Project Proposal – Catastrophic risks of AI in space

Advances in edge AI, distributed inference, and cognitive cloud computing are making it feasible to delegate the control of satellites, spacecraft, and telecommunications networks to AI systems operating in space rather than from ground control centres on Earth. Coupled with 6G non-terrestrial networks, this shift promises to remove the latency and bandwidth bottlenecks that currently constrain ground-based control, and will extend from Earth orbit to cislunar space as the Artemis programme establishes a lunar communications and navigation hub. This project wants to prove that the same architecture that delivers these efficiency gains also concentrates a novel and underexamined surface of catastrophic risk, precisely because it places increasingly autonomous, capable systems beyond the reach of timely human intervention.

The project investigates four interacting risk pathways. First, and most centrally, **loss of control**: as AI is delegated authority over critical orbital and cislunar infrastructure, physical inaccessibility, communication delay, and operational autonomy may erode the ability of human operators to correct, override, or shut down misbehaving systems, turning ordinary alignment and reliability failures into events that are difficult or impossible to control. Second, **cyberattacks**: space-resident AI models become high-value targets, where adversarial manipulation, model exfiltration, or compromise could be used to commandeer satellites and disrupt global communications. Third, **power concentration risks:** space is fundamentally unregulated and we currently lack a sufficiently strong governance structure to avoid power concentration of AI in space. This could happen in different ways, the most concerning one being an AI with secret loyalties capable of taking over after staying dormant for years. With the recent merger of xAI and SpaceX, these concerns appear to be motivated. Lastly, **the role of a potential AGI in space and the possible catastrophic risks on Earth**. A collective and edge AI framework in space could be impossible to turn off, and in case of a misaligned AGI this would cause new and unexplored risks.

Why this project?

Based on my previous work and my knowledge of the field, I know that these technologies are currently being implemented and represent the future of space governance. Autonomous AI is already present in many Earth observation satellites, and will progressively assume prominent roles in space exploration. The migration of control from ground stations to onboard, increasingly autonomous AI is not speculative: ESA's "Cognitive Cloud Computing in Space" programme, the PhiSat-2 mission, the AI-eXpress constellation, and JPL's space-edge-computing work are all building exactly this architecture today, and the Artemis programme will extend it to cislunar space through a Moon-orbit communications and navigation hub. The capabilities that create the risks in this proposal are being designed and launched now, which makes this a question about the near-term trajectory of a real system rather than a distant hypothetical.

The agencies and companies driving this work are explicitly motivated by latency reduction, autonomy, and commercialisation (e.g. ESA Agenda 2025's market-creation goals, the "satellite-as-a-service" model). Safety work that does exist, such as fault detection and collision avoidance, is narrow and reliability-focused rather than concerned with loss of control, adversarial compromise of space-resident models, or strategic stability.

According to my initial literature review, space-systems engineers, cybersecurity researchers, and space-policy analysts have each touched parts of this problem, but it is largely absent from the AI-safety and catastrophic-risk literature where loss-of-control and frontier-risk thinking lives. My intention is to bridge this gap applying conceptual tools of AI safety, such as loss of control, alignment problem, and shutdown resistance, to a domain where these tools have not yet been systematically applied.

I believe that my work can be of interest to practitioners from various fields: space researchers, AI safety scholars, policymakers, and space-related industry. Given the timing of the incoming Moon missions, ESA and NASA would be a specific target audience, which is unusual for AI safety. Moreover, SpaceX is currently to only private company merging space exploration and AI so they would be another target audience for the work.

Potential deliverables

I am aiming to publish this work either as a blog in a think tank or a preprint on arXiv. I am currently in contact with leading think tanks in the field, and they expressed great interest in this work. There is also the potential to present this work at conferences.

Who am I looking for?

I am looking for two or potentially three well-motivated mentees. Ideally, one of them would have a strong scientific background (physics, engineering, etc.), and the other(s) would come from a policy background (or be interested in), economics, political science etc. This would be a core project of mine, so I am looking for people bringing new ideas and who are very interested in the topic. Having space-related knowledge is not mandatory.

What I can offer

I have several years of experience supervising people. In physics, I was team leader for a paper on quantum computing which was published in a prestigious peer-reviewed journal, and another paper on edge AI for physics now submitted to peer review. I am currently principal investigator for a big project with 15 people.

I based my success as a team leader on the success of the work that I collaborated on. So far, I supervised four PhD students: all of them wrote the research they have done as a chapter of their PhD thesis. I supported one of them in obtaining a prestigious postdoctoral position in the UK. I also supervised four other 4 undergraduate students. One of them presented our work at an international conference, and I supported their application for a Master's in one of the most competitive programmes in the UK, which was successful. I intend to support the professional growth of the mentees even after the fellowship.

My mentoring style is one meeting of about one hour per week plus another two years divided between a potential second meeting and draft corrections. I am normally very reactive on slack/email. I set directions, but I support independent ideas.