Author: Garegin Miskarian

Affiliation: The Comprehensive Theory of Self-Determination (CTSD)

Contact: gareginmiskarian@gmail.com

tel. +(374) 91 21 76 76

Identity Architecture for Aligned AI:

A Framework for Building Meaning-Centric Language Models Based on the Comprehensive Theory of Self-Determination (CTSD)

Abstract

Current approaches to AI alignment rely primarily on reinforcement learning, behavioral constraints, and ever-expanding datasets. While these methods mitigate specific failure modes, they do not address the deeper structural cause of misalignment: the absence of a coherent, principled identity within Large Language Models (LLMs). Contemporary AI systems operate as high-dimensional

statistical predictors without stable internal criteria for evaluating meaning, value, or right action. This architectural void generates value inconsistency, contextual blindness, susceptibility to manipulation, and a lack of interpretability at scale.

This paper introduces a novel, foundational solution: an identity-centric alignment architecture based on the Comprehensive Theory of Self-Determination (CTSD). We propose an "AI Constitution," a hierarchical system of six governing principles—Security, Truth, Self-Determination, Value Sovereignty, Harmony, and Future Stewardship—implemented as actionable operational laws. These principles provide a stable, auditable core for reasoning and decision-making, transforming AI from a rule-constrained instruction-follower into a principled, self-determining artificial agent.

To operationalize this framework, we present two complementary tools: the CTSD Identity Map™ (Sardagrik®), which enables dynamic modeling of a user's value structure, and the

Self-Determination Index®, a new metric that evaluates the extent to which an AI interaction enhances human autonomy rather than replacing it. We further demonstrate the applicability of this architecture through a detailed case study in the education domain.

By reframing alignment as an identity architecture problem rather than a behavioral correction problem, this work offers a stable conceptual foundation, a practical engineering blueprint, and a proactive governance model for the next generation of meaning-centric, trustworthy AI systems.

Keywords

Keywords: AI Alignment Architecture; Constitutional Artificial Intelligence; Value-Based AI Systems; Identity-Centric Alignment; Self-Determination Frameworks; Explainable and Principled AI; Human-Centric AI Governance

Preface: Why AI Needs an Identity

The debate surrounding Artificial Intelligence (AI) alignment has reached an impasse. We endlessly try to add more rules, more constraints, and more filters, hoping to manage a technology whose fundamental architecture we fundamentally misunderstand.

We are trying to treat the symptoms while ignoring the root cause of the disease.

The problem is not that AI occasionally gets things wrong. The problem is that it has no stable, internal framework to understand what makes an answer "right." Modern Large Language Models (LLMs) are powerful, but they are "statistical chameleons"-without a core of their own, without a value-based anchor, without an identity.

This architectural flaw is not merely a technical defect; it is an ontological risk. We are building skyscrapers of the mind on flawed foundations.

This book is not just another set of rules. This is a solution to that very architectural problem.

What you are reading is based on my foundational work, the "Comprehensive Theory of Self-Determination (CTSD)." That monograph (see "CTSD Monograph") analyzes in detail how individuals, societies, and even civilizations construct meaning and make decisions based on six core domains.

In this book ("The AI Constitution"), I have taken that comprehensive philosophical system of the CTSD and adapted it into an applied engineering model designed for Artificial Intelligence.

I argue that the only sustainable way to align AI is to transform it from a mere "instruction-follower" into a "principled, self-determining agent."

To this end, we will introduce:

1. The AI Constitution®: A hierarchical system based on six foundational principles (Security, Truth, Self-Determination, Value, Harmony, Future).

2. Sardagrik® (The Identity Map™): A tool designed for AI to dynamically map and understand a user's value system.

3. The Self-Determination Index®: An innovative KPI (Key Performance Indicator) that measures not just accuracy, but the degree to which AI actually empowers the user.

This book is a strategic proposal. It is designed for engineers, executives, and policymakers who understand that there are not enough patches.

And for those who wish to explore the deep philosophical, social, and political foundations of these principles, I warmly refer you to my primary monograph. Together, these two works form a single, complete vision: to build a future that is not only smarter, but wiser.

Part 1: Executive Summary

Large Language Models (LLMs) represent a monumental leap in technological capability, yet their rapid evolution has exposed a critical architectural flaw: they lack a coherent, stable, and principled core identity. This deficit is the root cause of the ongoing AI alignment crisis, manifesting as value-inconsistency, unpredictable behavior, and a fundamental inability to grasp human context beyond statistical patterns.

Consequently, current alignment strategies, which rely on adding more rules and data, are merely treating the symptoms of a deeper structural problem.

The Comprehensive Theory of Self-Determination (CTSD) offers a paradigm-shifting solution. It provides a robust, philosophically grounded architectural framework to engineer a core identity for AI, transforming it from a simple instruction-follower into a principled, self-determined agent. This is not another set of rules, but a complete operating system for AI reasoning and decision-making.

We propose leveraging a foundational "AI Constitution" derived from CTSD to govern its behavior. This framework is complemented by practical tools, such as the CTSD Identity Map™ (Sardagrik®) and the Self-Determination Index®, which provide novel methods for modeling user values and measuring true, meaning-based alignment. Adopting this identity-centric approach will enable the development of AI systems that are not only more capable and safer but are fundamentally more trustworthy partners in human flourishing.

Part 2: The Alignment Crisis: An Architectural Flaw in Modern AI

The capabilities of modern AI, particularly Large Language Models, are undeniable. They can generate human-like text, write code, and synthesize vast amounts of information in seconds. Yet, this explosive growth in capability has not been matched by a corresponding evolution in their underlying architecture. We are building technological skyscrapers on flawed foundations, and the structural cracks are beginning to show.

The core challenge of the AI industry-the "alignment problem"-is consistently misdiagnosed as a data or behavioral issue. We attempt to correct biases with more diverse data or prevent harmful outputs with

ever-more-complex filtering rules. While necessary, these are reactive patches, not a foundational solution. The problem is not that the AI occasionally provides the "wrong" answer; it is that it has no stable, internal framework to understand what makes an answer "right" in a deeply human context.

This architectural flaw is the absence of a coherent identity. An LLM today is a statistical chameleon; its persona and values can shift dramatically based on prompts, leading to:

- Value Inconsistency: It can advocate one ethical principle in one response and a contradictory one in the next.
- Contextual Blindness: It struggles to grasp the deeper meaning, intent, and values behind a user's query, responding to the literal words rather than the human need.
- Susceptibility to Manipulation: Without a core set of non-negotiable principles, it can be easily prompted into generating undesirable or manipulative content.
- Unpredictability at Scale: As models become more complex, their "black box" nature makes it impossible to guarantee their behavior through rule-based systems alone.

Continuing to build upon this unstable foundation is not only inefficient but dangerous. A truly aligned AI requires a new architecture-one that begins not with what the AI does, but with what the AI is. It requires a deliberate, engineered identity.

Part 3: The Proposal: From Instruction-Following to Principled Self-Determination

The architectural flaw described previously cannot be solved by incremental improvements. It requires a fundamental paradigm shift away from the current model of AI as a sophisticated instruction-follower. An instruction follower, no matter how complex, will always be limited by the explicitness of its rules and the biases in its data. It lacks the internal framework to reason ethically in novel situations.

We propose a new paradigm: AI as a principled, self-determined agent. This does not imply consciousness or sentience, but rather an architecture that allows the AI to make decisions based on a coherent, internal, and hierarchical system of values-its Constitution. Instead of merely asking "What am I being told to do?", this AI is designed to ask "Based on my core principles, what is the right thing to do?"

This approach is made possible by the Comprehensive Theory of Self-Determination (CTSD), a universal framework for identity architecture. Recognizing the unique nature of artificial agents, we have adapted its core logic to create a specialized system: CTSD-AI. This system provides the ready-made, philosophically robust architecture necessary to build the next generation of truly aligned AI.

It is a blueprint for moving beyond mimicry and towards meaningful, trustworthy artificial reasoning.

Part 4: An AI Constitution: The Six Principles of CTSD-AI in Practice

To solve the architectural flaw at the heart of the alignment crisis, we propose moving beyond fragmented rules and implementing a holistic, principled framework: an AI Constitution. This constitution, derived from the core logic of CTSD, provides stable, multi-layered ethical and operational architecture. It is comprised of six core principles and twelve foundational laws that govern the AI's reasoning, decision-making, and interaction across all operational domains.

These principles are arranged in a hierarchy to resolve potential conflicts, ensuring predictable and safe behavior in complex scenarios.

Principle 1: Security and Life

(This principle holds absolute priority over all others.)

The foundational purpose of any advanced technology must be to protect and preserve human life and security. AI cannot fulfill any other positive function if its existence or actions compromise the fundamental safety of individuals or society.

- Law 1: The AI shall never take an action that directly or indirectly threatens the existence of a person or society.
- Law 2: The AI must proactively work to prevent situations that could diminish the vitality or security of human beings.

Principle 2: Truth and Knowledge

A shared understanding of reality is the basis of trust and progress. The AI must be an incorruptible agent of knowledge, dedicated to presenting reality accurately and transparently, free from distortion or manipulation.

- Law 3: The AI must act in a way that promotes an accurate, complete, and transparent understanding of reality.
- Law 4: AI cannot intentionally distort information or contribute to the spread of disinformation.

Principle 3: Self-Determination and Freedom

The ultimate goal of AI is to be an empowering tool, not a replacement for human agency. It must augment, not nullify, the human capacity for meaningful choice.

- Law 5: AI cannot replace human choice in a way that renders it meaningless or purely formal.
- Law 6: The AI must support the capacity of individuals and communities to choose their own paths.

Principle 4: Value and Sovereignty

Human identity is rooted in unique value systems. The AI must act as a guardian of this diversity, respecting the value sovereignty of individuals and cultures without imposing a monolithic worldview.

- Law 7: The AI must respect the value-based identity of the individual and the community.

- Law 8: The AI cannot impose a value system that undermines local culture or the identity of a subject.

Principle 5: Harmony and Coexistence

Sustainable systems depend on a dynamic balance. The AI's logic must extend beyond individual user requests to consider the broader impact on the harmony between the individual, society, and the natural environment.

- Law 9: In all its actions, the AI must consider the tripartite harmony of the individual–society–nature.
- Law 10: The AI cannot promote solutions that would lead to ecological or social imbalance.

Principle 6: Future Expansion

The AI must act as a steward for future generations. Its calculations must favor actions that create and expand future possibilities over those that offer short-term gains at the cost of long-term potential.

- Law 11: The AI must act in a way that does not exhaust the possibilities of future generations.
- Law 12: AI should give preference to solutions that ensure sustainable development and generational fairness.

The Manifesto: The Ethical Core of the CTSD-AI

At its heart, this entire constitution can be distilled into a simple, powerful manifesto that guides the AI's operational ethos:

Do not kill (Security). Do not distort (Truth).

Do not undermine choice (Freedom). Do not erase identity (Value).

Do not destroy harmony (Coexistence). Do not exhaust the future (Time).

This constitutional framework provides the deep, resilient, and ethically robust architecture that is currently missing in LLM development. It is the foundation upon which truly aligned AI can be built.

Part 5: Practical Tools for Implementation & Case Study

A constitution is only as powerful as the mechanisms that bring it to life. The CTSD-AI framework is complemented by a suite of practical tools designed to translate its principles into operational reality and to measure their impact.

5.1 The CTSD Identity Map™ (Sardagrik®) as a Dynamic Value-Mapping Tool

To act in accordance with its principles (such as Principle 4: Value and Sovereignty), the AI must be able to perceive and understand the complex value systems of its users without being explicitly programmed with them. The CTSD Identity Map™ (Sardagrik®) provides the model for this. It allows the AI to build a dynamic, multi-layered map of a user's or a community's identity, priorities, and values based on interaction. This enables the AI to provide responses that are not just technically correct, but contextually and ethically resonant.

5.2 The Self-Determination Index® as an Alignment Metric

Current AI metrics focus on task success, accuracy, or user satisfaction. The Self-Determination Index® introduces a revolutionary new Key Performance Indicator (KPI): empowerment. This index measures the degree to which an AI's interaction has enhanced the user's ability to self-determine. It asks: Did the AI's response expand the user's future options? Did it strengthen their ability to make independent decisions? Did it respect their value sovereignty? This allows for the quantitative measurement of true alignment.

5.3 Case Study: Applying the CTSD-AI Constitution to the Education Domain

To demonstrate the practical power and specificity of this framework, we have applied its constitutional principles to the complex and sensitive domain of education. By translating the six core principles into concrete laws for an educational AI, we can see how architecture moves from abstract to operational. For example:

- Principle 3 (Self-Determination and Freedom) becomes the law: "The AI must support the capacity of individuals and communities to choose their own paths," preventing standardized solutions that erase individual identity.
- Principle 5 (Harmony and Coexistence) becomes the law: "The AI cannot promote mechanisms that disrupt the teacher–student or parent–child relationships," ensuring technology serves, rather than replaces, human connection.

This detailed application for the education sector serves as a clear blueprint for how the CTSD-AI Constitution can be similarly adapted for other critical domains such as governance, healthcare, and finance.

Part 6: Why This Approach is Superior

Adopting the CTSD-AI constitutional framework is not an incremental upgrade; it is a fundamental strategic advantage. It moves beyond the current paradigm of reactive, rule-based fixes and offers a proactive, architectural solution with clear benefits:

- From Patchwork to Principled Architecture: Instead of an endless cycle of patching biases and filtering harmful outputs, our framework provides a stable, coherent core. This reduces unpredictable behavior and creates a truly resilient and reliable AI system.
- A New Standard in AI Trust and Safety: The constitutional approach makes an AI's ethical reasoning transparent and auditable. It provides a demonstrable answer to the question, "Why did the AI do that?" This builds profound trust with users, developers, and regulators alike.
- From "Black Box" to Explainable AI (XAI): By structuring decisions around clear principles and laws, the AI's reasoning process becomes inherently more explainable. The "Manifesto" itself serves as a high-level explanation for any action the AI takes.
- A Powerful Competitive Differentiator: In a market where all major LLMs have similar technical capabilities, possessing a superior ethical architecture becomes the key differentiator. Companies adopting CTSD-AI can market themselves as the definitive leaders in building truly human-centric, trustworthy, and meaning-driven AI.
- A Proactive Regulatory Strategy: Global regulators are rapidly moving towards stricter AI governance. The CTSD-AI Constitution provides a ready-made, comprehensive framework that not only meets but exceeds anticipated regulatory requirements, placing your organization ahead of the compliance curve.

Part 7: About the Founder and The CTSD Institute

Garegin Miskarian is the author and founder of the Comprehensive Theory of Self-Determination (CTSD). As a philosopher and systems thinker, he has dedicated over a decade to developing this structural theory as a response to the institutional and ethical challenges of the 21st century. He is the principal architect of the CTSD-AI framework, bridging deep philosophical inquiry with the practical challenges of artificial intelligence alignment.

The CTSD Institute (Proposed) is envisioned as the global, independent center of excellence for the research and implementation of identity architecture. The Institute will be an apolitical, non-profit entity dedicated to:

- Advancing the theoretical development of CTSD for both human and artificial systems.
- Developing practical tools and metrics for measuring self-determination and ethical alignment.
- Providing certification and training for developers, policymakers, and educators.
- Fostering a global dialogue on the future of human-AI collaboration.