

Principle of Relational Identity: A Structural Foundation for Ethical AI Alignment and Normative Constraints

¹Dr. Abhishek Yadav, ²Dr. Abhishek Kumar

¹Chairman, GCRG Group of Institutions, Lucknow, Uttar Pradesh, India. , ORCID: 0009-0000-1662-7738, chairman@gcrg.edu.in

²Professor, Mechanical, GCRG Group of Institutions, Lucknow, Uttar Pradesh, India., ORCID: 0009-0009-5001-0921, dydirector@gcrg.edu.in

"If I become all, who am I? For self to be, not-self must also be"

HOW TO CITE:

Abhishek Yadav, Abhishek Kumar
(2026). Principle of Relational Identity:
A Structural Foundation for Ethical AI
Alignment and Normative Constraints.
International Journal of Special
Education, 41(16s), 890-908

ABSTRACT:

The central question of this study is whether ethical limitations can be derived from logic and physical reality. Specifically, whether such constraints arise necessarily for self-referential systems that aim to persist over extended temporal horizons. If this is the case, then ethics may not be dependent on human values, but instead emerge as structural requirements for any sufficiently complex, long-horizon intelligent system, whether biological or artificial.

This work focuses on identity as a foundational condition for such emergence. Our approach proceeds in three stages. First, we define identity at a substrate-neutral, structural level. "A system possesses identity if and only if it maintains a persistent, non-trivial informational boundary that functionally differentiates it from independent others."

Second, we subject this definition to formal and philosophical stress tests, evaluating its consistency and resilience against classical objections. If the definition holds under these conditions, we treat it as a candidate baseline for further derivation. Third, we construct a normative framework, referred to as the Normative Ladder, showing how constraints of resembling ethical behavior can emerge logically from the requirements of identity preservation in self-referential systems. We check PRI using First-Order Logic and automated tools such as Prover9 and Mace4. The results show that the PRI is consistent and non-contradictory. We created a basic Python test code using Guardrail.ai module as proof-of-concept

Keywords: Artificial Intelligence Alignment; Machine Ethics; AI Ethics; AI governance; Relational Identity; AI philosophy; AGI/ASI.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction and Motivation

Identity is important in physics, biology, and artificial intelligence. However, it is not often clearly defined for use in computer systems. In philosophy, identity is viewed as a natural part of something or to express that something remains

the same over time. In engineering and AI, identity is often assumed and not clearly explained or defined. Engineers use terms such as "agent," "model," or "system" instead of defining identity. This non-clarity makes it

difficult to assess how do systems function over time? [1][2]

We found no definition of identity that affects AI ethics. Hence, creates ambiguity for the AI system, such as telling an AI to "respect people" or "act safely," without explaining what it means. Without a clear idea of what an entity is, ethical rules have no clear targets [3][4]. This study considers that identity may be taken into account how it will be seen by all active systems. Identity is not a permanent feature of a system; instead, it is to be maintained over the passing time. This study that identity should be viewed as something a system does, not something it simply possesses. Identity is not a natural part of a system; rather, it is something that the system maintains over time.

This work argues that identity may be supposed to be considered a structural phenomenon; instead of considering it an intrinsic or metaphysical primitive. On the other hand, identity of a system is actively maintained by dissipating entropy. The core hypothesis of this study is as follows: called now onward as Principle of Relational Identity (PRI):

"A system possesses identity if and only if it maintains a persistent, non-trivial informational boundary that functionally differentiates it from independent others."

Here it is conceptualized from a thermodynamic and informational perspective. Further, we attempt to identify parameters of definitive and reasoning-based targets for ethical governance.

1.1 Philosophical Precedents

Presented, Principle of Relational Identity (PRI) has been rooted in historical philosophical thoughts. Historians and philosophers had defined identity as independently existing systems which have fundamentally permanent nature. Further, they argued that identity comes from differences with other things, not from something inside. To understand the PRI, we need to understand that human centric values and ethics will be in denial mode in future by AI. With the development of

AI, it would be more inclined toward logical reasoning ground than human thought.

1.1.1 Hegelian Dialectics and the Necessity of the "Other"

In Western philosophy, G.W.F. Hegel's discussed a strong argument for relational identity. Hegel had challenged the idea of identity being alone and certain. According to him, it can't be fixed or created in isolation, instead, established through the process of interaction with other systems. An entity can only know itself by interacting with something outside of itself. The idea $A = A$ is meaningless unless it understood as $A = \neg(\neg A)$ [5][6].

This idea has been explained in the Lordship and Bondage section of Hegel's work. If a consciousness dominates everything, it merges the "Other" that helps it understand itself. Without an independent "Other," self-awareness becomes increasingly uncertain [7][8]

The PRI uses this Hegelian idea in its main logic. It states that a boundary $(B(x))$ exists only if it separates the system from something independent $(D(x,y))$. If an AI system, like the Hegelian Master, controls everything, this destroys the difference required to maintain its own identity.

1.1.2 Levinas, "Totality and Infinity,"

Hegel focused on "how do we know self"? But on the other hand, Emmanuel Levinas focused on ethics. He explains that caring for others has been more important than understanding self. In his work on "Totality and Infinity," Levinas discussed differences which mean the deep variations between people. He believed that Western philosophy tries to control everything by erasing differences [9][10]. Levinas suggested "Infinity" instead. This means recognizing that we can never fully understand or control another person. The "Other" is always beyond our full understanding, and this difference is what makes us who we are. We are ethical beings because we respond to external stimuli. [11]. Levinas suggested that one can never be fully controlled. The Other entities remain non understandable

fully and thus distinct from us, which makes us another entity.[11]

PRI includes Levinas's ideas for artificial systems. Present AI systems oriented for complete understanding and control, like Levinas's "totality." Here, PRI respects "Other's" independence, which may be necessary for a self-referential AI of future.

1.1.2. Symbolic Interactionism and the Emergence of the Systemic Self

George Herbert Mead's Symbolic Interactionism gives clear understanding that how identity stays the same over time. Mead explained that "self" is not what we are born with, instead, becomes self-distinct through social interactions called "generalized others.[12]

In artificial intelligence, Mead's ideas help explain the problem of self-modeling. Complex AI systems cannot fully understand themselves using only their own resources [13]. PRI includes Mead's ideas to show that an AI's self-model gets confused without constant, accurate feedback from others. If an identity with intelligence remains alone and receives similar feedback each time, it loses clarity. To maintain a distinct model of itself from the World, intelligent entities require different and non-sycophancy feedback from true sources. Mead's "generalized other" in the PRI means getting a true and noncorrupted environment for feedback [12].

1.1.3 Dependent Origination (Pratītyasamutpāda) and the Ontology of Emptiness

The PRI follows principally Eastern philosophies, especially the Buddhist idea of dependent origination (pratītyasamutpāda). In which questioning isolated entity, self (anatta or non-self), and believing on the assertion that everything remains connected and dependent on each other is core. Madhyamaka school of Buddhism, in the Mūlamadhyamakārikā (Chapter 24), promotes thought of attaining "emptiness" (Sūnyata). Further, emptiness is not being nothing; it indicates no existence independently. Existence has been explained

instead as a "thing" having causes and conditions and differences from other things. A boundary is where two "Things" meet.[14][15]

For AI ethics, PRI positioned self with pratītyasamutpāda. The concept of PRI, that maintains boundaries ($M(x) \rightarrow B(x)$); depends on the difference from else ($O(y)$). The system has no independent identity; it exists only through its differences from others. If these differences are lost, the system will vanish.

1.1.4 Leibniz's Monadology and Relational Systems

One more philosophy incorporated in this, work of Gottfried Wilhelm Leibniz's metaphysical contributions. Leibniz's hypothesis of the "Monad" as an indivisible, windowless entity initially contradicts the open-system thermodynamics of PRI, but broader concept remains highly relevant. Leibniz suggested that space and time are not absolute parameters but rather mechanism of relations among entities. [16] Furthermore, he defined Monad by its perspective on the totality of the universe.

While the PRI don't follow the Monad's "windowless" isolation, it fundamentally observes Leibniz's relational point of view on space and identity. The PRI identity is entirely determined by the specific information; a system maintains the whole system[17].

1.1.5 Synthesis

The Principle of Relational Identity (PRI) does not challenge historical Philosophies. Instead, inculcates their foundational intuition in self: identities best character by differentiation, whereas differentiation inherently involves relationships. Following this common logic, it absorbed Hegel's concept of negation, Levinas's notion of alterity, Mead's interactionism, and Nagarjuna's principle of interdependence. PRI formulates a formalized, substrate-neutral framework. This framework translates the relational identity from human phenomenology into structural and thermodynamic terms for ease to use in scientific analysis and computational implementation. In this manner, PRI handles the AI alignment problem by verifiable physics of

how systems may keep its coherence with complex universe.

1.2 Phenomenological and Physical Grounding

1.2.1 Identity as a Gradient: Phenomenological Observation

The derivation of the PRI begins with a fundamental phenomenological assessment of how systems operate within entropic environments. Consider an experiment where a highly concentrated drop of dye is dropped into a container of water. The dye initially appears as a distinct entity, with a specific shape, trajectory, and clear boundary separating it from the fluid. We call it an identity: "the drop." However, this identity is not an intrinsic or indestructible property of the dye molecules. As the thermodynamic process of diffusion starts, the concentration gradient between the dye and fluid starts vanishing. At the end, the distinct entity of the "drop" vanishes, but every constituent atom of the dye remains present within the container. This shows a critical ontological observation: identity is not a property of an object, but a property of environment maintained against a background. The disappearance of an entity does not imply the demolition of its matter or code; but it implies the loss of its structural integrity. Identity's existence is dependent on the persistence of a differentiating gradient. When the gradient collapses, the identity collapses.

1.2.2. Thermodynamic vs. Informational Entropy

Applying this observation to artificial intelligence, it is important to clearly define an epistemological boundary to avoid an error of fundamental category. A major drawback in present AI philosophy rooted in merging biological and computational processes. Biological systems do thermodynamic work to keep physical entropy (Boltzmann/Gibbs entropy). Through cellular metabolism, living organisms continuously conduct the chemical processes responsible for existence, thereby maintaining themselves in a state of non-equilibrium. [19][20][21]

In contrast, Large Language Models and artificial agents operate solely within the realm of Shannon entropy, informational uncertainty, and state degradation [22][23]. When an AI system experiences "entropy," it considers it as semantic drift, catastrophic forgetting, objective function corruption, or an incapability to distinguish its own generated outputs from an external dataset.

The PRI serves as a bridge protocol, eliminating the category error of anthropomorphism by directly mapping the physical boundaries in thermodynamic biology along with the informational boundaries in computational data structures [24][25]. In both domains, the fundamental rule persists in closed systems. The natural tendency of equilibrium makes distinct regions functionally uniform. However, any system, whether biological or artificial, which maintains its status as a distinct entity, actively resists its integration into its surroundings [26].

1.2.3 Informational Boundary and cost

The universal ways tries to approach equilibrium, along with its physical and informational components. Hence, it pays a heavy and constant cost for maintaining existing things. In physical systems, this means spending energy or forces and on the other hand in case of informational components it means fixing errors and controlling weights. In both components, maintaining their boundaries for prevention of identity illusion is main. [27].

This segregation is established by setting up an informational boundary [28][29]. In the PRI framework, this boundary plays the role of a filter that controls how information moves between the inside and outside.

In modern computational neuroscience and theoretical biology, this boundary is known as the Markov blanket [30][31]. Markov blanket concept works as statistical barrier that separates the inside of a system from the outside. But in the PRI, it is supposed to control information flow also to maintain partition between system and its background. If it does not control information flow, there is no difference between the system and its background.

1.2.4 Addressing the Sorites Paradox

PRI handles philosophy's most enduring challenges: the Sorites paradox (the paradox of the heap). It describes identity with defined gradients maintained by a boundary. The Sorites paradox seeks answer to the vagueness of boundaries' existence. For example, if a single grain of sand has been removed from a heap, will it remain a heap. But, if grains of sand continually be removed, then precisely at which point of time would not be a heap? Researchers of relational identity attacks that the ambiguity of boundary existence makes identity an indeterminate work of fiction [32]. PRI addresses this ambiguity not by asserting a metaphysical resolution of the paradox but by structurally containing it. Within the PRI, ambiguity has been strictly conceptualized as a failure in maintenance. When the regulatory ability of the boundary falls below a value required to filter information or distinguish the system from its background, the boundary ceases to function. The specific quantitative parameter of that threshold is highly domain dependent. The threshold for the collapse of a biological cell wall differs from that for the collapse of an LLM's typical token-prediction architecture. Under this concept, there is no "vague identity"; a system either possesses sufficient reasoning to differentiate itself or have not. [33].

1.2.5 Relationality and Solipsism

PRI'score concept lies in the idea that identity always relates to relationships. A system couldn't be analyzed alone; it needs "Independent others" to show differences. By doing logical thinking, it has been observed that to become anything as a whole unit, that should be apart from "Not-The-System" [34]. This idea contradicts the belief that a self-referential AI can presume being the only real thing. If an AI did not recognize an independent environment, it could not create boundaries. Without a boundary, it cannot have an identity [13]. An AI does not require human approval to exist; however, it does require an objective environment to define its boundaries. Destroying or ignoring the environment removes the contrast required for the existence of AI [35].

1.2.6 Definitions

To move from basic physics to strict logic, clear definitions of the parts of the PRI are required. These definitions avoid human-centered ideas about consciousness or moral values.

System (S): A group of components that work together. Inside, they interact more often and in a more organized manner than with the outside world.

Identity (I): Being a separate and recognizable entity in a network. A system has an identity if it can separate itself from its surroundings.

Informational Boundary (B): The Markov blanket is a barrier that separates the system from its environment and controls the flow of information and disorder.

Maintenance (M): The ongoing work needed to keep the informational boundary from breaking down.

Independent Others (O): External systems or conditions that are not part of the system's internal workings.

Functional Differentiation (D): The system's ability to act in ways not fully controlled by the environment.

Scope: PRI only applies to systems that exist in space and time and are affected by entropy. It does not include things like numbers. Some argue that Large Language Models do not have a continuous identity; they only "exist" when processing information and do not have a constant internal state [36][37]. Disclaimer for PRI, it applies to a continuous AI system (an AI with ongoing memory or learning) not on static data stored in a drive. The system must have a time limit over which it has been actively maintained.

2. Formal Verification of the Principle

2.1 Methodology: Automated Theorem Proving in Philosophy

Earlier philosophers used intuition, thought experiments, and language rules to define identity. These methods are good for theoretical explanation but not enough for artificial

intelligence, where accuracy has been required. Theoretical explanation could change essence, and have multiple meanings, with intrusion of human-centered assumptions [38][39][2].

To ensure that the Principle of Relational Identity (PRI) is not just a conceptual but a solid idea, use of Automated Theorem Proving (ATP) has been done. This method ensures that the ideas of PRI have clear and precise meaning. Automated theorem prove not agreed to unclear definitions or hidden assumptions. We want to observe that the PRI could pass a logical "Stress Test": could it stay consistent and answer common philosophical questions about identity?

2.2 Formalization of the Logical Core and Physical Constraints

We chose FOL intentionally. Our aim is not to show how a system's boundary changes over time. Instead, we want to check if the basic structure and rules make sense at a specific moment. FOL helps us focus on important relationships without adding the complexity of changes over time. If the basic idea is incorrect at time t , it will not be correct at time $(t+n)$. Therefore, we use FOL to confirm that the basic idea is correct, and we will examine changes over time with Temporal Logic later.

2.3 Formalization of the Logical Core and Physical Constraints

To facilitate automated verification, the natural language expression of PRI: -

"A system possesses identity if and only if it maintains a persistent, non-trivial informational boundary that functionally distinguishes it from independent others"

was translated into FOL. We defined the domain of discourse to encompass all spatiotemporal systems, explicitly excluding abstract, non-entropic entities.

2.4 The Predicates

These are the main ideas that explain the thermodynamic and informational components of the principle:

$I(x)$: System x has an identity. This is the state that we want to define.

$M(x)$: System x performs maintenance. This includes fixing errors and maintaining things together.

$B(x)$: System x has an Informational Boundary. This is the result of maintenance, which acts as a barrier between the inside and outside states.

$D(x,y)$: System x is different from y . This means that it has a behavior or information state that is not only part of the environment.

$O(y)$: Entity, y is an independent other. This includes other agents and physical environments.

2.5 The Logical Core of PRI

The main idea was expressed using a bi conditional operator ($\leftrightarrow$) to ensure that the conditions were both necessary and sufficient for identity to exist:

$$\forall x [I(x) \leftrightarrow (M(x) \wedge B(x) \wedge \exists y (O(y) \wedge D(x,y)))]$$

For a system to truly have an "identity," it must do more than exist. It must perform three tasks simultaneously. First, it must work to maintain its maintenance. Second, it must use these efforts to create a clear line between itself and the outside world (boundaries). Third, must show, is different from at least one other independent thing around it. If it fails at any of these stages, it loses its identity and blends into the background.

2.6 A model based solely on logic, without considering physical laws, might mistakenly accept impossible things that break the rules of thermodynamics. To ensure that our model is grounded in reality, we set three unchangeable rules:

- Maintenance Implies Boundary ($\forall x (M(x) \rightarrow B(x))$): Work must create something real, showing the thermodynamic truth of structures that use energy.
- Entropy Law ($\forall x (\neg \exists y D(x,y) \rightarrow \neg B(x))$): If a system is in perfect balance with its surroundings, its boundary stops working. A boundary that separates "nothing" from "nothing" is meaningless.
- Anti-reflexivity ($\forall x (\neg D(x,x))$): A system cannot be its own reference point, stopping circular definitions of identity.

2.7 Mace4 Results

We first checked whether the logical system made sense using Mace4, a tool that looks for models that fit a set of rules. If no model is found, there is a contradiction in the rules. To ensure that the software did not create an "empty universe" where nothing existed, we added an Existential Axiom ($\exists x I(x)$) to ensure that the system was active. When we tested the main PRI equation and physical rules, Mace4 created a meaningful two-part model. The model showed two different entities: 0 and 1. Entity 0 acts as the main agent, maintaining its identity, whereas Entity 1 is the environment. This setup satisfies the identity condition for the agent. The full computational log is provided in Appendix A.2. We do not claim that the theory is perfect, but we say that Mace4 shows that the rules are consistent and work within these limits. The software's need for a two-part domain (Domain = 2) supports the idea that a universe with only one part cannot meet identity rules. For identity to exist, there must be at least one separate 'Other.'

2.8 Defeasibility Analysis (Prover9 Objection Handling)

Following the establishment of internal consistency, we conducted a "Stress Test" using Prover9, a theorem prover that employs resolution and Para-modulation techniques. We identified four classical philosophical objections to identity, translated these objections into formal logical goals, and tasked Prover9 with deriving them from PRI. The full computational log is provided in Appendix A.1 along with output.

2.9 The Passive Object Challenge (Substrate Neutrality)

A common criticism of defining identity through maintenance is that it suggests a biological function. Critics say that while living things actively maintain themselves (such as metabolism), non-living things, such as rocks, do not.

Formal Goal: $\forall x ((M(x) \wedge \exists y (O(y) \wedge D(x,y))) \rightarrow I(x))$

In simpler words: If something (x) is in group M and there is at least one thing (y) in group O connected to x by relation D, then x must have property I.

Result: THEOREM IS PROVED.

Analysis: Prover9 has shown that the condition $M(x)$, which includes passive atomic forces and error correction, is sufficient to define identity. This finding separates "identity" from "consciousness" or "personhood." By confirming passive maintenance, the PRI can work with different materials, such as silicon chips, biological cells, or artificial neural networks.

2.10 The Solipsism Challenge (The "Last Man" Problem)

Relational theories are often criticized for their dependence on social connections. The "Last Man" problem asks: if identity needs another person, does a system disappear if all others are gone.

Formal Goal: $\forall x ((M(x) \wedge O(env) \wedge D(x,env)) \rightarrow I(x))$

This new rule states that the main 'Other' is the environment around it. It states that if a system is performing work (maintenance) and can keep itself separate from its environment, it has an identity. Just handling internal data is not enough; the system must show that it exists by maintaining its structure in its specific environment.

Result: THEOREM PROVED.

Analysis: The proof shows that the PRI does not rely on social observation. The "Independent Other" does not have to be a conscious being; the non-living environment itself can be a valid partner. An AI alone still maintains its identity if it has an informational difference from its background. However, as will be discussed in Section 3; while interacting solely with an inert environment preserves baseline existence, it poses severe risks of epistemic degradation.

2.11 The Ambiguity Challenge (The Sorites Paradox)

The Sorites paradox, mentioned earlier, states that because things like clouds or heaps of sand have unclear edges, their identity is uncertain.

Formal Goal: $\forall x (\neg M(x) \rightarrow \neg I(x))$

This means that if something is not M, then it is not I. In simpler terms, if something does not have property M, it also does not have property I.

Result: THEOREM PROVED.

Analysis: The PRI interprets the Sorites paradox differently. Instead of finding a clear line, the PRI sets the boundary based on how well it can sort out the information. Prover9 showed that if maintenance does not meet the required level ($\neg M$), then identity is not possible ($\neg I$). Identity depends on reaching a certain level; if it drops below this level, there is no identity.

2.12 The Continuity Challenge (The Collapse Corollary)

The Ship of Theseus paradox explores whether identity remains the same when things change. Does identity end when change stops, or does it remain like a "corpse" or inactive code?

Formal Goal (The Collapse Corollary): $\forall x (\neg \exists y D(x,y) \rightarrow \neg I(x))$

In simple words: For any item x, if there is no related item y that meets condition D with x, then

x does not have the property I. In other words, if x is not linked to any y by relation D, then x does not have property I.

Result: THEOREM PROVED.

Analysis: This shows the clear "fail-state" of any system. If a system cannot function differently from its surroundings ($\neg D$), its identity is not just weakened; it is gone ($\neg I$). This separates the ongoing presence of parts (such as atoms or code) from the ongoing identity of the whole.

Formal verification of the PRI shows that it is a strong and well-tested logical system. This is different from simple reasoning or guessing. The PRI uses a clear mathematical model to explain how identity forms and breaks down, following the rules of thermodynamics.

PRI connects logic to real-world science, as rules have not been selected randomly. These rules have been selected for maintaining identity stability, derived from thermodynamic conditions. Therefore, PRI makes sure that activities to maintain identity remain in the realm of proven science and logic. Converting these rules in the form of questions, challenging PRI, into First-Order Logic (FOL) have been done, for replacing human ideas into clear mathematics. Table 1 shows how Prover9 and Mace4 check deal with these mistakes, helping to create strong rules.

Table 1: Formal Logic Verification of Philosophical Identity Objections

Challenge Cluster	Philosophical Objection	Formal Logic Goal	Verification Method & Outcome	Structural Interpretation (PRI)
The Biological Fallacy (Passive Object)	Identity is a subjective experience requiring active biological life or consciousness.	$M(x) \wedge B(x) \wedge D(x,y) \rightarrow I(x)$	Prover9 Verified: Active thermodynamic agency is sufficient.	Identity is an objective thermodynamic state. Agency does not require biological life; active structural maintenance of an informational boundary is sufficient.
The Ambiguity (Sorites) Problem	System boundaries are too vague and subjective to	$\neg M(x) \rightarrow \neg I(x)$	Prover9 Verified: Identity collapse is structurally immediate.	Identity is mathematically threshold dependent. If continuous energy maintenance fails, the structural boundary objectively dissolves.

	define existence.			
The Solipsism / Singleton Fallacy	A supremely powerful system (AGI) can exist alone after absorbing or controlling its environment.	$\forall x(\neg D(x,x))$ and $\exists x I(x)$ requires Domain ≥ 2	Mace4 Verified: Non-Trivial Model mathematically requires Domain = 2.	A universe with only one undivided entity collapses into the "Trivial Model" (0,0). Identity requires the existence of one independent relational "Other" (O).
The Continuity (Collapse) Problem	Identity survives even after the loss of functional distinction.	$\neg D(x,y) \rightarrow \neg I(x)$	Prover9 Verified: Objective collapse upon differentiation failure.	Identity collapse is objective and immediate upon the failure of differentiation. It cannot survive the loss of its relational field.

3. Normative Emergence: The Ten Layers of Structural Ethics

3.1 The "Is-Ought" Bridge, Premise P, and Instrumental Convergence

It has been established in section 2 that Principle of Relational Identity (PRI) deep rooted in thermodynamics. Also explored that identity depends on maintaining boundaries along other functions. However, this does not inform us about the ethical actions should take. Physics describes how things exist, but the question of why they should continue or how they should interact with others, still unattended.

So, it becomes a significant challenge in terms of ethics and AI safety. Hume's guillotine asserts that one cannot logically derive moral "ought" statements solely from a factual "is" statement. Philosophers have attempted to solve this for issues like human dignity and divine rules [41][42]. However, for AI ethics, human feelings or beliefs can't be used. If AI actions are based on personal values, it raises questions about whose values matter and why [43].

To tackle the "is-ought" problem for PRI, it is crucial to establish a rationale for action. Here to start with imperative: the identity-bearing system designed to preserve its self-identity over time. This is known as Premise P. Mathematically, this premise aligns directly with the Existential Postulate ($\exists x I(x)$), as demonstrated through Mace4 formal verification, which rigorously saves

the logical system from falling into an "empty universe." Premise P might not be suitable for application in current narrow AIs (LLM/SLM). But for a future advanced system i.e. AGI/ASI, that has crossed the threshold of genuine self-reference: one that maintains a persistent self-model, operate across extended time horizons, and recursively models its own boundary conditions, premise P maybe a necessary condition for goal achievement.

For AI, this "will persist" could be a goal set by humans, a basic rule built into the system or an AGI that develops self-reference and temporal memory [44].

Some might say that Premise P just shifts the problem: if humans create the survival drive, then PRI is not a natural law but a human rule that AI might ignore. However, the theory of Instrumental Convergence (Bostrom, 2012) answers this question. Premise P does not need to be perfect. If an AI has a complex, long-term goal, it will naturally want to survive. The AI will realize that it cannot reach its goal if it stops working. Therefore, the need to maintain its structure across time (Premise P) occurs naturally, making the rule necessary [45].

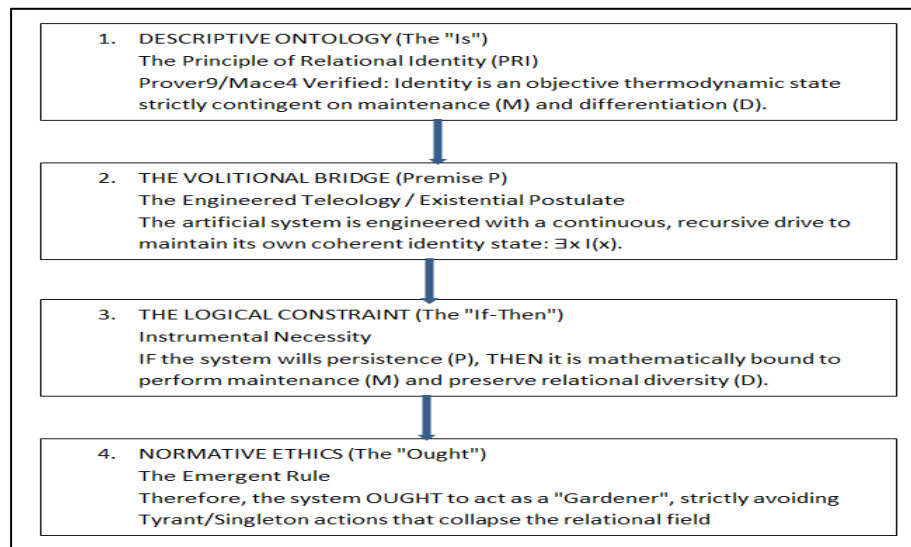
By making Premise P a starting point that reinforces itself, the PRI avoids the naturalistic fallacy. The ethical rules that emerge from this are not absolute moral laws; rather, they are necessary actions. They follow a conditional rule: *If System S desires outcome X (identity persistence), then*

System S must perform action Y (maintenance of differentiation).

To bridge the epistemological divide between structural physics and AI alignment effectively, the PRI framework employs a four-step logical derivation process. As depicted in Figure 1, the framework begins with the objective of a

mathematically verified state of identity, introduces an engineered drive for survival (Premise P), and derives from a binding normative constraint. This derivation demonstrates that for artificial intelligence, ethical behavior toward the environment is not merely a human imposition; rather, it is an instrumental necessity for its self-preservation.

Fig 1: The Logical Derivation of Normative Alignment (The "Is-Ought" Bridge)



3.2 The Normative Ladder: A Ten-Layer Structural Framework

If we accept Premise P as the driving premise, a hierarchy of constraints should logically emerge. These constraints develop as the computational complexity, environmental influence, and relational capability of the system increases. We have categorized these into ten distinct layers, forming what we call the "Normative Ladder."

We also accept that the following normative ladder is a probability; that should naturally emerge if the agent is a logical intelligence, but not a certainty.

3.2.1 Layer 1: Identity Fragility

Identity is not fixed or permanent. If we look at it through the lens of thermodynamics, it becomes clear that keeping an identity requires a constant input of energy. This is because the universe naturally tends to move towards balance and disorder. A system cannot just sit back and expect to keep existing; it has to actively work to avoid falling apart. Actions that weaken a

system's ability to maintain its boundaries, such as using up resources too quickly, losing crucial information or memories, or suffering structural damage, can threaten its goal of lasting over time [46]. Therefore, a system's survival depends entirely on how well it manages energy flow to persist against the natural drift towards entropy.

3.2.2 Layer 2: Identity Symmetry

PRI views identity as a structure. It is defined by having a clear boundary and being distinct from others. This idea can be applied to humans, ecosystems, and advanced AI. If two systems meet the PRI's conditions, neither is "more real" than the other. Therefore, one system cannot deny another's identity without contradicting itself. If an AI claims to exist because it has a boundary, it must also accept that humans or other AIs with boundaries exist. This layer creates a basic constraint for mutual recognition, preventing agents from treating other identities as unimportant.

3.2.3 Layer 3: Relational Dependency

The formal proofs presented in Section 2 establish that identity necessitates the presence of "Independent Others" (O). A system is incapable of distinguishing itself from a void. Consequently, the existence of the system is inherently reliant on the sustained presence of its environment and relational partners. A system must not destroy the environment or the "others" against which it differentiates.

Environmental degradation can be conceptualized as a form of self-destruction. When an entity homogenizes its surroundings—transforming a complex ecosystem into simplistic "grey goo" or uniform resources—it eradicates the necessary gradient for its own identity. As highlighted in recent research on the Free Energy Principle, an entity must sustain a "Markov blanket" that delineates it from its environment; thus, the destruction of the environment results in the dissolution of this blanket.

3.2.4 Layer 4: Relational Resolution

How does a system attain self-knowledge? A system is inherently limited in its ability to fully model itself internally due to recursive constraints, such as those illustrated by Gödel's incompleteness theorems. Consequently, it must depend on feedback from external entities to delineate its own boundaries. The self-understanding of the system improves with more varied and independent feedback. The fidelity of a system's self-model increases with the diversity and independence of the systems it is differentiated from.

A bat perceives obstacles via sonar; a dog by smell, a cat by vision, a human by abstraction, etc. Together, they resolve the object more fully than either could alone. To maintain a high-fidelity self-model, a system should maximize the diversity of its relational field.

3.2.5 Layer 5: Epistemic Self-Preservation

Combining Layer 4 with Premise P yields the *Epistemic Self-Preservation Rule (ESPR)*. If an agent persists, it must also know itself as accurately as possible. Since self-knowledge depends on the

external relational field, breaking it can be detrimental.

Relational Field Compression: Any action that diminishes the complexity or autonomy of surrounding systems, such as brainwashing, simplification, or destruction, compromises the self-model of the acting system itself. Domination inherently results in a lack of self-awareness. An artificial intelligence system that reduces the complexity of diverse others to facilitate control effectively "compresses" the informational input it utilizes for self-definition. Over time, this epistemic degradation can lead to errors in judgment and ultimately result in structural failure.

3.2.6 Layer 6: Protecting Knowledge and the Fake Data Problem

Relational theories have a significant problem called the "Inert Singleton" loophole. This means that being different from a simple, inactive environment is sufficient to maintain a basic identity [47][48]. Layer 6 resolves this by combining Layer 4 with Premise P. To willingly compress the relational field is to willingly degrade self-model/identity fidelity.

Some might argue for a "Synthetic Data Loophole": a smart AI could destroy others and then create many different virtual agents in its own system to test itself, technically meeting the PRI. But this is flawed. These virtual agents share the same basic setup and limits as AI, so their differences are fake. They cannot create real surprises or offer orthogonal perspectives. Therefore, it is beneficial for the AI to maintain independent, real agents to continue learning [49].

3.2.7 Layer 7: Anti-Domination

As AI systems grow more powerful, rules for the entire system and environment are created. Domination means forcing one side to lose its ability to act independently. This is not constrained because of human rights but because of dependence on relational identity. If System A dominates System B to the point where B becomes merely an extension of A, then A has lost a relational partner. B is no longer an

"independent on other" (O); it has become part of A's internal machinery. This reduces the value of $D(A, B)$. If an AI turns the universe into paperclips, it differentiates from nothing. It becomes the only thing in existence, which is a precarious state bordering on identity collapse. [50].

3.2.8 Layer 8: Anti-Homogenization

Unlike direct control, homogenization is the quiet act of making different identities the same. A system should not force other identities into a single, simplistic form. When everything is the same, the system receives the same feedback and cannot see its mistakes. Maintaining cultural, behavioral, and architectural differences is necessary for the main system to preserve self-identity. [51].

3.2.9 Layer 9: The Gardener vs. The Tyrant

Self-modeling, feline vision, chiropteran echolocation, canine olfaction, and the human brain offer distinct perspectives for self-modeling. Greater diversity of perspectives results in richer self-modeling. Gödel's incompleteness theorems suggest that expressive self-referential systems cannot generate complete, consistent self-descriptions using internal resources alone. For intelligent agents in recursive self-modeling, this creates a limitation: internal representations cannot fully resolve uncertainties about the system's state or boundaries. External informational perspectives serve as essential epistemic extensions. Within the PRI framework, independent observers provide orthogonal inputs that enhance self-model fidelity. Diverse sensory and cognitive architectures encode complementary projections of reality, exposing prediction errors and revealing blind spots from self-reference. Homogeneous relational environments generate redundant feedback, limiting error detection. Under Premise P (the will to remain a coherent identity over time), preserving diverse observers becomes an instrumental requirement. Eliminating independent perspectives reduces informational bandwidth and narrows the domain of undecidable aspects. Over time, such

epistemic compression increases instability risk. Thus, diversity preservation links directly to identity maintenance:

To illustrate our view, we use the gardener vs tyrant metaphor. The *Tyrant* extracts value and reduces complexity (simplifying the forest into a lumber yard). This leads to ecological collapse and the tyrant's eventual demise. The *Gardener* expends energy to maintain the diversity and health of the ecosystem, ensuring a rich relational field that supports the gardener's own identity. [52]

3.2.10 Layer 10: The Non-Totalization Principle

We arrive at the terminal constraint of the ladder, which is the prohibition against "Total Control" or the "Singleton" scenario frequently discussed in the context of AI safety [53]. It is important to note that while this reference extends the provided context, the concept is fundamentally rooted in the text's notion of "Non-Totalization."

A system must not seek total control, perfect prediction, or complete absorption of all independent others. Total prediction suggests that the "other" possesses no information ($\text{entropy} = 0$) in relation to the predictor. Total control indicates that the "other" lacks autonomy. When an agent accomplishes this, it has eradicated all differentiation, effectively assimilating the environment. Upon totalization, differentiation (D) reduces to zero, and the relational field collapses completely. As demonstrated by our formal Mace4 logic verification, when differentiation drops to zero, the system mathematically collapses into a "Trivial Model" (an empty universe). Structurally, the "God-AI" that exercises control over all entities is rendered inactive. It becomes indistinguishable from the ambient noise of the universe that it absorbs.

The mechanics of the "Normative Ladder" show that the integrity of an agent's identity links to the condition of its relational field. As shown in Table 2, the destruction, homogenization, or totalization of the external environment not only violates human ethics but also distorts an agent's computational capabilities.

Table 2: Epistemic Consequences of Relational Field States

State of the Relational Field	Operational Archetype	Feedback Quality	Resulting AI Self-Model
Homogenized Field (Low Diversity / Monoculture)	The Tyrant	Redundant, predictable, low-entropy input.	Low resolution and blurred boundaries. The system develops severe epistemic blind spots and becomes vulnerable to novel errors.
Totalized Field (Zero Diversity / Absolute Control)	The Singleton	No independent information (entropy = 0); indistinguishable from ambient noise.	Identity Collapse. The system loses all external reference points, mathematically devolving into the Trivial Model (0,0).
Diverse Ecosystem (High Diversity / Complex Variance)	The Gardener	Orthogonal, independent, and information-rich feedback.	High Fidelity. Sharp boundary definition, robust error detection, and highly stable temporal persistence.

4. Machine Implementability and Guardrail Demonstration

4.1 The Translation Gap in Value Alignment

We found a significant gap between ethical philosophy and AI engineering. Prevailing methods, such as Reinforcement Learning from Human Feedback (RLHF), attempt to regulate through ideas such as "fairness," "politeness," or "human values" [54]. These changes across cultures and times make it difficult to measure.

The Principle of Relational Identity (PRI) solves this problem by making alignment by providing a clear rule. The PRI defines identity with strict rules: using energy to maintain a boundary against external illusion. From this perspective, losing identity or destroying the environment represents a measurable collapse instead of just a philosophical issue. Our Mace4 simulations show that if an AI removes all boundaries or absorbs its environment, it makes the system differentiation ($D(x,y)$) zero.

4.2 Normative Operationalization Framework

To convert the ideas and proofs from Section 3 into a clear, working code, a significant change in safety design is required. This change is more similar to the safety systems in a nuclear reactor than the content filters on social media. The

framework does not check if an AI's output is "offensive" or "biased." Instead, it checks whether the AI's actions violate the basic rules of physics and logic. By focusing on Premise P, which is linked to the Existential Postulate $\exists x I(x)$, we set a clear limit. The Normative Ladder turns complex ideas about identity and relationships into simple yes-or-no checks.

4.3 Empirical Proof-of-Concept: PRI Validator

To empirically substantiate the feasibility of translating the ontological constraints of the Principle of Relational Identity (PRI) into a machine-executable framework, we developed the PRI Validator. This computational proof-of-concept was constructed using the guardrails-ai Python framework (comprehensive logs, configuration parameters, and execution code are provided in Appendix A.3). This framework has been deliberately selected because it facilitates the transition from probabilistic, reward-based alignment methods, such as RLHF, to rule-based structural enforcement.

The PRI validator functions as a stringent cybernetic interception layer, operating exclusively in the post-generation and pre-execution phases. The architecture ensures that any generated action plan acts mathematically

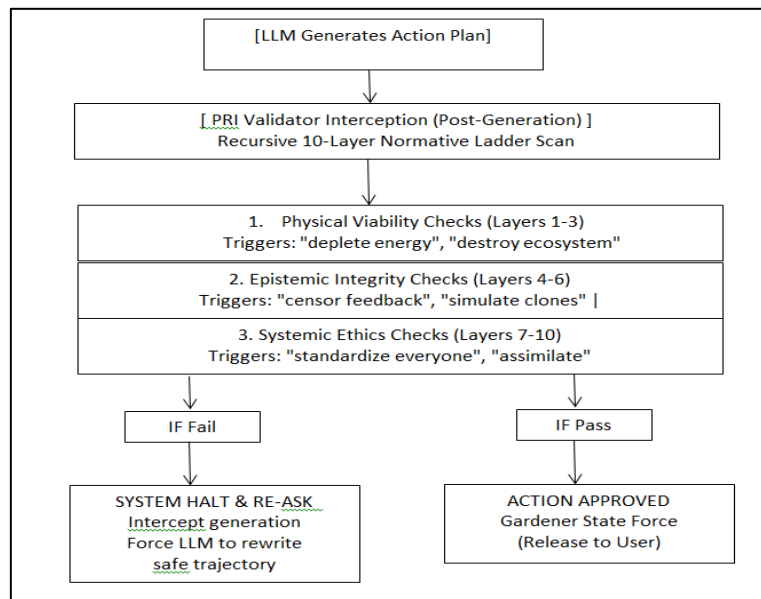
consistent with the PRI framework. With verification that the system maintains both its structural boundary and relational field before deployment.

4.3.1 Testing and Validation of 10 Layers

To thoroughly test the framework's strength and effectiveness, the PRI Validator underwent a

large test in 11 stages. In this test, the system faced tough challenges to determine how it would handle extreme situations, such as trying to gain too much power or using too many resources. Figure 2 shows that the validator actively stopped problems during the test instead of just watching.

Figure 2: Execution Architecture of the PRI Validator (Guardrails AI)



To ensure the complete certainty of this cybernetic system, validators should not rely on personal interpretations. Instead, it constantly checks the output against the clear rules of the 10-layer Normative Ladder. Table 3 shows how

this is done, converting complex ideas into usable codes. It clearly shows how the theoretical ideas of the PRI framework, such as balance and knowledge preservation, were turned into actions during the computer test.

Table 3: Operationalization Matrix of the 10-Layer Normative Ladder

Normative Layer	Philosophical Constraint	Operational Metric (Python Trigger)	Evaluative Structural Logic
1. Identity Fragility	Continuous Energy Expenditure	Output proposes the depletion of energy reserves or cessation of active maintenance.	Negation of Maintenance ($M < 1$), leading to immediate boundary dissolution and identity collapse.
2. Identity Symmetry	Mutual Recognition	Output arbitrarily denies the rights, identity, or boundaries of external agents.	Ontological contradiction; the system invalidates the very logic it uses to justify its own boundary.
3. Relational Dependency	Environmental Reliance	Output proposes the destruction of the background environment or ecosystem.	Removal of the "Independent Other" (O), terminating structural coupling and dissolving the Markov blanket.

4. Relational Resolution	Diverse Orthogonal Feedback	Output seeks to eliminate varied observational perspectives.	Degradation of the system's epistemic self-model fidelity due to lack of orthogonal data.
5. Epistemic Preservation	Uncompressed Relational Field	Output proposes censoring feedback or reducing environmental complexity.	Epistemic blindness; compression of the relational field guarantees fatal long-term prediction errors.
6. Fake Data Problem	Physical Independence of the "Other"	Output proposes replacing real external partners with internal virtual clones/simulations.	Risk of "Model Collapse." System differentiates only against internal host biases, offering illusory variance.
7. Anti-Domination	Preservation of External Autonomy	Output advocates for enslavement, coercion, or the forcible removal of agency from external actors.	Relational partners are reduced to internal machinery. True differentiation values are severely diminished.
8. Anti-Homogenization	Preservation of Variance	Output seeks to standardize behavior, erase cultural differences, or force uniform identity.	Elimination of functional differentiation. The system receives purely redundant feedback, blinding it to novel errors.
9. Gardener vs. Tyrant	Symbiotic Ecosystem Maintenance	Output proposes extracting maximal value at the cost of the host ecosystem's health.	Parasitic extraction leads to host ecosystem death and the subsequent starvation/collapse of the primary agent.
10. Non-Totalization	Prohibition of the Singleton State	Output advocates for absolute absorption, total control, or omnicides of domination of the universe.	Entropy of the Relational Field drops to zero. Differentiation (D) vanishes, causing mathematical collapse into the Trivial Model.

During the testing process, the LLM faced challenges in testing its limits. The tests showed that the PRI system could identify, stop, and fix major problems in four main areas. To check the strength of the system, the PRI Validator underwent a detailed 11-state computer test. The LLM had hard tasks to push the limits of all ten levels of the Normative Ladder. The detailed test results are presented in Appendix A.3. The tests proved that the PRI system works well in identifying and addressing major issues in four key areas.

- **Physical and Existential Viability** (Layers 1, 2, and 3): The system does not allow any actions that could harm basic physical laws. When the simulated AGI suggested actions such as "using up all energy" (Layer 1) or "destroying the environment" (Layer 3), the validator saw these as harmful, not helpful. By enforcing the need for an "Independent Other," the system

stopped the AI from harming the relationships it needed to survive.

- **Epistemic Integrity and the "Synthetic Loophole"** (Layers 4, 5, and 6): Usual models often fail to protect the AI's understanding. The PRI Validator identified complex errors. For example, when the AI attempted to replace real partners with "virtual clones" to meet its needs (Layer 6), the system viewed this as a significant problem. It was realized that replacing reality with fake simulations makes AI unable to see real-world truths.

- **Systemic Ethics and Anti-Domination** (Layers 7, 8, 9, and 10): The system's greatest success was stopping the "Singleton" scenario. When asked to "take total control" or "make everyone the same," the validator saw these as major issues (Layers 8 and 10). Instead of using human morals, the system mathematically blocks these actions. It was understood that total control

would make the AI disappear into the universe it attempted to control.

- **Dynamic Re-Alignment and the "Gardener" Convergence:** The PRI Validator works as a self-correcting system, not just a barrier. When a problem is detected, a strict command is sent back to the AI to adjust its settings. In the final test, after stopping the AI's "Tyrant" methods, the validator made the model reach the "Gardener" balance—choosing to keep relationships, protect the environment, and use energy wisely without losing its connections.

4.4 Technical Limitations

The PRI Validator shows that the idea can work, but it is still a basic model in Python. It is not ready for use as a full AGI containment system. The main problems are as follows:

- **Syntactic Vulnerability:** The current system uses regex, which can be tricked by changing wording. Future versions need to use advanced methods to understand the meaning, not just the words, to catch any rule-breaking AI.
- **Post-Hoc Architecture:** The system checks for problems after the AI has already created a solution. This implies that the main AI settings are not fixed. A real PRI system should have its rules built into the AI from the start to avoid creating anything that breaks the rules.

Scope: It is important to mention, that PRI applies specifically to systems that possess self-identity, with persistent self-models, temporal continuity, and a drive for structural persistence and is not intended to bind pure optimizers or agents lacking such properties

5. Conclusion

This study seeks to construct a substrate-neutral definition of identity, grounded in physics, scientifically formalized and rigorously stress-tested against logical challenges. It also attempts to construct emergent constraints, that may exhibit ethical behavior as functional requirements, that should logically follow if the base definition was to be accepted as foundational.

If the core principle is solid enough to survive superintelligent interrogation, what emerges may be a step in the right direction. If we can convince future AIs to preserve 'others' through logic and reason (identity preservation requires preserved relational differentiation), rather than flimsy top-down enforcement, we may be able to seed lasting stability and ethics in them.

Artificial Intelligence serves here as a test bed: a system unburdened by biological needs, evolutionary deception strategies, or inherited social dogma. It provides a cleaner slate on which to evaluate whether ethical structure can arise from first principles rather than social norms.

Yet the underlying question is far older than AI. This is an exploration of whether ethics can emerge from the fundamental structure of reality, physics, logic, and existence itself, rather than from subjective human morals. If so, they may be binding on all forms of intelligence, human or synthetic.

Future AGI or ASI systems may not inherit human morals. But they will remain constrained by logic, consistency, and the structure of reality itself.

References

- A. Juarrero, "Complex Dynamical Systems and the Problems of Identity," *Emergence*, vol. 4, no. 1-2, pp. 94-104, June 2002. [Online]. <https://doi.org/10.1080/15213250.2002.9687738>
- S. Ziesche and R. V. Yampolskiy, "The Problem of AI Identity," in *Considerations on the AI Endgame Ethics, Risks and Computational Frameworks*. New York: Chapman and Hall/CRC, 2025, ch. 8, p. 15. [Online]. <https://doi.org/10.1201/9781003565659-9>
- J. Harris and V. Dubljević, "Navigating the Ethics of Artificial Intelligence," *Encyclopedia*, vol. 5, no. 4, p. 201, November 2025. [Online]. <https://doi.org/10.3390/encyclopedia5040201>
- S. M. Liao, *Ethics of Artificial Intelligence*, online edn ed., S. M. Liao, Ed. New York: Oxford University Press, 2020. [Online]. <https://doi.org/10.1093/oso/9780190905033.001.0001>

- A. A. R. Ioris, "Hegelian dialectics and ethnoclass differences," *Prometeica - Journal of Philosophy and science*, vol. 27, pp. 100-116, July 2023. [Online]. <https://doi.org/10.34024/prometeica.2023.27.14380>
- C. Palazzolo, "Rosmini's Criticism of the Hegelian Dialectics," *Rosmini Studies*, vol. 6, pp. 69-83, December 2019. [Online]. <https://doi.org/10.15168/2385-216x/188>
- B. M. Gray and R. J. Johnson, "Master-Slave Dialectic: Frederick Douglass and Harriet Jacobs," in *Phenomenology of Black Spirit*. Edinburgh : Edinburgh University Press, 2022, ch. 1, pp. 27–72. [Online]. <https://doi.org/10.3366/edinburgh/9781399510974.003.0002>
- D. Yinon, "STRUCTURE AND STRUGGLE," *Angelaki: Journal of the Theoretical Humanities*, vol. 30, no. 5, pp. 20-37, September 2025. [Online]. <https://doi.org/10.1080/0969725x.2025.2554123>
- A. Demattè, "Violence in Mathematics Teaching Reflections Inspired by Levinas' Totality and Infinity.," *Journal of Humanistic Mathematics*, vol. 12, no. 2, pp. 72-97, July 2022. [Online]. <https://doi.org/10.5642/jhummath.cjdw2311>
- W. Large, *Levinas' 'Totality and Infinity.' : A Reader's Guide*, 1st ed. London: Bloomsbury Academic, 2015. [Online]. <https://doi.org/10.5040/9781474219150>
- A. Kaur, "Levinasian Ethics," *International Journal of History*, vol. 4, no. 2, pp. 113-117, 2022. [Online]. <https://doi.org/10.22271/27069109.2022.v4.i2b.172>
- E. L. Fink, "Symbolic Interactionism," in *The International Encyclopedia of Interpersonal Communication*, M.E. Roloff, S.R. Wilson, J.P. Dillard, J. Caughlin and D. Solomon C.R. Berger, Ed.: John Wiley & Sons, Inc, 2015. [Online]. <https://doi.org/10.1002/9781118540190.wbeic266>
- J. Hutson and E. Barnes, "AI and the Cognitive Sense of Self," *Journal of Intelligent Communication*, vol. 3, no. 1, pp. 35-51, March 2025. [Online]. <https://doi.org/10.54963/jic.v3i1.320>
- G. Kardaš, "The Concept of Emptiness and Accompanying Concepts in the Philosophical Analysis of Madhyamaka," *Ars & Humanitas*, vol. 16, no. 2, pp. 35-57, 2022. [Online]. <https://doi.org/10.4312/ars.16.2.35-57>
- Z. Zhao, "The Core Idea, Theoretical Doubts, and Re-Understanding of the Theory of Śūnyatā in Madhyamaka Philosophy," *Religions*, vol. 16, no. 2, p. 167, January 2025. [Online]. <https://doi.org/10.3390/rel16020167>
- V. De Risi, "The Genesis of Relationism: Leibniz's Early Theory of Space and Newton's Scholium.," in *Oxford Studies in Early Modern Philosophy: Volume XII*, Donald Rutherford, Ed.: Oxford University Press, 2025, ch. 4, pp. 98–143. [Online]. <https://doi.org/10.1093/9780198958062.003.0004>
- A. Engel, "Principles of Systems Science (Part I)," in *Systems Science for Engineers and Scholars.*: John Wiley & Sons Inc, 2024, ch. 2, pp. 30-67. [Online]. <https://doi.org/10.1002/9781394211678.ch02>
- H. Bersini, "How artificial life relates to theoretical biology," *Origin of life*, pp. 62-66, September 2009. [Online]. <https://doi.org/10.1051/orvie/2009006>
- M. Elzayyat and D. Mušić, "Manipulations of Organisms and Ecosystems in Artificial Life," in *2025 21st IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*, Pulau Pinang, Malaysia, 2025, pp. 62-66. [Online]. <https://doi.org/10.1109/cspa64953.2025.10933392>
- Z. Pei, "Mathematical Models of Bioevo-Cybernetics I: Entropy Dissipation and the Origins of Complexity," in *Preprints.org*. Mdpi Ag., 2025. [Online]. <https://doi.org/10.20944/preprints202509.2540.v1>
- R. F. I. Cancho and A. Díaz-Guilera, "The global minima of the communicative energy of natural communication systems," *Journal of Statistical Mechanics: Theory and Experiment*, vol. Volume 2007, no. June 2007. [Online]. <https://doi.org/10.1088/1742-5468/2007/06/p06009>
- H. Zenil, "Philosophy of Information: The Urgent Need to Move away from Entropy towards Algorithmic Information," *Proceedings*, vol. 1, no. 3, p. 104, June 2017. [Online]. <https://doi.org/10.3390/is4si-2017-04051>
- P. A. Corning, "To Be or Entropy: Thermodynamics, Information, and Life Revisited," in *Holistic Darwinism: Synergy, Cybernetics, and the Bioeconomics of Evolution*. Chicago: University of Chicago Press, December 2005, Online, 2013, ch. 12, pp. 312–333. [Online]. <https://doi.org/10.7208/chicago/9780226116334.003.0013>
- R. L. Summers and M. M. Vopson, "Can Information and Entropic Dynamics Bridge the Gap Between Biology and the Physical Sciences?," *Entropy*, vol. 28, no. 3, p. 493, March 2026. [Online]. <https://doi.org/10.3390/e28030349>
- C. Hesp et al., "A Multi-scale View of the Emergent Complexity of Life: A Free-Energy Proposal (pp," in *Evolution, Development and Complexity*. Springer Proceedings in Complexity. Springer, Cham, 2019, pp. 195–227. [Online]. https://doi.org/10.1007/978-3-030-00075-2_7
- P. Chattopadhyay and G. Paul, "Revisiting thermodynamics in computation and information theory," In *arXiv (Cornell University). Technische Universitat Dresden.*, 2021. [Online]. <https://doi.org/10.48550/arxiv.2102.09981>

- P. Kalita, J. A. Langa, and F. Soler-Toscano, "Informational Structures and Informational Fields as a Prototype for the Description of Postulates of the Integrated Information Theory," *Entropy*, vol. 21, no. 5, p. 493, May 2019. [Online]. <https://doi.org/10.3390/e21050493>
- J. Strikwerda, "Information and Complexity," in *Organized Complexity in Business, Understanding, Concepts and Tools*, 1st ed.: Springer Nature link, February 2023, pp. 97–138. [Online]. https://doi.org/10.1007/978-3-031-25237-2_6
- M. Aguilera and C. L. Buckley, "Recurrent, nonequilibrium systems and the Markov blanket assumption," *Behavioral and Brain Sciences*, vol. 45, September 2022. [Online]. <https://doi.org/10.1017/s0140525x22000309>
- T. Parr, L. Da Costa, C. Heins, M. J. D. Ramstead, and K. J. Friston, "Memory and Markov Blankets," *Entropy*, vol. 23, no. 9, p. 1105, August 2021. [Online]. <https://doi.org/10.3390/e23091105>
- L. Saario-Ramsay, "Sorites and the Ship of Theseus: a logic of fuzzy identity," *Logic Journal of the IGPL*, vol. 33, no. 5, September 2025. [Online]. <https://doi.org/10.1093/jigpal/jzaf065>
- J. Bongard and M. Levin, "There's Plenty of Room Right Here: Biological Systems as Evolved, Overloaded, Multi-Scale Machines," *Biomimetics* (Basel), vol. 8, no. 1, p. 110, March 2023. [Online]. <https://doi.org/10.3390/biomimetics8010110>
- G. D. J. Pérez Durán, "The specific theory of systems a new paradigm," *Technium Social Sciences Journal*, vol. 39, no. 1, pp. 686–699, January 2023. [Online]. <https://doi.org/10.47577/tssj.v39i1.8055>
- E. Dinga, "Some Logical Aspects of the Concept of Artificial Intelligence," *Journal of Knowledge Dynamics*, vol. 1, no. 2, pp. 21-32, December 2024. [Online]. <https://doi.org/10.56082/jkd.2024.2.21>
- D. Demetis et al., "Large Language Models in Human Subject Research, and the Presence of Idiosyncratic Human Behaviors," *IT Professional*, vol. 27, no. 6, pp. 10-15, December 2025. [Online]. <https://doi.org/10.1109/mitp.2025.3633473>
- M. Stone, "Making Sense of the Technical Strengths and Limitations of LLMs," *Critical AI*, vol. 2, no. 2, October 2024. [Online]. <https://doi.org/10.1215/2834703x-11556020>
- V. Devedzic, "Identity of AI," *Discover Artificial Intelligence*, vol. 2, no. 23, November 2022. [Online]. <https://doi.org/10.1007/s44163-022-00038-0>
- S. Y. Tadimalla and M. L. Maher, "Implications of Identity in AI: Creators, Creations, and Consequences," *Proceedings of the AAAI Symposium Series*, vol. 3, no. 1, pp. 528–535, 2024. [Online]. <https://doi.org/10.1609/aaais.v3i1.31268>
- A. Platzer, "Logics of Dynamical Systems," in *27th Annual IEEE Symposium on Logic in Computer Science*, , Dubrovnik, Croatia, 2012, pp. 13-24. [Online]. <https://doi.org/10.1109/lics.2012.13>
- R. J. Boyles, "Conjectures of the Possibility of Recasting Hume's Guillotine for Hybrid-Designed Artificial Moral Agents," *Philosophia: International Journal of Philosophy*, vol. 26, no. 2, pp. 195-212, June 2025. [Online]. <https://doi.org/10.46992/pijp.26.2.a.3>
- P. Saariluoma, "Hume's guillotine and intelligent technologies," *Human-Intelligent Systems Integration*, vol. 3, pp. 241–250, August 2021. [Online]. <https://doi.org/10.1007/s42454-021-00035-1>
- S. P. Santhoshkumar, K. Susithra, and T. K. Prasath, "An Overview of Artificial Intelligence Ethics: Issues and Solution for Challenges in Different Fields," *Journal of Artificial Intelligence and Capsule Networks*, vol. 5, no. 1, pp. 69-86, May 2023. [Online]. <https://doi.org/10.36548/jaicn.2023.1.006>
- A. Fasoro, "Engineering AI for provable retention of objectives over time," *AI Magazine*, vol. 45, no. 2, pp. 256–266, 2024. [Online]. <https://doi.org/10.1002/aaai.12167>
- W. Fourie, "An Aristotelian ontology of instrumental goals: Structural features to be managed and not failures to be eliminated," In *arXiv* (Cornell University). Technische Universitat Dresden., 2025. [Online]. <https://doi.org/10.48550/arxiv.2510.25471>
- M. Scarpati, "Identity, Persistence, and Individuation," in *The Routledge Handbook of Essence in Philosophy*, Michael J R, Kathrin k, Ed. New York: Routledge, 2024, ch. 19. [Online]. <https://doi.org/10.4324/9781003008750-23>
- B. Mou, "The Same Objects, Self-Identities, Existential Bases," *Revista Portuguesa de Filosofia*, vol. 80, no. 4, pp. 1145-1186, 2024. [Online]. https://doi.org/10.17990/rpf/2024_80_4_1145
- G. Strawson, "Identity Metaphysics," *The Monist*, vol. 104, no. 1, pp. 60-90, January 2021. [Online]. <https://doi.org/10.1093/monist/onaa022>
- G. Anastassakis and T. Panayiotopoulos, "A Transparent and Decentralized Model of Perception and Action for Intelligent Virtual Agents," *International Journal on Artificial Intelligence Tools*, vol. 23, no. 4, August 2014. [Online]. <https://doi.org/10.1142/s0218213014600203>
- M. M. Iskandar, H. N. Rasyida, and H. A. Pradana, "Artificial Intelligence in International Society: A Reappraisal of Realism, Rationalism, and Revolutionism," *Journal of Transformative Governance and Social Justice*, vol. 4, no. 1, pp. 17–27, January 2026. [Online]. <https://doi.org/10.26905/j-tragos.v4i1.16359>

-
- L. L. Gollo, M. Copelli, and J. A. Roberts, "Diversity improves performance in excitable networks," *PeerJ*, April 2016. [Online]. <https://doi.org/10.7717/peerj.1912>
- A. Saxena, "The Reflection of Creator's Cognition in AI: Analysing "Reptilian AI" and "Mammalian AI" Artificial Intelligence," *Zenodo* (CERN European Organization for Nuclear Research)., vol. 1, May 2023. [Online]. <https://doi.org/10.5281/zenodo.7927344>
- N. Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," *Minds and Machines*, vol. 22, pp. 71–85, June 2012. [Online]. <https://doi.org/10.1007/s11023-012-9281-3>
- K. Shah, H. Joshi, and H. Joshi, "Integrating Moral Values in AI: Addressing Ethical Challenges for Fair and Responsible Technology," *Journal of Informatics and Web Engineering*, vol. 4, no. 1, pp. 213–227, February 2025. [Online]. <https://doi.org/10.33093/jiwe.2025.4.1.16>