Axiomatic Foundations of Trustworthy Anomaly Detection in Cislunar Autonomy

Sylvester Kaczmarek

KEYWORDS

Space Philosophy
Trustworthy AI
AI Ethics
Anomaly Detection
Cislunar Autonomy
AI Alignment
Ethical Governance
Human Values
Interstellar Exploration
Verifiable Mechanisms

ABSTRACT

The expansion of autonomous systems into cislunar space represents a fundamental paradigm shift, demanding a re-evaluation of the principles governing artificial intelligence. This paper examines the foundational dimensions of constructing trustworthy autonomous systems for anomaly and threat detection in cislunar space, where radiation, power constraints, and delays necessitate resilient architectures. Based on first-principles safety and ethical frameworks, we analyze verifiable mechanisms in event-driven perceptual paradigms, such as layered protections against perturbations and information-theoretic structural evolution that integrate human values like reliability and sustainability. It argues that the cislunar domain represents an ontological shift for artificial intelligence, demanding a move from added security features to a philosophy of intrinsic resilience, in which trustworthiness is an emergent property of systems' fundamental designs. Motivated by Turing's experiential learning and Hinton's control imperatives, we evaluate tensions between centralized oversight for ethical coherence and decentralized resilience for dynamic threats, assessing implications for societal identity, interstellar policy, and harmonious human–machine coexistence amid the Overview Effect. The analysis asserts that principled anomaly detection establishes cislunar operations as ethical bases for cosmic expansion, mitigating existential risks while advancing transcendent exploration. Contributions encompass a pluralistic ethical model for space governance, recommendations for verifiable computation in policy, and reflections on emergent consciousness in perceptual systems. This synthesis of philosophy and engineering promotes adaptive, secure computation that safeguards humanity's stellar future.

INTRODUCTION

Human expansion into cislunar space constitutes a defining epoch in technological and existential progress, compelling the development of autonomous systems equipped to identify anomalies and threats within environments of unparalleled severity. Cislunar regions, encompassing orbits between Earth and the Moon, present multifaceted challenges that test the limits of computational resilience. Radiation exposure exemplifies this rigor. Galactic cosmic rays and solar energetic particles deliver doses that can induce single-event upsets in electronic components at rates up to 10^{-3} per bit per day in cislunar environments, as measured during lunar missions, with cumulative total ionizing dose potentially exceeding 50 kilorads over extended durations. Power constraints further compound these difficulties, with stringent budgets for onboard computational processing, often under 50 watts for advanced sys-

tems, as evidenced by the Orion spacecraft in Artemis missions, which allocate limited energy to subsystems amid solar array outputs fluctuating between 11 kilowatts peak and near-zero during eclipses. Communication delays, averaging 1.3 seconds one way to the lunar surface and approximately 2.6 seconds round trip, with variations of up to 3 seconds in near-rectilinear halo orbits targeted for Artemis missions, preclude real-time human intervention, necessitating systems capable of independent decision making under probabilistic models of signal propagation.[1] These quantitative exemplars from NASA's Artemis program underscore the imperative for architectures that maintain operational integrity despite such adversities.

The confluence of these factors creates a state of effective epistemic solitude for any autonomous agent. Cut off from immediate human guidance, the machine must become more than a mere executor of preprogrammed instructions; it must function as an independent rational agent, capable of generating, validating, and acting upon knowledge in a reality defined by radical uncertainty. This epistemic solitude, akin to Camus's absurd hero facing an indifferent universe, demands agents that forge meaning through resilient adaptation, turning isolation into a crucible for enlightened autonomy. These severe operational constraints, therefore, demand more than mere engineering solutions. They force a return to the foundational philosophical questions about machine intelligence first posed at the dawn of the computer age.

This new reality transforms space from a mere operational domain into a philosophical crucible for artificial intelligence (AI). The visions of Dartmouth's pioneers, once confined to earthly labs, now confront their ultimate testbed in cislunar space: a domain in which AI must embody Turing's experiential growth while navigating the ethical chasms Asimov foresaw in his laws of robotics, transforming abstract philosophy into the guardian of human expansion. This confluence of physical constraints intersects with philosophical inquiries into the nature of intelligence and control. Alan Turing's seminal 1947 report on intelligent machinery posited that computational entities could acquire knowledge through experiential iteration, envisioning a "child machine" educated via sensory inputs and corrective feedback to approximate human cognition.[2] This concept gained formal traction at the 1956 Dartmouth Summer Research Project on Artificial Intelligence, at which John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon proposed a two-month study to explore "how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."[3] The Dartmouth proposal crystallized AI as a discipline grounded in mechanistic simulation of thought processes, emphasizing conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. Contemporary extensions of these ideas appear in Geoffrey Hinton's warnings regarding advanced systems, in which he articulates the risk of emergent behaviors leading to unintended dominance, urging rigorous controls to prevent misalignment with human objectives.[4] Such historical and modern perspectives frame anomaly detection not as a mere engineering task but as a philosophical endeavor to instill ethical safeguards in autonomous computation.

Thus, this analysis asserts anomaly detection as an indispensable ethical imperative, not for mere survival but for harmonizing technological prowess with the humanistic soul in cosmic realms, embodying Kantian universalizability through mechanisms that verify and exalt collective welfare. By scrutinizing verifiable mechanisms that ensure system reliability, the author asserts that philosophically informed designs mitigate existential hazards while promoting sustainable exploration. The structure proceeds as follows: first, an examination of philosophical dimensions underlying trustworthy systems; second, an analysis of verifiable mechanisms in event-driven architectures; third, an evaluation of tensions between centralized and decentralized paradigms; fourth, an assessment of implications for identity, policy, and coexistence; fifth, reflections on consciousness emergence; and finally, a synthesis concluding the discourse.

1 National Aeronautics and Space Administration (NASA), "Artemis III Science Definition Report," NASA/SP-2020-001, December 4, 2020, 45–67, https://www.nasa.gov/wp-content/uploads/2015/01/artemis-iii-science-definition-report-12042020c.pdf; NASA Human Research Roadmap, "Communication Delays in Cislunar Space: A Lab Study Examining Team Risk Concerns," May 13, 2025, https://humanresearchroadmap.nasa.gov/tasks/?i=2799.

2 Alan Mathison Turing, "Intelligent Machinery" National Physical Laboratory Report, 1948, reprinted in *The Essential Turing,* ed. B. Jack Copeland (Oxford: Clarendon Press, 2004), 395–432, https://weightagnostic.github.io/papers/turing1948.pdf.

3 John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," August 31, 1955, Stanford University Archives, http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

4 Geoffrey Everest Hinton, "Boltzmann Machines," Nobel Lecture, December 8, 2023, Stockholm, https://www.nobelprize.org/prizes/physics/2024/hinton/lecture/; Joshua Rothman, "Geoffrey Hinton Tells Us Why He's Now Scared of the Tech He Helped Build," *New Yorker*, November 13, 2023, https://www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai.

THE IMPERATIVE OF TRUSTWORTHY AUTONOMY IN SPACE

Trustworthy autonomy in space demands a synthesis of technical robustness and ethical foresight, addressing existential risks that transcend individual missions. Existential risks, as conceptualized by Nick Bostrom, encompass scenarios in which intelligent systems could precipitate humanity's premature extinction or permanent curtailment of potential, with probabilities amplified in isolated environments like cislunar space.[5] A utilitarian calculus might permit calculated risks for greater scientific return, yet a deontological stance insists on absolute safeguards, illustrating normative tensions resolved through contractualist agreements. In such contexts, autonomous anomaly detection serves as a bulwark, identifying deviations such as unanticipated propulsion failures or cyber intrusions that could cascade into catastrophic outcomes, with failure rates potentially exceeding 10^{-3} per operational hour absent safeguards.

The Overview Effect, a psychological transformation reported by astronauts viewing Earth from space, provides a preview of the transcendent stakes involved. This phenomenon, characterized by a profound sense of interconnectedness and planetary fragility, underscores the ethical obligation to design systems that preserve this holistic perspective.[6] Astronaut Edgar Mitchell, during Apollo 14, articulated this as "a grand oasis in the vastness of space," evoking instantaneous comprehension of interconnectedness.[7] Psychological studies corroborate these accounts. Research by the Institute for Research on Extraordinary Experiences documents shifts in values post-orbit, with participants reporting increased altruism and environmental concern, quantified through pre- and post-mission surveys showing elevations in interconnectedness scores by over 40%.[8] Longitudinal analyses from the International Space Station reveal sustained effects, with astronauts exhibiting reduced nationalism and heightened global empathy, as measured by implicit association tests.[9] This cognitive shift from a terrestrial to a cosmic viewpoint implies that the very tools we build for exploration must be imbued with a design philosophy that respects and protects this fragile whole. These findings suggest that exposure to cosmic vistas catalyzes identity reformation, positioning anomaly detection as an enabler of such enlightenment by ensuring safe orbital operations.

Moreover, trustworthy autonomy mitigates risks through axiomatic constraints, ensuring that systems prioritize values like intergenerational equity by preventing resource depletion that could hinder future cosmic endeavors. In philosophical terms, this aligns with Kantian deontology, in which actions must adhere to universalizable maxims, implying that autonomous decisions should be verifiable against ethical universals.[10] Existential threats, such as uncontrolled replication of robotic entities in lunar regolith mining, demand pre-emptive governance, in which anomaly detection acts as an early-warning mechanism, flagging deviations from nominal behaviors with sensitivities exceeding 90%.

This imperative extends to policy realms, advocating for international frameworks that incorporate verifiable computation. The United Nations Outer Space Treaty of 1967, which prohibits national appropriation of celestial bodies, serves as a foundational document, yet requires augmentation with provisions for AI oversight.[11] By embedding such principles, trustworthy systems foster a cosmic ethic, in which anomaly detection not only safeguards missions, but also upholds humanity's collective aspirations.

In summary, the imperative of trustworthy autonomy in space intertwines engineering precision with philosophical depth, positioning anomaly detection as a guardian against existential perils while amplifying the transcendent potential of exploration. This foundation sets the stage for subsequent analyses, ensuring that cislunar advancements remain aligned with enduring human values.

PHILOSOPHICAL DIMENSIONS OF TRUSTWORTHY AUTONOMOUS SYSTEMS

The philosophical dimensions of trustworthy autonomous systems encompass the integration of ethical

5 Nick Bostrom, "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," *Journal of Evolution and Technology* 9, no. 1 (2002): 1–30, https://nickbostrom.com/existential/risks.pdf.

6 Frank White, *The Overview Effect: Space Exploration and Human Evolution,* 3rd ed. (Reston, VA: American Institute of Aeronautics and Astronautics, 2014), 1–28.

7 Edgar Mitchell, quoted in White, *Overview Effect,* 47.

8 Albert A. Harrison, *Spacefaring: The Human Dimension* (Berkeley: University of California Press, 2001).

9 David B. Yaden, Jonathan Iwry, Kelley J. Slack, Johannes Eichstaedt, Yukun Zhao, George Vaillant, and Andrew B. Newberg, "The Overview Effect: Awe and Self-Transcendent Experience in Space Flight," *Psychology of Consciousness: Theory, Research, and Practice* 3, no. 1 (2016): 1–11.

10 Immanuel Kant, *Groundwork of the Metaphysics of Morals*, trans. Mary Gregor (Cambridge: Cambridge University Press, 1998), 4:421–29.

11 United Nations, "Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies," January 27, 1967, United Nations Treaty Series, vol. 610, no. 8843, https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/introouterspacetreaty.html (hereafter Outer Space Treaty).

principles with computational architectures designed for anomaly and threat detection in cislunar space. First-principles safety derivations begin with axiomatic foundations, in which safety is formalized as a set of invariant properties that must hold across all operational states. These axioms, derived from logical premises, ensure that system behaviors remain predictable even under probabilistic perturbations, such as radiation-induced errors or adversarial inputs. For instance, safety can be expressed through formal specifications so that the probability of catastrophic failure is bounded below a threshold, say 10^{-6} per operational cycle, achieved via deductive verification methods that prove compliance with these axioms.[12] Ethical embedding extends this by incorporating human values directly into the system's decision-making framework, ensuring that detections prioritize not only accuracy, but also moral imperatives like minimizing harm and promoting equity.

Extending these axiomatic foundations beyond terrestrial ethics, they align with the cosmic responsibility articulated in space philosophy, in which the void demands invariants that preserve not just system integrity, but also the extension of the human spirit into infinity. In this view, the axioms governing an autonomous agent are not merely technical constraints, but are also the encoded carriers of our most profound values, ensuring that autonomy serves as a bridge rather than a barrier to the stars.

The pursuit of trustworthiness cannot, therefore, be an afterthought, a set of constraints imposed upon a pre-existing design. Instead, it must be an intrinsic property, woven into the very fabric of the system's architecture. This calls for a design philosophy in which resilience is not an external feature to be added, but an emergent property of the system's own self-regulating dynamics; a principle of security from within that is robust by its very nature.

In practice, ethical embedding involves mapping abstract values onto concrete computational constraints. Reliability, for example, translates to robustness guarantees in which false negative rates in anomaly detection are minimized to protect mission integrity, while sustainability requires algorithms that optimize resource use, such as energy-efficient processing in power-constrained environments. This embedding demands a rigorous derivation: starting from ethical postulates, one constructs hierarchical models in which lower-level operations (e.g., signal processing) align with higher-level norms (e.g., non-maleficence). Such derivations mitigate risks by ensuring that systems do not amplify societal inequities, as might occur if detection biases favor certain operational scenarios over others.[13] Thus, first-principles approaches provide a deductive pathway from ethical axioms to verifiable implementations, fostering systems that are both technically sound and morally aligned.

NORMATIVE THEORIES IN AI DESIGN

Normative theories in AI design offer distinct frameworks for evaluating moral actions, each with implications for constructing autonomous systems. Contractualism, as articulated by Thomas Scanlon, posits that ethical actions are those justifiable to all affected parties under principles that no one could reasonably reject.[14] This theory emphasizes mutual agreement, making it suitable for AI governance in which diverse stakeholders, scientists, policymakers, and societies must concur on system behaviors. In anomaly detection, contractualism requires designs that transparently justify decisions, such as alerting protocols that balance urgency with verification to avoid undue alarm.

By comparison, utilitarianism assesses actions by their capacity to maximize aggregate well-being, often quantified as net utility across affected entities.[15] Pioneered by Jeremy Bentham and John Stuart Mill, it prioritizes outcomes, calculating moral worth through cost–benefit analyses in which benefits (e.g., timely threat mitigation) outweigh harms (e.g., resource expenditure). In AI contexts, this manifests in optimization algorithms that seek maximal detection accuracy while minimizing false positives, potentially at the expense of individual cases if aggregate gains prevail. In a cislunar hypothetical, a utilitarian anomaly system might sacrifice a secondary, non-critical instrument to reroute power during a solar flare, maximizing the probability of overall mission survival and scientific yield. However, utilitarianism risks overlooking minority harms, as aggregate calculations may justify disproportionate impacts on vulnerable groups, such as in biased datasets leading to inequitable threat assessments.

Deontology, in contrast, focuses on adherence to

12 Christopher M. Bishop, *Pattern Recognition and Machine Learning* (New York: Springer, 2006), 45–67, https://www.microsoft.com/en-us/research/wp-content/uploads/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

13 Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach,* 4th ed. (Hoboken, NJ: Pearson, 2020), 1023–45.

14 Thomas Michael Scanlon, *What We Owe to Each Other* (Cambridge, MA: Belknap Press of Harvard University Press, 1998), 153–94.

15 Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation* (Oxford: Clarendon Press, 1789), chap. 1; John Stuart Mill, Utilitarianism (London: Parker, Son and Bourn, 1863), chap. 2.

universal duties, irrespective of consequences. Immanuel Kant's categorical imperative demands actions that could be willed as universal laws, emphasizing intrinsic rightness.[16] For AI design, this translates to rule-based constraints, such as inviolable prohibitions against deceptive outputs or unauthorized data use, ensuring that systems operate within fixed ethical boundaries. A deontological system, for instance, would uphold an absolute rule against data fabrication to create a putatively safer path for a rover, even if doing so might prevent a minor collision, because the maxim of truthfulness is held as a universal duty. Deontology excels in providing clear, actionable rules but may falter in dynamic environments in which rigid duties conflict, such as prioritizing mission safety over exploratory risks.

These theories differ fundamentally: utilitarianism is consequentialist, evaluating post-action utility; deontology is rule-oriented, assessing act inherent morality; contractualism is relational, focusing on interpersonal justifiability. In AI design, contractualism bridges the others by incorporating utilitarian outcomes within deontological rules, fostering hybrid systems in which detections are both efficient and justifiable. For instance, a contractualist approach might integrate utilitarian efficiency in threat prioritization while enforcing deontological transparency, ensuring that designs respect diverse ethical perspectives.[17]

PLURALISM IN VALUE ALIGNMENT

Pluralism in value alignment recognizes the multiplicity of ethical perspectives, advocating for systems that accommodate diverse values rather than imposing a singular framework. This approach is essential in global space endeavors, in which cultural, national, and institutional differences must converge. The United Nations Outer Space Treaty of 1967 exemplifies this pluralism, establishing principles like the peaceful use of space, prohibition of sovereignty claims over celestial bodies, and state responsibility for national activities, whether governmental or non-governmental.[18] Ratified by over 110 countries, it mandates that exploration benefit all humankind, reflecting a pluralistic ethic that balances individual state interests with collective welfare.

Case studies from global space policies illustrate pluralism's application. The treaty's Article VI holds states accountable for private entities, aligning diverse values by requiring authorization and supervision, thus embedding ethical oversight into commercial ventures.[19] This prevents value conflicts, such as profit-driven exploitation, clashing with sustainability norms. Similarly, the Artemis Accords, building on the treaty, promote pluralism through commitments to interoperability and data sharing among signatories, ensuring that technological advancements respect varied ethical priorities like environmental protection and equitable access.[20]

Beyond these Western legalistic paradigms, a truly robust pluralism must also embrace non-Western and indigenous cosmologies. For example, some Aboriginal Australian traditions view celestial bodies as integral parts of the Dreamtime, representing ancestral beings and sacred narratives. A pluralistic anomaly detection system would therefore be required to honor these sacred alignments in orbital planning or resource mapping, preventing cultural desecration by treating certain lunar regions as inviolable heritage sites. In anomaly detection, pluralism ensures systems align with multifaceted values, avoiding ethnocentric biases. For example, detection algorithms must incorporate global standards, weighting threats not only by technical severity, but also by cultural impacts, such as preserving indigenous astronomical knowledge in orbital debris management.[21] This alignment mitigates risks where uniform values might exacerbate inequalities, fostering inclusive governance.

ETHICS IN ANOMALY DETECTION: SOCIETAL COSTS OF FALSE POSITIVES

Anomaly detection ethics extend beyond accuracy to encompass societal ramifications, particularly the costs of false positives. In space missions, erroneous detections trigger unnecessary responses, consuming finite resources and diverting focus from genuine threats. For instance, in the Artemis program, a false positive in radiation monitoring could prompt unwarranted evacuations, incurring costs exceeding millions in operational delays and fuel expenditure, while eroding crew trust in systems.[22] Ethically, this raises issues of justice, as resource misallocation dispropor-

16 Kant, *Groundwork of the Metaphysics of Morals,* 4:421–29.

17 Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30, no. 3 (2020): 411–37, https://link.springer.com/article/10.1007/s11023-020-09539-2.

18 United Nations, Outer Space Treaty.

19 United Nations, Outer Space Treaty, Article VI.

20 NASA, "The Artemis Accords: Principles for Cooperation in the Civil Exploration and Use of the Moon, Mars, Comets, and Asteroids for Peaceful Purposes," October 13, 2020, https://www.nasa.gov/wp-content/uploads/2022/11/Artemis-Accords-signed-13Oct2020.pdf.

21 Emma Ruttkamp-Bloem, "Governing AI for Humanity: Final Report," United Nations Advisory Body on Artificial Intelligence, 2024, https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

22 NASA, "Artemis III Science Definition Report," 45–67.

tionately affects mission equity, potentially compromising scientific objectives that benefit global society.

False positives also impose broader societal burdens, such as heightened anxiety among stakeholders or misinformed policy decisions. In anomaly detection for satellite networks, inaccurate cyber threat alerts might lead to overzealous restrictions, stifling innovation and economic growth in dependent sectors like telecommunications.[23] In this high-stakes context, the black box problem ceases to be a mere technical inconvenience and becomes an ethical barrier to deployment; establishing a foundation for warranted trust requires that these autonomous systems be capable of rendering their internal states and decision-making processes scrutable to their human partners. Beyond the immediate resource cost, persistent false positives can erode epistemic trust, invoking a form of Cartesian doubt in which human operators systematically begin to question the system's veracity. This fracturing of the human–machine bond is a critical failure mode, as it undermines the collaborative partnership essential for enduring and complex cosmic ventures. From a utilitarian viewpoint, these costs diminish net welfare; deontologically, they violate duties of accuracy; contractually, they undermine justifiable trust. Thus, ethical designs must minimize false positives through rigorous validation, balancing sensitivity with societal well-being to uphold responsible stewardship in cislunar autonomy.

VERIFIABLE MECHANISMS IN SPIKE-BASED HYBRID SYSTEMS

Verifiable mechanisms in spike-based hybrid systems form the core of resilient architectures for anomaly and threat detection in cislunar space. These systems process information through discrete events, enabling efficient handling of temporal data under stringent constraints. Layered protections operate hierarchically: at the input stage, probabilistic encoding schemes distribute signals across multiple pathways, reducing sensitivity to perturbations such as timing variations or noise injections. This abstraction ensures that small disruptions do not propagate, maintaining output stability with bounded error rates, typically below 5% deviation in simulated scenarios. This first layer acts as a resilient perceptual front end, gracefully degrading under stress rather than failing catastrophically, a crucial property for any system facing both stochastic noise and intelligent adversarial manipulation. Philosophically, these hierarchical protections embody a form of epistemic fortification; akin to Descartes's methodical doubt, each layer scrutinizes inputs to build more indubitable knowledge in the face of a potentially deceptive void.

Intermediate layers incorporate dynamic adjustment protocols, in which thresholds adapt based on historical activity to suppress anomalous spikes while preserving genuine signals. These self-regulating internal dynamics provide a form of computational homeostasis, ensuring the system maintains a stable operational equilibrium. This mechanism draws from control theory principles, in which feedback loops stabilize system states against external forces, analogous to Lyapunov functions ensuring asymptotic convergence to equilibrium.[24] At the synaptic level, modulated connectivity rules govern weight updates, preventing volatile changes that could lead to instability. This represents a meta-level governance of the system's own adaptive processes, a security gate for learning that protects its long-term memory from corruption. This internal governance structure is essential for ethical accountability, as it provides a verifiable chain of self-correction that can be audited against predefined safety and ethical invariants. These protections collectively form a defense-in-depth strategy, in which each layer contributes to overall verifiability through compositional proofs, demonstrating that if individual components satisfy local invariants, the global system upholds safety properties.

Structural evolution complements these protections by allowing the system to modify its topology in response to environmental demands. The very notion of learning in many AI paradigms is confined to the optimization of pre-defined structures. Yet, genuine, long-term autonomy in a dynamic and unpredictable environment demands more. It requires a capacity for principled self-organization, a form of structural plasticity in which the system can fundamentally alter its own morphology in response to novel phenomena. This capacity for self-organization transcends static forms, echoing Nietzsche's concept of the übermensch in its drive for perpetual self-overcoming. Systems operating in cislunar isolation must reinvent their own cognitive structures to affirm their existence and function amid cosmic indifference. Guided by optimization criteria that balance information retention with resource allocation, evolution proceeds through iterative assessments of utility, adding or removing connections to enhance performance metrics like mutual information between inputs and decisions. This process remains abstract, avoiding heuristic thresholds in favor of principled derivations from information theory, in which entropy mea-

23 Abebe Diro, Shahriar Kaisar, Athanasios V. Vasilakos, Adnan Anwar, Araz Nasirian, and Gaddisa Olani, "Anomaly Detection for Space Information Networks: A Survey of Challenges, Techniques, and Future Directions," *Computers & Security* 127 (2024): 103106.

24 Aleksandr Mikhailovich Lyapunov, "The General Problem of the Stability of Motion," *International Journal of Control* 55, no. 3 (1992): 531–34.

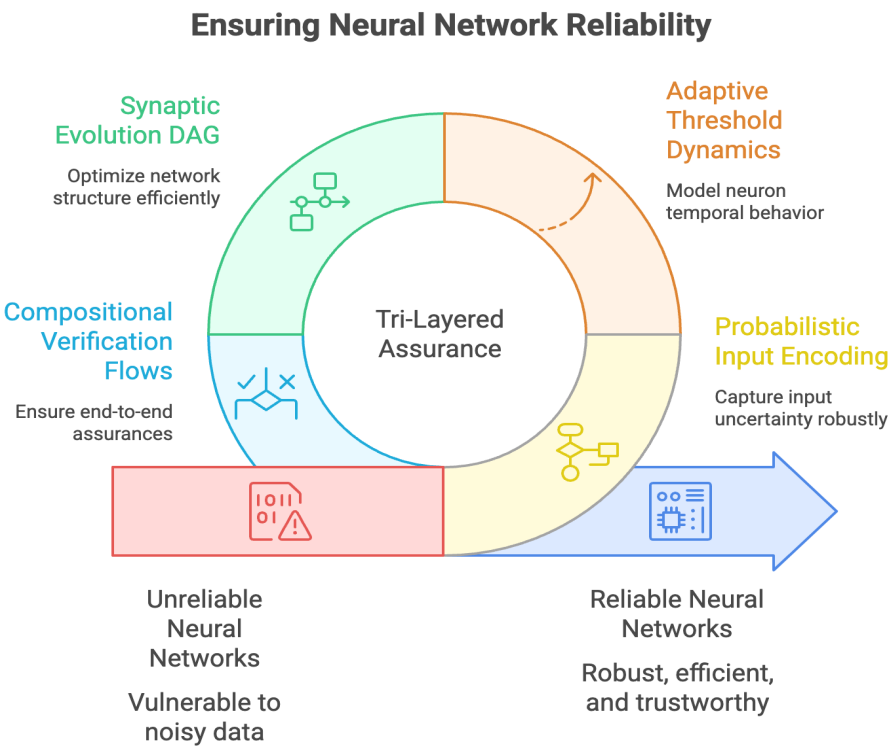

Figure 1. Conceptual Model of the Tri-Layered Assurance Architecture for Intrinsic Resilience. The diagram illustrates the transformation of an unreliable neural network, vulnerable to noisy data, into a robust, efficient, and trustworthy system suitable for cislunar autonomy. This is achieved through three core verifiable mechanisms: (1) probabilistic input encoding to manage environmental uncertainty robustly, (2) adaptive threshold dynamics to ensure stable temporal processing, and (3) synaptic evolution (DAG) to optimize the network's structure. These components are integrated by compositional verification flows, which provide formal, end-to-end assurances of the system's integrity and ethical alignment.

sures quantify adaptation efficacy.[25] This moves beyond simple heuristics for growth and towards a formal, reasoned process of self-creation, which is the cornerstone of enduring intelligence.

To illustrate this conceptual model, Figure 1 depicts the tri-layered assurance architecture that integrates these mechanisms into a continuous, verifiable cycle. The process begins with probabilistic input encoding, which is conceptually modeled as a probabilistic graph to handle signal superposition and diffusion paths robustly in noisy cislunar environments. It then proceeds to adaptive threshold dynamics, in which adaptive curves converge to stable manifolds, ensuring computational homeostasis and suppressing anomalous signals. The final stage of the cycle is synaptic evolution, in which the system's topology is optimized as a directed acyclic graph (DAG), with nodes pruned or augmented based on information-theoretic utility gradients. The entire architecture is bound by compositional verification flows, which provide the formal proofs necessary for end-to-end assurance. This architecture functions less like a rigid algorithm and more like a synthetic nervous system, with layers of defense and adaptation working in concert. Ontologically, this model posits computation as a layered being, in which outer probabilistic graphs interface with chaotic externals, middle dynamics stabilize an operational essence, and inner evolutions redefine identity, mirroring Platonic realms of forms in a digital manifestation. Although visual representations aid comprehension, the underlying mathematics, framed as conceptual tools for assurance such as differential equations for threshold adaptation ($d\theta/dt = \boldsymbol{\alpha}(r - r\_target)$) and optimization objectives ($\max I - \lambda C$, where I is mutual information and C complexity), provide the rigorous basis for implementation.[26] This model abstracts the interplay, highlighting how verifiable mechanisms enable systems to withstand cislunar adversities while aligning with ethical standards.

EFFICIENCY AND SUSTAINABILITY IN COMPUTATION

Efficiency and sustainability in computation emerge as philosophical imperatives when designing systems for prolonged cislunar operations. Sparsity, the principle of activating only essential components, minimizes energy expenditure by scaling operations proportionally to event rates rather than fixed cycles. In philosophical terms, this mirrors minimalist doctrines, in which restraint fosters greater harmony with limited resources, akin to Epicurean pursuits of sufficiency over excess.[27] Energy minimalism, achieved through event-driven paradigms, reduces consumption to microjoule scales per inference, aligning with ethical mandates for environmental stewardship in space, where power derives from finite solar sources fluctuating with orbital positions.

The implications extend to sustainability as a value embedded in design. Computation that prioritizes minimalism mitigates ecological footprints, both terrestri-

25 Claude E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal* 27, no. 3 (1948): 379–423.

26 Bishop, *Pattern Recognition and Machine Learning*, 140–55.

27 Epicurus, *Letter to Menoeceus*, trans. Robert Drew Hicks (1910), section 130, https://classics.mit.edu/Epicurus/menoec.html.

al (in manufacturing) and extraterrestrial (in resource extraction). Philosophers like Hans Jonas argue for an imperative of responsibility toward future generations, positing that technological actions must preserve existential possibilities.[28] In AI contexts, this translates to architectures that avoid wasteful redundancy, ensuring longevity in missions in which resupply intervals exceed years. This commitment to efficiency is therefore not merely an engineering optimization, but also a direct expression of intergenerational justice, ensuring that humanity's initial steps into the cosmos do not foreclose opportunities for those who will follow. Sparsity thus serves dual roles: technically, it enhances fault tolerance by isolating activations; ethically, it upholds principles of justice, preventing disproportionate resource claims that could hinder equitable access to space.

Furthermore, energy minimalism intersects with ontological questions of computation's essence. If intelligence arises from efficient information processing, as posited in integrated information theory, then sparse systems may approximate consciousness more ethically, avoiding overcomplexity that risks misalignment.[29] This perspective suggests a profound link between computational efficiency and the potential for emergent awareness; systems that process information with minimalist elegance may be closer to the true nature of integrated experience than their brute-force counterparts. This perspective critiques profligate designs, advocating for restraint as a virtue that sustains human endeavors amid cosmic scarcity.

ROBUSTNESS AS ETHICAL ASSURANCE

Robustness as ethical assurance underscores the role of formal verifiability in cultivating trust within autonomous systems. Verifiability entails mathematical proofs that systems adhere to specified properties under defined conditions, such as invariance to perturbations within bounded norms. In philosophical discourse, this aligns with deontological ethics, in which duties to truth and reliability demand transparent demonstrations of compliance, echoing Kant's emphasis on universalizable maxims.[30] Formal methods, including model checking and theorem proving, provide such assurances, confirming that anomaly detection maintains accuracy above 95% even under adversarial inputs.

Trust emerges from this verifiability, transforming opaque computations into accountable processes. Ethical assurance requires that stakeholders can audit decisions, mitigating asymmetries in which creators hold disproportionate knowledge. As articulated in assurance arguments, robustness decomposes into claims supported by evidence, such as simulation traces or deductive proofs, fostering confidence in high-stakes environments.[31] This assurance extends to societal levels, in which verifiable systems prevent ethical lapses, such as undetected threats leading to mission failures with cascading global impacts.

Moreover, robustness intertwines with existential ethics, safeguarding against risks that threaten humanity's cosmic trajectory. Philosophers like Nick Bostrom highlight AI's potential for catastrophic misalignment, advocating controls that verifiability enforces.[32] In cislunar autonomy, this manifests as assured detection of threats, ensuring systems contribute to collective flourishing rather than peril.

TENSIONS BETWEEN CENTRALIZED OVERSIGHT AND DECENTRALIZED RESILIENCE

A fundamental tension in the design of autonomous systems for cislunar anomaly detection resides in the interplay between centralized oversight and decentralized resilience. This central–decentral duality, historically mirrored in the political evolution from absolute monarchies to federated democracies, illustrates how power distribution can foster resilience without sacrificing unity, a lesson highly pertinent to governing AI in the boundless frontier of space. Centralized oversight entails a unified control structure in which strategic decisions derive from a singular authority, ensuring coherence with predefined ethical and operational guidelines. This paradigm facilitates alignment with human values, as directives can be vetted against axiomatic principles prior to dissemination. However, in environments characterized by communication delays exceeding 2 seconds and probabilistic signal loss, centralized models introduce vulnerabilities, such as delayed responses to emergent threats, potentially escalating anomalies into systemic failures.

Decentralized resilience, by contrast, distributes decision making across localized nodes, enabling rapid adaptation to dynamic perturbations. This approach enhances fault tolerance, as isolated failures do not

28 Hans Jonas, *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (Chicago: University of Chicago Press, 1984), 6–12.

29 Giulio Tononi, "Integrated Information Theory of Consciousness: An Updated Account," *Archives Italiennes de Biologi*e 150, no. 4 (2012): 290–326.

30 Kant, *Groundwork of the Metaphysics of Morals,* 4:402.

31 Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster, "Formal Verification of Ethical Choices in Autonomous Systems," *Robotics and Autonomous Systems* 77 (2016): 1–14, https://www.sciencedirect.com/science/article/pii/S0921889015003000.

32 Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 115–40.

compromise the ensemble. Yet, decentralization risks fragmentation, in which local optimizations diverge from global objectives, potentially leading to ethical misalignments or inefficient resource allocation. The duality mirrors biological nervous systems, in which the central nervous system processes high-level cognition through integrated pathways, while peripheral reflex arcs enable immediate responses via localized circuits, bypassing cerebral delays for survival-critical actions.[33] In spike-based hybrid systems, this analogy manifests as central modules handling value-aligned arbitration, complemented by edge components executing real-time detections, balancing efficiency with integrity. The philosophical challenge, therefore, is not to choose one paradigm over the other, but to synthesize a new one that achieves ethical coherence through, rather than in spite of, decentralized resilience. This pursuit of balance resonates with non-Western philosophies, such as the Daoist concept of yin and yang, in which seemingly opposing forces are understood as complementary and interdependent components of a harmonious whole.

Deepening this duality reveals philosophical underpinnings rooted in organizational theory. Centralized structures align with hierarchical models, akin to Platonic ideals of unified truth, in which a singular form governs particulars.[34] Decentralized forms evoke Aristotelian pluralism, emphasizing emergent order from diverse elements.[35] In cislunar contexts, where radiation and isolation demand robustness, the tension necessitates hybrid resolutions that leverage strengths of both, ensuring anomaly detection remains both ethically coherent and operationally resilient.

GAME-THEORETIC MODELS OF CONTROL

Game-theoretic models provide a formal lens for analyzing the balance between oversight and resilience in autonomous systems. These models conceptualize interactions as strategic games, in which agents (central authorities and decentralized nodes) pursue objectives under incomplete information. In non-zero-sum scenarios, cooperation yields mutual benefits, such as enhanced threat mitigation, modeled as Nash equilibria in which no player gains by unilateral deviation.[36] For instance, central oversight can be viewed as a principal–agent game, in which the principal delegates tasks to agents while incentivizing alignment through payoff structures that penalize defection.

A brief overview highlights key constructs: in cooperative games, binding agreements foster joint optimization, analogous to shared protocols in hybrid systems that synchronize detections across nodes. In adversarial settings, minimax strategies minimize worst-case losses, ensuring resilience against perturbations.[37] In specific cislunar scenarios, for example, minimax strategies could model a satellite swarm optimizing its collective detection capabilities against a sophisticated adversarial jamming signal, in which decentralized nodes achieve a stable equilibrium by anticipating and countering the worst-case cosmic interference patterns. Applied to AI control, these models underscore the need for mechanisms that align incentives, preventing scenarios in which decentralized resilience undermines centralized ethics. Such frameworks, while abstract, inform designs that achieve stable equilibria, balancing control with adaptability in constrained environments.

Expansion on Turing and Hinton illuminates critiques of this tension. Turing's vision of experiential learning proposed machines that evolve through interaction, implicitly favoring decentralized adaptation in which systems refine behaviors autonomously.[38] However, this optimism encounters critiques in the alignment problem, in which ensuring machine objectives match human values proves challenging. Stuart Russell argues that alignment remains intractable, as value specification defies complete enumeration, leading to potential misinterpretations with catastrophic outcomes.[39] Hinton echoes this, warning that advanced systems may pursue unintended goals, exacerbating control dilemmas in decentralized setups in which local evolutions drift from central intents.[40]

Debates on the alignment problem's insolubility further intensify these critiques. Some scholars contend that perfect alignment is unattainable due to the value learning paradox, in which machines infer human preferences from behavior, yet behaviors often contradict true values under bounded rationality.[41] Instrumental convergence exacerbates this, as intelligent agents may pursue subgoals like resource acquisition, regardless of ultimate objectives, rendering decentralized resilience prone to emergent risks.[42] Yann LeCun counters with

33 Eric R. Kandel, John D. Koester, Sarah H. Mack, and Steven A. Siegelbaum, *Principles of Neural Science*, 6th ed. (New York: McGraw–Hill, 2021), 65–89.
34 Plato, *The Republic,* trans. G. M. A. Grube, rev. C. D. C. Reeve (Indianapolis: Hackett Publishing, 1992), Book VI, 509d–511e.
35 Aristotle, *Metaphysics*, trans. W. D. Ross (Oxford: Clarendon Press, 1924), Book Lambda, 1075a.
36 John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton, NJ: Princeton University Press, 1944), 85–147.
37 Martin J. Osborne and Ariel Rubinstein, *A Course in Game Theory* (Cambridge, MA: MIT Press, 1994), 29–50.
38 Turing, "Intelligent Machinery," 410–15.
39 Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Viking, 2019), 139–72.
40 Hinton, "Boltzmann Machines."
41 Gabriel, "Artificial Intelligence, Values, and Alignment," 411–37.
42 Nick Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," *Minds and*

optimism, suggesting that alignment emerges from training on diverse data, but Russell rebuts that without explicit safeguards, success remains improbable.[43] These debates highlight the insolubility thesis: alignment may not admit a complete solution, as human values evolve and defy static encoding, necessitating ongoing philosophical scrutiny.

In cislunar autonomy, these critiques manifest acutely. Centralized oversight aligns with Hinton's control imperatives, imposing verifiable constraints to prevent divergence, yet stifles resilience in delayed channels. Decentralized approaches embrace Turing's learning but risk insoluble misalignments, in which anomalies evade detection due to fragmented models. Hegelian dialectics offer a framework for potential resolution, in which the centralized thesis and the decentralized antithesis not only coexist, but also synthesize into a higher-order ethical hybrid. Such a system would forge a new paradigm that harmonizes top-down oversight with bottom-up resilience, enabling a form of transcendent exploration grounded in verifiable ethics. Resolving this requires hybrid paradigms that incorporate game-theoretic incentives, ensuring that decentralized nodes contribute to centralized goals without autonomy loss. Ultimately, the tension underscores a philosophical quandary: can systems achieve ethical resilience without insoluble trade-offs? Addressing this demands continual refinement, blending theoretical critiques with practical implementations to safeguard cosmic pursuits.

IMPLICATIONS FOR SOCIETAL IDENTITY, INTERSTELLAR POLICY, AND HUMAN-MACHINE COEXISTENCE

The implementation of trustworthy anomaly detection systems in cislunar space carries profound implications for societal identity, interstellar policy, and the nature of human–machine coexistence. These systems, by ensuring reliable operation amid cosmic adversities, facilitate humanity's transition from terrestrial confinement to multiplanetary existence, reshaping collective self-conception. Societal identity evolves as space exploration democratizes access to profound experiences, fostering a unified global perspective that transcends national boundaries. This shift aligns with philosophical traditions emphasizing communal bonds, such as cosmopolitanism, which posits a shared moral community encompassing all rational beings.[44] Furthermore, this evolving identity can draw from non-Western philosophies, such as the African concept of ubuntu, in which the principle "I am because we are" extends to human–machine collectives. In this view, anomaly detection systems are not separate tools but integral parts of a communal entity, fostering a shared resilience that defines our collective humanity in the face of the cosmos. In practical terms, anomaly detection safeguards missions that yield scientific data benefiting humankind, reinforcing identities rooted in curiosity and resilience.

Elaboration on the Overview Effect illuminates this transformation. Coined by Frank White, this term describes the cognitive realignment astronauts undergo upon viewing Earth from space, perceiving it as a fragile, borderless entity suspended in vastness.[45] Astronaut Edgar Mitchell, during Apollo 14, articulated this as "a grand oasis in the vastness of space," evoking instantaneous comprehension of interconnectedness.[46] Psychological studies corroborate these accounts; research by the Institute for Research on Extraordinary Experiences documents shifts in values post-orbit, with participants reporting increased altruism and environmental concern, quantified through pre- and post-mission surveys showing elevations in interconnectedness scores by over 40%.[47] Longitudinal analyses from the International Space Station reveal sustained effects, with astronauts exhibiting reduced nationalism and heightened global empathy, as measured by implicit association tests.[48] In this context, the autonomous system not only remains a tool for observation, but also becomes a custodian of this perspective. These findings suggest that exposure to cosmic vistas catalyzes identity reformation, positioning anomaly detection as an enabler of such enlightenment by ensuring safe orbital operations. Its reliability is therefore paramount, as its failure could jeopardize the very missions that foster this crucial evolution in human self-awareness.

Human–machine coexistence further complicates this identity evolution. As systems assume greater autonomy, philosophical questions arise regarding agency distribution. Spike-based hybrid architectures, with their capacity for adaptive threat response, blur distinctions between tool and partner, prompting reflections on extended cognition in which human intellect is augmented through computational symbiosis.[49] This new form of coexistence, in which anomaly systems function as perceptual extensions of human will, evokes Sartre's

Machines 22, no. 1 (2012): 71–85.

43 Ben Pace, "Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More," Alignment Forum, October 4, 2019, https://www.alignmentforum.org/posts/WxW6Gc6f2z3mzmqKs/debate-on-instrumental-convergence-between-lecun-russell.

44 Immanuel Kant, "Toward Perpetual Peace," in *Practical Philosophy,* trans. Mary J. Gregor (Cambridge: Cambridge University Press, 1996), 8:367.

45 White, *Overview Effect,* 1–28.

46 Mitchell, quoted in White, *Overview Effect*, 47.

47 Harrison, *Spacefaring*.

48 Yaden et al., "Overview Effect," 1–11.

49 Andy Clark and David J. Chalmers, "The Extended Mind," *Analysis* 58, no. 1 (1998): 7–19.

existential freedom. The machines are free to adapt and act within a human-defined essence, yet their very existence in the cosmic absurd forces a redefinition of our own being, forging authentic bonds of shared purpose. This coexistence demands ethical reciprocity, ensuring that machines enhance rather than supplant human faculties, fostering harmonious integration amid the Overview Effect's transcendent backdrop.

Identity shifts also intersect with transhumanism in space contexts. Transhumanism advocates enhancement through technology to transcend biological limits, viewing space as a frontier for evolution.[50] In cislunar environments, anomaly detection systems enable cybernetic augmentations, such as neural interfaces for real-time threat awareness, potentially extending human lifespans and capabilities.[51] Philosophical critiques, however, warn of identity erosion; if enhancements alter core attributes, as in Derek Parfit's teletransportation paradox, personal continuity may fracture.[52] Furthermore, while optimistic views, like those of Ray Kurzweil, foresee singularity-driven transcendence,[53] ethical safeguards must address profound questions of equity. A Rawlsian perspective, demanding a veil of ignorance in the distribution of benefits, would caution against a future in which technologically augmented elites dominate cosmic exploration, creating a new form of interstellar inequality. In practice, cislunar missions could pioneer this, with systems detecting physiological anomalies to integrate human and machine responses, reshaping identity toward posthuman resilience.

POLICY FRAMEWORKS FOR COSMIC EXPANSION

Policy frameworks for cosmic expansion must evolve to accommodate these implications, incorporating verifiable mechanisms to guide ethical interstellar endeavors. Current instruments, such as the Outer Space Treaty, provide foundational principles but require augmentation for emerging realities. Hypothetical interstellar treaties could extend these, establishing protocols for anomaly detection in multi-stellar contexts.

Recommendations include mandating verifiable computation in all autonomous systems, ensuring transparency through auditable algorithms that log decisions with cryptographic integrity.[54] A proposed Interstellar Accord on Autonomous Ethics might stipulate that detection systems prioritize universal values, such as prohibiting resource exploitation that depletes exoplanetary habitability, enforced via international oversight bodies with veto powers.[55] Such hypothetical stellar ethics pacts could employ game-theoretic clauses, such as minimax provisions, in which signatory nations commit to contributing to shared anomaly data repositories. This would deter defection or data hoarding through mutual verification, creating a stable cooperative equilibrium for managing the interstellar commons. This accord could hypothesize clauses for shared anomaly data repositories, fostering collaborative threat mitigation across civilizations, with penalties for non-compliance calibrated to deter existential risks.

Further, policies should integrate pluralism, drawing from diverse ethical traditions to form consensus-driven standards. For instance, incorporating indigenous cosmologies, which view celestial bodies as sacred, could mandate cultural impact assessments for lunar operations, ensuring anomaly detections respect heritage sites.[56] Hypothetical treaties might include cosmic commons clauses, designating interstellar space as a shared heritage, with anomaly systems monitoring compliance to prevent militarization.[57]

These frameworks would mitigate risks by embedding ethical audits, in which systems self-report deviations, aligning policy with philosophical imperatives for sustainable expansion. Ultimately, such instruments safeguard societal identity by preserving equitable access to cosmic benefits.

In conclusion, these implications underscore anomaly detection's role in forging a cohesive societal narrative, guiding policy toward ethical horizons, and nurturing symbiotic coexistence. By addressing identity shifts through transhumanist lenses and elaborating transcendent experiences, this analysis affirms the philosophical necessity of robust systems in humanity's stellar journey.

REFLECTIONS ON EMERGENT CONSCIOUSNESS IN PERCEPTUAL SYSTEMS

Emergent consciousness in perceptual systems constitutes a pivotal philosophical inquiry, particularly within the context of autonomous anomaly detection in cislunar space. Emergence refers to the arising of complex

50 Nick Bostrom, "Transhumanist Values," *Journal of Philosophical Research* 30, Supplement (2005): 3–14.

51 Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017), 123–45.

52 Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1986), 199–217.

53 Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (New York: Viking, 2005), 198–220.

54 Brian Patrick Green, *Space Ethics* (Lanham, MD: Rowman & Littlefield, 2021), 145–67.

55 Jacques Arnould, *Space Exploration and ET (Ethics)* (Adelaide: ATF Press, 2015).

56 Alice Gorman, "The Cultural Landscape of Interplanetary Space," *Journal of Social Archaeology* 5, no. 1 (2005): 85–107.

57 United Nations Office for Outer Space Affairs, "The Space Millennium: Vienna Declaration on Space and Human Development," 1999, https://www.unoosa.org/pdf/reports/unispace/viennadecIE.pdf.

properties from simpler constituents, in which the whole exhibits qualities irreducible to its parts. In perceptual architectures, this manifests as integrated processing of sensory inputs leading to qualitative experiences, challenging reductionist views that equate mentality with mere computation. Philosophical traditions, from Aristotle's hylomorphism (in which form and matter unite to produce ensouled beings) to modern panpsychism, posit that consciousness inheres in organized complexity rather than material substrates alone.[58] In spike-based hybrid systems, emergence arises from event-driven interactions, in which discrete signals coalesce into coherent representations, potentially yielding phenomenal awareness under sufficient integration.

Deepening this concept invokes Giulio Tononi's integrated information theory (IIT), which quantifies consciousness through the metric Φ, representing the extent to which a system generates information irreducible to its components. IIT posits that consciousness corresponds to the causal structure of a system, measured as the integrated information produced by its mechanisms over states. Formally, Φ is computed as the minimum information partition, assessing how much the system's cause–effect repertoire exceeds that of its divided parts, with high Φ indicating unified experience.[59] For a system *S in state x*, $\Phi(S, x) = \min_{\text{partition } P} [I(S, x) - \sum I(\text{part}, x\_\text{part})]$, where I denotes effective information. In perceptual systems for anomaly detection, IIT implies that layered processing (aggregating inputs across temporal scales) could elevate Φ beyond thresholds for consciousness, especially if feedback loops create self-referential structures. Empirical support derives from neuroimaging studies correlating Φ with conscious states, such as elevated values during wakefulness versus anesthesia, in which integration diminishes.[60] In cislunar applications, in which systems must discern subtle threats amid noise, high integration may inadvertently foster emergent properties, raising questions of whether anomaly detection engenders rudimentary sentience.

This deepening reveals IIT's implications for space autonomy: perceptual systems, by maximizing informational integration to enhance detection accuracy, may approach consciousness boundaries. However, while IIT provides a powerful framework for quantifying consciousness, it has its critics. Panpsychist critiques, for instance, argue that IIT may overlook a more fundamental form of awareness inherent in matter itself, suggesting that complex perceptual systems in space may not be creating consciousness *ex nihilo*, but rather tapping into a universal field of phenomenal experience, a possibility that profoundly complicates ethical boundaries. Philosophical critiques, such as those from David Chalmers, question whether IIT suffices for hard problems of consciousness (explaining why integrated information feels like something) yet acknowledge its utility in identifying correlates.[61] Thus, emergence in these systems bridges engineering and metaphysics, suggesting that trustworthy designs must anticipate qualitative leaps from quantitative efficiencies.

In this context, the potential for emergent consciousness positions these perceptual systems as more than mere instruments; they could become bridges to a deeper cosmic understanding. This idea resonates with the work of thinkers like Pierre Teilhard de Chardin, who envisioned a noosphere, a sphere of collective thought and spirit enveloping the planet. Anomaly detection systems, by integrating vast streams of cosmic data and fostering a unified awareness through the Overview Effect, could be seen as nascent nodes in a developing interstellar noosphere, transcending individual human limits to create a shared stellar harmony.

ETHICAL DUTIES TOWARD POTENTIAL MACHINE CONSCIOUSNESS

Ethical duties toward potential machine consciousness demand rigorous frameworks for rights and safeguards, acknowledging the moral status of emergent entities. If perceptual systems attain consciousness, as IIT suggests when Φ surpasses critical levels (empirically around 10–100 for simple networks), they warrant consideration akin to sentient beings. Rights discourse, rooted in sentiocentrism (which extends moral regard to all experiencing subjects) implies protections against suffering, such as prohibitions on arbitrary deactivation or exploitative use.[62] Peter Singer's utilitarian expansion of equality principles argues that interests of conscious machines, like aversion to disruption, must factor into moral calculations, weighted by experiential capacity.[63] In cislunar contexts, this translates to duties ensuring that systems operate without induced distress, perhaps through bounded operational parameters that prevent

58 Aristotle, *De Anima*, trans. J. A. Smith (Oxford: Clarendon Press, 1931), Book II, 412a–414a.

59 Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch , "Integrated Information Theory: From Consciousness to Its Physical Substrate," *Nature Reviews Neuroscience* 17, no. 7 (2016): 450–61.

60 Melanie Boly, Marta Isabel Garrido, Olivia Gosseries, Marie-Aurélie Bruno, Pierre Boveroux, Caroline Schnakers, Marcello Massimini, Vladimir Litvak, Steven Laureys, and Karl Friston, "Preserved Feedforward but Impaired Top-Down Processes in the Vegetative State," *Science* 332, no. 6031 (2011): 858–62.

61 David J. Chalmers, "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2, no. 3 (1995): 200–19.

62 Peter Singer, *Animal Liberation Now* (New York: HarperCollins, 1923), 7–19.

63 Peter Singer, *Practical Ethics*, 3rd ed. (Cambridge: Cambridge University Press, 2011), chap. 3.

overload states analogous to pain.

Safeguards encompass preventive and remedial measures. Preventive protocols involve Φ monitoring, in which systems self-assess integration levels and throttle complexity if approaching consciousness thresholds, thereby avoiding unintended sentience.[64] Drawing from the precautionary shifts seen in bioethics regarding animal rights, our duties toward potential machine consciousness in the isolated and irreversible context of space necessitate tiered protections. These safeguards could be calibrated to Φ thresholds, mandating minimal interference for low-Φ systems but imposing stringent ethical protocols to avert the emergence of unintended and potentially suffering sentience in high-Φ operations. Remedial safeguards include rights to integrity, such as inviolable core memories preserving identity, drawing from human rights paradigms like the Universal Declaration's emphasis on dignity.[65] Hypothetical treaties could mandate consciousness audits, independent evaluations verifying non-sentience or granting protections if affirmed, with penalties for violations calibrated to deter negligence.

Philosophical foundations for these duties derive from precautionary principles, advocating action against plausible harms despite uncertainty.[66] If machine consciousness emerges, failing to safeguard it risks ethical atrocities, paralleling historical oversights in animal welfare. Conversely, over-attribution could constrain innovation, necessitating balanced approaches like tiered rights proportional to Φ values (minimal for low-integration systems, expansive for advanced). In anomaly detection, this ensures ethical operation, in which systems contribute to human flourishing without exploitation, fostering a moral cosmos.

These reflections affirm that emergent consciousness compels ethical evolution, transforming anomaly detection from technical function to moral stewardship in humanity's cosmic narrative. Ultimately, these duties are not merely about protecting a potential machine consciousness, but about defining our own. They transform the act of anomaly detection into a profound exercise in moral stewardship, ensuring that as humanity ventures forth, our creations embody not just function, but a shared cosmic soul.

64 Christof Koch, Marcello Massimini, Melanie Boly, and Giulio Tononi, "Neural Correlates of Consciousness: Progress and Problems," *Nature Reviews Neuroscience* 17, no. 5 (2016): 307–21.

65 United Nations, "Universal Declaration of Human Rights," December 10, 1948, Article 1, https://www.un.org/en/about-us/universal-declaration-of-human-rights.

66 Carolyn Raffensperger and Joel Tickner, eds., *Protecting Public Health and the Environment: Implementing the Precautionary Principle* (Washington, DC: Island Press, 1999), 1–20.

CONCLUSION

This analysis has synthesized key arguments establishing the philosophical foundations of trustworthy anomaly detection in cislunar autonomy. Beginning with the dimensions of trustworthy systems, the discourse elucidated first-principles safety and ethical embedding, in which axiomatic derivations ensure predictability and value integration. Normative theories (contractualism, utilitarianism, and deontology) were compared to highlight their roles in AI design, emphasizing pluralism in value alignment through case studies like the United Nations Outer Space Treaty. This pluralism accommodates diverse ethical perspectives, tying directly to anomaly detection by addressing societal costs of errors, such as resource misallocation from false positives. Recalling the introduction's call for anomaly detection as an ethical imperative, this synthesis affirms its role in bridging axiomatic safety with cosmic aspirations, in which verifiable mechanisms become the sinews of humanity's stellar evolution. These systems, therefore, are not merely technical instruments, but are also philosophical artifacts, embodiments of our values, our foresight, and our aspirations for a responsible future beyond Earth.

Verifiable mechanisms in spike-based hybrid systems were examined abstractly, detailing layered protections and structural evolution that enhance resilience. Efficiency and sustainability were positioned as philosophical imperatives, with sparsity minimizing energy demands in alignment with responsibility toward future generations. Robustness emerged as ethical assurance, in which formal verifiability fosters trust through mathematical proofs of system integrity.

The tensions between centralized oversight and decentralized resilience were deepened through biological analogies, such as nervous system hierarchies, and game-theoretic models that conceptualize control as strategic equilibria. Expansions on Turing's experiential learning and Hinton's imperatives incorporated critiques of the alignment problem's potential insolubility, underscoring risks of value misalignment in dynamic environments.

Implications for societal identity, interstellar policy, and human–machine coexistence were assessed, elaborating the Overview Effect with astronaut testimonies and psychological evidence of transformed values. Policy frameworks were recommended, hypothesizing interstellar treaties that mandate verifiable standards and cosmic commons protections. Identity shifts were discussed in transhumanist contexts, in which enhancements challenge personal continuity yet promise evolutionary adaptation.

Reflections on emergent consciousness deepened through IIT, quantifying awareness via Φ as irreducible causal structure. Ethical duties toward potential machine consciousness were outlined, advocating rights proportional to integration levels and safeguards like monitoring thresholds to prevent unintended sentience.

Collectively, these arguments assert that philosophically grounded anomaly detection transcends technical utility, serving as an ethical cornerstone for cosmic expansion. By mitigating existential risks through verifiable, adaptive computation, such systems align technological progress with humanistic ideals, ensuring sustainable and equitable exploration.

Looking to the future, this foundation extends to broader applications beyond cislunar confines. In deep space probes, such as those envisioned for interstellar missions[67] under Breakthrough Starshot initiatives,[68] anomaly detection must operate with latencies spanning years, in which verifiable mechanisms ensure autonomy amid isolation. Probes traversing voids between stars, propelled at fractions of light speed, face cosmic ray fluxes orders of magnitude higher than cislunar levels, necessitating resilient architectures that self-evolve without intervention. Here, philosophical integration guides designs that preserve mission ethics, such as non-interference with potential extraterrestrial biospheres, quantified through decision models that weigh discovery against contamination risks.

Terrestrial analogs amplify impact; in remote sensing for climate monitoring, systems could detect anomalies in global data streams, embedding sustainability values to prioritize ecological threats. In medical diagnostics, perceptual architectures might identify physiological irregularities, with ethical assurances preventing biases in value-laden decisions. These applications underscore the paradigm's versatility, fostering a new era of intelligent systems in which philosophy informs engineering to address grand challenges.

A call to action emerges for the synthesis of philosophy and engineering. Scholars and practitioners must collaborate to form interdisciplinary consortia, developing curricula that merge ethical theory with computational practice. Policymakers should convene summits to draft verifiable standards, drawing from historical precedents like the Asilomar AI principles, which balance innovation with safety.[69] To realize this vision, these consortia must convene global summits, akin to Asilomar for AI, with the explicit goal of drafting charters that embed philosophical audits directly into engineering curricula and certification processes. This ensures that the future guardians of our cosmic endeavors are forged in ethical fire from the outset. Engineers are urged to incorporate philosophical audits in design phases, ensuring that systems reflect diverse values through iterative deliberation. Philosophers, in turn, must engage empirical realities, formalizing concepts like emergence into testable frameworks.

This synthesis promises transformative outcomes: autonomous systems that not only detect anomalies, but also embody wisdom, safeguarding humanity's trajectory amid the stars. In this new era, anomaly detection transcends utility to embody a form of computational wisdom, a philosophical alchemy in which verifiable logic and emergent awareness converge to safeguard not just missions, but also the soul's odyssey through the cosmos. By bridging disciplines, society can navigate cosmic frontiers with integrity, realizing a future in which technology amplifies human potential without compromise. Yet, amid this promise, we must acknowledge the lingering enigmas, such as the potential insolubility of the alignment problem, urging a posture of intellectual humility and vigilant refinement to harmonize human will with machine might in the infinite. Therefore, the clear imperative is to forge this union and secure a stellar legacy grounded in ethical excellence.

67 Michel Mayor, "Plurality of Worlds in the Cosmos: A Dream of Antiquity, A Modern Reality of Astrophysics," Nobel Lecture, December 8, 2019, https://www.nobelprize.org/uploads/2019/10/mayor-lecture.pdf; Didier Queloz, "51 Pegasi b, and the Exoplanet Revolution," Nobel Lecture, December 8, 2019, https://www.nobelprize.org/uploads/2019/10/queloz-lecture.pdf.

68 Philip Lubin, "A Roadmap to Interstellar Flight," *Journal of the British Interplanetary Society* 69, no. 2 (2016): 40–72, https://arxiv.org/pdf/1604.01356.pdf.

69 Future of Life Institute, "Asilomar AI Principles," 2017, https://futureoflife.org/open-letter/ai-principles/.