# Adding Reflective Governance to LLMs

**Parisa Salmani , Peter R. Lewis**
Ontario Tech University
{parisa.salmani, peter.lewis}@ontariotechu.ca

## 1 Introduction

With the rapid advancement of artificial intelligence (AI) systems and large language models (LLMs), these technologies are increasingly used to generate content, provide answers, and engage in meaningful conversations. However, a significant gap remains: AI systems often lack the ability to reflect on their own behavior or reason critically about their responses. This paper introduces and explores the concept of reflective governance in LLM-based systems, presenting an architecture designed to enable these models to evaluate and refine their outputs, promoting more thoughtful and trustworthy interaction.

## 2 Reflective Artificial Intelligence

Reflection in humans is a complex process involving various cognitive activities, such as thinking critically, analyzing actions, and evaluating potential outcomes [Colley *et al.*, 2012]. This complexity makes it difficult to define reflection within a single framework for artificial intelligence [Pitt, 2014]. Nonetheless, a definition can be articulated: reflective architecture can leverage techniques and tools to help artificial intelligence adapt, consider context, and align with human values. The goal is to make AI systems sensitive to societal needs and capable of ethical decision-making.

Margaret Boden [Boden, 2016] characterizes artificial intelligence as "computers that do the sorts of things that minds can do." This perspective emphasizes the transfer of human cognitive capabilities to machines, which has driven the continuous evolution of innovative AI technologies. Reflective AI extends this vision by not only automating cognitive tasks but also equipping AI systems with the ability to refine their behavior. Such advancements suggest a future where machines are not merely tools but participants in ethical reasoning and complex decision-making processes.

To provide AI systems with reflective capabilities—thus expanding Boden's "sorts of things"—we must first establish a suitable architecture [Lewis and Sarkadi, 2024]. This involves decoupling reflection from decision-making and action to ensure a clear description of cognitive functions. Next, reflective processes must be developed, tailored to the specific type of reflection needed. A reflective agent could incorporate one or more of these processes, guided by theoretical framework that adapts to varying requirements and contexts. At their core, reflective processes operate as reflective loops, iteratively evaluating and improving outputs to align with predefined goals, constraints, ethics, and norms.

### 2.1 Reflective Loops

Reflective systems often utilize various types of feedback loops, each serving different purposes in adapting and improving system behavior. Notable categories, as discussed by Lewis and Sarkadi [Lewis and Sarkadi, 2024], include *Integrating Experience and External Factors Loops*, which can incorporate new design goals into existing reflective models and operational objectives, enhancing adaptability and alignment with evolving requirements. Another category is the *Critique and Imagination Loops*, such as processes like active experimentation aimed at refining potential behaviors through iterative learning and creative problem-solving. Additionally, *Reflective Thinking Loops* emphasize meta-cognitive evaluation, enabling systems to assess their reasoning processes for accuracy and improvement.

This paper, however, focuses on a more fundamental category of reflective loops: the *Governance Loop*. Governance loops play a critical role in overseeing and regulating the actions of reflective systems to ensure that their behavior follows norms, ethical standards, and predefined objectives. This loop acts between an AI's decision-making system (Reflective Reasoning) and its actions (Actuators).

It evaluates decisions in context without requiring the AI to constantly re-learn its decision rules. Instead, the system can compare its behavior to predefined ethical goals or rules and decide, "This isn't right—try a different option." [Lewis and Sarkadi, 2024]

For example, if the AI's environment or ethical priorities change, it can adapt by reassessing its actions rather than starting over. This approach builds on earlier research into ethical AI, where robots or agents are given frameworks to evaluate the morality of their behavior (like Winfield's "consequence engines" [Winfield *et al.*, 2019]).

By continuously monitoring system outputs and decision pathways, governance loops maintain alignment with intended goals, prevent deviations, and enforce accountability within the system's operational framework. These loops are essential for ensuring that reflective systems remain trustworthy and robust in dynamic environments.

The focus of this paper is to introduce the reflective gover-

nance architecture by adding the reflective governance layer to the language models. The proposed architecture is primarily designed for language models but can be generalized to other AI systems in future work.

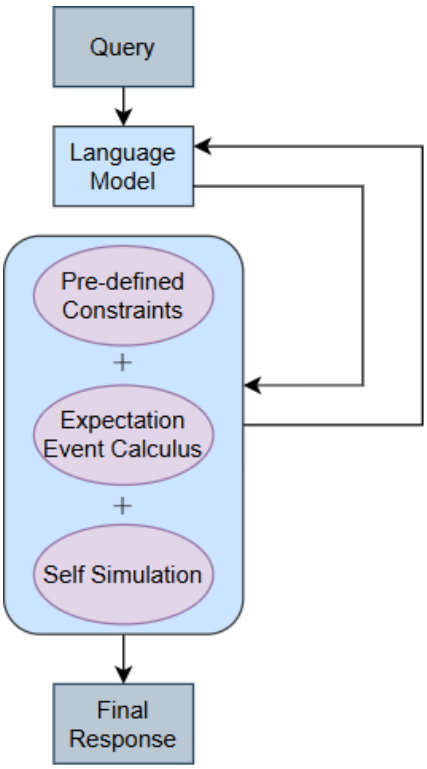


Figure 1: Reflective governance for LMs ensures responses meet safety and expectation standards. Reflective modules evaluate the response for potential consequences and compliance with predefined criteria. If met, the response is delivered; otherwise, the query is revised, and the process repeats until all conditions are satisfied.

## 3 Governance Loop on LLMs

The core idea revolves around a continuous feedback loop that governs the model's actions and outputs before they are delivered to the end user. At the heart of this system lies the reflective reasoning module—a cognitive filter designed to evaluate potential actions or responses before they are executed. This module assesses the alignment of the model's outputs with predefined behavioral constraints, ethical guidelines, or contextual requirements. Essentially, it can function as a safeguard, enabling the system to self-regulate and prevent undesirable behaviors.

Although current literature primarily focuses on surface-level checks—like [Inan *et al.*, 2023] ensuring responses are free from overtly offensive content—this is only the first step in creating a robust governance framework. Deeper investigations into the underlying biases and implicit assumptions [Bai *et al.*, 2024] embedded in outputs are crucial for trustworthy AI systems. For example, aligning responses with social norms [Lloyd and Lewis, 2023], cultural expectations [Wong, 2020], or ethical principles can reduce the risk of harmful consequences.

Advanced modules could enhance this governance loop by employing techniques such as expectation calculus, which evaluates the logical consistency of outputs with predefined expectations, or self-simulation, where the model anticipates potential consequences of its outputs. These additional layers of scrutiny can prevent harmful, biased, non-normative behaviors, or misleading outcomes, pushing the boundaries of what is currently achievable in reflective AI systems.

## 4 Reflective Governance in Practice

To design such a framework, one could utilize expectation-based event calculus [Cranefield, 2014] as a foundation for encoding the logical rules, constraints, and expectations. This approach provides a structured reasoning mechanism to model events and their causal relationships over time. These logical rules can be dynamically adjusted to account for context, enabling greater adaptability and precision.

Self-simulation can complement this setup by allowing the model to generate hypothetical scenarios and predict the consequences of its actions or responses before finalizing them. By simulating potential outcomes within the framework, the system can identify and correct errors, inconsistencies, or violations of constraints.

Moreover, an LLM-based classification module can evaluate generated responses against the predefined rules. This module operates as an intermediary, verifying outputs for issues such as the use of forbidden words or predefined unsafe contexts. As shown in 1, when a module identifies unmet constraints, it can refine the prompt to address the specified criteria in the next iteration.

Integrating these layers to an LLM-based AI system can reinforce the system's reasoning processes, aligning them more closely with the intended objectives. Additionally, reflective reasoning layers may be supported by integrated knowledge-bases and further reasoning models as well. Such approach can enhance the reliability and trustworthiness of AI systems operating in complex, real-world applications.

## 5 Conclusion

Despite the extensive body of literature on large language models and the diverse approaches to enforcing safety taxonomies, a critical gap remains: these systems lack mechanisms for self-reasoning and reflection on their behavior. This paper can address this limitation by proposing a reflective governance architecture for language models, recognized as prominent examples of AI systems.

The proposed architecture introduces modular components that enable LLMs to iteratively assess and refine their outputs, incorporating reflective reasoning layer for governance into the response generation process. By embedding reflection into AI workflows, this framework can enhance the system's ability to self-correct, follow the constraints, and maintain safety and ethical standards.

Future work will focus on implementing the proposed framework and architecture for LLMs to develop systems that are not only robust and reliable but also capable of continuous improvement in alignment with societal norms.

## References


[Bai *et al.*, 2024] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024.

[Boden, 2016] Margaret A Boden. Ai: Its nature and future. 2016.

[Colley *et al.*, 2012] Binta M Colley, Andrea R Bilics, and Carol M Lerch. Reflection: A key component to thinking critically. *The Canadian Journal for the Scholarship of Teaching and Learning*, 3(1), Sep. 2012.

[Cranefield, 2014] Stephen Cranefield. Agents and expectations. pages 234–255, 2014.

[Inan *et al.*, 2023] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[Lewis and Sarkadi, 2024] Peter R Lewis and Ştefan Sarkadi. Reflective artificial intelligence. *Minds and Machines*, 34(2):1–30, 2024.

[Lloyd and Lewis, 2023] Nathan Lloyd and Peter R Lewis. Towards reflective normative agents. In *Conference of the European Social Simulation Association*, pages 587–599. Springer, 2023.

[Pitt, 2014] Jeremy Pitt. Computer after me, the: Awareness and self-awareness in autonomic systems. Aug 2014.

[Winfield *et al.*, 2019] Alan F. Winfield, Katina Michael, Jeremy Pitt, and Vanessa Evers. Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue]. *Proceedings of the IEEE*, 107(3):509–517, 2019.

[Wong, 2020] Pak-Hang Wong. Cultural differences as excuses? human rights and cultural values in global ethics and governance of ai. *Philosophy & Technology*, 33(4):705–715, 2020.