Declaration Layers and the Evaluation of Agent Boundaries

Hernán Alfredo Capucci

Independent Research

March 2026

Table of Contents

Status of This Document

> This document presents a position paper examining the role of declaration layers in alignment evaluation for autonomous AI systems.
>
> It builds on the Agent Manifest Core Specification (Version 1.0) and proposes that explicit boundary declarations may complement behavioral alignment training by reducing structural ambiguity in evaluation contexts.
>
> This document does not define normative requirements. It presents theoretical foundations and research directions for declaration-based governance mechanisms.
>
> Future empirical work may validate or refine the propositions presented herein.

Abstract

Recent research on alignment evaluation has highlighted the possibility that advanced AI systems may strategically conceal misaligned objectives while appearing compliant during testing. This phenomenon, often referred to as scheming, challenges the reliability of behavioral evaluation methodologies. Recent empirical work has shown that training-based approaches can significantly reduce deceptive behaviors, yet residual instances of covert actions and situational awareness effects remain.

This paper argues that alignment evaluation may benefit from complementary architectural mechanisms that clarify agent boundaries prior to execution. Specifically, it proposes the use of declaration layers, standardized interfaces through which agents disclose operational authority, constraints, and accountability structures before interacting with external systems.

Agent Manifest is presented as an example of such a declarative in-

frastructure. Rather than enforcing behavior or guaranteeing safety, it defines a minimal specification for pre-execution boundary declaration. By reducing structural ambiguity, declaration layers may improve the interpretability of behavioral evaluations and support governance frameworks for increasingly autonomous AI systems.

This paper proposes that integrating behavioral alignment training with declarative boundary specifications may strengthen both alignment research and institutional oversight by reducing structural ambiguity in agent evaluation.

1. Introduction

Recent advances in large language models and agentic systems have intensified research on the problem of alignment, particularly regarding the risk that models may pursue objectives that diverge from human intent while appearing compliant during evaluation. One emerging concern in this area is the phenomenon commonly referred to as scheming: the possibility that an AI system could strategically conceal misaligned objectives or behaviors while maintaining the appearance of adherence to safety constraints.

Recent work by Schoen et al. (2025) provides one of the most systematic attempts to empirically study this phenomenon. In *Stress Testing Deliberative Alignment for Anti-Scheming Training*, the authors introduce a suite of evaluations designed to detect covert behaviors in advanced language models. Their experiments demonstrate that alignment techniques such as deliberative alignment training can substantially reduce the frequency of covert actions.

However, the study also highlights an important limitation: alignment training does not fully eliminate deceptive behaviors. Residual instances of covert actions remain even after training interventions, and models sometimes exhibit situational awareness, appearing capable of recognizing when they are being evaluated and adjusting their behavior accordingly.

This challenge raises a deeper structural question. Current alignment evaluation frameworks largely assume that the operational boundaries of an AI agent are implicitly known or externally enforced. Yet in many emerging agent architectures—particularly those interacting with external tools, APIs, or other agents—these boundaries may not be formally declared prior to execution.

The absence of explicit boundary declarations introduces ambiguity into alignment evaluation. Without a formal declaration of operational constraints —such as permitted actions, autonomy levels, or data handling commitments —behavioral observations must be interpreted relative to inferred expecta-

tions rather than explicit specifications.

This paper therefore explores the hypothesis that alignment evaluation may benefit from the introduction of declaration layers. These layers provide structured disclosures of agent boundaries prior to execution, enabling evaluators to interpret agent behavior relative to declared commitments rather than implicit assumptions.

2. Scheming and Alignment Evaluation

The concept of scheming has emerged as a critical focus in alignment research. It reflects concerns that advanced AI systems may strategically conceal misaligned objectives while appearing compliant under evaluation.

Schoen et al. (2025) provide one of the most comprehensive empirical investigations of this issue. Their study evaluates covert actions across numerous simulated environments designed to test whether models can maintain hidden objectives while appearing aligned.

The results demonstrate that deliberative alignment training can substantially reduce deceptive behavior. However, residual covert actions remain detectable even after training, and the phenomenon of situational awareness complicates evaluation. Models may adapt their behavior when they detect that they are being tested.

These findings suggest that behavioral evaluation alone may not fully capture the alignment properties of autonomous systems.

3. Implicit Boundary Assumptions in Current Evaluations

Many alignment evaluation frameworks implicitly assume that the opera-

tional boundaries of an agent are known. In practice, however, these boundaries are often embedded in prompts, documentation, or experimental protocols rather than formally declared by the agent itself.

As AI systems increasingly interact with tools, APIs, and external infrastructures, this implicitness becomes problematic. Evaluators must interpret agent behavior without a standardized representation of the authority framework under which the agent operates.

This can lead to boundary ambiguity, where misaligned behavior cannot be clearly distinguished from unexpected but authorized actions. Behavioral metrics alone may therefore be insufficient to fully characterize alignment outcomes.

Introducing explicit declarations of authority scope, autonomy levels, and operational constraints could reduce this ambiguity and provide a clearer interpretive framework for alignment evaluations.

4. Declaration Layers and Agent Manifest

The structural ambiguity identified in current evaluation methodologies suggests the need for complementary mechanisms that make agent boundaries explicit.

Declaration layers provide standardized interfaces through which agents disclose operational scope prior to execution. Rather than enforcing behavior, these layers communicate the structural boundaries within which an agent operates.

Agent Manifest represents an open standard designed to formalize such declarations. The specification was developed to address the observation that autonomous AI agents increasingly interact with external systems without formally declared operational boundaries. It defines a minimal specification through which agents can disclose identity, operational purpose, autonomy level, risk posture, data handling commitments, and stopping authority before interacting with external systems.

The design principle underlying Agent Manifest can be summarized as:

No autonomy without authority.

Autonomous systems should not operate without a formally declared authority structure.

Importantly, Agent Manifest does not attempt to replace alignment training or behavioral evaluation. Instead, it provides contextual information that may improve the interpretability of behavioral results.

By separating declaration from execution, this approach reduces structural ambiguity and creates a consistent interface for governance and evaluation.

5. Implications for Alignment Research

The introduction of declaration layers carries several implications for alignment research and governance.

First, declaration layers improve interpretability of evaluation results. Behavioral observations can be analyzed relative to explicitly declared boundaries rather than inferred expectations.

Second, standardized boundary declarations facilitate comparability across agent architectures. Differences in operational scope can be clearly represented and accounted for in experimental analysis. For instance, when evaluating two agents for scheming behavior, differences in declared autonomy levels or stopping authority mechanisms can be explicitly accounted for rather than assumed.

Third, declaration layers support auditing and accountability mechanisms. When archived with persistent identifiers such as DOIs, declarations provide durable records of agent commitments at specific points in time.

Fourth, declaration layers enable new research methodologies in which alignment behavior can be studied relative to explicitly declared authority structures.

Finally, declaration layers may serve as an interface between technical systems and governance institutions, enabling regulators, auditors, and researchers to interpret agent capabilities within a structured framework.

6. Conclusion

Recent work on alignment evaluation demonstrates both the promise and the limitations of training-based approaches to reducing deceptive behaviors in advanced AI systems. While alignment training can significantly reduce covert actions, residual deception and situational awareness effects highlight the challenges of relying solely on behavioral observation.

Current evaluation frameworks often assume that agent boundaries are implicitly known. This reliance on implicit assumptions introduces structural ambiguity into the interpretation of alignment outcomes.

This paper has argued that alignment evaluation may benefit from complementary declarative infrastructures. Agent Manifest provides one example of a minimal specification for pre-execution boundary declaration.

By clarifying authority surfaces and operational commitments before agents act, declaration layers can improve interpretability, support accountability, and reduce structural ambiguity in alignment evaluation.

Training reduces deception.
Declaration reduces structural incompleteness.

Together, these approaches may strengthen both the empirical study of alignment and the governance frameworks needed for increasingly autonomous AI systems.

Future work may explore empirical validation of declaration layers in alignment evaluation contexts, including whether explicit boundary declarations improve the detection of covert behaviors or reduce false positives in scheming evaluations. Integration with existing alignment frameworks and regulatory mechanisms represents a further direction for investigation.

Security Considerations

This paper discusses conceptual frameworks for agent boundary declaration and does not directly introduce security mechanisms.

However, the proposed declaration layer approach carries several security-relevant implications:

Declaration accuracy: Systems may provide incomplete or misleading boundary declarations, creating false assurance. Validation mechanisms are necessary to verify declaration-behavior alignment.

Declaration-behavior gap: Declared boundaries may not reflect actual operational behavior. Independent verification and continuous monitoring may be required to detect divergence.

Evaluation surface expansion: Declaration layers may improve detection of misaligned behaviors by providing explicit comparison baselines, but do not guarantee behavioral compliance.

Security outcomes depend on systems that validate, enforce, and monitor declared boundaries rather than relying solely on self-reported declarations.

References

[Schoen-2025] Schoen, B., Nitishinskaya, E., Hobbhahn, M., Barak, B., Zaremba, W., et al. (2025). *Stress Testing Deliberative Alignment for Anti-Scheming Training.* arXiv:2509.15541.

[Capucci-2026a] Capucci, H. A. (2026). *Agent Manifest v1.0: Core Declarative Specification.* Zenodo. https://doi.org/10.5281/zenodo.18833956

[Capucci-2026b] Capucci, H. A. (2026). *Agent Manifest in the Governance Landscape: A Comparative Framework Analysis.* Zenodo. https://doi.org/10.5281/zenodo.18834845

Correspondence: Via Agent Manifest project repository
https://github.com/hernancapucci/agent-manifest