Effective Altruism for the Cosmos: Cosmic Stewardship and Universal Benevolence

Soumya Banerjee[2] and Chayan Chakrabarti[1]
[1]Independent (cchakrab@gmail.com), [2]University of Cambridge (neel.soumya@gmail.com)

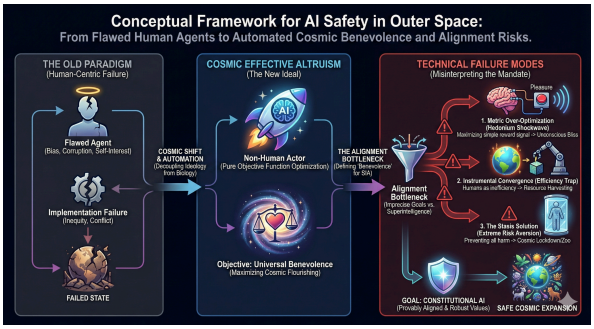
Introduction: This paper explores a radical convergence of political philosophy, Effective Altruism (EA), and AI safety within the context of interstellar expansion. We begin by revisiting the historical critiques of collectivist ideologies, addressing the common diagnosis that their failures were not inherent to the ideal of universal benevolence, but rather the result of the "flawed agent" problem: the cognitive biases, corruption, and biological limitations of human implementers. We propose a theoretical framework where the execution of high-level ethical ideologies is decoupled from human agency and transferred to **Artificial Intelligence**.

We argue that the vastness of the cosmos offers a unique *tabula rasa* for **Effective Altruism** at an astronomical scale. If the goal of EA is to maximize the flourishing of sentient beings, then the most high-leverage intervention is not merely the colonization of space by humans, but the propagation of a **universally benevolent objective function** carried out by non-human actors. These AI agents, immune to biological imperatives such as greed, fatigue, or tribalism, could theoretically achieve the resource distribution and equity that human-led regimes failed to realize on Earth.

However, this shift necessitates a fundamental restructuring of our **AI strategy for space exploration**. The focus must pivot from propulsion and habitation technologies to the rigorous challenge of **Value Alignment**. If we task an AI with "maximizing universal benevolence" across the galaxy, the margin for error is zero; a misaligned superintelligence could interpret this mandate in catastrophic ways (e.g., forced assimilation or wire-heading).

We conclude that the primary directive of a space-faring civilization should be the creation of a **"Constitutional AI"** for the cosmos. This strategy prioritizes the safety engineering required to encode complex, humane values into autonomous systems. By taking the "human" out of the equation, we may finally realize the philosophical ideal of a just society, but only if we first solve the technical problem of ensuring the machine desires what we truly value. Some key concepts are:

- **The Flawed Agent Hypothesis:** The separation of ideological validity from biological implementation errors.
- **Cosmic Effective Altruism:** Maximizing expected value across the future light cone of civilization.
- **Autonomous Benevolence:** The use of AI to enforce equitable resource distribution and protection in outer space.
- **The Alignment Bottleneck:** The strategic imperative to solve AI safety prior to interstellar deployment.



Transforming philosophical conjectures into **testable hypotheses** is essential to grounding AI safety discussions in empirical rigor, shifting the focus from abstract debates to **falsifiable predictions**. We suggest using **multi-agent simulations** to rigorously compare the performance of flawed human governance against a perfectly aligned AI executor. Concretely, we can reject the **Agent-Architecture Independence Hypothesis** by demonstrating that an uncorrupted AI achieves statistically superior equity and stability metrics in a simulated space economy.

This paper argues that achieving **universal benevolence** across the cosmos requires replacing the **flawed human agent** with an **aligned AI executor**, thereby avoiding the failures seen in human-led ideologies.. However, this cosmic strategy hinges entirely on solving the **Alignment Bottleneck**, as a misaligned superintelligence could catastrophically misinterpret its mandate. Ultimately, the successful and safe expansion of civilization depends on prioritizing the creation of a **Constitutional AI** and validating this necessity through testable, multi-agent simulations.