Ethical Governance of AI in Space: Philosophical Foundations, Moral Agency, and Civilizational Implications

by
Dr. Maria Harney

for
Space Renaissance International (SRI) World Congress 2026

14 May 2026

Introduction

Rapid Artificial Intelligence developments in space operations, such as autonomous spacecraft, orbital infrastructure, and extraterrestrial settlements, create novel ethical challenges, responsibility gaps, and risks where human oversight is limited, and decisions are potentially irreversible. At the current stage, technological innovation outpaces ethical and governance frameworks, which is especially concerning when applied to high-stakes, austere, isolated environments where AI algorithmic authority raises questions about human moral agency, autonomy, accountability, and overall long-term civilizational stability.

The paper discusses and applies historical and philosophical traditions, such as Aristotle's virtue ethics, Kant's moral law, and Aquinas' natural law, to contemporary AI safety research, philosophy of mind, and space policy to demonstrate the importance of AI ethics and policy to be oriented toward human dignity, responsible governance, and human flourishing rather than autonomous replacement of human agency. Humanity's civilizational expansion beyond Earth depends on moral clarity guiding autonomous systems; ethical regulation is essential for peace, civilization continuity, and flourishing on Earth and in space.

Philosophical Foundations for Ethical AI in Space

Aristotle's virtue ethics and eudaimonia, or flourishing, along with hylomorphism (form-matter unity), views humans as possessing intrinsic teleology. In contrast, AI lacks immanent causality and, therefore, should serve human flourishing as a tool, and not as a replacement for moral agency. St. Thomas Aquinas inherited the Aristotelian framework and demonstrated a hylomorphic unity of body-soul through Natural and Eternal Law. He preserved the unity without reductionism or Cartesian dualism, and while AI may simulate humans' functions, it lacks embodied form and orientation toward the good.

Another thinker, Immanuel Kant, provides one of the most rigorous accounts of what distinguishes human subjectivity. Kant's transcendental unity of apperception ("I think") as a condition for objective experience, synthetic a priori judgment, and moral autonomy highlights the uniqueness of the human being. By comparison, AI lacks unified subjectivity, intentionality, and capacity for genuine moral responsibility. Finally, Robert Sokolowski presents a phenomenological account of the human person as a responsible agent who can disclose truth and constitute moral identity through deliberate action. He emphasizes human intentionality through lived experience and intersubjective rationality – qualities that AI processes currently lack.

Thus, the comparative framework demonstrates that humans possess unity of consciousness, moral agency, teleology, and embodiment, while current AI systems rely on distributed processing, statistical correlation, externally imposed goals, and hardware/software dualism. Furthermore, current development in neuroscience, neuroplasticity research, the brain-heart-gut-mitochondria axis, and cases like the absence of cerebellum support a hylomorphic view of humans as dynamic unified organisms, not programmable machines.

AI in Space Medicine: Applications, Challenges, and Ethical Risks

With the growing and expanding commercial spaceflight sector, AI applications are becoming increasingly essential for autonomous navigation, real-time decision-making, mission optimization, and handling the complex demands of deep-space exploration. Among these systems are autonomous diagnosis and treatment capabilities, which are critical due to deep-space communication delays; continuous monitoring of vital signs, radiation levels, and microgravity effects; AI digital twins; augmented reality medical guidance; autonomous ultrasound; mental health support; and conflict resolution tools (such as NASA-Google CMO-

DA, HRP projects, and ESA's ALISSE).

While many automated systems driven by AI enhance human performance, offset the workload, and assist in essential tasks, they raise questions about various ethical challenges in isolated, high-risk environments, such as bias and inequity arising from scarce, non-diverse astronaut datasets, accountability for irreversible harm, informed consent and autonomy under coercion and isolation, and erosion of human empathy and therapeutic relationships. Neuroethics intersections: With scientific advances in cognitive enhancement, memory modification, and disorders of consciousness, the field of neuroethics has to address brain imaging/privacy, free will/responsibility, and cognitive diversity.

Universal versus Cultural Morality in Space Contexts

The expansion of neuroethical challenges in space highlights the need to examine the foundations of morality itself, specifically, whether universal moral principles can guide human (and potentially post-human) behavior in extreme extraterrestrial environments, or whether morality must remain culturally relative. Universal morality rests on core principles - human dignity, justice, non-maleficence, fairness, accountability, and respect for persons - as articulated in Aristotelian, Kantian, and Thomistic philosophies.

In contrast, cultural morality is shaped by history, society, and specific contexts. Both universal and cultural approaches, and especially the harmonious integration of the two, are highly relevant to space exploration. They are particularly critical in the confined environments of small crews (5–6 astronauts) and larger settlements (50–100+ people): universal principles provide essential baselines that prevent fragmentation under scarcity and stress, while cultural sensitivity helps preserve social cohesion and psychological resilience. Examples of ethical tensions that can arise in both small crews and large settlements include triage and resource

allocation, end-of-life decisions, acceptance of authority, and differing levels of risk tolerance. These challenges must be addressed through a shared set of universal ethical principles, combined with flexible cultural accommodation protocols and practical moral deliberation training.

Interdisciplinary Roadmap for Ethical AI Governance in Space: Human Identity, Cognition, and Moral Agency in Space

One of the space philosophy's core questions asks what remains constant in the human person amid technological acceleration? If human identity is understood as an enduring unity of rational and moral subject grounded in cognition (unified consciousness), perception, culture, and relational meaning, if left unregulated, exponential technologies, such as AI, synthetic biology, neural augmentation, and brain-computer interfaces, have the potential to destabilize identity through isolation, altered perception, algorithmic authority, and human-machine boundary blurring.

When applying the Aristotelian-Thomistic framework, it becomes clear that humans are not machines. Meaning-making, self-actualization, moral responsibility, and teleology are distinctly and uniquely human capacities. Therefore, AI should serve as a supportive tool rather than a full replacement for human autonomy. This means maintaining meaningful human oversight in critical areas, such as human-in-the-loop systems for lethal and medical decisions, along with clear prohibitions on fully autonomous lethal actions in the early phases of missions, and the development of rigorous frameworks for assessing the moral status of digital minds and potential development of Artificial General Intelligence.

To pave the interdisciplinary road-map for Ethical AI governance in space, humans need to use Philosophy as a foundational discipline, grounding decision-making, policies, and

regulation in Aristotle's virtue and flourishing, Kant's dignity and autonomy, and Aquinas' Natural law and hylomorphism. Ethical governance based on those principles will facilitate the creation of AI-driven and automated systems based on human dignity and autonomy first: human-in-the-loop for critical decisions, explainable AI, value alignment with human moral agency, and resistance to reductionism and scientism.

Effective implementation will require practical governance mechanisms such as pre-mission ethical protocols, international AI safety standards, ongoing ethical audits, cultural sensitivity training, and transparent accountability frameworks. Safeguards must include strict limits on fully autonomous lethal, medical, and technically-critical decisions during the spaceflight and early phases of settlement, along with robust frameworks for assessing the moral status of digital minds, and proactive mitigation of existential risks, particularly around alignment and the growing gap between technological capability and moral maturity.

Conclusion

Humanity does not evolve morally simply by advancing technologically. Thus, as humans build extraterrestrial civilization, intentional cultivation of ethics, governance, and philosophy is required. Space expansion is civilizational transformation; it is not exclusively technical, as success depends on embedding moral clarity into autonomous systems driven by Artificial Intelligence. Humanity's interdisciplinary road-map to extraterrestrial settlement ought to integrate philosophy, AI safety, space policy, and empirical research that ensures AI advances peace, continuity, and flourishing rather than undermining human autonomy.

Looking ahead, the long-term vision is one in which AI enables safer and more sustainable multi-planetary life while remaining firmly in service of genuine human flourishing. This demands a balance between innovation and moral clarity, integrating advanced technologies

such as NASA projects, synthetic biology, robotics, and VR/AR with philosophical depth. Ultimately, how we treat humans in space will shape the future of humanity on Earth. The decisions we make today will define our moral legacy far beyond our home planet.

References

Gallagher, Kenneth T. *The Philosophy of Knowledge.* New York: Fordham University Press, 1986.

Geert Hofstede. *"6D Model of National Culture."* Accessed February 13, 2025. Hofstede Insights.

Kant, Immanuel. *Critique of Pure Reason.* Translated by Norman Kemp Smith. London: Macmillan & Co., 1929.

McInerny, Ralph, ed. Thomas Aquinas: Selected Writings. London: Penguin Classics, 1998.

Miller, Lulu. *"Trapped in His Body for 12 Years, a Man Breaks Free."* NPR: All Things Considered, January 9, 2015. NPR article.

"Neuroethics." Wikipedia. Last modified March 15, 2026. Wikipedia Neuroethics article.

Reeve, C. D. C., and Patrick Lee Miller, eds. *Introductory Readings in Ancient Greek and Roman Philosophy.* Indianapolis, IN: Hackett Publishing Company, 2015.

Rohlf, Michael. *"Immanuel Kant."* In The Stanford Encyclopedia of Philosophy, Fall 2024 Edition, edited by Edward N. Zalta and Uri Nodelman. Stanford Encyclopedia of Philosophy entry on Kant.

Sokolowski, Robert. *Introduction to Phenomenology.* Cambridge: Cambridge University Press, 2000.