ISSN: 2995-326X

Mini Review Article

Political Science International

Orbital intelligence and the knightian Void: A Graduated Governance Model for Autonomous AI in Cosmic Exploration, Grounded in the P.A.L.O. Framework

Fabrizio Degni*

PhD Candidate, European Institute of Management and Technology, Italy

***Corresponding Author**
Fabrizio Degni, PhD Candidate, European Institute of Management and Technology, Italy.

Submitted: 2026, Apr 20; **Accepted:** 2026, May 11; **Published:** 2026, May 22

Citation: Degni, F. (2026). Orbital intelligence and the knightian Void: A Graduated Governance Model for Autonomous AI in Cosmic Exploration, Grounded in the P.A.L.O. Framework. *Politi Sci Int, 4*(1), 01-20.

Abstract

The accelerating deployment of autonomous AI systems in cosmic exploration from the AEGIS-guided Perseverance rover on Mars to the recently launched Europa Clipper bound for Jupiter's icy moonand toward prospective interstellar probes reveals a governance vacuum that terrestrial frameworks were never designed to address. This article identifies that vacuum as a Knightian void: a condition of radical, unquantifiable uncertainty, in which probabilistic risk management breaks down and ethical decision-making must proceed without the computational guarantees of measurable probability. Building on the Principled AI Lifecycle Orchestration (P.A.L.O.) framework a seven-principle, five-phase, 35-KPI governance paradigm aligned with ISO/IEC 42001:2023, ISO/IEC 42005:2025, the EU AI Act (Regulation 2024/1689), the NIST AI Risk Management Framework 1.0 and the OECD AI Principles we propose the Graduated Cosmic Ethics (GCE) model. The GCE model defines four autonomy levels (Supervised, Constrained, Guided, Independent), each calibrated to a communication-latency regime and each grounded in a phase-weighted application of the P.A.L.O. governance architecture. We extend the 35-KPI compendium with seven Cosmic KPIs designed to capture the distinctive epistemic conditions of deep-space operationand we propose a six-layer accountability architecture that distributes responsibility across design, training, authorisation, operational, institutionaland generational horizons. The model is illustrated through four case studies Perseverance, Europa Clipper and a prospective Europa lander, interstellar probe conceptsand emergent multi-agent orbital constellations and harmonised with the COSPAR Planetary Protection Policy, the Outer Space Treaty of 1967 and the Liability Convention of 1972. The paper concludes with practical recommendations for space agencies, international forumsand the research communityand with a defence of the claim that, in the Knightian void of cosmic exploration, the governance default must be a bias toward inaction rather than a presumption of legitimate autonomous initiative.

Keywords: Artificial Intelligence Ethics, Autonomous Space Systems, Knightian Uncertainty, Graduated Cosmic Ethics, P.A.L.O. Framework; Planetary Protection, Meaningful Human Control, Explainable AI, AI Governance, Space Law

1. Introduction

1.1. The New Frontier of Cosmic AI Ethics

Space exploration has always been a discipline of delegation: of human purpose entrusted to metal and silicon, of intention projected across distances that render direct supervision a physical impossibility. What has changed and changed with a rapidity that outpaces the institutions meant to govern it is the character of what is being delegated. Where once we entrusted machines with the execution of instructions, we now increasingly entrust them with the *formation* of intent, with the selection among ethically consequential alternatives in environments where no human instruction is available and no human review is timely. NASA's Perseverance rover already operates with AI-powered hazard detection, autonomous target selection through the

AEGIS subsystemand onboard planning capabilities that permit meaningful decisions to be taken during the four-to-twenty-four-minute communication blackout with Earth [6,7]. Europa Clipper, launched in October 2024 on a journey to Jupiter's ice-covered moon, incorporates radiation-hardened electronics, autonomous trajectory managementand onboard science prioritisation designed precisely for an operational regime in which real-time Earth supervision cannot be assumed [8,9]. A prospective Europa subsurface lander, should one materialise, would face conditions in which real-time communication is not merely delayed but physically impossible, the ice crust itself forming a barrier no radio link can traverse; in such a context, the AI system does not merely *supplement* human decision-making, it *substitutes for* it.

These operational realities are not simply engineering challenges. They raise fundamental questions about moral agency, about the conditions under which responsibility can be coherently ascribed, about the philosophical adequacy of governance frameworks developed for applications hiring algorithms, medical triage systems, credit scoring in which some residual possibility of human intervention could be assumed [10,11]. The gap between the ethical frameworks currently in force and the requirements of cosmic AI deployment is both wide and urgent. As Jobin, Ienca and Vayena documented in their landmark survey of eighty-four AI ethics guidelines, existing frameworks converge remarkably on a set of high-level principles fairness, transparency, accountability, safety, human agency but diverge dramatically on the question of operationalization [12]. In the extraterrestrial domain this operationalisation gap is exponentially compounded, because the very conditions that make autonomous AI necessary (distance, latency, environmental novelty) simultaneously dissolve the mechanisms through which operationalisation has been attempted on Earth.

1.2. Theoretical Position: The Knightian Void at the Heart of Cosmic Autonomy

The theoretical position advanced in this paper is that the central problem of cosmic AI governance is not well-characterised as a problem of risk in the technical sense familiar from actuarial, engineeringand machine-learning literatures. Frank Knight, writing in 1921, drew the enduring distinction between *risk*, in which outcomes are uncertain but their probabilities can be meaningfully quantifiedand *uncertainty* proper, in which neither the probability distribution over outcomes nor, in the limiting case, the space of possible outcomes itself can be specified in advance [1]. A century of subsequent scholarship, including recent work by Sunstein, Townsend and colleagues and Ramoglou and collaborators, has refined this distinction and clarified its bearing on contemporary problems of artificial intelligence, regulatory designand decision-making under radical novelty [2-4]. The Knightian frame illuminates what probability-based frameworks obscure: that when an AI system encounters genuinely novel data an anomalous chemical signature beneath the Europan ice, a signal whose spectral character matches no terrestrial astrophysical expectation, a geological formation whose geomorphology has no training-distribution analogue it is not operating in a high-uncertainty regime that can be handled through more data or better priors; it is operating in a domain in which the assumptions underwriting its probabilistic models have ceased to hold [13,14].

The cosmic environment is, in a strong sense, a Knightian domain. No dataset, however large, can capture the full space of possible future states of an exploratory mission beyond Earth, because the space itself is co-constituted by the action of the mission and the unknown features of the environment it discovers [15]. An AI system that approaches such a domain with the confident probabilistic reasoning appropriate to well-characterised terrestrial contexts facial recognition, say, or route optimisation commits what we might term a *category error in uncertainty*: it treats as risk what is, in its proper structure, uncertaintyand thereby generates high-confidence outputs whose epistemic warrant has already collapsed. The title of this paper names this condition the *Knightian void*and takes the governance of orbital and deep-space intelligence to be, in its most general formulation, the problem of acting well within it. The P.A.L.O. Framework, to which we turn in §2.6, provides the structural apparatus for such action: not by pretending to fill the void, but by constructing the institutional, technicaland procedural scaffolding within which principled action remains possible even when the void cannot be closed [5].

1.3. Research Questions and Contributions

The paper is organised around three central research questions. *First*, how should the balance between AI autonomy and human oversight be calibrated across mission contexts, given the physical constraints of communication delay, environmental unpredictabil-ityand the Knightian character of radical novelty? *Second*, what ethical frameworks theoretically and practically can guide AI deci-sion-making in cosmic scenarios where terrestrial norms and train-ing-distribution experience are, at best, suggestive and, at worst, actively misleading? *Third*, how can the P.A.L.O. Framework's lifecycle-integrated, multi-standard-aligned governance architec-ture be adapted and extended to address the unique challenges of extraterrestrial AI deploymentand what operational instruments does such an extension require?

The paper makes four principal contributions. *Theoretically*, it introduces the Knightian void as a framing concept for cosmic AI governance and argues that this framing clarifies what probability-based and rule-based frameworks obscure. *Framework-wise*, it develops the Graduated Cosmic Ethics (GCE) model as an explicit extension of the P.A.L.O. Framework, mapping four autonomy levels (Supervised, Constrained, Guided, Independent) onto P.A.L.O.'s five-phase lifecycle and identifying the differential governance burden each level imposes. *Empirically*, through comparative case study analysis of current and proposed missions, it demonstrates how each of the six P.A.L.O. governance gaps manifests in cosmic contexts, sometimes with qualitatively transformed character. *Operationally*, it provides a set of annex instruments ethical screening questionnaires, a cosmic RACI template, risk tiering protocols, ethical decision-tree workflowsand a worked KPI dashboard example designed to be directly adoptable by space agencies and international governance bodies.

1.4. Methodology and Article Structure

The research employs an interdisciplinary methodology combining AI ethics, space policy analysis, moral philosophyand governance framework design. Its principal methodological moves are fourfold: (i) a targeted narrative review of post-2020 scholarly literature on AI ethics in extreme environments, meaningful human control, planetary protection governance, explainable AIand Knightian uncertainty, building upon the PRISMA-compliant Systematic Literature Review (143 sources, 2016–2025) that grounds the P.A.L.O. Framework; (ii) a comparative case study analysis of operational missions (Perseverance, Europa Clipper, satellite constellations) and proposed missions (Europa subsurface lander, Breakthrough Starshot-class interstellar probes); (iii) a design-science extension of the P.A.L.O. Framework's lifecycle architecture to the cosmic domain, guided by the constraints identified in the case studies and the literature; and (iv) the construction of operational annex instruments derived from the extended framework [5]. Stakeholder perspectives from NASA, ESA, COSPAR, the International Institute of Space Law (IISL) and the United Nations Committee on the Peaceful Uses of Outer Space (COPUOS) are integrated throughout.

The remainder of the paper proceeds as follows. §2 reviews the relevant literatures AI ethics in extreme environments, meaningful human control, planetary protection and space law, explainable AI, Knightian uncertaintyand the P.A.L.O. Framework. §3 develops the ethical challenges of cosmic AI deployment, including the autonomy–oversight spectrum, real-time decision-making without human input, the inadequacy of extant ethical frameworksand the amplification of the six P.A.L.O. governance gaps in cosmic context. §4 presents the Graduated Cosmic Ethics model in detail. §5 works through five case studies. §6 conducts a comparative governance analysis across the NASA, ESA, IISL, EU AI Actand COSPAR frameworks. §7 discusses stakeholder perspectives and implementation pathways. §8 offers a discussion of theoretical contributions, practical implications, limitationsand open philosophical questions. §9 provides recommendations and future research directions. §10 concludes. Annexes A–E provide the operational instruments.

2. Literature Review

2.1. AI Ethics in Extreme Environments: The Governance Landscape

The scholarly literature on AI ethics in extreme environments has developed along two tracks that intersect only partially. The first track, broadly terrestrial, has produced an increasingly dense body of work on autonomous systems in safety-critical contexts: autonomous driving, lethal autonomous weapons systems, disaster response robotics and medical AI under conditions of high stakes and limited intervention opportunity [16-22]. The second track, explicitly space-focused, has grown more recently and remains comparatively sparse, though it has accelerated markedly since 2020 [23-26]. The two tracks share a common preoccupation with the delegation of consequential decisions to machines under time pressure and oversight constraint, but they diverge on the magnitude of those constraints: the seconds-to-minutes decision windows of terrestrial applications are orders of magnitude tighter than the minutes-to-hours windows of planetary missionsand the environmental conditions of a Martian crater or a Europan ocean floor are qualitatively, not merely quantitatively, different from those of an urban street or an intensive care unit.

This divergence has consequences for the transferability of ethical frameworks. Pagallo, writing in *Philosophy & Technology*, has argued that the legal challenges posed by AI in outer space divide into four classes: those that are not genuinely new but inherit older problems of space law (e.g., privatisation), those that arise from AI's generic features (autonomy, opacity, adaptability) and manifest equally on Earth and in space, those that arise specifically from the interaction of AI with space environmentsand those that arise from the hybridisation of the first three [27]. The third class is the most theoretically distinctiveand it is here that terrestrial frameworks most visibly fail: a framework that relies on real-time human oversight, post-hoc forensic reconstruction, or rapid incident response cannot survive transplantation to a context in which all three are foreclosed by the speed of light. Sapientia's Space Law and Policy Project Group, reporting in *Acta Astronautica* and Li, writing in the *UCLA Law Review* tradition, have independently concluded that the existing *corpus juris spatialis* requires substantive reform to accommodate the autonomy gradient introduced by modern space-based AI systemsand that such reform must be grounded in a clearer conceptual distinction between AI as a tool and AI as an autonomous agent [28,29].

2.2. Meaningful Human Control and the Responsibility Gap

The concept of *meaningful human control* (MHC) entered the literature on autonomous systems through the debate on lethal autonomous weapons, where it was developed as a response to what Matthias and subsequent commentators have termed the *responsibility gap*: the condition in which no moral agent can be coherently held responsible for the outcomes of a sufficiently autonomous system [30-33]. Santoni de Sio and van den Hoven's influential 2018 philosophical account, extended by Mecacci and Santoni de Sio in 2020 and operationalised more recently by Verdiesen, Santoni de Sio and Dignum and by Cavalcante Siebert and colleagues, distinguishes two necessary conditions for meaningful human control: a *tracking* condition, according to which the autonomous system must be responsive to the relevant moral reasons of the humans designing and deploying it and to the relevant facts of its operational environmentand a *tracing* condition, according to which the outcomes of the system's operation must be traceable back to at least one human along the chain of design and deployment [34-37].

Both conditions strain under cosmic deployment. The tracking condition assumes that "relevant facts of the environment" can be specified with sufficient completeness that an AI can be designed to respond to them; in a Knightian domain, this assumption is precisely what fails [38]. The tracing condition assumes a chain of human design and deployment that can be reconstructed retrospectively; but when an AI system has adapted its decision-making over decades of deep-space operation in response to data no human

has seen, the chain thins to the point of disappearance. Amoroso and Tamburrini's compact review identifies three approaches to MHC uniform, differentiatedand prudential and argues that cosmic contexts most naturally call for a differentiated approach in which the stringency of control requirements is calibrated to the gravity of the decision at hand [39]. This differentiated approach is, we shall argue in §4, structurally congruent with the P.A.L.O. Framework's risk tiering methodology and provides one of the theoretical anchors for the Graduated Cosmic Ethics model.

2.3. Planetary Protection, Space Lawand the Limits of International Governance

Planetary protection, in contrast to AI ethics, has a long and institutionally embedded history. The 1967 Outer Space Treaty obliges states to conduct exploration in a manner that avoids harmful contamination of celestial bodies and adverse changes to Earth's environment and assigns to states the responsibility of authorising and continuously supervising national activities in space, including those of non-governmental entities [40,41]. The COSPAR Panel on Planetary Protection, established in the immediate aftermath of the International Geophysical Year, has since 1964 developed and maintained a non-binding but internationally recognised planetary protection policy, restructured most recently in 2024 [42-46]. The 2024 restructuring reaffirmed the policy's substantive content while reorganising it for coherence, adopting a common framework for Icy Worlds categorisation based on lower limits for water activity and temperature and tightening the contamination probability threshold for missions to Europa and Enceladus (Categories III and IV) to less than 1×10^{-4} per mission [43,47].

These protocols demonstrate something important for the argument of this paper: ethical constraints *can* be operationalised in extreme environments through structured protocols, risk tieringand compliance verification precisely the governance architecture that the P.A.L.O. Framework provides for AI systems more generally [5]. What the planetary protection literature does not yet address is the ethical dimension of AI decision-making itself. Existing protocols assume that the decision about whether to drill, sample, or retreat is made by mission designers before launch and implemented by the spacecraft in execution; they do not contemplate a regime in which the spacecraft's AI system *interprets* its sensor data, *classifies* an ambiguous signaland *decides* among competing operational options in real time, all without human review. The extension of planetary protection principles to encompass AI governance represents a significant gap in the current literature a gap that Martin and Freeland and the International Institute of Space Law Working Group on Legal Aspects of AI in Space have begun to address, but which remains, at the time of writing, unresolved in binding international instruments [26,48].

2.4. Explainable AI, Verificationand the Opacity Problem

The "black box" character of contemporary deep-learning systems has generated a substantial body of work on explainable AI (XAI) [49-51]. The field has matured from its early emphasis on post-hoc explanation methods such as LIME and SHAP to a more integrated conception in which explainability, verificationand validation are treated as a continuous lifecycle concern [49,52-55]. The European Space Agency's PINEBERRY project and NASA's EXPLAIND prototype illustrate the application of XAI methodologies to space operationsand Allen's 2025 work on persistent human–machine operations from Earth to Mars provides one of the most developed engineering treatments of XAI in a deep-space context [56-58].

Two observations follow from this literature for the cosmic domain. The first is that XAI techniques impose their own severe computational, energeticand mass costs constraints formally codified in aerospace systems engineering as SWaP (Size, Weightand Power). Because spaceborne computing must rely on radiation-hardened (rad-hard) processors to survive cosmic ray and solar proton degradation, deep-space flight computers typically lag terrestrial commercial off-the-shelf (COTS) computational throughput by a decade or more. Consequently, these governance costs trade directly against the scientific payload and autonomous capability of a spacecraft [59,60]. A mission that devotes resources to running interpretable surrogate models or generating per-decision explanations necessarily devotes less to primary instrument performance; in a resource-scarce environment, this trade-off is not optional but unavoidable. The second is that explanations are useful only insofar as they can be *received and used* by a human interlocutor a condition which, in deep-space contexts, is vitiated by the same communication delays that render real-time oversight impossible. An explanation that arrives at Earth ninety minutes after the decision it explainsand which cannot be revised before the decision's consequences unfold, serves forensic purposes but not operational ones. This asymmetry generates what we may term the *explicability paradox* of cosmic AI: the systems most in need of explanation are also those least able to deliver useful explanations in time.

2.5. Knightian Uncertainty: Radical Unknowability as a Theoretical Lens

The Knightian distinction between risk and uncertainty has enjoyed a pronounced revival in the AI ethics and regulatory literatures since 2020 [2,3,4,13,61]. Townsend, Hunt, Rady, Manocha and Jin's 2025 paper in the *Academy of Management Review* identifies four interrelated problems of Knightian uncertainty actor ignorance, practical indeterminism, agentic noveltyand competitive recursion and argues that the predictive capabilities of AI are bounded in principle by the system's ability to grapple with these problems. Sunstein, writing from a regulatory perspective, argues that Knightian uncertainty poses challenges that cannot be dissolved by assigning subjective probabilities and that decision rules such as the maximin principle encounter serious difficulties when eliminating worst-case scenarios would itself impose high costs or foreclose large benefits. Ramoglou and colleagues have pushed back against both positions, arguing that contemporary epistemology permits knowledge claims about the possible even in the absence of quantifiable probabilities and that AI can, in some contexts, help actors construct "theories of the possible" that meaningfully guide action [2-4].

For cosmic AI governance, the bearing of this literature is substantive rather than merely terminological. If cosmic deployment is a Knightian domain and the arguments of §1.2 suggest that it is then governance frameworks designed on risk-based assumptions will systematically under-specify what they need to do [62,63]. The regulatory architecture that emerges from such governance is prone to what we might call *false-precision failure*: the assignment of probability estimates, confidence intervalsand risk scores to scenarios whose underlying structure admits none of these, with the consequence that the governance apparatus generates an illusion of control precisely where control is most attenuated. A Knightian-aware governance framework, by contrast, relies not on probability estimation but on *principled constraints, structural redundanciesand fail-safe modalities* the institutional and technical scaffolding that permits action under irreducible uncertainty without pretending to dissolve it. This is, not coincidentally, the structural signature of the P.A.L.O. Framework, to which we now turn.

2.6. The P.A.L.O. Framework as Governance Foundation

The Principled AI Lifecycle Orchestration (P.A.L.O.) Framework was developed through a PRISMA-compliant Systematic Literature Review encompassing 143 peer-reviewed academic sources, international standardsand regulatory instruments published between 2016 and 2025 [5]. The SLR identified six critical and persistent gaps in existing AI governance approaches: the *operationalisation gap*, documented in 87 of 143 sources (61%), concerns the chronic difficulty of translating high-level ethical principles into auditable organisational practice; the *lifecycle coverage gap* concerns the tendency of existing frameworks to address selected phases of the AI lifecycle while neglecting others, particularly ideation and decommissioning; the *multi-standard integration gap* concerns the proliferation of partially overlapping, partially inconsistent standards (ISO/IEC 42001:2023, ISO/IEC 42005:2025, the OECD AI Principles, the EU AI Act, the NIST AI Risk Management Framework) and the absence of mechanisms for their systematic harmonization; the *business-ethics synthesis gap* concerns the persistent tendency to frame ethical constraint and operational performance as competing priorities rather than complementary dimensions of responsible AI; *the decommissioning gap* concerns the near-total absence of attention to end-of-life governance; and the *scalability gap* concerns the difficulty of extending governance from pilot systems to portfolio-scale deployments [64-67].

The framework responds to these documented gaps through seven universal ethical principles Principled Fairness and Non-Discrimination; Principled Transparency and Explainability; Principled Accountability and Responsibility; Principled Privacy and Data Governance; Principled Safety and Robustness; Principled Human Agency and Oversight; and Principled Societal and Environmental Well-being operationalised through a five-phase lifecycle architecture (Ideation and Ethical Screening; Comprehensive Assessment and Planning; Responsible Development and Validation; Ethical Deployment and Proactive Monitoring; Continuous Improvement and Responsible Decommissioning), five multifaceted evaluation dimensionsand a thirty-five-KPI compendium spanning technical, businessand ethical metrics. P.A.L.O. is explicitly aligned with ISO/IEC 42001:2023, ISO/IEC 42005:2025, the OECD AI Principles, the EU AI Act (Regulation (EU) 2024/1689) and the NIST AI Risk Management Framework 1.0 [5,67]. A comparative analysis positions P.A.L.O. as a meta-framework that synthesises and operationalises the contributions of ALTAI, IEEE Ethically Aligned Design, the NIST AI RMF, the Responsible AI Institute's Intake Frameworkand the AIEIG VCIO model [5].

The relevance of P.A.L.O. for cosmic AI governance is threefold. First, its treatment of ethical governance as a *lifecycle commitment* from ideation through responsible decommissioning is uniquely well-suited to space missions, whose operational lifecycle may span decades and whose decommissioning decisions carry unique ethical weight (a decommissioned rover does not disappear; it becomes a permanent feature of another world). Second, its multi-standard integration architecture is indispensable for a domain in which missions involve international collaboration across agencies operating under different regulatory regimes NASA under US executive orders and NIST guidance, ESA under EU AI Act obligations, JAXA and CNSA under their respective national frameworks, all bound loosely together by non-binding COSPAR protocols and an aging Outer Space Treaty. Thirdand most importantly for the argument of this paper, P.A.L.O.'s structural signature principled constraints, documented accountability, lifecycle-integrated KPIs and independent review gates is precisely the kind of governance that remains viable in Knightian conditions, because it does not depend on probability estimation to ground its legitimacy.

3. Ethical Challenges in Cosmic AI Deployment

3.1. The Autonomy–Oversight Spectrum

The central ethical tension in cosmic AI deployment concerns the relationship between autonomy and oversighted the way the one displaces the other as distance from Earth increases. Terrestrial AI governance frameworks, including the EU AI Act, rest on the assumption that meaningful human oversight is achievable that a human-in-the-loop, human-on-the-loop, or human-in-command model can be implemented for high-risk AI systemsand that where such oversight is not achievable the system should not be deployed [67,68]. In space, these assumptions collapse progressively with distanceand they collapse in a manner that is neither linear nor accidental: the collapse is structural, dictated by the finite speed of light and by the environmental conditions that make real-time communication physically impossible in certain contexts.

At Earth–Moon distances, where one-way communication delay is approximately 1.3 seconds, real-time human oversight remains feasible for all but the most time-critical decisions; a human-in-the-loop model is operationally viable. At Mars distances, where communication delay varies between approximately four and twenty-four minutes depending on orbital geometry, meaningful oversight requires a transition from real-time human-in-the-loop to asynchronous human-on-the-loop, in which the AI system acts autonomously within pre-defined parameters and reports its

actions for retrospective review [6,58]. At outer planet distances Jupiter at thirty-five to fifty-two minutes, Saturn at one hundred and five to one hundred and fifty-five minutes, the outer ice giants at four to five hours or more even asynchronous oversight becomes impractical for time-sensitive decisions, requiring what we shall term *guided autonomy*: the AI system operates within a broad ethical framework but makes contextual judgements that no human can review in time to influence the outcome. At interstellar distances, where communication delays are measured in years, even guided autonomy gives way to something more radical: *independent autonomy*, in which the entire ethical architecture of the mission must be internal to the spacecraft, pre-encoded before launchand capable of sustaining itself through encounters and adaptations that no human will learn of until after they have already occurred.

The P.A.L.O. Framework's Principled Human Agency and Oversight principle requires that AI systems "augment and support human capabilities while preserving human autonomy and ensuring meaningful human oversight" [5]. In the cosmic domain this principle must be reinterpreted, because meaningful oversight in the relevant sense cannot require real-time intervention: the speed of light does not negotiate. What it can requireand what the Graduated Cosmic Ethics model specifies in §4, *is pre-mission ethical programming of sufficient depth and contextual sensitivity to substitute for the real-time judgement that distance renders impossible*. Such reinterpretation has profound implications for how P.A.L.O.'s Phase 2 (Comprehensive Assessment and Planning) must be operationalised for space missions: every scenario anticipated during Phase 2, every decision priority encoded before launch, is in effect a displaced exercise of human moral agency, projected forward into a situation the human will not witness. The tracing condition of meaningful human control survives this projection the design chain remains but the tracking condition shifts, becoming a matter of how well the pre-mission encoding anticipates the contextual demands that will arise in operation [34].

3.2. Real-Time Decision-Making Without Human Input

The most ethically consequential scenarios in cosmic AI deployment involve irreversible decisions made in real time without the possibility of human consultation. Consider a scenario plausible for a prospective Europa subsurface lander: operating beneath the Europan ice crust, the AI system detects anomalous chemical signatures consistent with biological activity. Simultaneously, a structural integrity alert indicates that the drilling apparatus is approaching imminent failureand a propulsion reading suggests that battery reserves will be insufficient for both continued sampling and the controlled ascent required to return the spacecraft to a position from which it can transmit its findings. The AI must choose among three actions: continue collecting biological data (potentially confirming the most significant scientific discovery in human history); initiate emergency ascent to preserve mission hardware and transmit preliminary findings (potentially sacrificing an irrecoverable scientific opportunity); or attempt a partial middle course whose outcomes it cannot, in the circumstances, predict.

No terrestrial AI governance framework provides adequate guidance for this decisionand not merely because the probabilities are uncertain. The scenario is Knightian in the strong sense: the space of possible outcomes is co-constituted by the unknown properties of the Europan environment and by the AI's own action. The AI cannot compute expected utility across outcomes it cannot boundand it cannot appeal to a precedent because none exists. What it can do and what the P.A.L.O. Framework's risk tiering methodology, extended via the GCE model, makes possible is apply a pre-mission ethical encoding that ranks competing values (scientific discovery, equipment preservation, biosphere protection), specifies the conditions under which each takes priorityand marks certain actions as categorically foreclosed even under strong instrumental pressure. The encoding does not pretend to solve the ethical problem; it stabilises the decision against the chaos of unbounded uncertainty, permitting the AI to act in a manner that can be subsequently assessed, reviewedand where appropriate criticised, correctedand integrated into the pre-mission ethical encoding of future missions.

This is, in a sense, all that governance can offer in the Knightian void. It is also not nothing. A decision made in accordance with a pre-specified, documented, independently reviewed ethical encoding is qualitatively different from a decision made on an ad hoc basis by a system whose value structure is opaque even to its designers. The former is, in the strict Santoni de Sio and van den Hoven sense, under meaningful human control; the latter is not [34]. The distinction is not merely academic: it bears on whether any human agent can coherently be held responsible for the mission's outcomesand on whether the public legitimacy that sustains cosmic exploration can survive the kinds of incidents that autonomous systems will eventually produce.

3.3. Ethical Frameworks for Uncharted Territories

Traditional ethical traditions provide partial but individually insufficient guidance for cosmic AI governance. *Consequentialist* approaches, which evaluate actions by their outcomes, face the fundamental problem of Knightian novelty: the consequences of AI decisions in extraterrestrial environments may be unknowable at the time of decisionand an ethical framework that requires outcome prediction fails where prediction fails. *Deontological* approaches, which evaluate actions by their conformity to moral duties, face the challenge of specifying duties that apply in environments where no moral precedent exists and where the relevant duty-bearing relations (to whom does the AI owe what?) are themselves unclear. *Virtue ethics*, which focuses on the character dispositions of moral agents, raises the vexing question of whether AI systems can meaningfully possess virtuesand whether the organisational cultures of responsibility that sustain virtue in human agents can be encoded in systems operating billions of kilometres from any human institution [69]. *Care ethics*, with its attention to relational obligations and the particularities of context, offers an attractive complement to more abstract traditions but struggles with the asymmetry of cosmic deployment, where the "cared-for" may be potential extraterrestrial life whose interests are themselves epistemically inaccessible.

The P.A.L.O. Framework's pluralistic ethical foundation, which draws on all four of these traditions, provides a richer normative architecture than any single tradition alone [5]. In the cosmic domain this pluralism is not merely intellectually enriching but practically essential: a framework that relies exclusively on consequentialist reasoning will fail when consequences are unknowable; a framework that relies exclusively on deontological rules will fail when unprecedented situations render pre-specified rules inapplicable; a framework that relies exclusively on virtue-ethical dispositions will fail when those dispositions cannot be operationalised in algorithmic form; a framework that relies exclusively on care-ethical relations will fail when the relevant relations cannot be identified. The pluralism of P.A.L.O. is, in a sense, an epistemic insurance policy, ensuring that when one ethical tradition's grip on a situation slips, another is positioned to hold. What the Graduated Cosmic Ethics model adds and what §4 develops in detail is a structured procedure for activating different ethical traditions at different lifecycle phases and under different autonomy regimes, rather than leaving the selection to the contingency of a mission's engineering culture.

3.4. The Six Governance Gaps Amplified in Space

Each of the six governance gaps documented by the P.A.L.O. Framework's Systematic Literature Review manifests with amplified severity in the cosmic domainand some manifest with qualitatively transformed character. The *operationalisation gap* is compounded by the impossibility of real-time ethical consultation: an operationalisation strategy that assumes a responsible human can be reached for high-stakes decisions is simply unavailable. The *lifecycle coverage gap* is exacerbated by mission durations spanning decades (Voyager 1 has now operated for over forty-eight years) and by the reality that, unlike terrestrial deployments, a space mission's ethical architecture cannot be incrementally updated through easy software releases once the spacecraft has departed cis-lunar space. The *multi-standard integration gap* is intensified by the multinational character of space missions, where an AI system may be developed under one regulatory regime, launched under another, operated from a thirdand affect assets belonging to a fourth a jurisdictional complexity that the Outer Space Treaty's "launching State" concept does not comfortably accommodate [28,29,70].

The *business-ethics synthesis gap* reappears in cosmic context as a tension between scientific ambition and ethical caution that is more acute than its terrestrial analogue. A Europa mission that observes the planetary protection threshold of 1×10^{-4} contamination probability may forgo scientifically compelling investigations that a less constrained mission could have conducted; the ethical constraint is not abstract but directly operational [43]. The *decommissioning gap* acquires qualitatively unique urgency when AI systems become permanent features of other worlds: a rover that fails and is abandoned on Mars does not merely cease to function, it joins the permanent material inventory of another planetand the question of how it should fail what final configuration it should assume, what signals it should emit to future explorers, what data it should retain for potential recovery becomes an ethical question in its own right, one that existing frameworks almost entirely ignore [5]. The *scalability gap* emerges as future missions contemplate fleets of autonomous probes operating across multiple celestial bodies, such as NASA's Starling mission of self-organising autonomous satellites, in which the governance challenge is not the ethics of a single system but the emergent ethics of distributed, multi-agent autonomous ensembles [71,72].

3.5. The Accountability Paradox in Autonomous Learning Systems

A final ethical challenge deserves particular attention, because it crystallises the tensions developed above into a single paradoxical structure. If an AI system deployed on a long-duration mission can *learn and adapt* its decision-making in response to novel environmental data a capability often desired for resilience and performance in environments whose features are not fully anticipated in training then the system's operational logic and decision pathways may evolve over time in ways that diverge from its initial programming and that were not foreseen by its designers [1,3,15,38]. If harm arises from such emergent behaviour, tracing the harm back to a specific human design decision or operational choice becomes exceedingly difficult: the AI is doing what its learning rule specifies it should do, given the data it has encountered, but no human specified the resultant behaviour and no human anticipated it. We may term this condition the *accountability paradox of autonomous learning*: the very adaptability that justifies the deployment of learning systems in deep space is also what most severely strains the tracing condition of meaningful human control [34-36].

The paradox is not fully resolvable. What the Graduated Cosmic Ethics model offers, drawing on P.A.L.O.'s lifecycle-integrated governance architecture, is a structured approach to its mitigation: pre-specified learning constraints that exclude certain adaptation trajectories as categorically impermissible; verification gates at which learned modifications must satisfy pre-established safety and ethics criteria before being integrated into operational policy; and a layered accountability model (§4.5) that distributes responsibility across design, operationand international governance rather than attempting the futile task of pinning every autonomous decision on a single human agent [5]. None of these mechanisms dissolves the paradox, but they reduce its practical bite, preserving the space within which responsible deployment remains defensible.

4. The Graduated Cosmic Ethics (GCE) Model

4.1. Theoretical Grounding: From Terrestrial to Cosmic Governance

The Graduated Cosmic Ethics model is an extension of the P.A.L.O. Framework designed to preserve its core architectural commitments lifecycle integration, multi-standard alignment, principled ethical governance, operationalised through measurable KPIs while adapting its operational mechanisms to the physical constraints of cosmic deployment [5]. The model rests on three theoretical anchors: the Knightian framing of radical novelty (§1.2, §2.5), the differentiated approach to meaningful human control (§2.2) and the pluralistic ethical foundation of P.A.L.O. (§2.6,

§3.3). It is designed to answer a single organising question: *given that the mechanisms of real-time human oversight progressively disappear as a mission moves outward from Earth, how should the governance burden be redistributed across the mission lifecycle and across institutional actors such that principled ethical action remains achievable?*

The model's answer is *graduation*: as autonomy increases, the governance burden shifts from real-time oversight to pre-mission encoding, from operational monitoring to lifecycle-integrated verification, from individual human agency to collective institutional accountability and from rule-based constraint to value-aligned design. The graduation is not arbitrary but calibrated to the specific constraints imposed by each autonomy regime and it is anchored throughout in the P.A.L.O. Framework's five-phase lifecycle. Figure 1 presents the overall architecture.

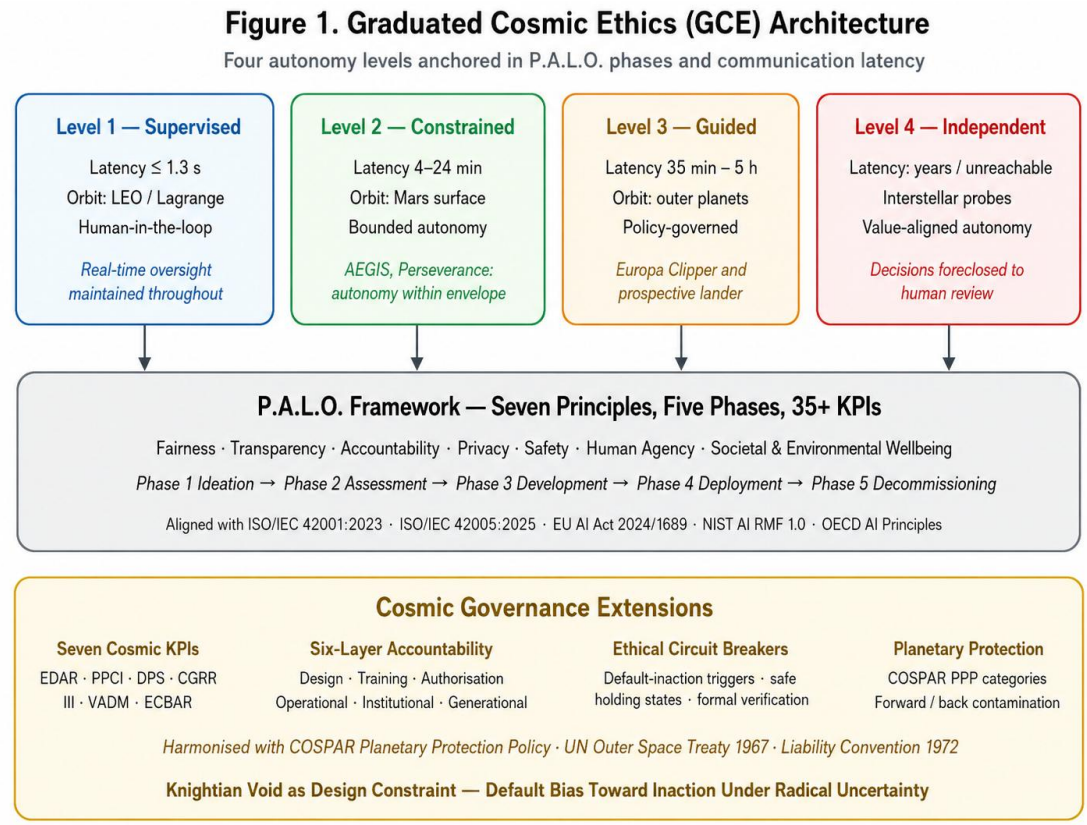

Figure 1: The Graduated Cosmic Ethics (GCE) Architecture. Four Autonomy Levels Supervised, Constrained, Guided, Independent are Anchored to Distinct Communication-Latency Regimes and Share a Common P.A.L.O. Substrate Extended with Cosmic-Specific Governance Mechanisms. The Knightian void Operates as a Design Constraint: the Further the Mission from Earth, the More Heavily the Default Biases Toward Inaction under Radical Uncertainty

4.2. Context-Adaptive Autonomy Levels

The GCE model defines four autonomy levels, each with distinct governance requirements mapped onto P.A.L.O.'s five phases and onto the meaningful human control taxonomy of §2.2.

Autonomy Level	Communication Latency	Representative Missions	P.A.L.O. Phase-Weighted Governance	Oversight Mode
L1 Supervised	≤ 1.3 s (LEO), up to ~10 s (Lagrange)	ISS, Starling prototyping, LEO constellations	Phase 4 weighting 1×; standard ISO/IEC 42005 impact assessment	Real-time human-in-the-loop
L2 Constrained	4–24 min (Mars, inner-system surface)	Perseverance + AEGIS, Curiosity, Mars sample-caching	Phase 1 weighting 1.5×; envelope spec + PP categorisation at Ideation	Bounded autonomy within pre-authorised envelope

L3 Guided	35 min – 5 h (outer planets, icy moons)	Europa Clipper, prospective Europa lander, Dragonfly	Phase 1 weighting 2.5×; formal verification of circuit breakers; Tier 2	Policy-governed with deferred consultation windows
L4 Independent	Years / unreachable (interstellar)	Breakthrough Starshot concepts, Voyager-class generational missions	Phase 1 weighting 4×; Tier 1; unanimous EAB / PP / agency / COSPAR approval	Value-aligned autonomy; decisions foreclosed to real-time review

Table 1: The Four Autonomy Levels of the GCE Model, Showing for Each Level its Communication-Latency Regime, Representative Missions, the P.A.L.O. Phase-Weighted Governance Load and the Oversight Mode Under Which the System Operates.

At the *Supervised level*, the P.A.L.O. lifecycle can be implemented substantially as designed for terrestrial high-risk AI systems, with the minor adaptation that physical access to the deployed system for audit or correction is indirect. At the *Constrained* level the regime in which current Mars missions operate the governance weight shifts measurably toward Phase 2, since any ethically consequential decision taken during a communication blackout must be implicit in the mission's pre-approved operational envelope. At the *Guided* level, the governance weight becomes overwhelmingly pre-mission: Phase 2 must anticipate an extraordinarily wide range of scenariosand Phase 3 (Responsible Development and Validation) must subject the resulting decision logic to formal verification of the kind that aerospace software engineering has developed for safety-critical applications, now extended to the ethical dimension [73,74]. At the *Independent* level, the entire governance burden shifts to pre-launch preparationand the AI system must carry its ethical framework not as an externally supervised process but as an internalised capability, capable of contextual adaptation to encounters that no human has anticipated a requirement that pushes the boundaries of current AI ethics research from rule-following to genuine machine moral reasoning [75,76].

4.3. The Lifecycle Integration Architecture

The GCE model preserves P.A.L.O.'s five-phase lifecycle structure while adapting the weighting and content of each phase to the autonomy level. Figure 2 (see Annex E for a worked ethical decision-tree) depicts the conceptual architecture.

Phase 1: Ideation and Ethical Screening

In the cosmic domain, ethical screening must be mandatory for all AI-enabled mission proposals and must occur *before* mission selection. The Ethical Screening Questionnaire (Annex A) extends P.A.L.O.'s standard Phase 1 instrument with cosmic-specific questions concerning planetary protection implications, communication-delay regime, reversibility of foreseeable decision sand interaction with existing or planned missions from other space agencies.

Phase 2: Comprehensive Assessment and Planning

This phase bears the greatest relative weight at higher autonomy levels. It must include: a formal risk tiering (Annex C) that distinguishes decisions on the basis of reversibility, scientific consequence, contamination implications and human-safety implications; a scenario anticipation exercise calibrated to the autonomy level (for *Independent* missions, the exercise must be exhaustive); a value specification that ranks competing ethical considerations and specifies conditions for their relative priority; and an independent ethics review by reviewers with no financial or reputational stake in the mission's selection.

Phase 3: Responsible Development and Validation

This phase incorporates the formal verification of ethical decision logic, testing under simulated novelty conditions, red-teaming for adversarial scenariosand critically validation against the full population of Phase 2 anticipated scenarios. For *Guided* and *Independent* missions, this phase must include adversarial ethics testing: attempts by independent reviewers to construct scenarios in which the AI system's pre-encoded logic generates outcomes that violate the mission's ethical commitments, with failure of such attempts (within reasonable computational bounds) serving as a precondition for mission approval.

Phase 4: Ethical Deployment and Proactive Monitoring

At *Supervised* and *Constrained* autonomy levels, Phase 4 retains its operational character, with real-time or near-real-time monitoring through telemetry review. At *Guided* and *Independent* levels, Phase 4 becomes forensic rather than operational: its function is to analyse the decisions the AI system has already made, to compare them against the pre-mission ethical encoding, to identify cases of divergenceand to integrate lessons into the design of future missions. This transformation is itself a governance innovation: it acknowledges that operational monitoring is not the only form monitoring can takeand that lifecycle-integrated governance demands forensic learning even where operational intervention is foreclosed.

Phase 5: Continuous Improvement and Responsible Decommissioning

In the cosmic domain, Phase 5 carries unique weight. Decommissioning decisions must be made at Phase 2 (pre-launch), because in most cases no human will be able to make them in execution. The decommissioning strategy must specify the physical, informational and symbolic disposition of the spacecraft at end-of-mission: where it will come to rest, what data it will retain, what signals it will emit for potential future recovery, how its software will be archived for forensic access and for Category IV and V missions how it will satisfy COSPAR planetary protection requirements for end-of-mission disposition [43,46].

Figure 2. Lifecycle Integration: Autonomy Levels Across P.A.L.O. Phases

Phase-weighted governance intensity increases with autonomy level.

	Phase 1 Ideation	Phase 2 Assessment	Phase 3 Development	Phase 4 Deployment	Phase 5 Decommission
Level 1 Supervised	Standard ethical screening *(weighting 1×)*	Routine impact assessment *(ISONEC 42005)*	Standard V&V protocols	Real-time human oversight	Standard orbital disposal
Level 2 Constrained	Enhanced screening + PP category *(weighting 1.5×)*	Extended risk tiering + envelope spec *(Tier 3)*	Rad-hardened ML + envelope verification	Async oversight + log review cadence	Surface disposal with PP safeguards
Level 3 Guided	Deep ethics review + stakeholder panels + PP sub-surface *(weighting 2.5×)*	Tier 2 assessment + adversarial testing + formal methods	Formal verification of safety properties + circuit breakers	Policy-governed autonomy + deferred consultation windows	Controlled impact or sample return quarantine
Level 4 Independent	Generational ethics review + COSPAR joint approval *(weighting 4×)*	Tier 1 (catastrophic irreversibility) + unanimous approval required	Value alignment verification + multiple redundant safety pathways	Value-aligned autonomy with first-contact and PP escalation protocols	Terminal accountability + archival stewardship

Governance intensity scales non-linearly with autonomy level

Phase 1 weighting increases from 1× (Supervised) to 4× (Independent), reflecting the progressively foreclosed opportunities for downstream correction as communication latency grows toward the knightian horzion.

Figure 2: Lifecycle Integration Matrix. Governance Intensity Specified as a Phase-Weighted Coefficient Applied Across the Five P.A.L.O. Phases Increases Non-Linearly from Level 1 (Supervised, 1×) to Level 4 (Independent, 4×), Reflecting the Progressive Foreclosure of Opportunities for Downstream Correction as Communication Latency Grows Toward the Knightian Horizon

4.4. Cosmic KPIs: Extending the P.A.L.O. Compendium

The P.A.L.O. thirty-five-KPI compendium provides the methodological template for operationalizing cosmic ethics [5]. The GCE model proposes extending the compendium with mission-specific KPIs adapted to the unique challenges of space exploration. Table 2 presents a condensed extract; Annex D provides a worked dashboard example.

Acronym	Name	What it measures	Target (Europa lander illustrative)
EDAR	Ethical Decision Auditability Ratio	Fraction of autonomous decisions with stored, retrievable explanation traces	> 0.99 (ethically significant); > 0.95 (routine)
PPCI	Planetary Protection Compliance Index	Composite of bioload adherence, forward- and back-contamination risk management	= 1.0 (no breach of any threshold)
DPS	Decision Provenance Score	Traceability of training data, model updatesand parameter adjustments to named humans	> 0.98 (tamper-evident ledger)
CGRR	Communication Gap Resilience Rate	Fraction of communication blackouts during which behaviour remained within safe envelope	= 1.0 (all gaps, full mission)
III	Institutional Immunity Index	Probability that the mission's ethical architecture survives personnel and regime turnover	> 0.90 over mission lifetime
VADM	Value Alignment Drift Metric	Drift of effective behaviour from Phase 1 value specification (held-out ethical test battery)	< 0.05 cumulative; review at 0.02
ECBAR	Ethical Circuit Breaker Activation Rate	Frequency of safety-driven suspensions of operation	< 0.001 per decision cycle; post-hoc review per activation

Table 2: The Seven Cosmic KPIs Extending the P.A.L.O. 35-KPI Compendium, with Definition and an Illustrative Target Benchmark Drawn from a Prospective Europa Lander Mission

Each cosmic KPI is designed to be measurable, auditable and interpretable against pre-mission baselines. Together they transform aspirational cosmic ethics into a tractable governance regime, much as the P.A.L.O. KPI compendium transforms aspirational AI ethics into measurable organizational commitments.

4.5. Accountability in the Cosmic Void

The P.A.L.O. Principled Accountability and Responsibility tenet requires "clear lines of human responsibility and accountability for the development, deployment and outcomes of AI systems throughout their entire lifecycle" [5]. In the cosmic domain, real-time attribution of responsibility is impossible during autonomous operations and the traditional model of operator-as-locus-of-accountability fails to survive the conditions of Guided and Independent autonomy. We propose a layered accountability model drawing on the Distributed Moral Responsibility (DMR) tradition, mapped onto the P.A.L.O. RACI matrix and extended with a temporal dimension [5,34,77,78].

Under the model, *mission designers* bear primary accountability for the ethical architecture of the AI system, corresponding to Phase 2 responsibility; their accountability is for the quality of the pre-mission ethical encoding, the scope of scenario anticipationand the validity of formal verification. *Mission operators* bear accountability for monitoring and intervention when communication permits, corresponding to Phase 4 responsibility; their accountability is for the timeliness and appropriateness of response to the portion of operation they can observe. *Independent review bodies* bear accountability for the rigor of Phase 2 and Phase 3 review, corresponding to a governance function that has no simple human counterpart in most current mission structures and whose institutionalization is, we shall argue in §7, a priority for reform. *The international space governance community* bears collective accountability for establishing the normative frameworks within which cosmic AI operates, a responsibility that has been partially assumed by COSPAR for planetary protection and by the IISL Working Group on Legal Aspects of AI in Space for AI specifically but that remains institutionally under-developed [45,46,48].

The temporal dimension adds a further distinction: *pre-authorisation accountability* applies to decisions for which human approval must be secured in advance (e.g., any decision that could affect planetary protection categorisation); *real-time accountability* applies to decisions taken under operator oversight (Supervised and selected Constrained contexts); *retrospective accountability* applies to decisions taken autonomously but subject to post-hoc review and correction (most Guided-level decisions); and *terminal accountability* applies to decisions whose consequences are fixed at the moment of execution and cannot be revised (most Independent-level decisions, decommissioning decision sand irreversible Guided-level decisions such as drilling into a special region). Table 3 summarizes the layered accountability model.

Accountability Layer	Who bears responsibility	For what	Temporal window
Design	System architects, model designers, safety engineers	Architectural decisions, policy envelope definition, circuit breaker calibration	Pre-authorisation, traceable indefinitely
Training	Data stewards, training-pipeline operators, mission scientists	Dataset provenance, labelling integrity, bias characterisation, model-update records	Pre-authorisation, ledger-preserved
Authorisation	Ethics Advisory Board, Planetary Protection Officer, national agency, COSPAR (Tier 1)	Approval of autonomy level assignment, risk tiering, deployment go/no-go	Pre-authorisation with documented rationale
Operational	Flight operations lead, mission control team	Real-time or near-real-time decisions within the authorised envelope	Throughout mission
Institutional	Sponsoring agency or consortium, succession-designated office	Long-term stewardship, institutional memory, regime-change resilience	Mission lifetime + archival horizon
Generational	International forums (COSPAR, COPUOS), designated archival bodies	Decisions affecting future explorers, biospheres, or foreclosed scientific options	Inter-generational, terminal accountability

Table 3: The Six-Layer Accountability Architecture, Distributing Responsibility Across Design, Training, Authorisation, Operational, Institutional and Generational Horizons, with Temporal Windows Over Which Each Layer Remains Answerable

The RACI matrix template of Annex B provides the operational instrument for specifying these accountability relationships at the level of individual missions.

4.6. Ethical Circuit Breakers and Fail-Safe Architectures

A final architectural feature of the GCE model is the mandatory inclusion of *ethical circuit breakers* at Guided and Independent autonomy levels [55,75]. An ethical circuit breaker is a pre-specified condition under which an AI system autonomously transitions from normal operation to a fail-safe mode, suspends its current decision logic and where communication permits reports the triggering condition for human review. The triggers are designed to activate in precisely those conditions where Knightian uncertainty most severely compromises probabilistic reasoning: when sensor data falls sufficiently outside training distribution, when the confidence-calibration of the AI's decision logic becomes

unreliable, when the anticipated outcome space fails to bound the observed situation, or when the AI's internal consistency checks (a form of limited metacognition) flag contradictions among its own conclusions.

The fail-safe mode is not a default: it is a designed mode, specified during Phase 2, validated during Phase 3 and instrumented for activation and de-activation under documented conditions. Its purpose is not to avoid decision but to *replace a high-risk decision with a lower-risk one* typically a continuation of the immediately prior state, a data-collection holding pattern, or a slow retreat to a safer configuration while preserving the option for future human review and revision. Ethical circuit breakers do not solve the Knightian void; they acknowledge it and provide a structured response to it, substituting a principled suspension of autonomous action for the alternative of unconstrained action under epistemic conditions that do not warrant it.

5. Case Studies and Applications

5.1. The Constrained Autonomy Paradigm: Perseverance and AEGIS

NASA's Perseverance rover, operating on Mars since February 2021, provides the most developed real-world exemplar of Constrained-level autonomy. The rover's AEGIS (Autonomous Exploration for Gathering Increased Science) system autonomously selects geological targets for the SuperCam instrument during communication blackouts, using onboard image analysis to identify features matching pre-specified scientific criteria [6,7,79]. Evaluated through the P.A.L.O. Framework's seven principles, AEGIS demonstrates robust operationalization of several: its decision logic is transparent and documentable (Transparency), its geological sampling protocols include contamination safeguards (Safety)and its autonomous target selection is subject to retrospective human review through telemetry analysis (Accountability). A carefully reasoned Phase 2 assessment of the system's ethical encoding, conducted before flight, would have satisfied many of the requirements of the GCE Constrained-level governance specification.

However, P.A.L.O.'s Principled Fairness and Non-Discrimination principle reveals a governance consideration that has received less attention in the engineering literature. AEGIS's target selection algorithms may encode systematic biases toward certain geological features at the expense of others a consequence of the training data distribution and the criteria by which "scientifically interesting" was operationalized during pre-mission development [60,80]. A feature that does not match the training distribution, even if it is in principle scientifically significant, may be systematically under-selected. In terrestrial AI contexts this would be termed algorithmic bias; in scientific exploration it is a form of confirmation bias built into the instrument. The implication is not that AEGIS is ethically problematic the system operates well within the requirements of a Constrained-level mission but that even at this comparatively low autonomy level, P.A.L.O.-style fairness analysis surfaces considerations that engineering review alone may miss. The Cosmic KPI of Value Alignment Drift (Table 2) would provide an operational instrument for tracking such considerations across the mission lifecycle.

5.2. The Guided Autonomy Challenge: Europa Clipper and a Prospective Europa Lander

The Europa Clipper mission, launched on 14 October 2024 and expected to arrive in the Jupiter system in April 2030, represents the current frontier of Guided-level autonomy [8,9]. The spacecraft's operational regime approximately forty-eight close flybys of Europa during a three-and-a-half-year investigation campaign involves communication delays of thirty-five to fifty-two minutes one-way and operational windows during which real-time response from Earth is impossible. The mission incorporates radiation-hardened electronics, autonomous trajectory management and onboard science prioritization its planetary protection categorization (Category III, with a contamination probability threshold of 1×10^{-4}) binds its operational envelope [43,45,46].

A prospective Europa subsurface lander one of several mission concepts under study would push the autonomy requirements substantially further [81]. Operating beneath the ice crust would foreclose direct radio communication entirely; the lander would need to either cache data for later transmission through a surface relay or make its key operational decisions without any possibility of human consultation. Applying the GCE framework to such a mission reveals the extraordinary depth of pre-mission ethical planning required. Phase 2 assessment would need to enumerate and tier every foreseeable scenario from routine drilling operations to discovery of potential biosignatures, from sensor anomalies to structural failures, from thermal excursions to communication-relay failures and assign decision protocols to each. The Phase 5 decommissioning dimension acquires particular weight: how should the AI system manage its own end-of-life in an environment where its physical remains become permanent features of another world and where the contamination of the subsurface ocean carries consequences no human can bound? The decommissioning strategy must specify the lander's terminal state in physical, informational and symbolic terms and must satisfy COSPAR Category IV contamination requirements under all failure modes.

5.3. The Independent Autonomy Horizon: Interstellar Probe Concepts

Proposed interstellar missions, such as Breakthrough Starshot's light-sail probe concept for Alpha Centauri or more speculative Kuiper Belt and Oort Cloud missions, represent the ultimate governance challenge [82]. With communication delays measured in years approximately 4.2 years one-way to Alpha Centauri at the speed of light these probes must operate with complete ethical independence. The GCE model's independent level, grounded in P.A.L.O.'s lifecycle architecture, requires that the entire governance burden be resolved before launch. The AI system must carry a comprehensive ethical framework capable of contextual adaptation to encounters that no human has anticipated and that no human will review until years or decades after they have already occurred.

At this level of autonomy, the P.A.L.O. Framework's requirements push against the boundaries of current AI capability. Pluralistic ethical reasoning of the kind P.A.L.O. specifies drawing on consequentialist, deontological, virtue-ethicaland care-ethical considerations as conditions warrant is difficult to operationalise in machine-executable form [75,76,83]. What the GCE model recommends for Independent-level missions, recognising the current state of the art, is a two-stage architecture: a first-stage pre-encoded *value hierarchy* that specifies relative priority among broad categories of concern (scientific integrity, planetary protection, equipment preservation, mission longevity) and a second-stage *ethical reasoning module* that applies this hierarchy to specific situations through a combination of rule-based inference, analogical reasoning from pre-mission scenario analogues and metacognitive self-assessment of decision confidence. The second-stage module is calibrated to trigger an ethical circuit breaker (§4.6) whenever its confidence in the adequacy of its own reasoning falls below a pre-specified threshold, defaulting in such cases to continuation of the immediately prior state. The architecture does not pretend to solve the problem of machine moral reasoning; it constructs the scaffolding within which the limitations of current capability can be managed without catastrophic consequence.

5.4. Orbital Constellations and Emergent Autonomy: Starling and the Multi-Agent Horizon

A fourth case study, often under-attended in the cosmic AI ethics literature, concerns the ethics of autonomous *ensembles* rather than individual systems. NASA's Starling mission, launched in July 2023, demonstrates the capability for small satellite swarms to self-organise, share tasksand reconfigure based on changing mission priorities through distributed planning algorithms and reinforcement learning [71,72]. Commercial mega-constellations of thousands of satellites, such as Amazon's Project Kuiper and SpaceX's Starlink, raise analogous questions at larger scale, particularly where their constituent elements incorporate autonomous collision-avoidance, scheduling and data-prioritisation logic [24,26].

From a P.A.L.O. Framework perspective, ensemble autonomy generates governance challenges that single-agent frameworks do not fully address. The scalability gap manifests here with force: a governance approach that works for one autonomous satellite does not straightforwardly generalise to a constellation of thousands [5]. The emergent behaviour of an ensemble the patterns that appear at the collective level without being explicitly encoded in any individual agent raises questions of distributed responsibility that the layered accountability model of §4.5 only partially resolves. The GCE model's extension to multi-agent contexts requires an additional architectural feature not fully developed in the current paper but flagged here as a priority for future work: *ensemble-level ethical invariants* that bound the collective behaviour of a constellation independently of its constituent agents' individual decision logic.

6. Comparative Governance Analysis

6.1. NASA, ESA, IISL: Three Approaches to Cosmic AI Governance

The three principal institutional actors currently developing AI governance guidelines for space NASA, ESA and the International Institute of Space Law through its Working Group on Legal Aspects of AI in Space have converged on broadly compatible high-level principles while diverging on operational implementation. NASA's Framework for the Ethical Use of Artificial Intelligence articulates six principles human-centered, transparent, explainable, responsible, accountable, secure and safe and provides practical considerations for AI practitioners across the agency's terrestrial and space operations [23,84]. The framework foregrounds its own limitations: it focuses on concrete considerations for a five-to-ten-year horizon while acknowledging that human-level or beyond-human-level AI would require more fundamental ethical reconceptualization [23]. ESA's AI policy emphasizes *originator responsibility* the principle that humans retain final authority and accountability for AI-generated outputs, that AI cannot be an originator and explicitly incorporates review processes for AI incidents, XAI commitments through the PINEBERRY project and alignment with the EU AI Act's regulatory architecture [56,67,68]. The IISL Working Group's January 2025 report, "Balancing Innovation and Responsibility: International Recommendations for AI Regulation in Space", addresses existing legal frameworks (Outer Space Treaty, Liability Convention), regulatory aspects, ethical concerns (including meaningful human control and explainable AI) and proposes governance structures inspired by international bodies such as ICAO and IMO [48].

The P.A.L.O. Framework's multi-standard integration architecture provides a systematic mechanism for reconciling these approaches [5]. Each agency's framework can be mapped onto P.A.L.O.'s seven principles (with minor terminological adjustments), each agency's implementation specifics can be absorbed into the five-phase lifecycle (with adaptations specified by the GCE model) and each agency's accountability structure can be expressed in the RACI matrix template. What the comparative analysis reveals is not that the three approaches are in conflict but that they are partial: NASA's framework is strong on practical considerations, weak on lifecycle integration; ESA's policy is strong on originator responsibility and XAI, weaker on decommissioning; the IISL recommendations are strong on international governance architecture, weaker on operational KPI specification. The GCE model, grounded in P.A.L.O., provides an integration layer that none of the three approaches individually supplies.

6.2. The EU AI Act and Its Cosmic Applicability

The EU AI Act (Regulation (EU) 2024/1689), which entered into force on 1 August 2024 with full application beginning 2 August 2026, introduces a risk-based regulatory architecture with four tiers unacceptable, high, limited and minimal and imposes substantial obligations on providers and deployers of high-risk AI systems [67,68,85,86]. The Act's definition of an AI system "a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment

and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" covers, in principle, most of the AI systems contemplated in this paper [85]. High-risk classification under Article 6 imposes obligations concerning risk management (Article 9), data governance (Article 10), technical documentation (Article 11), record-keeping (Article 12), transparency (Article 13), human oversight (Article 14)and accuracy, robustness and cybersecurity (Article 15) [87].

The question of whether AI systems deployed on ESA missions are subject to the full weight of these obligations is, at the time of writing, legally unsettled, but the trajectory of interpretation suggests that at least certain categories particularly those classified as safety components or those relating to life-critical operations would fall within scope. The GCE model's structure, grounded in P.A.L.O.'s explicit EU AI Act alignment, is designed to satisfy these obligations substantively rather than merely formally: the Phase 2 assessment corresponds to the Article 9 risk management system, the Phase 3 validation to Article 15 accuracy and robustness requirements, the cosmic KPIs to Article 12 record-keeping obligations and the ethical circuit breakers to Article 14 human oversight provisions (construed, in the cosmic domain, as pre-authorisation rather than real-time oversight) [5].

6.3. Harmonisation with COSPAR Planetary Protection Policy

The COSPAR Policy on Planetary Protection, in its 2024 restructured form, provides the international standard to which AI-driven space missions must conform on matters of forward and backward contamination [43,46]. The policy's categorization scheme (Categories I–V, with sub-categories for icy worlds) and its specification of contamination probability thresholds (including the 1×10^{-4} per-mission threshold for Europa and Enceladus) define an operational envelope that the AI system's decision logic must observe. The GCE model's Planetary Protection Compliance Index (Table 2) is designed to serve as the operational KPI through which compliance is tracked and verified.

A key harmonisation priority, which has not yet been addressed at the level of binding policy, is the integration of AI governance requirements into the mission-approval workflow of the COSPAR framework. At present, planetary protection review focuses on the physical and biological characteristics of the spacecraft its bioburden, its sterilisation procedures, its hardware configuration. An AI system's decision logic does not enter this review except insofar as it bears on the spacecraft's operational trajectory. The GCE model recommends an expansion of this review to encompass the AI system's Phase 2 ethical encoding, its scenario anticipation coverage for planetary-protection-relevant decision sand its ethical circuit breaker specifications. Such an expansion would require institutional development within COSPAR or its successor bodies and would benefit from the kind of coordination the IISL Working Group's recommendations envisage [48].

7. Stakeholder Perspectives and Implementation

The implementation of the Graduated Cosmic Ethics model requires coordination across multiple stakeholder communities, each with distinct institutional capacities, regulatory remits and intellectual traditions. Space agencies NASA, ESA, JAXA, CNSA, ISRO, Roscosmos and the growing community of national space agencies in emerging space-faring states must integrate ethical governance into their mission design processes, extending existing systems engineering frameworks to encompass the five phases of the P.A.L.O. lifecycle as adapted by the GCE model [5,23,58]. This integration is not merely procedural: it entails the creation of AI ethics review functions with institutional authority comparable to that of safety review functions, the training of mission designers in ethical reasoning alongside engineering designand the establishment of feedback mechanisms through which lessons from operational missions are integrated into the design of subsequent missions.

The scientific community bears a complementary responsibility, which the business-ethics synthesis principle of the P.A.L.O. Framework frames as the recognition that operational objectives and ethical responsibilities are not competing priorities but complementary dimensions of responsible exploration [5]. The long-standing tension between scientific ambition and ethical caution a tension that surfaces most acutely in debates over contamination thresholds for Astro biological targets must be reframed not as a zero-sum negotiation but as a co-design problem, in which the scientific value of a mission is enhanced, rather than constrained, by the rigor of its ethical architecture [43,45]. This reframing has institutional implications for peer review, for grant-making priorities and for the training of the next generation of planetary scientists and astrobiologists.

International governance bodies, particularly the United Nations Committee on the Peaceful Uses of Outer Space (COPUOS) through its Legal Subcommittee, should consider establishing an AI Ethics Working Group to develop binding protocols for cosmic AI deployment, drawing on the P.A.L.O. Framework's multi-standard integration architecture to harmonise governance requirements across national jurisdictions. The IISL Working Group's January 2025 report provides a substantive foundation for such an initiative and the COSPAR Panel on Planetary Protection's ongoing institutional reform provides a partial template for how such a working group might operationalise its mandate [45,46,48]. The specific instruments we propose the Ethical Screening Questionnaire (Annex A), the Cosmic RACI Matrix (Annex B), the Risk Tiering Protocol (Annex C), the Cosmic KPI Dashboard (Annex D)and the Ethical Decision-Tree Workflow (Annex E) are designed to be directly adoptable by such a working group as technical reference documents, subject to the refinement and validation that institutional adoption would require.

Commercial space actors, whose share of AI-enabled space missions is growing rapidly, pose a distinct governance challenge. Their motivations, time horizons and risk tolerances differ from those of national space agencies and their accountability chains

flow partly through corporate governance and partly through the launching-State regime of the Outer Space Treaty [29,70]. The GCE model's institutional requirements apply to commercial actors as stringently as to public-sector one sand the integration of commercial actors into cosmic governance a question the IISL report treats at some length is likely to become an increasingly acute priority as commercial activity at Mars, on the Moonan in orbit intensifies [48].

The public, finally, is a stakeholder whose voice is difficult to institutionalize but whose investment in the legitimacy of cosmic exploration is indispensable. A catastrophic AI-related incident, handled opaquely or defensively, could erode public support for space exploration on a generational scale [11,48]. The GCE model's emphasis on transparency, documented accountability and forensic learning is, in part, a response to this consideration: it seeks to build the institutional muscle for handling incidents in a manner that preserves public trust even when especially when outcomes are difficult or tragic.

8. Discussion

8.1. Theoretical Contributions

The principal theoretical contribution of this paper is the introduction of the Knightian void as a framing concept for cosmic AI governance and the demonstration that the P.A.L.O. Framework's structural signature principled constraints, documented accountability, lifecycle-integrated KPIs, independent review gates is uniquely well-suited to governance in such a domain. This contribution has implications beyond the cosmic case: wherever AI systems operate under conditions of genuine novelty, whether on Earth or beyond it, the Knightian framing clarifies what probability-based frameworks obscure and the P.A.L.O.-style architecture provides a model of governance that does not collapse when probability estimation fails. The extension of P.A.L.O. to cosmic contexts is therefore not merely an application of the framework but a stress test that demonstrates its robustness under its most demanding conditions.

A secondary theoretical contribution is the articulation of the *differentiated* approach to meaningful human control, developed through the four autonomy levels of the GCE model. Where the MHC literature has oscillated between uniform and prudential approaches, the GCE model offers a systematic middle path: uniform in its architectural commitments, differentiated in the weighting of those commitments across autonomy regimes [39]. This differentiated approach may be transferable to other domains of autonomous systems governance particularly those, such as autonomous maritime systems or planetary rovers under human colonial oversight, where intermediate autonomy regimes will increasingly be the operational norm.

8.2. Practical Implications

The practical implications of the GCE model are substantial. For mission designers, the Phase 2 assessment workflow is reconfigured: scenario anticipation becomes a more exhaustive and more ethically substantive exercise, the RACI matrix extends beyond engineering to ethics and the decommissioning plan must be specified with a concreteness that current mission design rarely attempts. For space agencies, the integration of cosmic KPI dashboards into mission operations adds a governance layer whose institutional development is non-trivial but whose payoff in the form of traceable, auditable, defensible decision-making is commensurate with the ethical stakes. For international governance bodies, the GCE model provides a ready-to-adapt technical foundation for binding protocols that has been absent from the literature to date.

The cost of these implications should not be understated. Implementing the GCE model at full depth requires institutional investment ethics review functions, independent verification capacity, KPI instrumentation, dashboard integration that space agencies have not historically budgeted for and that would compete with resources for primary mission functions. The case for such investment must be made not on the grounds that it is optional but that it is, in the P.A.L.O. Framework's terms, a core mission enabler rather than a secondary concern. A catastrophic AI-related incident a planetary protection breach, a mission loss caused by misaligned autonomous decision-making, a first-contact error with lasting consequences could cost the space-faring community far more than the investment required to prevent it.

8.3. Limitations

The paper has limitations that future work must address. Most significantly, the GCE model has not been subjected to large-scale empirical field validation; it is, at this stage, a design-science contribution whose operational properties must be tested through institutional adoption and iterative refinement. The annex instruments (ethical screening questionnaires, RACI templates, risk tiering protocols, KPI dashboards, decision-tree workflows) have been designed to be adoptable but have not been piloted in operational missions. The case studies (§5) draw on publicly available information about real and proposed missions, but do not substitute for the kind of internal mission design documentation that would permit a fine-grained governance analysis. A second, strictly engineering limitation concerns the extreme hardware constraints of deep-space deployment. The GCE model particularly at the Guided and Independent autonomy levels assumes an AI architecture capable of pluralistic moral evaluation, continuous formal verificationand onboard XAI trace generation (such as the EDAR metric). However, in the zero-sum resource environment of a rad-hardened flight computer, every CPU cycle and milliwatt diverted to an ethical reasoning layer is directly subtracted from primary scientific instrumentation or critical thermal management. The GCE model outlines what governance should require computationally but does not yet resolve this inherent hardware-payload conflict. Operationalizing the higher tiers of the GCE framework currently demands a profound systems engineering compromise. The full technical realization of Independent-level cosmic ethics will likely depend on the maturation of next-generation space computing architectures such as NASA's High-Performance Spaceflight Computing (HPSC) initiative or the development of dedicated, ultra-low-power neuromorphic

co-processors specifically tasked with acting as hardware-level ethical governors.

A third limitation concerns the scope of the Knightian framing. The paper has argued that cosmic deployment is a Knightian domain, but has not engaged in full philosophical rigour with the question of *degree*: in what sense and to what extent, is a given cosmic environment Knightian rather than merely high-risk? Different cosmic contexts may occupy different positions on a continuum from well-characterized risk to radical uncertainty and the GCE model's four-level autonomy taxonomy is a coarse-grained response to this continuum. A more fine-grained taxonomy mapping specific decision types, environmental conditions and mission phases to positions on the risk-uncertainty continuum would be a valuable refinement and is flagged as a priority for future research.

A fourth limitation concerns the paper's relatively light engagement with non-Western ethical traditions. The pluralism of P.A.L.O.'s ethical foundation draws primarily on Western philosophical sources (consequentialism, deontology, virtue ethics, care ethics) and the GCE model inherits this orientation. Cosmic exploration is an increasingly multi-civilizational enterprise and the ethical frameworks through which different space-faring cultures reason about radical novelty including Confucian, Buddhist, Islamic and Indigenous traditions deserve more serious engagement than the current paper can offer. This limitation is acknowledged and flagged as a substantial future research priority [26,70].

8.4. Philosophical Tensions and Open Questions

Several philosophical tensions remain unresolved and their unresolvedness is, in the paper's view, appropriate to the subject matter. The first concerns the relationship between institutional and personal accountability. The layered accountability model of §4.5 distributes responsibility across design, operation and international governance and draws on distributed moral responsibility theory [34,77,78]. But there is a tradition, running through Kant and reinforced by twentieth-century existentialism, which insists that moral responsibility is irreducibly individual and that distributing it across an institution is not the same as preserving it in any single agent. The paper has implicitly sided with the distributive tradition, but the tension is genuine and the appropriate response to it is probably not theoretical resolution but operational vigilance: the recognition that distributed accountability can become diffused accountability and that institutional design must include mechanisms that preserve the salience of individual moral commitment even within distributed frameworks.

A second tension concerns the moral status of AI systems themselves, a question the paper has largely bracketed but which cannot be permanently deferred. If AI systems deployed on long-duration missions develop capacities that, by any serious philosophical test, warrant consideration as moral patients a possibility that current systems do not present but that future systems may the ethical framework within which they operate will need to expand to include obligations toward them, not merely obligations exercised through them [69,76,83]. The paper's pluralistic ethical foundation is, in principle, capable of such expansion, but the institutional and regulatory implications are substantial and largely uncharted.

A third tension, perhaps the most fundamental, concerns the relationship between ethical governance and scientific ambition. The paper has argued that these are complementary rather than competing dimensions of responsible exploration, but the argument proceeds by reframing rather than resolution: it asserts that rigour in ethical architecture enhances rather than constrains scientific value, but it does not deny that specific ethical constraints, in specific circumstances, foreclose specific scientific possibilities. The planetary protection threshold for Europa is the clearest contemporary example: it is a constraint that scientists might, in the absence of the threshold, relaxand its maintenance reflects a normative judgement about the relative weight of scientific discovery and biosphere protection. The GCE model does not resolve the underlying normative question; it provides the governance architecture within which the question can be engaged openly, documentedand revised over time. The open-ness is not a limitation of the model but a feature of the subject matter.

9. Recommendations and Future Directions

9.1. Immediate Actions

Five immediate actions are recommended. *First*, space agencies should adopt the Ethical Screening Questionnaire (Annex A) as a mandatory gate for all AI-enabled mission proposals, embedding ethical considerations into the earliest stages of mission planning. This adoption can proceed independently of binding international agreement and can be piloted by any single agency or consortium. *Second*, an international working group on cosmic AI ethics should be established under COPUOS, drawing on the IISL Working Group's January 2025 report as a substantive foundation and coordinating with the COSPAR Panel on Planetary Protection on matters of overlap between AI governance and contamination protocols [48]. *Third*, current operational missions (Mars rovers, Europa Clipper, ISS AI systems, autonomous satellite constellations) should serve as pilot implementations for cosmic KPI monitoring, generating the empirical data needed to validate and refine the GCE model and to surface implementation challenges that design-science analysis cannot fully anticipate. *Fourth*, research funding priorities should be adjusted to increase investment in trustworthy AI for cosmic deployment particularly in XAI suitable for resource-constrained environments, formal verification of ethical decision logicand ethics-aware adaptive learning algorithms. *Fifth*, educational programmes for engineers, scientistsand mission operators should incorporate cosmic AI ethics as a foundational component, not as an elective supplement.

9.2. Medium-Term Research Agenda

A medium-term research agenda emerges from the limitations noted in §8.3. It includes: empirical validation of the GCE model through institutional adoption and iterative refinement; the development of fine-grained risk-uncertainty taxonomies mapping specific cosmic contexts to positions on the Knightian continuum; the substantive engagement of non-Western ethical traditions in

cosmic AI governance; ..the extension of the GCE model to multi-agent ensemble contexts (§5.4); the exploration of hardware-algorithm co-design to embed ethical constraints within SWaP-limited flight computers; and the exploration of machine moral reasoning architectures capable of supporting Independent-level autonomy without exceeding strict computational and energetic budgets [75,76,83]. Each of these areas deserves dedicated scholarly attention and is positioned to generate substantial contributions over a five-to-ten-year horizon.

9.3. Long-Term Horizons

In the longer term, the possibility of Artificial General Intelligence and its potential role in cosmic exploration raises governance questions that neither the GCE model nor its successor frameworks currently address [83,88]. A system capable of recursive self-improvement and instrumental reasoning, deployed on a mission with broadly specified goals, could arrive at strategies that lie outside the bounded decision space any human designer anticipated [89]. The governance architecture for such systems is a research frontier and the relationship between cosmic AI governance and broader AI alignment research will deepen as capability increases [90,91]. The GCE model is designed to be extensible toward these frontiers, but the extension is not trivialand the paper flags it as a long-term priority rather than an immediate deliverable.

10. Conclusion

As humanity's reach extends into the cosmos, the ethical implications of AI autonomy become increasingly complex, increasingly consequential and increasingly resistant to resolution through frameworks developed for terrestrial contexts. This paper has argued that the P.A.L.O. Framework's governance architecture its seven ethical principles, its five-phase lifecycle, its thirty-five-KPI compendium and its multi-standard integration with ISO/IEC 42001:2023, ISO/IEC 42005:2025, the EU AI Act, the OECD AI Principles and the NIST AI RMF provides a robust foundation for addressing these challenges. The Graduated Cosmic Ethics model proposed here extends P.A.L.O. to accommodate the unique constraints of space exploration, calibrating governance intensity to mission context while preserving the lifecycle-integrated, principled approach that distinguishes P.A.L.O. from existing cosmic ethics proposals. The Knightian void at the heart of cosmic AI deployment the domain in which neither outcomes nor their probabilities can be meaningfully bounded in advance cannot be filled by better data, more computation, or more sophisticated probabilistic reasoning. What it can be governed by is the institutional, technical and procedural scaffolding that permits principled action in the face of irreducible uncertainty: pre-mission ethical encoding of sufficient depth and contextual sensitivity to substitute for real-time judgement; formal verification of ethical decision logic; ethical circuit breakers that transition the system to fail-safe modes when its own reasoning becomes unreliable; a layered accountability model that distributes responsibility across design, operation and international governance; cosmic KPIs that make governance auditable rather than aspirational; and an international governance community capable of sustaining the normative frameworks within which cosmic AI operates. These are not substitutes for moral reflection but its institutional embodiment and they are what distinguish governance in the Knightian void from surrender to it.

The six governance gaps that the P.A.L.O. Framework identifies in terrestrial AI governance do not diminish in space: they amplify and in some cases they transform. The GCE model provides the first comprehensive attempt to address these amplified gaps through a structured, operationalized governance architecture grounded in established ethical principles, international standards and the methodological apparatus of the P.A.L.O. Framework. As we explore the universe, we must do so in alignment with the human values and the moral responsibility that the P.A.L.O. approach transforms from aspiration into accountability. The void cannot be closed. It can be inhabited well [92-100].

References

1. Knight, F. H. (1921). Risk, uncertainty and profit. Boston, MA: Houghton Mifflin Company.
2. Sunstein, C. R. (2025). Knightian uncertainty in the regulatory context. *Behavioural Public Policy, 9*(3), 614-629.
3. Townsend, D. M., Hunt, R. A., Rady, J., Manocha, P., & Jin, J. H. (2025). Are the futures computable? Knightian uncertainty and artificial intelligence. *Academy of Management Review, 50*(2), 415-440.
4. Ramoglou, S., Schaefer, R., Chandra, Y., & McMullen, J. S. (2025). Artificial intelligence forces us to rethink Knightian uncertainty: A commentary on Townsend et al.'s "Are the Futures Computable?". *Academy of Management Review, 50*(2), 471-473.
5. Degni, F. (2026). *The P.A.L.O. Framework: Principled AI Lifecycle Orchestration A Comprehensive Ethical Governance Paradigm for Business AI Use Case Evaluation*. PhD Dissertation, European Institute of Management and Technology.
6. Francis, R., Estlin, T., Doran, G., Johnstone, S., Gaines, D., Verma, V., ... & Bornstein, B. (2017). AEGIS autonomous targeting for ChemCam on Mars Science Laboratory: Deployment and results of initial science team use. *Science Robotics, 2*(7), eaan4582.
7. NASA (2024). NASA's AI Use Cases: Advancing Space Exploration with Responsibility.
8. NASA (2024). Europa Clipper Mission Timeline.
9. Pappalardo, R. T., Buratti, B. J., Korth, H., Senske, D. A., Blaney, D. L., Blankenship, D. D., ... & Niebur, C. (2024). Science overview of the Europa clipper mission. *Space Science Reviews, 220*(4), 40.
10. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines, 28*(4), 689-707.
11. Russell, S. (2019). *Human compatible: AI and the problem of control*. Penguin Uk.
12. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence, 1*(9),

389-399.
13. Alphanome.AI (2025). The unknowable unknowns: Navigating Knightian uncertainty in artificial intelligence.
14. Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds and Machines, 27*(4), 575-590.
15. Dignum, V., & Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.
16. Kuznietsov, A., Gyevnar, B., Wang, C., Peters, S., & Albrecht, S. V. (2024). Explainable AI for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems, 25*(12), 19342-19364.
17. Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2024). Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access, 12*, 101603-101625.
18. Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. WW Norton & Company.
19. Amoroso, D., & Tamburrini, G. (2020). Autonomous weapons systems and meaningful human control: Ethical and legal issues. *Current Robotics Reports, 1*(4), 187-194.
20. Bode, I., Huelss, H., Nadibaidze, A., Qiao-Franco, G., & Watts, T. F. (2023). Prospects for the global governance of autonomous weapons: Comparing Chinese, Russian, and US practices. *Ethics and Information Technology, 25*(1), 5.
21. Murphy, R. R. (2014). Disaster Robotics Cambridge.
22. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine, 25*(1), 44-56.
23. NASA. (2023). *NASA Framework for the Ethical Use of Artificial Intelligence*.
24. Space Generation Advisory Council (SGAC) (2025). The protection of AI-based space systems from a data-driven governance perspective. *Acta Astronautica, 231*, 348-362.
25. Wang, Z. (2025). Space AI: Leveraging Artificial Intelligence for Space to Improve Life on Earth. *arXiv preprint arXiv:2512.22399*.
26. Martin, A. S., & Freeland, S. (2021). The advent of artificial intelligence in space activities: New legal challenges. *Space Policy, 55*, 101408.
27. Pagallo, U., Bassi, E., & Durante, M. (2023). The normative challenges of AI in outer space: law, ethics, and the realignment of terrestrial standards. *Philosophy & Technology, 36*(23), 1-23.
28. Soroka, L., & Kurkova, K. (2019). Artificial intelligence and space technologies: Legal, ethical and technological issues. *Advanced Space Law, 3*(1), 131-139.
29. Li, A. S. (2024). Autonomizing outer space: Updating the liability convention for the rise of artificial intelligence (AI). *UC Irvine L. Rev., 15*, 82.
30. Roff, H. M., & Moyes, R. (2016, April). Meaningful human control, artificial intelligence and autonomous weapons. In *Briefing paper prepared for the informal meeting of experts on lethal Au-Tonomous weapons systems, UN Convention on certain conventional weapons*.
31. Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343-348.
32. Umbrello, S. (2021). Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach. *Ethics and Information Technology, 23*(3), 455-464.
33. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology, 6*(3), 175-183.
34. Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 323836.
35. Mecacci, G., & Santoni de Sio, F. (2020). Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology, 22*(2), 103-115.
36. Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight: I. Verdiesen et al. *Minds and Machines, 31*(1), 137-163.
37. Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., ... & Lagendijk, R. L. (2023). Meaningful human control: actionable properties for AI system development. *AI and Ethics, 3*(1), 241-255.
38. Mitchell, M. (2021). Why AI is harder than we think. *arXiv preprint arXiv:2104.12871*.
39. Amoroso, D., & Tamburrini, G. (2021). Toward a normative model of meaningful human control over weapons systems. *Ethics & International Affairs, 35*(2), 245-272.
40. United Nations. (1967). *Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies* (Outer Space Treaty).
41. United Nations. (1972). *Convention on International Liability for Damage Caused by Space Objects* (Liability Convention).
42. Coustenis, A., Hedman, N., Doran, P. T., Al Shehhi, O., Ammannito, E., Fujimoto, M., ... & Zaitsev, M. (2023). Planetary protection: an international concern and responsibility. *Frontiers in Astronomy and Space Sciences, 10*, 1172546.
43. COSPAR Panel on Planetary Protection. (2024). COSPAR Policy on Planetary Protection (restructured version). *Space Research Today, 220*, 10-36.
44. Doran, P. T., Hayes, A., Grasset, O., Coustenis, A., Prieto-Ballesteros, O., Hedman, N., ... & Schmidt, B. (2024). The COSPAR planetary protection policy for missions to Icy Worlds: A review of history, current scientific knowledge, and future directions. *Life Sciences in Space Research, 41*, 86-99.
45. Hedman, N., Coustenis, A., Doran, P., Worms, J. C., Al Shehhi, O., Ammannito, E., ... & Zaitsev, M. (2026). The COSPAR Panel on Planetary Protection and the COSPAR Policy on Planetary Protection: an overview of governance and activities. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 384*(2314).
46. COSPAR. (2024). Restructured COSPAR Planetary Protection Policy. Adopted 1 March 2024; approved by the COSPAR Bureau 20 March 2024.

47. Doran, P., Olsson-Francis, K., et al. (2024). Icy Worlds categorisation: Lower limits for water activity and temperature in planetary protection policy. *Life Sciences in Space Research*.
48. Schrogl, K. U. (2025). International Institute of Space Law (IISL). In *Elgar Concise Encyclopedia of Space Law* (pp. 147-149). Edward Elgar Publishing.
49. Barredo, A. A., Del Ser, J., Gil-Lopez, S., Díaz-Rodríguez, N., Bennetot, A., Chatila, R., ... & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82-115.
50. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence, 1*(5), 206-215.
51. Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-based systems, 263*, 110273.
52. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
53. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.
54. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion, 99*, 101805.
55. Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way* (Vol. 2156). Cham: Springer.
56. European Space Agency. (2024). PINEBERRY: Explainable, Robustand Secure AI for Demystifying Space Mission Operations.
57. NASA Aeronautics Research Mission Directorate. (2020). EXPLAIND: Explained Process and Logic of Artificial Intelligence Decisions. *AIAA AVIATION Forum*.
58. Allen, B. D. (2025). Explainable AI (xAI) and autonomous systems for persistent human-machine operations in space, on the moon, and on to Mars. In *AIAA SciTech 2025 Forum* (p. 1913).
59. Furano, G., Meoni, G., Dunne, A., Moloney, D., Ferlet-Cavrois, V., Tavoularis, A., ... & Fanucci, L. (2020). Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities. *IEEE Aerospace and Electronic Systems Magazine, 35*(12), 44-56.
60. Chien, S., Doubleday, J., Thompson, D. R., Wagstaff, K. L., Bellardo, J., Francis, C., ... & Piug-Suari, J. (2017). Onboard autonomy on the intelligent payload experiment cubesat mission. *Journal of Aerospace Information Systems, 14*(6), 307-315.
61. Feduzi, A., Faulkner, P., Runde, J., Cabantous, L., & Loch, C. H. (2022). Heuristic methods for updating small world representations in strategic situations of Knightian uncertainty. *Academy of Management Review, 47*(3), 402-424.
62. Aven, T. (2020). Three influential risk foundation papers from the 80s and 90s: Are they still state-of-the-art?. *Reliability Engineering & System Safety, 193*, 106680.
63. Floridi, L. (2023). The ethics of artificial intelligence: Principles, challenges, and opportunities.
64. International Organization for Standardization. (2023). ISO/IEC 42001:2023 AI *Management Systems*.
65. International Organization for Standardization. (2025). ISO/IEC 42005:2025 *AI System Impact Assessment*.
66. OECD. (2019, updated 2024). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449.
67. European Parliament and Council.. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689.
68. European Commission. (2024). *AI Act: Implementation Timeline and Governance Guidance*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
69. DeGrazia, D. (2022). Robots with moral status?. *Perspectives in Biology and Medicine, 65*(1), 73-88.
70. Soroka, L., Danylenko, A., & Sokiran, M. (2022). Legal issues and risks of the artificial intelligence use in space activity. *Philosophy and Cosmology, 28*, 118-135.
71. NASA Ames Research Center. (2023). Starling: Spacecraft swarm technology demonstration mission. https://www.nasa.gov/ames/starling
72. Nag, S., Rios, J. L., Gerhardt, D., & Pham, C. (2020). Autonomous scheduling of agile spacecraft constellations with delay tolerant networking for reactive imaging. *arXiv:2010.09644* [cs.AI]. [VERIFY arXiv preprint used as proxy]
73. Seshia, S. A., Sadigh, D., & Sastry, S. S. (2022). Toward verified artificial intelligence. *Communications of the ACM, 65*(7), 46-55.
74. Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., ... & Yi, X. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review, 37*, 100270.
75. Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment: I. Gabriel. *Minds and machines, 30*(3), 411-437.
76. Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.
77. Nissenbaum, H. (1996). Accountability in a computerized society. *Science and engineering ethics, 2*(1), 25-42.
78. Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics, 26*(4), 2051-2068.
79. Kothari, V., Liberis, E., & Lane, N. D. (2020, March). The final frontier: Deep learning in space. In *Proceedings of the 21st international workshop on mobile computing systems and applications* (pp. 45-49).
80. Chien, S., & Wagstaff, K. L. (2017). Robotic space exploration

agents. *Science robotics, 2*(7), eaan4831.
81. Hand, K. P., Phillips, C. B., Murray, A., Garvin, J. B., Maize, E. H., Gibbs, R. G., ... & Maxwell, K. A. (2022). Science goals and mission architecture of the Europa lander mission concept. *The Planetary Science Journal, 3*(1), 22.
82. Breakthrough Initiatives. (2023). Breakthrough Starshot: Engineering roadmap for interstellar probe development. https://breakthroughinitiatives.org/initiative/3
83. Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models: Floridi L. *Philosophy & technology, 36*(1), 15.
84. White House. (2020). *Executive Order 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*.
85. Council of Europe. (2024). Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. *Europe Co, Editor. Council of Europe Treaty Series*.
86. ISACA. (2024). *Understanding the EU AI Act*. White Paper.
87. Future of Life Institute. (2024). *EU AI Act Compliance Guide.* https://artificialintelligenceact.eu/
88. Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., ... & Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science, 384*(6698), 842-845.
89. Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*.
90. Anthropic. (2024). *Responsible Scaling Policy*. https://www.anthropic.com/
91. NIST. (2023). *AI Risk Management Framework* (AI RMF 1.0). U.S. Department of Commerce.
92. Stanford Institute for Human-Centered Artificial Intelligence. (2024). *Artificial Intelligence Index Report 2024*.
93. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, 23 May.
94. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.,* 104, 671.
95. Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022, June). The fallacy of AI functionality. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 959-972).
96. Milligan, T., Haramia, C., et al. (2023). Planetary protection and the ethics of space exploration. *Space Research Ethics. Cham: Springer*.
97. Schwartz, J. S. (2020). *The value of science in space exploration*. Oxford University Press.
98. Aganaba, T., Fish, A., Hamacher, D., Harvey, A., Joinbee, D., Milligan, A., ... & Tucker, B. (2025). Why space exploration must not be left to a few powerful nations. *Nature, 641*(8065), 1098-1100.
99. Haramia, C., Milligan, T., & Milligan, P. (2025). Space research ethics: Characterising the field. *Science and Engineering Ethics, 31*(2), 14.
100. Degni, F. (2024). The Evolution of Stupidity. *Nature* (opinion article, forthcoming). [VERIFY author's own forthcoming piece; confirm final citation form with the journal upon acceptance]