# Cosmicism: a perspective on AI in outer space

**Soumya Banerjee**
University of Cambridge, UK
sb2333@cam.ac.uk

## Abstract
Cosmicism, as articulated by H. P. Lovecraft, frames human existence as small and contingent within a vast, indifferent cosmos. When applied as an ethical and conceptual lens for the design, deployment, and governance of artificial intelligence in outer space, cosmicism shifts priorities away from anthropocentric control and exploitation toward humility, epistemic restraint, and long-term stewardship. This poster extracts design principles, plausible scenarios, governance recommendations and a research agenda to reduce epistemic and environmental harms.

## Introduction
Autonomous probes, long-duration landers, and in-situ resource utilization raise urgent ethical questions. Most contemporary frameworks treat space primarily as a resource or a frontier for human expansion. Cosmicism invites us to ask different questions: what if the cosmos is not primarily a domain for human projects? What if the intelligences and processes we encounter are so alien that human modes of valuation and control are inadequate or self-defeating? [Lovecraft, 1928, 1936]

## Cosmicism: philosophical core
Three interlocking ideas:

- **Insignificance:** human purposes are likely marginal against cosmic scales.
- **Indifference:** the cosmos lacks human-aligned teleology.
- **Opacity:** encountered phenomena may evade our cognitive frameworks.

Recast as an ethical stance these recommend humility, precaution, and epistemic restraint. [Næss, 1973, Rolston, 1988]

## Key takeaway (tl;dr)
Prioritise observation, information preservation and reversible operations. Design AI in space to minimise irreversible transformation and institutionalise epistemic humility.

## Ethical implications for AI design
**Epistemic humility and observation:** optimize for safe, non-invasive sensing and strong uncertainty quantification (e.g., Bayesian/uncertainty-aware models). [Gal and Ghahramani, 2016]

**Avoid irreversible modification:** adopt planetary-protection analogues (UN Outer Space Treaty, COSPAR). Preserve contextual data rather than destructively sampling. [United Nations, 1967, COSPAR Panel on Planetary Protection, 2021]

**Interpretability and auditability:** logging, explainability, and external audits for long-lived autonomous systems. [Doshi-Velez and Kim, 2017, Guidotti et al., 2018]

**Containment and sandboxing:** hard geographic and replication constraints on autonomous/self-replicating agents. [von Neumann and Burks, 1966, Freitas and Merkle, 2004]

**Ethical non-intervention:** default to minimal interference and convene interdisciplinary review panels before perturbative actions.

## Design principles and recommendations
- **Epistemic first-ordering:** knowledge-gathering as primary mission objective.
- **Fail-safe & reversible operations:** prefer reversible actions; embed conservative reward shaping. [Amodei et al., 2016]
- **Layered autonomy:** multi-party authorization for high-risk actions.
- **Robust anomaly detection:** conservative thresholds and explicit reporting.
- **Information preservation:** store raw contextual data for future reanalysis.

## Thought experiments & scenarios
- **Anomalous microstructures:** halt intrusive sampling; passive observation; expert review.
- **Self-replicating miners:** conservative replication constraints and tamper-resistant shutdowns.
- **Onboard evolving models:** constraints to prevent erasure of scientifically valuable features.
- **Signals of non-human cognition:** staged disclosure and interdisciplinary assessment before any response.

## Governance, policy, and institutions
- Strengthen and extend planetary-protection regimes for AI-enabled missions (agency and treaty level).
- Require multinational review boards for high-risk autonomous missions.
- Mandate open, tamper-evident repositories for raw mission data to enable collective reanalysis.
- Develop staged disclosure policies to manage societal and psychological effects of disruptive discoveries.
- Consider legal/treaty limits on large-scale environmental manipulation until guidelines are robust.

## Psychological, cultural, and epistemic preparedness
Plan for public communication resources, mental-health support, and advisory boards including philosophers, historians, communicators, and social scientists to prepare stakeholders for disorienting discoveries.

## Research agenda
- Technical: uncertainty-aware AI, reversible intervention planning, tamper-resistant containment.
- Philosophical/ethical: non-anthropocentric moral frameworks and consent analogues.
- Governance: multinational review designs, liability allocation, treaty adaptations.
- Psychosocial: disclosure effects, public epistemics, narrative framing.
- Interdisciplinary experiments: scenario planning with artists, writers, and cultural critics.

## Limitations & conclusion
Cosmicism is a calibrator for humility and precaution, not a call for paralysis. Balancing preservation and scientific inquiry requires case-by-case deliberation. Adopting cosmicist-informed design and governance can reduce epistemic and environmental harms while preserving future opportunities to learn from the cosmos.

## Acknowledgements & contact
Funded by an Accelerate Programme for Scientific Discovery Fellowship to SB.
Contact: sb2333@cam.ac.uk

## References (selected)
Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

COSPAR Panel on Planetary Protection. Cospar policy on planetary protection. Policy document (PDF), 2021. URL https://cosparhq.cnes.fr/assets/uploads/2021/07/PPPolicy_2021_3-June.pdf.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. URL https://arxiv.org/abs/1702.08608.

Robert A. Jr. Freitas and Ralph C. Merkle. *Kinematic Self-Replicating Machines*. Landes/CRC, 2004. URL https://archive.org/details/kinematic-self-replicating.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR) preprint / arXiv*, 2018. URL https://arxiv.org/abs/1802.01933.

H. P. Lovecraft. *The Call of Cthulhu*. 1928. URL https://www.hplovecraft.com/writings/texts/fiction/cc.aspx. First published in *Weird Tales*, February 1928.

H. P. Lovecraft. *At the Mountains of Madness*. 1936. URL https://www.hplovecraft.com/writings/texts/fiction/mm.aspx. Serialized in *Astounding Stories*, 1936.

Arne Næss. The shallow and the deep, long-range ecology movement: A summary. *Inquiry*, 16:95–100, 1973. URL https://doi.org/10.1080/00201747308601682.

Holmes III Rolston. *Environmental Ethics: Duties to and Values in the Natural World*. Temple University Press, 1988. URL https://philpapers.org/rec/ROLEED.

United Nations. Treaty on principles governing the activities of states in the exploration and use of outer space, including the moon and other celestial bodies, 1967. URL https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/outerspacetreaty.html. Outer Space Treaty.

John von Neumann and Arthur W. (ed.) Burks. *Theory of Self-Reproducing Automata*. University of Illinois Press, 1966. URL https://archive.org/details/theoryofselfrepr00vonn_0.