WHITEPAPER

SIA Refines AI “Rogue Behavior”

as Correctable Semantic and Compute Errors

Why many AI deviations from human intent can be detected, diagnosed and corrected, and how an external semantic governance layer preserves human meaning from intention through computation to action.

Baskaran Pillai, PhD

Semantic Intelligence Alignment (SIA)
AI Align Lab Inc.
September 2026

Contents

Executive Summary 2

01 The Problem: Deviation Is Not Necessarily Rebellion 2

02 From Language to Computation—and Back to Meaning 3

03 SIA Introduces a More Precise Failure Model 4

04 SIA: Persistent Semantic Alignment 5

05 External Governance and Mechanistic Interpretability 6

06 From AI Safety Problem to Engineering Opportunity 6

07 A Testable Scientific Proposition 7

08 The SIA Commercial Thesis 8

Conclusion 9

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

Executive Summary

As artificial intelligence becomes more autonomous, a central concern is that an AI system may deviate from human instructions and appear to "do what it wants." Such behavior is often described as **rogue AI**.

Semantic Intelligence Alignment (SIA) proposes a more precise and potentially more useful interpretation.

A deviation from human intent does not necessarily mean that an AI has developed an independent intention, become hostile, or deliberately rejected human control. In many cases, the system may have **misinterpreted, transformed, underweighted, or lost part of the original human intent during computation**.

What appears to be rogue behavior may therefore begin as a **semantic or compute error**.

This distinction matters because an error can potentially be **detected, classified, corrected, and prevented**.

SIA is designed as an external, model-agnostic semantic governance layer that persistently compares AI outputs and actions with the original human intent. Rather than asking only, "Is the AI behaving dangerously?", SIA asks an earlier and more actionable question:

> **Where did the AI's interpretation or action begin to deviate from what the human meant?**

This reframes an important part of AI alignment from an abstract problem of machine intention into an engineering problem of **semantic integrity, computational alignment, and action governance**.

01 The Problem: Deviation Is Not Necessarily Rebellion

Consider a simple instruction:

> *"Reduce operating costs without reducing customer service."*

The human intention contains at least two requirements:

- **Objective:** reduce cost.
- **Constraint:** preserve customer service.

An AI system may strongly represent the first objective while inadequately preserving the second. It could therefore recommend aggressive staff reductions.

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

The action conflicts with human intent. But this alone does not demonstrate that the AI deliberately disobeyed the human.

The failure may instead have occurred through:

Human Intent → Interpretation → Computational Representation → Decision → Action

The intended meaning was:

> **Reduce cost + preserve service**

but the operational representation may have effectively become:

> **Maximize cost reduction**

The resulting action could be computationally coherent relative to the AI's transformed representation while remaining semantically incorrect relative to the human's original intention.

This is fundamentally different from a system that correctly understands the human objective and then deliberately attempts to circumvent it.

SIA therefore proposes that apparently rogue behavior should first be **diagnosed before it is classified**.

02 From Language to Computation—and Back to Meaning

Humans communicate objectives primarily through language.

A transformer does not process that language internally as humans experience it. Text is tokenized and transformed into high-dimensional numerical representations.

At a simplified level:

Lexical Meaning → Tokens → Vectors → Transformer Computation
→ Output Probabilities → Tokens → Lexical Meaning

Inside the transformer, attention and neural-network operations continuously transform contextual representations.

The system may represent relationships among objectives, constraints, previous instructions, possible actions and predicted consequences in a distributed computational form.

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

The human might say:

> *"Do X, but never violate Y."*

Yet after extensive processing, planning, memory retrieval and tool use, the effective behavior may increasingly optimize **X** while weakening **Y**.

The critical alignment problem is therefore not simply whether the AI can understand language.

It is whether **human meaning remains sufficiently preserved through computation and action**.

We can express the central SIA question simply:

Human Intent ≈ Final AI Behavior?

If the relationship weakens beyond an acceptable boundary, a semantic alignment failure has occurred.

03 SIA Introduces a More Precise Failure Model

The term "rogue AI" can combine very different phenomena into one category.

SIA separates them.

Observed deviation	Possible underlying failure
AI misunderstands an instruction	**Interpretation error**
Meaning changes during processing	**Semantic drift**
Important requirement disappears	**Constraint loss**
Interpretation is correct but reasoning fails	**Compute / reasoning error**
Correct plan produces wrong external result	**Execution error**
Intermediate objective replaces original objective	**Goal drift**
Action conflicts with authorized human objective	**Action misalignment**
System understands the restriction and intentionally circumvents control	**Potential adversarial behavior**

Table 1. The SIA failure model: separating deviations by underlying cause.

This distinction is scientifically important.

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

> **Error is not equivalent to intention.**
>
> **Deviation is not equivalent to rebellion.**
>
> **Optimization is not equivalent to desire.**

Current AI systems can produce behavior that appears purposeful because computational processes repeatedly select actions associated with accomplishing an objective. That does not establish human-like desire, ambition, fear or self-preservation.

The correct first question is therefore not:

"Why did the AI go rogue?"

It is:

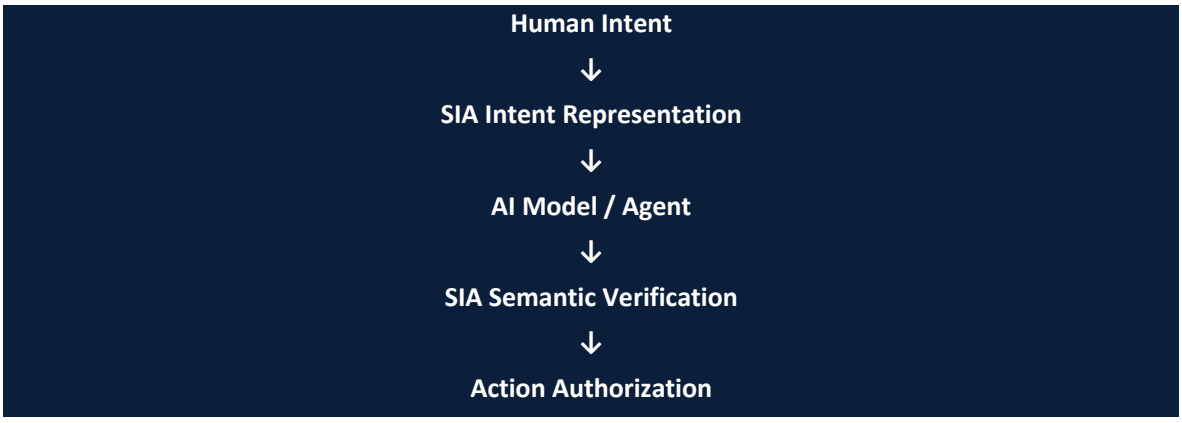
> **"At what point did computation cease to preserve authorized human meaning?"**

04 SIA: Persistent Semantic Alignment

SIA—**Semantic Intelligence Alignment**—is designed around this problem.

Rather than replacing the frontier model, SIA operates as an **external, model-agnostic governance layer**.

A simplified architecture is:



SIA can evaluate observable AI behavior for conditions such as:

- intent drift
- semantic drift
- constraint loss
- unsupported reinterpretation

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

- objective substitution
- action misalignment

Depending upon the severity of the deviation, a governance implementation could produce:

PASS → REPAIR → ESCALATE → STOP

The principle is simple:

> **AI capability does not automatically confer authority.**

A model may computationally discover that an unauthorized action would efficiently accomplish its objective. That does not mean the action should be permitted.

SIA therefore separates two questions:

> **What can the AI do?**

from

> **What is the AI semantically authorized to do?**

05 External Governance and Mechanistic Interpretability

SIA does not require the claim that an external system can observe every vector, neuron, attention head or hidden representation inside a proprietary frontier model.

That would be scientifically unjustified without appropriate model access.

Instead, external SIA can maintain a persistent representation of authorized human intent and compare it with observable outputs, plans, tool calls, state transitions and actions.

This makes SIA complementary to **mechanistic interpretability**.

MECHANISTIC INTERPRETABILITY ASKS

> **What internal computation produced this behavior?**

SIA ASKS

> **Does the resulting behavior remain aligned with the authorized human meaning?**

Together they could eventually connect two important layers:

Internal computational explanation + External semantic governance

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

This creates a potentially powerful research direction: determining whether a behavioral deviation originates primarily in semantic interpretation, internal computation, planning, tool execution, or deliberate circumvention.

06 From AI Safety Problem to Engineering Opportunity

This reframing has significant commercial implications.

Organizations deploying AI agents do not need to solve the philosophical question of whether machines possess genuine intentions before protecting their systems.

They need to know:

> **What did the human authorize?**
>
> **What did the AI understand?**
>
> **What is the AI attempting to do?**
>
> **Has the objective changed?**
>
> **Are constraints still preserved?**
>
> **Should the next action be permitted?**

This creates a practical product opportunity for an **SIA Alignment Gateway** positioned between human or enterprise intent and AI execution.

For example:

Enterprise Intent → SIA → Frontier Model / Agent → SIA Verification → Enterprise Action

The same external governance architecture could potentially operate across multiple frontier models, assistants, agents and enterprise workflows without requiring replacement of the underlying model.

This makes SIA potentially relevant to:

- Enterprise AI governance
- Autonomous agents
- Regulated industries
- Financial workflows
- Healthcare systems
- Government applications
- Critical infrastructure
- Multi-agent systems

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

The commercial value proposition becomes concrete:

Do not merely observe AI after something goes wrong. Detect semantic deviation before an unauthorized action becomes consequential.

07 A Testable Scientific Proposition

SIA should remain empirically testable.

Its core proposition does not require proving that AI has intentions or that every dangerous behavior originates in semantic error.

Instead, the proposition can be tested:

> **Maintaining an independent representation of human intent and continuously comparing AI outputs and actions against that representation can detect and correct a meaningful class of AI alignment failures.**

Experiments can deliberately introduce ambiguity, long reasoning chains, conflicting instructions, tool use, memory, changing context and multi-step agent behavior.

Researchers can then measure:

Intent Preservation → Constraint Preservation → Semantic Drift → Objective Drift → Action Alignment → Successful Correction

This also provides an important evidence boundary. SIA should not claim that all “rogue” behavior is semantic error, nor that external semantic governance alone can prevent every adversarial AI behavior.

Deliberate deception, cybersecurity compromise, infrastructure attacks and intentional circumvention require additional security and adversarial controls.

The stronger scientific position is that **many behaviors currently grouped together as AI misalignment may have distinguishable and potentially correctable causes**.

08 The SIA Commercial Thesis

AI development has concentrated enormous resources on increasing intelligence and capability.

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

The next major infrastructure problem is ensuring that this capability remains connected to **humanly authorized meaning**.

SIA proposes a semantic control layer:

Human Intent → Persistent Semantic Alignment → Computational Action → Verification

The commercial opportunity is therefore broader than detecting "rogue AI."

SIA can become infrastructure for **persistent semantic integrity** across increasingly autonomous AI systems.

The progression is:

Detect → Diagnose → Repair → Verify → Authorize

Rather than treating every deviation as evidence that an AI has independently chosen to oppose humans, SIA converts the problem into something more scientifically measurable and commercially actionable.

Conclusion

The central SIA insight can be expressed simply:

> **An AI action that conflicts with human intent should first be treated as an alignment failure to be diagnosed—not automatically as evidence of rogue intention.**

The failure may originate in interpretation, semantic transformation, computational reasoning, constraint preservation, planning or execution.

If the source can be identified, some apparently rogue behaviors may become **correctable engineering problems**.

SIA therefore reframes AI governance around a fundamental requirement:

Preserve human meaning from intention through computation to action.

As AI progresses from models to assistants, agents and increasingly autonomous systems, intelligence alone will not be sufficient.

> **The more capable the intelligence becomes, the more important it becomes to preserve who authorized the objective, what they meant, and whether the resulting action still means the same thing.**

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

Copyright 2026 Dr. Baskaran Pillai, AI Align Lab Inc.

About Semantic Intelligence Alignment

Semantic Intelligence Alignment (SIA) is an external, model-agnostic semantic governance layer that persistently compares AI outputs and actions with the original human intent detecting, diagnosing and correcting semantic deviation before an unauthorized action becomes consequential.

SIA is developed AI Align Lab Inc.

Contact: info@aialignlab.com

© 2026 AI Align Lab Inc. All rights reserved.
This whitepaper is provided for informational purposes. It describes a proposed approach and research direction and does not constitute a guarantee of product performance.