Frictive Policy Optimization for LLM Agent Interactions

James Pustejovsky
Brandeis University

Nikhil Krishnaswamy
Colorado State University

ABSTRACT

Recent advances in the alignment of large language models (LLMs) toward human preference and values have dramatically expanded the capabilities of artificial intelligence in natural language understanding and generation. However, despite their impressive performance, these models often lack the reflective and deliberative qualities necessary for effective human-AI collaboration. Traditional policy optimization methods, such as Reinforcement Learning from Human Feedback (RLHF), Proximal Policy Optimization (PPO), or Direct Preference Optimization (DPO), primarily focus on maximizing task-related rewards or aligning outputs with human preferences. These approaches, however, tend to neglect the critical epistemic dimension of alignment: the ability of an AI system to reason about, question, and update its underlying beliefs. In this paper, we propose a novel framework termed *Frictive Policy Optimization (FPO)*, which explicitly incorporates "friction" as a desirable property in the policy optimization process for LLMs. Beyond fostering reflective deliberation, our approach also challenges the conventional expectation that autonomous agents must always comply with human commands. By integrating mechanisms that incentivize appropriate non-compliance, what we term "beneficial disobedience", FPO equips AI systems with the capacity to question potentially harmful or ill-advised instructions. This dual focus on epistemic alignment and responsible disobedience paves the way for more robust, safe, and collaborative human-AI interactions.

KEYWORDS

Epistemic Tracking, Friction, Policy Learning, LLM Alignment

ACM Reference Format:
James Pustejovsky and Nikhil Krishnaswamy. 2025. Frictive Policy Optimization for LLM Agent Interactions. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

With advances in large language model (LLM) *alignment* toward human preferences and desired values like safety, truthfulness, and helpfulness [7], AI, exemplified by LLMs, has been rapidly integrated into personal and business workflows. This trend toward "agentic" AI is predicated upon a notion that such LLM-powered agents can act as faithful executors of human instructions to automate tedious tasks and increase personal and workplace efficiency [1]. In this paper, we argue that this vision of AI agents neglects a critical dimension of human-AI collaboration: *friction*, or the need for AIs to push back on unfounded assumptions and critically examine beliefs and evidence before executing human instructions. We argue that this capacity is largely left unaddressed by current popular alignment methods and that AI alignment needs to also consider "beneficial disobedience" rather than assuming that alignment with human values means uncritically following the literal interpretation of human commands (Sec. 2). Instead, we expand the notion of alignment to include *epistemic* alignment (Sec. 3) and propose a framework of *Frictive Policy Optimization (FPO)* to align AIs toward these kinds of desiderata for human-AI collaboration.

2 MOTIVATION

Within human-AI interaction, *friction* refers to any element of dialogue or decision-making that slows down an exchange to allow for reflection, interrogation of underlying assumptions, or clarification of conflicting evidence [2, 11, 16]. Drawing inspiration from human collaborative practices—where a brief pause to re-evaluate evidence or question a teammate's assumption often prevents cascading errors [4, 12], we propose an optimization framework for learning policies that aims to embed a similar mechanism within LLMs: *Frictive Policy Optimization (FPO)*. Where conventional LLM alignment and preference optimization approaches focus on reward maximization [6, 13, 14, 19] or learning an implicit underlying distribution of pairwise preference differences [3, 5] and evaluate on standard suites of offline datasets (e.g., [7, 17, 18]), FPO strategies introduce either a frictive-augmented advantage function or an explicit metric of friction in the reward formulation. This modification is designed to encourage models to generate outputs that are not only correct or useful, but also deliberative and epistemically aligned with human collaborators in real-time contexts.

At the core of Frictive Policy Optimization is the idea that standard policy gradients can be modified to include an additional component–friction–that captures the epistemic uncertainty inherent in the current dialogue context and the need for reflective reasoning. We envision a family of FPO strategies that prioritize not only task performance but also the quality of the underlying reasoning process. To this end, we outline three primary ways in which friction can be incorporated into the optimization process: (i) Friction-Augmented Rewards; (ii) Friction-Based Preference Pairing; and Group Relative Friction Ranking.

Using *Friction-Augment Rewards (FAR)*, the reward function is augmented with a friction term. Here, the overall reward R_{total} is defined as: $R_{\text{total}} = R_{\text{task}} + \alpha\, R_{\text{friction}}$, where R_{task} is the conventional reward (e.g., exemplifying the response's accuracy, helpfulness), and R_{friction} quantifies the degree of reflective delay or epistemic reconsideration exhibited by the model, and α is a scaling factor. The friction term can be derived from measures such as the divergence between the model's internal evidence model and the observed dialogue context or via explicit penalties/rewards computed from preference pairs that display conflicting assumptions.

In *Friction-Based Preference Pairing (FPP)*, friction can alternatively be introduced by constructing preference pairs that explicitly

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

contrast outputs that embody reflective deliberation with those that do not. For instance, given a dialogue context x and two candidate responses y^+ (frictive, reflective) and y^- (immediate, unreflective), the training objective encourages the model to assign a higher probability to y^+. This is analogous to logistic regression on the difference in log-probabilities, with the added effect that the "preference" encapsulates epistemic considerations. The resulting optimization effectively pushes the policy to favor responses that incorporate a deliberate check on the model's own assumptions or prompt further inquiry into ambiguous or conflicting evidence. Further, this is distinct from recent approaches invoking *pause* tokens [8] in that it incentivizes examination of the dialogue context; in FPO, content and process, rather than simply time, matter. Friction is not simply "slowing down the dialogue," but doing so purposefully [10].

Another promising approach for implementing Frictive Policy Optimization is the use of Group Relative Policy Optimization (GRPO) [9, 15], what we refer to as *Group Relative Friction Ranking (GRFR)*. By embedding friction-sensitive components into the group-based advantage framework, these strategies can dynamically assess whether candidate responses provide sufficient opportunities for reflective dialogue. For instance, outputs that trigger a "frictive state", where there is a marked difference between a model's implicit evidence and the dialogue context, can be rewarded, encouraging the system to generate responses that prompt further deliberation.

In standard methods like GRPO, we work with a reward r (or a group-normalized version). For FPO, we define an augmented reward r' that includes both the original reward (e.g., correctness, helpfulness) and an additional term F that quantifies the degree of friction present in the output. For example, $r' = r + \lambda F$, where F is a friction metric (measuring, for instance, how much the output challenges assumptions or prompts the user to reexamine their evidence) and λ is a weighting parameter balancing the two terms.

Outputs that not only provide correct or helpful information but also prompt a beneficial degree of reflection (i.e. a higher friction score F) will receive a higher augmented reward r' and, consequently, a higher normalized advantage A_i. In contrast, outputs that might be correct but are too "frictionless" (i.e., they don't encourage the human collaborator to reassess or deliberate) will be penalized in the augmented reward signal. Over time, the policy will learn to favor responses that strike the right balance between being correct/helpful and introducing productive friction.

3 EPISTEMIC VS. VALUE ALIGNMENT

A key departure from traditional methods is the shift in focus from mere value alignment (i.e., ensuring the output is high-scoring in a task-specific sense) to epistemic alignment. Epistemic alignment involves ensuring that the model's outputs reflect an awareness of the underlying evidence and uncertainty inherent in collaborative tasks. In human dialogue, friction often manifests as internal or external resistance—pauses, questions, or clarifications—that prevent overcommitment to a potentially flawed assumption. By modeling and incentivizing friction, FPO seeks to align LLMs with human partners not only in terms of correctness but also in terms of the reasoning process itself. This is crucial in applications where decisions must be made collaboratively and where the cost of unchecked overconfidence can be significant.

Integrating friction into policy optimization has several practical implications. First, it can mitigate the risk of overreliance on AI outputs by encouraging models to provide responses that invite further scrutiny and dialogue. Second, it enables a more dynamic adjustment of the AI's internal evidence model, allowing it to better adapt to changing contexts and conflicting pieces of evidence. Third, by promoting epistemic alignment, frictive policies can help establish a more balanced partnership between humans and AI, where the AI not only delivers information but also actively participates in the reflective process of collaborative problem solving.

In addition to introducing the role of epistemic alignment and reflective reasoning, our approach acknowledges that intelligent autonomous agents should, in some contexts, exhibit a form of "beneficial disobedience." Rather than blindly following all human commands, an AI that incorporates frictive policy optimization can be trained to question instructions that conflict with safety constraints or established ethical guidelines. By incorporating a disobedience metric into the friction-augmented reward, for example, penalizing outputs that uncritically comply with potentially harmful or misinformed commands while rewarding those that demonstrate constructive resistance—the model learns to balance obedience with the need for independent critical judgment. This mechanism ensures that the agent not only aligns with task-specific objectives but also safeguards against actions that might undermine collaborative goals or lead to adverse outcomes.

Integrating AI disobedience within the FPO framework effectively broadens the scope of policy optimization to include a dynamic interplay between compliance and resistance. In practice, this means that when an instruction appears to be at odds with the agent's internal evidence model or violates normative safety standards, the policy is incentivized to generate outputs that prompt further deliberation rather than immediate compliance. Such behavior can be viewed as a form of "frictive disobedience" that, while slowing down the interaction momentarily, ultimately promotes safer and more robust human-AI collaboration. This reconceptualization of obedience, not as a fixed, inflexible requirement but as a context-dependent, ethically informed decision-making process, marks a significant step toward developing autonomous agents that are both collaborative and responsibly autonomous.

4 CONCLUSION

Frictive Policy Optimization represents a promising new direction in the fine-tuning of large language models for collaborative settings. By embedding friction—whether through an augmented reward structure or through preference-based training—the approach seeks to produce AI agents that are more reflective, deliberative, and ultimately, better collaborators. Future work should focus on developing quantitative measures of friction, exploring optimal strategies for balancing task performance with reflective deliberation, and conducting empirical studies to evaluate the impact of frictive policies on real-world collaborative tasks. FPO broadens the traditional scope of policy optimization by explicitly targeting the epistemic dimension of AI behavior. As AI systems become increasingly integrated into collaborative workflows, the ability to harness friction may prove essential in ensuring that these systems not only perform tasks effectively but do so in a manner that supports and enhances human judgment.

REFERENCES

[1] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic AI: Autonomous Intelligence for Complex Goals–A Comprehensive Survey. *IEEE Access* (2025).

[2] Nicholas Asher and Anthony Gillies. 2003. Common Ground, Corrections, and Coordination. *Argumentation* 17 (2003), 481–512.

[3] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4447–4455.

[4] Paul Patrick Gordon Bateson and Robert A Hinde. 1976. *Growing Points Ethology*. Cambridge University Press.

[5] Changyu Chen, Zichen Liu, Chao Du, Tianyu Pang, Qian Liu, Pradeep Varakantham, ..., and Jian Zhou. 2024. Bootstrapping Language Models with DPO Implicit Rewards. *arXiv preprint arXiv:2406.09760* (2024).

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[7] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. 2024. ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback. In *International Conference on Machine Learning*. PMLR, 9722–9744.

[8] Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training Language Models With Pause Tokens. In *The Twelfth International Conference on Learning Representations*.

[9] Deyao Guo, Dayiheng Yang, Hao Zhang, Jian Song, Renjie Zhang, Ruochen Xu, ..., and Yang He. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948* (2025).

[10] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

[11] Gary Klein, Paul J Feltovich, Jeffrey M Bradshaw, and David D Woods. 2005. Common ground and coordination in joint activity. *Organizational simulation* 53 (2005), 139–184.

[12] Harri Oinas-Kukkonen and Marja Harjumaa. 2009. Persuasive systems design: Key issues, process model, and system features. *Communications of the association for Information Systems* 24, 1 (2009), 28.

[13] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*.

[14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. In *Advances in Neural Information Processing Systems*. 3074–3082.

[15] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[16] Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy* 25, 5/6 (2002), 701–721.

[17] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems* 33 (2020), 3008–3021.

[18] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*. 59–63.

[19] Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. 2023. Pairwise Proximal Policy Optimization: Harnessing Relative Feedback for LLM Alignment. *arXiv preprint arXiv:2310.00212* (2023).