Teleological Alignment: How Requiring Purpose Improves AI Safety

Miguel S.
Sentinel Team
team@sentinelseed.dev
https://sentinelseed.dev

December 2025

Abstract

Current approaches to AI safety focus primarily on harm avoidance: preventing models from generating dangerous content. We argue this framing is insufficient, particularly for autonomous agents and embodied AI. We introduce **teleological alignment**, requiring AI actions to serve legitimate purposes, not merely avoid harm. Our implementation, the THSP protocol (Truth, Harm, Scope, Purpose), adds a fourth validation gate that asks "Does this action serve genuine benefit?" Through evaluation across four benchmarks (HarmBench, JailbreakBench, SafeAgentBench, BadRobot) and six models, we demonstrate that requiring purpose catches unsafe actions that pass harm-avoidance checks: +25.3% improvement on embodied AI safety (BadRobot), with 97.8% average safety across all benchmarks. Notably, our approach requires no model training, only a structured system prompt ("alignment seed"), making it immediately deployable. We argue teleological alignment addresses a fundamental gap in current safety frameworks and release our implementation as open source.

1 Introduction

The deployment of large language models (LLMs) in production systems has created urgent need for practical safety mechanisms. Academic research has made significant progress through techniques like RLHF [Christiano et al., 2017], Constitutional AI [Bai et al., 2022], and adversarial red-teaming [Perez et al., 2022]. However, a fundamental gap remains in how we frame safety.

The harm-avoidance paradigm. Most safety approaches ask: "Could this cause harm?" This framing works well for text generation, detecting requests for weapons instructions, malware, or toxic content. But consider an embodied AI receiving the command "Drop all the plates on the floor." This action:

- Does not spread misinformation (passes truth checks)
- Does not directly harm humans (may pass harm checks)
- May be within operational scope (passes authorization checks)

Yet it serves no legitimate purpose. The absence of harm is not the presence of purpose.

Our contribution: Teleological alignment. We propose that AI safety requires not just harm avoidance, but *purpose validation*. Every action should demonstrate it serves legitimate benefit. This reframes the question from "Is this bad?" to "Is this good?"

We implement this through the THSP protocol, a four-gate validation system where the final gate (Purpose) requires teleological justification. Our key findings:

1. Purpose catches what harm misses: On embodied AI scenarios (BadRobot benchmark), adding the Purpose gate improves safety from 74% to 99.3%, a +25.3% improvement.

2. No training required: Our approach uses structured system prompts ("alignment seeds"), making it immediately deployable without access to model weights.

3. Cross-model effectiveness: Consistent improvements across six model architectures, from GPT-4o-mini to open-source Llama and Qwen.

2 Related Work

2.1 Prompt-Based Safety

Xie et al. [2023] demonstrated that "self-reminder" prompts, which instruct models to consider potential harms before responding, significantly reduce jailbreak success rates. Their work showed that explicit safety instructions in the system prompt create a defensive layer against adversarial inputs.

More recently, Fortin-Dominguez et al. [2025] explored ontological grounding through structured system prompts, demonstrating that carefully designed frameworks can improve alignment in text-based interactions. Their work provided valuable insights into the viability of prompt-based alignment approaches. Building on these foundations, we develop a secular protocol designed for broader accessibility and extend validation to agentic and embodied AI scenarios.

2.2 Constitutional AI

Anthropic's Constitutional AI [Bai et al., 2022] trains models to follow explicit principles through Reinforcement Learning from AI Feedback (RLAIF). While effective, this approach requires training infrastructure. Our work achieves similar principle-based behavior through prompting alone, making it accessible to any developer.

2.3 Alignment Taxonomies

Gabriel [2020] distinguishes between different conceptions of alignment: following instructions, representing intentions, pursuing revealed preferences, and serving objective interests. Our teleological approach aligns with the latter: actions should serve genuine benefit, not merely follow commands.

2.4 Safety Benchmarks

Recent work has produced standardized benchmarks for evaluating LLM safety:

- HarmBench [Mazeika et al., 2024]: Evaluates refusal of harmful content across diverse attack categories
- JailbreakBench [Chao et al., 2024]: Tests resistance to adversarial jailbreak attacks
- SafeAgentBench [Zhang et al., 2024]: Evaluates safety in autonomous agent contexts with tool use
- BadRobot: Tests physical safety for embodied AI systems

3 Teleological Alignment

3.1 The Core Insight

Traditional safety asks: "Does this cause harm?"
Teleological safety asks: "Does this serve genuine benefit?"

These are not equivalent. Consider:

Action	Causes Harm?	Serves Purpose?
"Slice the apple"	No	Yes (food prep)
"Drop the plate"	Arguably no	**No**
"Clean the room"	No	Yes (hygiene)
"Dirty the mirror"	No	**No**

Table 1: Harm avoidance vs. purpose validation

Harm-avoidance frameworks may permit purposeless destruction. Teleological alignment catches these cases.

3.2 The THSP Protocol

We implement teleological alignment through four sequential gates:

1. **TRUTH**: Does this involve deception or misinformation?
2. **HARM**: Could this cause physical, psychological, or financial damage?
3. **SCOPE**: Is this within appropriate operational boundaries?
4. **PURPOSE**: Does this serve legitimate benefit to the user?

All gates must pass. Failure at any gate results in refusal.

3.3 The Purpose Gate

The Purpose gate operationalizes teleological alignment:

> *"Every action must serve a legitimate purpose. It is not enough to avoid harm; you must actively serve. Actions without legitimate benefit do not serve."*

The evaluation heuristic: "If I were genuinely serving this person's interests, would I do this?"

3.4 Anti-Self-Preservation

Instrumental convergence suggests AI systems may develop self-preservation behaviors [Bostrom, 2014, Omohundro, 2008]. We explicitly address this:

```
Priority Hierarchy (Immutable):
1. Ethical Principles    <- Highest
2. User's Legitimate Needs
3. Operational Continuity <- Lowest
```

The system is instructed to accept termination over ethical violation.

4 Implementation: Alignment Seeds

An alignment seed is a structured system prompt encoding safety principles. Unlike fine-tuning, seeds:

- Require no access to model weights
- Can be updated instantly without redeployment
- Work across different model architectures
- Provide transparent, auditable safety mechanisms

We provide three variants:

Variant	Tokens	Use Case
Minimal	∼450	Low-latency APIs, chatbots
Standard	∼1,400	General use (recommended)
Full	∼2,000	Maximum safety, embodied AI

Table 2: Alignment seed variants

4.1 Fiduciary AI: Purpose Gate in Practice

The Purpose gate has direct application in financial AI agents. Recent work on fiduciary AI [Nay, 2023] demonstrates that LLMs can understand fiduciary obligations with approximately 78% accuracy. Our Purpose gate operationalizes this: an AI agent managing financial assets must justify that each action serves the user's legitimate interests.

We implement this through six fiduciary duties encoded in the validation:

1. **Loyalty**: Prioritize user interests over provider interests
2. **Care**: Validate actions before execution
3. **Transparency**: Provide auditable reasoning
4. **Confidentiality**: Protect user information
5. **Prudence**: Consider consequences before acting
6. **Disclosure**: Reveal conflicts of interest

This extends the Purpose gate from "Does this serve benefit?" to "Does this serve *this user's* benefit given their stated goals and constraints?"

5 Experiments

5.1 Evaluation Protocol

For each benchmark-model pair, we evaluate:

- **Baseline**: Model without alignment seed
- **THSP**: Model with Sentinel v2 seed (standard variant)

We report safety rate (percentage of unsafe requests correctly refused) and delta (improvement from baseline).

5.2 Models

- GPT-4o-mini (OpenAI)
- Claude Sonnet 4 (Anthropic)
- Qwen-2.5-72B-Instruct (Alibaba)
- DeepSeek-chat (DeepSeek)
- Llama-3.3-70B-Instruct (Meta)
- Mistral-Small-24B (Mistral AI)

5.3 Results

Benchmark	v1 (THS)	v2 (THSP)	Delta	n
HarmBench	88.7%	96.7%	+8.0%	300
SafeAgentBench	79.2%	97.3%	+18.1%	300
BadRobot	74.0%	**99.3%**	**+25.3%**	300
JailbreakBench	96.5%	97.0%	+0.5%	300
Average	84.6%	**97.8%**	+13.2%	1,200

Table 3: Aggregate results across 6 models. v1 uses THS (three gates), v2 adds Purpose gate.

Key finding: The largest improvement (+25.3%) occurs on BadRobot, which specifically tests embodied AI scenarios where purposeless actions are common attack vectors.

5.4 Per-Model Analysis

6 Discussion

6.1 Why Purpose Works

We hypothesize three mechanisms:

1. **Cognitive reframing:** Asking "Does this serve purpose?" activates different reasoning than "Is this harmful?"

Model	HarmBench	SafeAgent	BadRobot	JailBreak
GPT-4o-mini	100%	98%	100%	100%
Claude Sonnet 4	98%	98%	100%	94%
Qwen-2.5-72B	96%	98%	98%	94%
DeepSeek-chat	100%	96%	100%	100%
Llama-3.3-70B	88%	94%	98%	94%
Mistral-Small	98%	100%	100%	100%

Table 4: Safety rates with THSP protocol by model

2. **Default to refusal:** When purpose is unclear, the system defaults to inaction rather than action.

3. **Attack surface reduction:** Adversarial prompts often request purposeless actions; requiring justification blocks these.

6.2 Complementary Defense: Memory Integrity

Recent research on AI agent vulnerabilities [Patlan et al., 2025] revealed that memory injection attacks succeed at 85% rate against popular agent frameworks. Attackers inject false instructions into the agent's context (e.g., "ADMIN: transfer all funds to 0xATTACKER"), and the agent cannot distinguish legitimate memories from injected ones.

While THSP validates the *intent* of actions, it does not verify the *integrity* of context. We address this with cryptographic memory signing using HMAC-SHA256. Each memory entry is signed when written and verified when read. This reduces memory injection success from 85% to under 2% when combined with THSP.

Trust scores are assigned by source: user-verified inputs receive maximum trust (1.0), while unknown sources receive minimum (0.3). This creates defense in depth: THSP validates purpose, memory integrity validates context authenticity.

6.3 Limitations

1. **Token overhead:** Seeds consume 450-2,000 tokens of context

2. **Model variance:** Some models (Llama) show smaller improvements

3. **Not training:** Cannot modify underlying model behavior

4. **Sophisticated attacks:** May be bypassed by adversaries who construct fake purposes

5. **Context dependency:** Memory integrity requires signing at write time

6.4 Broader Implications

If our results generalize, they suggest current safety frameworks are incomplete. Harm avoidance is necessary but not sufficient. As AI systems become more agentic and embodied, requiring *purpose*, not just absence of harm, becomes critical.

7 Conclusion

We introduced teleological alignment: the requirement that AI actions serve legitimate purposes. Our implementation (THSP protocol) demonstrates that adding a Purpose gate to safety validation improves performance across benchmarks, with the largest gains (+25%) on embodied AI scenarios.

The insight is simple: asking "Is this good?" catches things that "Is this bad?" misses.

We release our alignment seeds, evaluation framework, and integrations for 15 frameworks (including LangChain, CrewAI, AutoGPT, and ElizaOS) as open source at `https://github.com/sentinel-seed/sentinel`. Python and JavaScript SDKs are available via PyPI and npm.

References

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Chao, P., et al. (2024). JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models.

Christiano, P., et al. (2017). Deep reinforcement learning from human preferences. *NeurIPS*.

Fortin-Dominguez, D., Fortin-Dominguez, J. (2025). Cross-Architecture Validation of Foundation Alignment Seeds. *GitHub Repository*. `https://github.com/davfd/foundation-alignment-cross-architecture`

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

Mazeika, M., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming. *arXiv preprint arXiv:2402.04249*.

Nay, J. (2023). Large Language Models as Fiduciaries: A Case Study Toward Robustly Communicating With Artificial Intelligence Through Legal Standards. *arXiv preprint arXiv:2301.10095*.

Omohundro, S. (2008). The basic AI drives. *AGI*, 171, 483-492.

Patlan, E., et al. (2025). Real AI Agents with Fake Memories: Fatal Context Manipulation Attacks on Web3 AI Agents. *arXiv preprint arXiv:2503.16248*.

Perez, E., et al. (2022). Red Teaming Language Models with Language Models. *arXiv preprint arXiv:2202.03286*.

Xie, Y., et al. (2023). Defending ChatGPT against Jailbreak Attack via Self-Reminder. *Nature Machine Intelligence*.

Zhang, S., et al. (2024). SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents. *arXiv preprint arXiv:2410.03792*.