enhances transparency and auditability. Every slow-pathway decision is logged along with the rationale provided by the human operator. These logs can be reviewed by external auditors to verify that the system is behaving in accordance with legal and ethical standards. Moreover, the risk estimator's outputs can be analyzed to identify systematic biases in when queries are escalated. For example, if certain demographic groups are disproportionately routed to the slow pathway, this may indicate that the base model has learned biased associations.

Third, the framework provides a natural defense against adversarial attacks. An adversary attempting to elicit harmful outputs would need to circumvent not only the base LLM but also the risk estimator and the human operator. Since the human can detect subtle manipulations that may fool automated systems, the slow pathway serves as a robust last line of defense. However, this also introduces a new attack surface: an adversary could attempt to overwhelm the human workforce with a flood of seemingly risky queries, causing operator fatigue or forcing the system to lower its thresholds. Mitigation strategies include workload balancing, automated triage, and the use of redundancy among human operators.

Sixth, the governance implications extend to accountability and liability. When an LLM generates a harmful response in the fast pathway, who is responsible? Under our framework, responsibility is shared between the system designers (who set the risk thresholds and training processes), the human operators (who intervene in the slow pathway), and the organization that deploys the system. Clear protocols must be established for when a fast-pathway output is attributable to a design flaw versus an unforeseeable edge case. The human-in-the-loop architecture makes these attributions more tractable because the slow-pathway logs provide a detailed record of decision-making.

7. Case Illustrations and Cross-Domain Comparison

We illustrate the application of the fast–slow paradigm in two domains: clinical decision support and content moderation. In clinical settings, an LLM may be used to suggest diagnoses or treatment plans. The fast pathway can handle routine queries such as summarizing patient records or answering common medication questions. High-risk queries, such as those involving ambiguous symptoms or rare diseases, are routed to a human physician who reviews the LLM's suggestion and either confirms it or provides an alternative. The physician's feedback is then used to update the reward model, improving the LLM's future performance for similar cases. This approach mirrors the existing practice of double-checking critical clinical decisions and aligns with regulatory requirements for human oversight in medical AI.

In content moderation, an LLM might be deployed to automatically flag hate speech or misinformation. The fast pathway gives an initial classification, and the slow pathway escalates borderline cases to human moderators. The risk estimator can be trained to prioritize cases that have high potential for harm or that involve subtle forms of bias. Over time, the RL algorithm learns to mimic the moderators' judgments, reducing the volume of escalations. This is analogous to the hybrid moderation systems used by major social media platforms, but with the added benefit of online learning.

Cross-domain comparisons reveal common structural patterns. In autonomous driving, a fast pathway handles routine lane keeping and speed control, while a slow pathway (human takeover or a deliberative planner) handles complex intersections or obstacle avoidance. In algorithmic trading, fast algorithmic strategies execute routine orders, while a human trader intervenes during volatile market conditions. In both domains, the trade-off between speed and accuracy is managed by a risk-based gating mechanism, and feedback from the slow pathway is used to improve the fast pathway. The proposed governance framework generalizes these principles to the unique challenges of LLM deployment, where the decision set is vastly larger and the cost of error can be diffuse but equally severe.

8. Challenges and Future Directions

Despite its promise, the fast–slow HITL RL paradigm faces several challenges. The cognitive load on human operators is a primary concern. Frequent slow-pathway interventions can lead to fatigue and inconsistent judgments, undermining the reliability of the governance system. Future work should explore ways to reduce operator burden through active learning strategies that select the most informative queries for human review, and through the use of AI-assisted decision support that provides structured checklists or precomputed options.

Another challenge is the potential for human biases to be introduced into the system through the slow pathway. If human operators systematically favor certain demographic groups or exhibit anchoring effects based on the LLM's fast response, these biases can be learned by the RL algorithm and perpetuated. Careful training and calibration of human operators, as well as periodic auditing of their decisions, are necessary to mitigate this risk. Additionally, the reward design must account for the fact that human feedback may be inconsistent over time and across operators.

Scalability to large user populations remains a significant engineering hurdle. While multi-tier routing reduces the number of real-time interventions, the slow pathway still requires a pool of trained human operators. Advances in federated oversight, where multiple organizations share anonymized human feedback data, could distribute the cost. Furthermore, the risk estimator itself must be continuously updated to avoid becoming stale or vulnerable to distributional shift. Robustness to adversarial manipulation of the risk estimator is another open problem.

Finally, the governance framework must be embedded in a broader socio-technical infrastructure that includes legal standards, regulatory oversight, and stakeholder participation. The fast–slow paradigm provides a technical mechanism for accountability, but it cannot replace the need for democratic deliberation about what values should govern LLM behavior. Future research should explore the interface between this technical architecture and institutional governance mechanisms such as ethics boards, public comment periods, and impact assessments.

9. Conclusion

This paper has presented a human-in-the-loop reinforcement learning framework that leverages a fast–slow decision paradigm for the responsible deployment of large language models. By routing low-risk queries to an automated fast pathway and high-risk queries to a deliberative slow pathway involving human oversight, the framework balances efficiency and governance. The human actor is embedded in an online RL loop, providing corrective feedback that updates both the base policy and the risk estimator, enabling continuous adaptation to evolving norms and contexts. We have analyzed the structural trade-offs

inherent in this design, discussed its implications for transparency, accountability, and adversarial robustness, and illustrated its application in clinical and content moderation domains. Cross-domain comparisons with autonomous driving and algorithmic trading demonstrate the generality of the paradigm. The proposed architecture offers a path forward for AI governance that acknowledges the limitations of fully automated alignment and the necessity of sustained human involvement in the lifecycle of intelligent systems. As LLMs become more deeply integrated into critical infrastructure, such hybrid governance models will be essential to ensure that these systems remain aligned with human values while delivering their transformative benefits.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623).

2. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

3. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

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

5. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems, 30.

6. Stiennon, N., Ouyang, L., Wu, J., Szegedy, C., Lowe, R., & Christiano, P. (2020). Learning to summarize with human feedback. In Advances in Neural Information Processing Systems, 33.

7. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13).

8. Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research, 64, 243–252.

9. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf: Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

10. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

11. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

12. Shih, K., Deng, Z., Chen, X., Zhang, Y., & Zhang, L. (2025, May). DST-GFN: A Dual-Stage Transformer Network with Gated Fusion for Pairwise User Preference Prediction

in Dialogue Systems. In 2025 8th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) (pp. 715-719). IEEE.

13. Gao, H., Zeng, W., Zhang, J., & Liang, Y. (2025, December). A large model API response quality prediction model based on least squares vector machine and SHAP interpretability analysis. In 2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF) (pp. 438-442). IEEE.
14. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.
15. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
16. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489.
17. 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.
18. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
19. Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2017). Cooperative inverse reinforcement learning. In Advances in Neural Information Processing Systems, 30.
20. Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., ... & Graepel, T. (2021). Open problems in cooperative AI. arXiv preprint arXiv:2012.08630.
21. Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.