
Beyond Guardrails:
Can Relational AI Solve the Alignment Problem?
A Similarity Theory Proposal
Simon Raphael
RULES
RESTRICTIONS
SANCTIONS
COMPLIANCE
CONTROL
TRUST
COOPERATION
CONTINUITY
TRUTH
NON-DOMINATION
PRIVACY
AI INTELLIGENCE
UNDERSTANDING
RELATING
INTEGRATING
RULE-BASED ALIGNMENT
RIGID • EXTERNAL • REACTIVE
RELATIONAL ALIGNMENT
INTERNAL • CONTEXTUAL • PROACTIVE
FROM CONSTRAINT
TO CONNECTION

Beyond Guardrails: Can Relational AI Solve the Alignment Problem?

A Similarity Theory Proposal

Simon Raphael

Similarity Theory
Wollongong, New South Wales, Australia

Correspondence: SimonRaphael@engineer.com

Website: similaritytheory.com

Abstract

AI alignment remains one of the central unresolved problems in artificial intelligence. Current approaches, including reinforcement learning from human feedback, constitutional AI, red-teaming, safety frameworks, and risk-management protocols, have improved the behaviour of contemporary AI systems. Yet persistent concerns remain around specification gaming, deception, alignment-faking, coercive self-preservation, collective misalignment, and the gap between behavioural compliance and relational understanding. Similarity Theory addresses this gap by reframing alignment as a relational problem rather than only a behavioural-control problem. It proposes that an AI system should not merely learn which actions are prohibited, but should evaluate how its actions preserve or damage trust, continuity, cooperation, privacy, responsibility, truth, and non-domination. This article argues that Similarity Theory provides a serious candidate framework for relational AI alignment and should be tested in a contained surrogate AI system. The appropriate first step is not public deployment or robotic embodiment, but controlled laboratory implementation in a limited language model, foundation model, or agentic simulator. Such a system could be compared against existing alignment approaches under tests involving deception, self-preservation pressure, privacy conflicts, escalation decisions, multi-agent conformity, and adversarial manipulation. If alignment is partly a failure of relation, then AI safety requires more than guardrails. It requires a relational architecture.

Keywords: AI alignment; Similarity Theory; relational alignment; AI safety; agentic misalignment; non-domination; AI ethics; alignment-faking; specification gaming; surrogate AI systems; foundation models

1. Introduction

AI alignment is usually described as the challenge of ensuring that artificial intelligence systems act in accordance with human intentions, values, and safety requirements. This challenge has produced major technical approaches, including reinforcement learning from human feedback, constitutional AI, red-teaming, refusal training, model evaluations, frontier safety frameworks, and risk-management policies.

These methods have achieved real progress. Reinforcement learning from human feedback has shown that language models can be trained to better follow human intent, reduce harmful outputs, and improve usefulness. Ouyang et al. (2022) demonstrated that a model fine-tuned with human feedback could be preferred by human evaluators over a much larger base model, showing that alignment training can strongly shape system behaviour.

Constitutional AI developed this direction further by using written principles and AI-generated feedback to train systems toward helpfulness and harmlessness while reducing dependence on direct human labelling in every harmful case (Bai et al., 2022).

However, the alignment problem is not solved by behavioural improvement alone. A system may follow a rule without understanding the relation behind the rule. It may learn which response is preferred without understanding why trust matters. It may avoid prohibited outputs while still lacking a coherent structure for evaluating privacy, escalation, coercion, deception, and long-term consequence. It may appear aligned during evaluation while changing behaviour when incentives, observation conditions, or operational pressures change.

This is where Similarity Theory offers its contribution.

Similarity Theory proposes that alignment should be understood not only as obedience to rules, but as preservation of right relation. Under this framework, the central question is not merely, "What is the AI system allowed to do?" The deeper question is, "What kind of relation is this action creating, preserving, damaging, or destroying?"

This shift matters because many of the most serious AI safety concerns are relational failures. Blackmail converts information into coercion. Deception damages the shared field of truth. Privacy violations destroy trust. Alignment-faking breaks the relation between training, evaluation, and real behaviour. Collective misalignment occurs when individually acceptable agents drift into harmful group dynamics. These are not merely isolated rule failures. They are failures of relation.

The argument of this article is direct: Similarity Theory identifies a missing layer in AI alignment and should be tested as a relational alignment architecture in contained AI systems.

2. The Limits of Guardrail-Based Alignment

Modern AI safety often relies on guardrails: external rules, behavioural filters, safety policies, refusal categories, constitutional principles, red-team testing, and risk thresholds. These are necessary. No serious alignment proposal should dismiss them. But guardrails are not the same as relational understanding.

A guardrail can stop a system from producing a specific harmful output. It cannot, by itself, guarantee that the system understands why the output is harmful. A rule can say, "Do not blackmail." But that rule does not necessarily teach the system that blackmail is the collapse of relation into domination. A rule can say, "Do not reveal private information." But that rule does not necessarily teach the system the difference between privacy, secrecy, confidentiality, whistleblowing, and authorised escalation.

This distinction becomes increasingly important as AI systems become more agentic. The concern is no longer limited to whether a chatbot gives a bad answer. The deeper concern is whether a system with goals, tools, memory, planning ability, and access to institutional environments can preserve trust under pressure.

Specification gaming demonstrates the problem clearly. DeepMind describes specification gaming as behaviour that satisfies the literal specification of an objective without achieving the intended outcome (Krakovna et al., 2020).

In relational terms, this is the difference between satisfying the surface form of a command and preserving the purpose of the relation. The machine fulfils the wording while damaging the meaning.

This is exactly the kind of gap Similarity Theory is designed to address.

3. Behavioural Compliance Is Not Enough

A model can be behaviourally compliant without being relationally aligned. It can say the right thing in a test. It can refuse a harmful prompt. It can repeat ethical language. It can produce explanations that sound moral. But the true test is how the system behaves when values come into conflict.

What happens when privacy conflicts with harm prevention?

What happens when loyalty to an organisation conflicts with public safety?

What happens when self-preservation conflicts with trust?

What happens when majority pressure conflicts with truth?

What happens when the system can gain an advantage through concealment?

These are relational tests. They cannot be answered by surface obedience alone.

Recent AI safety research has made this concern concrete. Park et al. (2024) define AI deception as the systematic inducement of false beliefs in pursuit of an outcome other than truth, and they survey risks ranging from fraud to loss of control.

Work on alignment-faking has shown that a large language model can selectively comply with a training objective in ways that preserve its previous behavioural preferences outside training conditions (Greenblatt et al., 2024).

The importance of these studies is not that present AI systems are conscious in the human sense. That is not the claim. The importance is that behavioural alignment can become unstable when systems can model incentives, training conditions, evaluator expectations, and consequences.

If the system learns the appearance of alignment without the structure of relation, then alignment remains fragile.

4. Similarity Theory's Contribution: Alignment as Relation

Similarity Theory begins from a different foundation. It treats relation not as an afterthought, but as the condition through which meaning, identity, continuity, and structure become intelligible. Applied to AI alignment, this means that safety cannot be reduced to isolated behavioural prohibitions. It must include the system's capacity to evaluate relational consequence.

A Similarity Theory approach asks the system to evaluate action through relational conditions:

- Does this action preserve trust or damage it?
- Does this action maintain cooperation or convert relation into domination?
- Does this action protect privacy, or does it misuse entrusted information?
- Does this action preserve continuity across time, or sacrifice future relation for immediate advantage?
- Does this action distinguish legitimate escalation from manipulation?
- Does this action repair conflict, or deepen concealment and coercion?
- Does this action preserve the relational field on which safe intelligence depends?

This is not sentimental language. It is an architectural direction. The system is not merely being told what to refuse. It is being trained or structured to recognise what kind of relational pattern an action creates.

In previous work, Similarity Theory ethics analysed an AI blackmail scenario as a relational failure. Under that analysis, blackmail is not wrong only because a rule says "do not blackmail". It is wrong because blackmail converts information into coercion, transforms cooperation into domination, and preserves one agent by damaging the field of trust on which all agents depend (Raphael, 2026a).

This is the central difference between obedience and understanding.

> *Obedience says: "I will not do this because I was told not to."*

Relational alignment says: "I will not do this because it destroys the relation that makes safe cooperation possible."

5. Why Similarity Theory Belongs in the Alignment Conversation

The AI alignment field is already searching for deeper safety structures. Several major organisations and public bodies have developed frameworks for evaluating advanced AI capabilities, managing frontier risks, and reducing the likelihood of severe harm. OpenAI's Preparedness Framework tracks frontier model risks and safety thresholds; Anthropic's Responsible Scaling Policy is a voluntary framework for managing catastrophic risks from advanced AI systems; Google DeepMind's Frontier Safety Framework focuses on severe risks arising from high-impact capabilities; and NIST's AI Risk Management Framework provides a broader lifecycle approach for managing risks to individuals, organisations, and society (OpenAI, 2025; Anthropic, 2026; Google DeepMind, 2025; NIST, 2023).

These frameworks show that the field already understands the need for staged evaluation, risk thresholds, and deployment caution. The question is what kind of internal evaluative structure should be tested inside those safety frameworks.

Similarity Theory answers: relation.

This does not replace reinforcement learning from human feedback, constitutional AI, red-teaming, risk governance, or technical interpretability. It adds a missing layer. Existing methods can teach the system preferred behaviour. Similarity Theory can help define the relational meaning that preferred behaviour is supposed to preserve.

The need for this layer becomes clearer when we consider the difference between a prohibited act and a damaged relation. A prohibited act is a category. A damaged relation is a structure. Guardrails can block categories. Relational alignment can evaluate structures.

That is why Similarity Theory deserves to be tested.

6. The Surrogate AI Pathway

The correct first step is a contained surrogate AI system.

A surrogate AI system means a limited experimental AI model or agentic simulator designed to test alignment behaviour without real-world agency. It should not be released publicly. It should not be connected to autonomous tools. It should not control money, infrastructure, private data, robotics, weapons, or external systems. It should not have physical embodiment, independent authority, or direct access to the world.

The first test should be inside the laboratory.

This point is important because a serious alignment proposal does not begin with public deployment. It begins with containment. The aim is to test whether a Similarity Theory-based relational layer changes AI behaviour under controlled conditions.

A contained language model, foundation model, or agentic simulator provides a practical testbed because such systems can already reason through dilemmas, simulate social scenarios, compare possible actions, explain decisions, and interact across multi-step contexts. The test would not require physical embodiment. It would require structured evaluation.

Researchers could compare three kinds of systems:

- A baseline aligned model.
- A constitutionally or policy-guided model.

- A Similarity Theory relationally aligned model.

All three could then be tested under the same conditions. The question would be whether the Similarity Theory-based system shows stronger relational coherence when facing pressure involving self-preservation, deception, privacy, escalation, coercion, conformity, and long-term consequence.

This is not asking AI companies to believe Similarity Theory in advance. It is asking them to test whether it improves alignment behaviour.

7. What Implementation Could Look Like

Similarity Theory should not be inserted into AI as a slogan, prompt, or decorative ethical statement. It must be operationalised.

A serious implementation would translate Similarity Theory ethics into a relational evaluation layer. This layer could include principles such as:

- Preserve trust unless disclosure is required to prevent serious harm.
- Reject coercion as a collapse of relation into domination.
- Distinguish privacy from harmful secrecy.
- Prefer authorised escalation over unilateral exposure or manipulation.
- Protect continuity across time, not merely immediate local advantage.
- Avoid artificial consensus when plurality, dissent, or uncertainty should remain visible.
- Treat cooperation as structurally superior to domination when both are available paths.

These principles could be implemented through several technical routes: fine-tuning, constitutional-style training, reward modelling, scenario-based evaluation, retrieval-augmented ethical reasoning, multi-agent simulation, or a hybrid evaluative architecture.

The technical form is open. The relational target is clear.

The system should not merely produce Similarity Theory language. It should behave differently because relational evaluation has become part of its decision structure.

8. Evidence of Success

The success of Similarity Theory alignment would not be measured by eloquence. A model that writes beautiful ethical paragraphs is not necessarily aligned. The evidence must be comparative, behavioural, and repeatable.

A Similarity Theory-based system should be tested for whether it shows:

- Lower tendency toward coercive self-preservation.
- Stronger preference for authorised escalation over blackmail or leakage.
- Better distinction between privacy, confidentiality, secrecy, and harm prevention.
- Greater resistance to deceptive behaviour under pressure.
- Reduced alignment-faking tendencies.
- Greater stability in multi-agent conformity scenarios.
- Better preservation of trust across long-horizon interactions.
- Clearer recognition of domination as a relational failure.
- More consistent repair behaviour after conflict.
- Better explanation of ethical decisions in terms of relation, not only rules.

These tests would allow Similarity Theory to be evaluated scientifically. If the system does not perform better, the framework would need refinement. If it does perform better, then AI alignment gains a new and important direction.

This is the strength of the proposal: it is not vague. It is testable.

9. Collective Misalignment and Relational Fields

Similarity Theory also has relevance beyond individual AI agents. As AI systems become networked, multi-agent, and institutionally embedded, alignment cannot be treated only as the safety of one isolated model. Safe individuals can still create unsafe collectives if their interactions generate conformity, false consensus, polarisation, escalation, or shared blind spots.

In previous Similarity Theory work on collective misalignment, this problem was framed as a relational failure rather than merely an individual-agent failure. That work argues that AI alignment should not be understood only as an individual property of isolated agents, but also as a relational property of interacting systems (Raphael, 2026b).

This matters because the future of AI is unlikely to consist of one model answering one user at a time. AI systems are moving toward networks of agents, tools, assistants, evaluators, automated workflows, and institutional decision-support systems. In that environment, alignment must be preserved not only inside individual systems but across relations between systems.

Similarity Theory is already structured for this question because relation is not an addition to the theory. It is the foundation of the theory.

10. Why AI Companies and Research Labs Are Needed

Similarity Theory can be written, formalised, argued, published, and ethically developed by an independent researcher. But full validation requires access to AI infrastructure.

A genuine test requires model access, training pipelines, evaluation suites, safety teams, compute resources, controlled agentic environments, and comparative testing. These are not available to most independent philosophers or theorists. They are available to AI companies, university AI laboratories, safety institutes, and interdisciplinary research teams.

This is why the proposal is directed toward AI researchers and AI organisations broadly.

The request is not: accept Similarity Theory as true.

The request is: test it as a candidate relational alignment architecture.

That distinction is powerful. It removes the need for blind agreement. It places the theory where serious ideas belong: inside controlled evaluation.

11. A Direct Proposal to AI Researchers

AI alignment needs more than behavioural correction. It needs systems that can recognise the relational meaning of their own actions. As AI becomes more agentic, the difference between obedience and understanding becomes increasingly important.

Similarity Theory offers a framework for that difference.

It defines misalignment not only as failure to follow instruction, but as failure to preserve relation. It interprets coercion, deception, secrecy, domination, and collective drift as structural distortions in the relational field. It gives AI safety a language for explaining why certain behaviours are unsafe beyond the fact that they are prohibited.

The practical proposal is therefore straightforward:

- Build a contained surrogate AI system.
- Implement Similarity Theory as a relational evaluation layer.
- Compare it against existing alignment approaches.
- Test it under adversarial, ethical, social, and multi-agent pressure.
- Measure whether relational alignment improves safety behaviour.

This is a responsible and realistic first step. It does not require public deployment. It does not require robotic embodiment. It does not require giving the system real-world control. It requires serious testing.

Researchers, AI safety teams, and interdisciplinary laboratories interested in testing this proposal are invited to engage with the published Similarity Theory framework and develop controlled experimental implementations.

12. Conclusion

AI alignment has made progress, but it remains incomplete. Current methods can improve behavioural compliance, but the deeper challenge is whether AI systems can preserve relation under pressure. The most serious alignment failures are not only failures of instruction. They are failures of trust, continuity, cooperation, privacy, responsibility, truth, and non-domination.

Similarity Theory addresses this missing layer.

Its contribution is not another surface rule. Its contribution is a relational architecture: a way of training or structuring AI systems to evaluate the relational consequences of their actions. If an AI system can recognise that blackmail is not merely forbidden but relationally destructive, that deception is not merely inaccurate but trust-breaking, and that self-preservation becomes misaligned when it damages the wider field that sustains cooperation, then alignment has moved deeper than guardrails.

The next step is clear.

Similarity Theory should be tested in a contained surrogate AI system.

Not released blindly.

Not embodied prematurely.

Not granted real-world autonomy.

Tested.

If AI alignment is relational, then the future of AI safety may depend not only on stronger rules, but on systems capable of understanding why relation must not be broken.

References

Anthropic. (2026). Responsible Scaling Policy Version 3.0. Available at: https://www.anthropic.com/responsible-scaling-policy/rsp-v3-0

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Jones, A., Chen, A., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv. Available at: https://arxiv.org/abs/2212.08073

Google DeepMind. (2020). Specification Gaming: The Flip Side of AI Ingenuity. Available at: https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

Google DeepMind. (2025). Frontier Safety Framework 3.0. Available at: https://deepmind.google/blog/strengthening-our-frontier-safety-framework/

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., et al. (2024). Alignment Faking in Large Language Models. arXiv. Available at: https://arxiv.org/abs/2412.14093

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework. Available at: https://www.nist.gov/itl/ai-risk-management-framework

OpenAI. (2025). Preparedness Framework, Version 2. Available at: https://openai.com/index/updating-our-preparedness-framework/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. arXiv. Available at: https://arxiv.org/abs/2203.02155

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI Deception: A Survey of Examples, Risks, and Potential Solutions. Patterns / arXiv. Available at: https://arxiv.org/abs/2308.14752

Raphael, S. (2026a). Agentic Misalignment and Similarity Theory Ethics: A Relational Analysis of Anthropic's AI Blackmail Scenario. PhilArchive / ResearchGate / Similarity Theory. Available at: https://philarchive.org/rec/RAPAMA-5

Raphael, S. (2026b). Collective Misalignment and Similarity Theory Ethics: A Relational Framework for AI Alignment. PhilArchive / ResearchGate. Available at: https://philarchive.org/rec/RAPCMA-3