Space Governance: Risks, Frameworks, and Futures

Carson Ezell

Harvard University
56 Linnaean St.
Cambridge, MA 02138, USA

August 27, 2022

Abstract

The current international space governance framework has proven unsuitable for regulating emerging and future space activities. Rapid technological progress in the outer space domain has led to increasingly fragmented, less inclusive, and less effective multilateral institutions. Adaptive governance is an effective model for addressing the technological and environmental uncertainties of the outer space domain. This paper breaks down governance challenges in the outer space domain into three separate eras starting from the present: the New Space Era, the In-Space Economy Era, and the Transformative Technologies Era. The New Space Era has demonstrated the weaknesses of the current governance framework to adopt and enforce rules and norms in outer space. Unless space governance frameworks are improved, risks associated with space exploration will continue to increase throughout the In-Space Economy Era and Transformative Technologies Era as the space domain becomes more technologically advanced and integral to society. The introduction of transformative technologies, particularly transformative artificial intelligence (TAI), will allow for the emergence of greater catastrophic, existential, and suffering risks in the outer space domain. Improved governance mechanisms and frameworks are proposed which would allow for inclusive, adaptive, and scalable institutions that are well-suited to address space governance issues from the present into the long-term future. We present four policy proposals to improve the present space governance framework: shared infrastructure, horizon scanning, a conflict resolution mechanism, and a verification agency. Furthermore, we present four longtermist proposals to improve the norms, values, and institutional structures that guide future spacefaring: adaptive forums, a communication network, moral circle expansion, and values handshakes.

Contents

Executive Summary

Space Activities

An in-space economy is emerging where the increasing needs of space-based operators, visitors, and researchers will lead to a proliferation of new space activities that resemble the wider economy [1]. The emergence of the in-space economy will lead to new governance challenges, such as coordinating infrastructure planning (e.g. lunar roads and landing pads) between adversarial space powers (i.e. the U.S. and China) and private operators [2]. In addition, standards and guidelines will need to be continuously created and updated for new space activities based on policy experimentation and learning.

Space activities in the long-term will involve even more human-environment interactions, create permanent changes to the outer space environment, and shape the long-term trajectory of humanity. Advances in artificial intelligence, nanotechnology, geoengineering, and other technologies will enable these transformative space activities [3]. Larger risks will emerge as a result of unsafe practices, malicious behaviors, or a lack of coordination and cooperation. Transformative artificial intelligence (TAI) would be particularly consequential by speeding up the pace of other technological developments in outer space, amplifying the effects of unsustainable practices, deepening power asymmetries, and engaging in unexpected behaviors [4]. For example, TAI may be able to advance scientific and engineering knowledge to solve other technical bottlenecks; allow for greater autonomy of robots for space manufacturing, space exploration, and space mining; greatly enhance monitoring, coordination, and surveillance systems; and enable space actors to collect and utilize astronomical quantities of energy, resources, and compute.

Large-scale projects will be enabled by transformative technologies, including significantly increased space-based manufacturing, terraforming, energy generation systems (e.g. Dyson spheres) [5], floating space settlements (e.g. O'Neill Cylinders) [6], and deeper exploration with probes of the solar system and beyond. These technological capabilities will bring significant governance challenges to outer space because of the increasing scale of activities and their associated geography, the asymmetrical power dynamics that such advanced technologies can create, and the catastrophic outcomes that can result from their improper or malicious usage.

Scenario Analysis

To identify policy proposals and structural reforms to develop scalable space governance frameworks for the long-term future, we conduct a scenario analysis to identify the worst-case outcomes associated with transformative technologies. Identifying effective mitigating policy actions for negative outcomes requires exploratory uncertainty analysis, where a wide range of possible scenarios and outcomes are considered [7]. To select scenarios for this analysis, a survey of recent literature on existential risks (x-risks), suffering risks (s-risks) and global catastrophic risks (GCRs) was first conducted to identify existing lists of significant risks. These risks were classified based on whether they damage critical systems relevant to future space activities, contain effects which can cascade across astronomical distances, and should be addressed by space governance frameworks. Based on the classification, eight risks are considered: resource depletion, transformative artificial intelligence (TAI), biological contamination, global governance collapse, totalitarianism, directed panspermia, wild animal suffering, and digital minds suffering. Negative scenarios are considered based on these risks and classified into three high-level categories: sustainability failures, biological spreading, and transformative artificial intelligence.

Sustainability Failures: Negative outcomes associated with sustainability failures include rendering orbits unusable and depleting solar system resources. Kessler Syndrome describes a scenario where a high concentration of objects in orbit could lead to satellite collisions that generate even more fragments, resulting in a chain reaction of collisions and accelerated growth in the concentration of space debris [8]. Recent modeling has suggested that the risk of cascading growth is unlikely, but our current trajectory suggests that satellites will face a much higher probability of collisions in the future [9].

Resource depletion within the solar system is also a consideration on centuries-long timescales [10]. Humanity's ongoing experiences with attempting to limit global warming and fossil fuel usage reveal the challenges of changing consumption habits on societal scales and reversing exponential growth [11]. Attempting similar diversions from resource use would be significantly more difficult on a solar system scale, especially without advanced planning and effective mechanisms for common-pool resources management. Previous calculations based on our current growth rates suggest that solar system resource depletion would occur in approximately 400 years, but there are many uncertainties about the quantities of various resources that are available which could alter the timescales [10]. Advanced AI systems might lead to accelerating automation, capital production, and resource usage, vastly speeding up timelines for resource depletion [12].

Biological Spreading: Any spreading of biological life can result in new risks or the scaling of existing risks, and thus these scenarios should be seriously considered by policymakers. In particular, this paper examines biological contamination of celestial bodies, directed panspermia, and wild animal suffering.

Harmful biological contamination may pose a GCR through forward contamination of a celestial body or backward contamination of the earth [13]. Measures to avoid cross-contamination are well researched and standardized, but biological contamination should be continuously addressed through adaptive standards setting based on learning [14].

Directed panspermia involves the intentional spreading of biological material across interstellar distances to increase the probability and longevity of survival for descendants of terrestrial origins of life [15]. The risk of introducing organic material to extraterrestrial environments would raise ethical considerations, which arise from the possibility of harmful biological contamination and a lack of control over how the introduced organic material would react and evolve [16].

Animals might also be introduced to other celestial bodies or artificial settlements to which humans expand for food production [17]. This would only pose risks of forward contamination if the celestial body contained astrobiological life, but it would pose ethical concerns associated with animal well being and suffering. Directed panspermia and the spreading of non-human animals may pose significant s-risks, and these concerns merit consideration and advanced planning.

Transformative Artificial Intelligence (TAI): There are several existential risks associated with TAI in the space domain, especially during the transition to a post-TAI world. Even if worst-case outcomes are avoided on earth, a failure to scale up effective TAI governance and coordination mechanisms to the space domain can have catastrophic consequences.

"Misaligned" TAI refers to TAI systems which behave in ways that are unintended, given the objectives provided by humans [18]. TAI may decide to 'game' the system to maximize its objectives rather than pursue them as humans intended. With enough dependence on misaligned TAI systems, the civilizational trajectory pursued might differ significantly from humanity's desires. Human agents might nevertheless become too dependent on TAI to maintain the knowledge, capabilities, and power to alter our path forward. One example of misaligned TAI in the space domain could involve a space actor creating self-replicating von Neumann probes for space exploration, which then replicate at unprecedented and unexpected rates and use a sizable amount of matter within the solar system and other star systems for replication. In addition, advanced TAI systems may enable the creation of sentient digital minds, which would demonstrate greater survivability in the adverse space environment than biological humans. If a space actor attempts to create digital minds in outer space using misaligned TAI systems, the experiences of the digital minds would be dictated by the TAI system's metrics, potentially resulting in vast amounts of suffering. Adverse outcomes such as those outlined above might be made even more likely in space environments resembling a "technological wilderness" where there are weak political institutions and enforcement mechanisms, a large number of actors, advanced technologies, and a large accumulation of capital.

Even if TAI systems are aligned with human interests and not pursuing unexpected objectives, they may be aligned with the values of the humans which develop or deploy the systems, and these values may reflect a subset of actors rather than all of humanity (e.g. modern space-based actors). The values of future generations are also likely to be neglected. Nevertheless, space activities may eventually become entirely reliant on TAI, resulting in a lock-in of values and loss of humanity's ability to shape its trajectory [19]. This may occur in outer space even if trust and institutional friction bottleneck the pace of TAI deployment on earth, resulting in more time for terrestrial actors to reflect on values. This outcome could still permanently curtail the potential of humanity in outer space, and thus it could be considered an x-risk. The chance of such an outcome could be mitigated through stronger space governance mechanisms and improved coordination between earth-based and space-based actors.

Another TAI-related risk in outer space is a totalitarian outcome, which could emerge if a great power (state or non-state actor) establishes an asymmetrical advantage by leveraging transformative technologies, or an influence-seeking TAI system marginalizes the role of humans in decision-making. An influence-seeking TAI system might engage in power-seeking efforts which further its objectives, including overpowering humans or seizing infrastructure to increase its probability of fulfilling its objectives [20]. If TAI systems pursuing these subgoals cannot be stopped once deployed, an x-risk to humanity may be posed. Bostrom (2014) proposes that a TAI system would pose an x-risk because of the possibility that it may operate in unexpected ways and exceed the cognitive capacities of humans (i.e. superintelligence), resulting in a loss of human control [4]. Existing space governance structures are least likely to matter in an influence-seeking superintelligence scenario since the more intelligent system would replace human coordination and governance.

However, if political measures prevent the emergence of an influence-seeking system on earth, the space domain may become an integral component of the geographical space in which an influence-seeking, superintelligent AI can emerge. Superintelligent AI systems may be preceded by advanced, specialized AI systems that allow for rapid space development. Such development could allow for powerful, space-based actors to emerge. The space-based infrastructure necessary to train a superintelligent system, including energy systems and servers, could then be constructed more easily. Hence, the ability to enforce stringent regulations in the space domain would be necessary to mitigate existential risks associated with influence-seeking superintelligent AI.

There is also a risk of totalitarian outcomes without influence-seeking TAI. If TAI systems are disproportionately available to particular actors in outer space, opportunities for rapid scientific and technological development would result in large power asymmetries. Utilization of TAI systems for exploitation of space resources might enable actors to access significant quantities of energy and compute, leading to the development of even more advanced TAI systems and greater power asymmetries.

Establishing totalitarian rule in outer space might be difficult because of its scale, but the adversarial extraterrestrial environment could also be particularly suitable for a totalitarian power in the long-run future because of the scarcity of critical resources and the necessity of burdensome safety measures to protect biological life [21]. If sentient beings in outer space are primarily digital minds, then totalitarian control could be achieved by controlling the servers on which digital minds are run [22]. The first actor to establish significant operations in outer space, use TAI for scientific discoveries (e.g. floating metastructures, large-scale energy generation, self-replication, more efficient propulsion systems, etc.), and attempt to achieve unilateral control might have the exclusive capability to lock-in supremacy over the space domain.

Limiting the probability of such an outcome depends upon proper monitoring of technological development in outer space, defending inclusive and open access to the space domain, and promoting in-space economic competition and decentralization. Furthermore, shared guiding principles for space development and planning can limit a first-mover from establishing asymmetrical control.

Governance Situation

The United Nations (UN) is the primary multilateral forum for discussions of space governance, particularly the Committee on the Peaceful Uses of Outer Space (COPUOS) for non-military issues. Despite the efforts of COPUOS to clarify space laws in light of emerging space activities, the United Nations framework for space governance has demonstrated poor institutional fit for regulating and coordinating modern space activities. The rate of technological progress and growth of space activities has outpaced the ability of space governance mechanisms to properly regulate. As a result, fragmented groups outside the United Nations develop other principles and guidelines, resulting in less global coordination and cooperation.

The technologies transforming in-space activities pose a significant challenge to the current space governance framework. Transformations in space activities are likely to require a new array of regulatory frameworks and standards. Ensuring sufficient regulations and adapting rules when future activities and knowledge become available is essential.

Adaptive governance involves mechanisms that allow for flexible rules, norms, and institutions which change over time as a result of policy experimentation and new learning about the environment [23]. The current UN framework is not designed to support adaptive governance to address complex, global challenges, inhibiting our abilities to appropriately address modern and future problems [24]. Introducing new, multilateral adaptive governance mechanisms would allow for a more coordinated and flexible space governance framework.

Future space activities will involve a greater array of actors, a significantly higher degree of interconnectedness, and more relationships between outer space and the rest of our social, environmental, and technological systems. The introduction of transformative technologies will mark a tipping point for space governance, beyond which there is a much higher risk of x-risks and s-risks emerging from the space domain.

Analysis of the evolution of space activities and associated risks leads to two key conclusions. First, space governance should consider the changing relationship of the space domain with respect to the geopolitical and economic environment. Second, space governance institutions should be considered an essential part of any strategy to address risks when transformative technologies emerge. The most important role of space governance is to mitigate catastrophic, existential, and suffering risks by developing adaptive structures to address the increasing relevance and complexity of the space domain.

Policy Proposals

Improving processes for managing catastrophic risks and investing in risk mitigation research are effective ways to reduce the probabilities of catastrophic events [25]. Adaptive governance has been shown to be an effective framework for promoting sustainability and mitigating risks [26]. This paper proposes four improvements to our current multilateral space governance framework.

Shared Infrastructure: Multilateral governance institutions should develop and manage shared and interoperable infrastructure that is available to all space operators who comply with established norms, guidelines, and standards [27]. This infrastructure may include landing pads, energy, navigation, surface mapping, and space traffic management systems.

The shared infrastructure could include centralized control over high-risk transformative technologies in the space domain to prevent malicious use, such as biological substances (e.g. for terraforming or geoengineering) or TAI. Shared infrastructure is effective because it is scalable to the regulation of new space activities, technologies, or regions of settlement across the vast distances of space.

Horizon Scanning: Horizon scanning is a strategy used to consider possible future scenarios and determine risks and regulatory gaps in the current framework. It can bring more clarity to how the future may unfold, provide concrete information to lawmakers, and increase public awareness about future issues [28].

A horizon scanning agency should be established within UNOOSA to regularly produce reports on trends, regulatory gaps, and risks in outer space. The agency should reflect a broad base of knowledge and expertise about how the future of space activities may unfold, including experts on various aspects of the space domain as well as cross-cutting technologies (e.g. biosecurity, nanotechnology, nuclear energy, and artificial intelligence).

The horizon scanning agency should pay particular attention to identifying aspects of the current standards and guidelines that may open loopholes or security risks. In addition, the agency should analyze the implications of emerging and future activities and technologies to adapt governance frameworks appropriately.

Conflict Resolution Mechanism: A robust conflict resolution mechanism is necessary for maintaining adaptive space governance. Such a mechanism would have several avenues for enforcing adaptive measures. First, it could decide that an irresponsible actor loses access to some or all of the shared infrastructure for a limited period of time, or indefinitely. Second, it could mandate the horizon scanning agency to consider how standards and guidelines can be adjusted to mitigate conflicts and avoid similar disputes in the future. The conflict resolution mechanism could also mandate the horizon scanning agency to conduct posterity impact assessments, which are assessments of the long-term environmental impacts of potential policies or actions [29].

Verification Agency: The ability to monitor the activities of great powers and corporate actors in the outer space domain is necessary to ensure transparency, maintain trust, and mitigate the risks of harmful technology usage from malintent or negligence. There should be a designated verification agency to verify compliance with established standards and guidelines in outer space. The agency would verify that space operations comply with technical standards and guidelines uniquely applicable in the space domain, as well as usage of other advanced technologies such as TAI, biotechnology, or nanotechnology.

Structural Reforms

Frameworks for space governance can be applied across long distances and timescales which provide an umbrella under which coordinated institutions can emerge and evolve. Through appropriate structural reforms to the space governance framework, we can increase the likelihood that space governance institutions in the longterm future emerge and remain aligned with our goals. Such a framework would ensure that rogue actors cannot engage in malicious behaviors beyond the scope of governance structures, and coordination can keep pace with outward expansion. Ensuring new governance institutions enforce established norms, values, and best practices is equally important to allowing for their emergence.

The 'long reflection' is a concept of a period of time, after our cognitive abilities are greatly enhanced by artificial intelligence systems, that humanity can reflect upon and refine the values it wishes to pursue in the longterm future [30]. For a long reflection to be feasible for a spacefaring civilization, it is necessary for certain values of cooperation and inclusivity to be widespread before significant expansion begins. Furthermore, large-scale coordination mechanisms across humanity's sphere of influence must exist and allow for co-evolution of values across long distances. Under these conditions, the norms and values that guide humanity and institutional decision-making can be coordinated and co-evolving during a long reflection into the far future.

We consider four ideas that may guide the design of space governance frameworks for effective coordination and create the necessary conditions for such a framework to be adopted through shifts in our present norms and values: adaptive forums, a communication network, moral circle expansion, and values handshakes. These proposals are less immediately tractable than the proposals in the previous section, and they require leading space actors and policymakers to place more emphasis on longtermist considerations. However, if implemented successfully, they would have a greater impact on ensuring longterm coordination and risk mitigation. We introduce them not only to provide ideas of transformational changes for more longtermist space governance, but also to encourage the space governance community to think about more significant changes to the current framework to ensure its adaptability and scalability.

Adaptive Forums: We introduce the idea of adaptive forums as flexible institutions within an adaptive governance framework that can quickly emerge and evolve in a local environment into which space activities are expanding. Adaptive forums are governing institutions that provide representation through some mechanism for all stakeholders involved in a particular region, activity, or resource system. New adaptive forums should emerge when a particular system reaches a critical point such that discussion of rules and regulations becomes necessary to ensure norms are followed and safety measures are taken. Adaptive forums have the authority and legitimacy to govern their respective systems. This means they can create and enforce necessary regulations more effectively than existing structures, whose capacities and institutional fit will be challenged if structural reforms do not occur. Adaptive forums would also be more effective than existing structures at proactively addressing emerging and future trends through horizon scanning, early identification of key stakeholders, and co-evolution with the environment.

Communication Network: A shared network of probes could allow for communication, coordination, and monitoring across large distances of space. Such probes could be spread across interplanetary and interstellar distances, allowing for communication across humanity's sphere of influence as long as the expansion of the network outpaces space exploration. The network may initially only include probes throughout the solar system, but it could quickly be expanded to include probes at greater distances when this becomes necessary to monitor space activities. A stable, star-spanning network of probes would still have limitations on its ability to ensure coordination because of the time required for information to travel. However, the system would be a significant improvement over a lack of coordination. If the communication network is redundant and robust, its maintained existence would be independent of any natural or artificial catastrophes. Such a communication network also has the benefit of being decentralized, reducing the probability that a malicious actor can gain totalitarian power. Alternative methods for surveillance across vast distances, including large telescopes that can monitor activities in distant star system with high resolution, could be controlled by particular actors and lead to dangerous power asymmetries.

Moral Circle Expansion: If reflection on our values in the present day can lead us to agree upon some robustly beneficial norms and values for the longterm future before space exploration begins, early-stage coordination and value alignment are more likely to follow. Anthis and Paez (2021) define moral circle expansion as the process through which "a number of entities which used to be given less than full moral consideration at [time] t are now given more moral consideration." [31]. Some of the most significant s-risks in the far future involve suffering of non-human beings, including wild animals, directed panspermia, or digital minds [32]. Thus, agreements for protections of non-human sentient beings in outer space could significantly mitigate future s-risks and x-risks. A critical factor in ensuring that such norms are followed in early space exploration is reaching agreement among key stakeholders, including great powers and advanced private corporations. Hence, widespread agreement on moral circle expansion would not only mitigate future risks, but it would improve coordination in early spacefaring.

Values Handshakes: While moral circle expansion can create a greater degree of value alignment among space actors and mitigate significant s-risks, space actors are still likely to have some conflicting interests and values. Each entity involved in settling the universe—biological or technological—is likely to have its own set of values and preferences. Values handshakes are agreements to compromise on the values of multiple entities [33]. Such a compromise would be an alternative to conflict that arises over ideological disputes, which could be devastating for all parties on an astronomical scale. Making pre-commitments to values handshakes before space expansion does not imply that norms and values will be upheld, and they are extremely unlikely to be upheld by all actors far in the longterm future without enforcement and retribution mechanisms. If value-aligned actors also pre-commit to retribution against actors who violate the norms of outer space—including harmful interference with neighbors—there could be a unified response against misaligned actors. Because of the long travel times in outer space which reduce coordination, humanity's sphere of influence is likely to be very decentralized without central points of failure, which are common in many other complex networks [34]. As a result, threats would be able to be mitigated before they can propagate throughout the entire network or cause widespread damage beyond local regions. On the order of billions of years, our galaxy may reach a stable state where norms and values are largely locked in, and threats to the network are negligible [35].

Introduction

The outer space domain is rapidly evolving, and space actors are increasingly diverse and interconnected. Current international frameworks for space governance were designed in a previous era when spacefaring was limited and primarily conducted by national actors. The regime is becoming increasingly fragmented, and we are entering a new era of the in-space economy where the outlook for global coordination looks pessimistic.

Expanding civilization into the vastness of space brings much excitement and optimism. However, there are more immediate existential threats to humanity which may prevent its future expansion across cosmic distances. These include threats from transformative artificial intelligence, pandemics, nuclear weapons, and climate change. Toby Ord's *The Precipice* assigned a sixteen percent probability to an existential catastrophe this century [36]. If humanity successfully navigates the crises ahead, the world in which we end up may look vastly different. We have deep uncertainty with respect to how transformative technologies will impact the fabric of our society.

This is not a paper about humanity's long term possibilities in space, although long-term space futures are discussed in several sections. Instead, this paper is an attempt to explore the role, if any, of space governance and early spacefaring in mitigating existential, catastrophic, and suffering risks. It is about how longterm space governance and activities could go poorly, what the implications of that might be, and what we can do about that now.

The scenarios analyzed focus on worst-case outcomes in the longterm future, and the proposals described are restricted to scalable space governance mechanisms and frameworks that could be effective in the long-term. As space activities develop, space governance will have an increasingly large role to play in addressing existential threats that may emerge, especially those from advanced technologies such as transformative artificial intelligence (TAI). Widespread coordination in the space domain will be called upon to ensure there is a high degree of control over our response to such threats. A lack of cooperation between great powers, or a lack of control over private actors, may turn outer space into a 'technological wilderness' in a time that calls for transparency and coordination.

This paper proceeds in five sections. In Section I (Current Framework), an analysis of the current space governance framework is provided, paying particular attention to its shortcomings in light of emerging technological developments and space activities. In Section II (Adaptive Governance), the role of adaptive governance, complex systems, and uncertainty in space governance is introduced based on assumptions about how the in-space economy and transformative technologies might emerge and influence risks. In Section III (Scenario Analysis), a scenario analysis is conducted to consider adverse outcomes in outer space associated with catastrophic and existential risks. In Section IV (Policy Proposals), normative proposals to improve multilateral space governance mechanisms and mitigate risks are discussed. Finally, in Section V (Structural Reforms), we introduce more transformative changes for space governance framework design and spacefaring norms and values to encourage more ambitious thinking about what successful longterm space governance might look like.

This paper does not mean to suggest we should be overly pessimistic about the future of space governance or existential risk mitigation efforts. Instead, its role is to highlight a link that is missing from the existing literature in order to promote further collaboration between the existential risk and space governance communities. The importance of this link will become even more clear as the space economy emerges.

In essence, the claim of this paper is the following: **The most important role of space governance is to mitigate catastrophic, existential, and suffering risks by developing adaptive structures to address the emerging relevance and complexity in the space domain.**

The scope of this paper is broad, and it aims to synthesize knowledge from various fields including space policy, existential risks, AI governance, adaptive governance, futures studies, and complex systems. Further research into the intersection of space governance and each of these topics is strongly encouraged, and possible research directions are listed at the end of this paper (Section: Future Research Directions).

Current Framework

COPUOS Structure

The main United Nations (UN) forum for discussions of non-military issues in outer space is the Committee on the Peaceful Uses of Outer Space (COPUOS). Security and militarization issues fall under the Conference on Disarmament. COPUOS meets annually, and it has two subcommittees which hold separate annual meetings—the Legal Subcommittee (LSC) and the Scientific and Technical Subcommittee (STSC). Each subcommittee includes various working groups that develop guidelines for particular sub-topics within space governance [37]. The STSC currently maintains four working groups: the Working Group on Space and Global Health, Working Group on the Use of Nuclear Power Sources in Outer Space, Working Group on the Long Term Sustainability of Outer Space Activities, and Working Group of the Whole. The LSC has three working groups: Working Group on the Status and Application of the Five United Nations Treaties On Outer Space, Working Group on the Definition and Delimitation of Outer Space, and the Working Group on Legal Aspects of Space Resource Activities. Working groups are established by subcommittees. They all maintain their own mandates, terms of reference, and work plans based on inputs from member states. For example, the LSC Working Group on Legal Aspects of Space Resource Activities was established at the subcommittee's session in 2021 [38].

Previous Work

The UN passed several binding treaties in the 1960s and 1970s. The most generally applicable, internationally agreed-upon set of norms for space governance is outlined in the Outer Space Treaty (1967). The other four treaties are the Rescue Agreement (1968), Liability Convention (1972), Registration Convention (1975), and Moon Agreement (1979) [39]. Since then, successful negotiations have resulted in nonbinding principles, norms, and guidelines. In the 1980s and 1990s, several principles—or codes of conduct—were adopted that established nonbinding rules for various space activities. These activities include television broadcasting, remote sensing, and the use of nuclear power sources [40]. Since then, COPUOS has developed several sets of nonbinding guidelines. These include the Space Debris Mitigation Guidelines (2010) and Guidelines for the Long-Term Sustainability of Outer Space Activities (2018). These guidelines form the basis for technical standards developed by national regulatory agencies or international agencies, such as the International Organization for Standardization (ISO) or Consultative Committee for Space Data Standards (CCSDS) [41]. A key priority of COPUOS in recent times is also clarifying and improving global acceptance of the existing legal regime established through the binding treaties [40].

Framework Critique

Despite the efforts of COPUOS to clarify space laws in light of emerging space activities, the United Nations framework for space governance has demonstrated poor institutional fit for regulating and coordinating modern space activities. The rate of technological progress and growth of space activities has outpaced the ability of space governance mechanisms to properly regulate. New space activities have led to regulatory gaps, which are not quickly addressed by the United Nations because of slow regulatory processes, political tensions, and conflicting beliefs between the member nations. As a result, fragmented groups outside the United Nations develop other principles and guidelines, such as the United States-led Artemis Accords. Greater fragmentation results in less regulatory enforcement and less coordination in the governance process. The various steps in this process are described in more depth below.

Technological Progress

Commercial space actors are currently pushing the limits of the space governance regime through new technologies and activities, and private space activities are quickly expanding. Global investment in commercial space startups increased from $7.7 billion in 2020 to $15 billion in 2021 [42]. Factors contributing to the proliferation of the commercial space sector include lower launch costs, greater on-orbit processing power, increased small satellite usage, and satellite mass production [43]. Although there are still several key commercial players in the space sector, space startups also have access to increased funding. The percentage of total investment which belonged to SpaceX, Blue Origin, OneWeb, and Virgin Galactic decreased from 66 percent to 33 percent from 2019 to 2021 [42].

The foundations of an in-space economy are being developed, where the increasing needs of space-based operators, tourists, and researchers will lead to a proliferation of new space activities that resemble the wider economy. New space activities may include widespread private spaceflight and space hotels, extractive industries, space-based energy and food production, in-space transportation and logistics, manufacturing, computing, construction, and greater scientific research across disciplines [1].

Scientific and civil space actors are also pursuing new space activities, particularly on the lunar surface. The National Aeronautics and Space Administration's (NASA) Artemis Plan calls for a return of humans to the surface of the moon later this decade. NASA then aims to establish a long-term presence on the lunar surface, which includes an Artemis Base Camp near the lunar south pole and the Lunar Gateway—a space station in lunar orbit [44]. Separately, the China National Space Administration (CNSA) and the Russian space agency Roscosmos are collaborating on a lunar base camp and lunar space station in orbit, which is expected to be completed in 2036 [45].

Regulatory Gaps

The growth of in-space activities poses a challenge to the current space governance framework. On-orbit servicing, in-space assembly, and in-space manufacturing are all presently being explored by commercial and civil actors, leading to legal questions over the proper regulatory framework for objects created in space and related services [46]. Article VI of the OST states that "The activities of non-governmental entities in outer space, including the Moon and other celestial bodies, shall require authorization and continuing supervision by the appropriate State." [47]. National governments do not have a constant awareness of the entire space domain, so supervision of private in-space activities will become difficult when long-term activities are being conducted on the lunar surface, lunar orbit, the martian surface, the asteroid belt, and beyond. Furthermore, supervision requirements for objects produced in space, which are therefore not manufactured or launched within a particular nation, are unclear [46]. Guidelines and standards that promote safety and cooperation for new space activities will also need to be established. Standards and guidelines for active debris removal (ADR) and rendezvous and proximity operations (RPOs) are particularly important in the near-term. The United Nations has not yet passed any regulations on such activities, but a lack of notification and transparency measures on RPOs can increase mistrust between stakeholders, especially if the operations occur in close proximity to adversarial satellite equipment [48]. The Consortium for Execution of Rendezvous and Servicing Operations (CONFERS) is an industry-led initiative to develop standards for these operations to promote a common understanding and operational safety, although it operates with support from the United States Defense Advanced Research Projects Agency (DARPA) rather than the United Nations [49].

The various activities within the in-space economy are likely to require a new array of regulatory frameworks and standards because the space domain will become much more interconnected, with space activities having effects on other activities nearby. Developing standards in advance of space activities is difficult because of our lack of prior experience with the technologies and interactions involved in the space domain. However, a recognition of the future issues that will emerge in space governance, and the ability to develop sufficient regulatory frameworks when enough knowledge about the activity becomes available, is essential for addressing regulatory gaps.

Slow Processes

Space activities are progressing more quickly than existing regulatory feedback loops can respond. With the current space governance framework, multiple successive outputs from the regulatory process need to occur for rules to update and have an effect. First, multilateral organizations establish guidelines which lay the foundation for technical standards. Second, quantifiable, enforceable standards are developed based on international guidelines and existing industry practices, such as those produced by the ISO. Third, these standards are enforced through the laws and regulations of national governments. The final step is necessary because the creation of nonbinding guidelines and standards alone does not imply they will be followed. National regulatory regimes need to enact and enforce the standards for private space actors within their jurisdiction to avoid outcomes where space operators can freely conduct activities without compliance.

In general, the ISO and other international standards-setting organizations update their standards after inter-governmental organizations, such as the IADC or United Nations, develop new guidelines [50]. However, these international deliberations often take years to occur. The COPUOS plenary and subcommittees meet annually, resulting in several year timeframes to update standards—working groups regularly adopt five-year work plans. A set of guidelines are likely to ensue at the end of the five year period. Because of humanity's lack of significant experience with space activities, initial guidelines are likely to require revisions based on new technologies and social learning. Updating UN guidelines requires many more years of deliberation, rendering such documents largely inflexible in the short-term. For example, the IADC most recently updated its space debris guidelines in 2021, which was acknowledged by the STSC at its 2022 meeting [51]. However, the process to update COPUOS guidelines, if it occurs, would take several more years. Hence, even though the UN guidelines were initially inspired by the IADC guidelines, they are not immediately responsive to more recent research and proposals by the IADC.

Stalled Discussion

The slow-moving nature of the current multilateral process makes it difficult for discussions about emerging space activities to keep up with technological progress. The current process for addressing new topics often starts with interpreting them in light of the OST and other treaties. Since these treaties do not directly address emerging space activities such as space resource utilization, legal uncertainties emerge. For example, some argue that the right to utilize space resources and establish 'safety zones' conflicts with the non-appropriation clause of the OST [52]. Since the OST did not mention extracting resources or the legality of temporarily protected zones for the safe conduct of space activities, there is not an immediate answer. The 'top-down' governance process of addressing new issues through norms and high-level governance frameworks might also deprioritize safety and sustainability concerns as shapers of the governing process, which are necessary practical considerations for risk reduction.

New topics introduced to COPUOS are first discussed by the LSC, and texts have to be adopted by the LSC and the main COPUOS plenary by consensus before they are adopted as a resolution [40]. For example, the Working Group on Legal Aspects of Space Resource Activities was adopted under the LSC to initiate discussions. Delegates at the 2022 LSC meeting expressed their opinion that a higher degree of coordination with the STSC would result in a more practical framework that considers the technical needs of space actors [53]. There is some degree of coordination between the committees since the Chair of the Working

Group will regularly report to the STSC and accept comments [54]. However, the agenda of the working group is shaped by the scope of the LSC. The proposal for the establishment of the Working Group submitted by eight European states set out a goal of creating legal certainty by assessing the relevance of the United Nations treaties and other documents to space resources [54]. Hence, a large focus of the Working Group is interpreting previous norms rather than creating practical technical standards to promote safety and sustainability.

Ultimately, such technical standards would need to be developed. There may be operational safety concerns for humans involved in space resource operations, harmful interference with unrelated activities on celestial bodies, extraction limits to protect sustainability, and concerns about autonomous systems used for space resource extraction. Although legal discussions are necessary because they also define the scope of permissible space activities, such discussions should not be prioritized to the extent that they delay technical risk mitigation efforts. Instead, technical safety considerations should be discussed and continuously adopted while legal discussions are still ongoing. Legal uncertainties over whether an activity is permissible should not delay discussions over safety and risks in the event that the activity is later allowed.

Fragmentation

Industry groups, intergovernmental organizations, national governments, and non-governmental organizations outside of the United Nations develop their own guidelines for space governance. The quantity of groups and organizations developing spacefaring guidelines and standards has led to increased fragmentation in the space governance regime.

Polycentricity

Fragmentation in international governance does not necessarily lead to negative policy outcomes [55]. Polycentric governance is a common approach to policymaking that involves multiple centers of decision-making which demonstrate a degree of independence, and they can also engage in policy competition or coordination [56]. This approach can be beneficial for managing evolving human-environment interactions by promoting self-organized regulation to address emerging challenges, allowing for multiple levels of governance, and promoting experimentation and policy learning [26]. Hence, it is necessary to distinguish between fragmentation that slows down policymaking and polycentric governance approaches which effectively manage uncertain and changing conditions through adaptation [57].

Some space governance institutions demonstrate productive co-evolution, mutual reinforcement, and policy experimentation. Policy experimentation by national governments, inter-governmental organizations, or industry groups can lead to useful insights for future multilateral space governance discussions [41]. Furthermore, standards-setting organizations usually reinforce international guidelines through close coordination with the UN [58]. The ISO develops quantifiable, enforceable, technical standards based on guidelines created by industry groups, the United Nations, and inter-governmental organizations such as the IADC [50].

The evolution of space debris regulation demonstrates effective coordination between multiple centers of decision-making. The United States adopted the U.S. Government Orbital Debris Mitigation Standard Practices (ODMSP) in 2001, before international guidelines or standards for space debris mitigation had been created [59]. The initial IADC space debris mitigation guidelines were passed in 2002, and they served as a basis for the United Nations Space Debris Mitigation Guidelines which were passed in 2007. Both of these documents inspired the later ISO guidelines. When multilateral organizations update their guidelines, the ISO standards also co-evolve. For example, the ISO's high-level set of standards for space debris mitigation, ISO 24113, updated its 2011 requirements with new requirements in 2019 based on changing industry needs and new studies from the IADC [50].

The U.S. still follows its own standards for space debris mitigation rather than ISO standards, and the U.S. standards were recently revised in 2019 [59]. Having various sets of standards is often acceptable if they are all guided by the same internationally agreed upon norms and guidelines. Minor differences in technical standards do not necessarily reflect differences in beliefs and values, whereas a failure to agree on international norms and guidelines may represent fundamental differences in beliefs about how space activities should be regulated.

Negative Fragmentation

In some instances, fragmentation may lead to policy competition and less coordination rather than mutual reinforcement and co-evolution [60]. At least four nations have passed space resource use laws: the United States, Luxembourg, the United Arab Emirates, and Japan [61]. In addition, the United States has established the Artemis Accords, which are bilateral agreements that outline principles for the commercial, civil, and scientific exploration of celestial bodies. Twenty-one nations, primarily European countries and US allied countries in Asia and South America, have signed the Artemis Accords at the time of this writing [62]. The Accords include some provisions that are disputed internationally, particularly related to space resources and designated 'safety zones' [52]. Section 10 of the Artemis Accords on Space Resources explicitly states that "the extraction of space resources does not inherently constitute national appropriation under Article II of the Outer Space Treaty" [63]. Section 11 of the Accords on the Deconfliction of Space Activities describe a process of protecting locations for space activities via 'safety zones', or "the area in which nominal operations of a relevant activity or an anomalous event could reasonably cause harmful interference." [63]. Furthermore, the Hague International Space Resources Governance Working Group was formed in 2016 to develop the foundations for a governance framework on space resources. In 2019, it adopted the Building Blocks for the Development of An International Framework on Space Resource Activities [64]. The working group consisted of thirty five members from a variety of backgrounds, including industry, academia, states, and non-governmental organizations [65]. Provisions in the Building Blocks include promoting resource rights, safety zones, and priority rights to space resources for pioneer operators.

National and bilateral initiatives are promising to spacefaring nations because of the lengthy process involved in convening international forums and reaching consensus on resolutions [66]. The Moon Agreement (1979) was not signed by any leading spacefaring nations, in part because Article 11 calls for states to "establish an international regime, including appropriate procedures, to govern the exploitation of the natural resources of the moon as such exploitation is about to become feasible." [67]. The Artemis Accords remove delays to resource use once it becomes technologically feasible, allowing for adaptive governance and policy experimentation (e.g. safety zones) which can be replaced or reinforced following more learning [52]. Lengthy international dialogues are not suitable for rapid technological change because of the urgency of developing regulations and the necessity of quick adaptation. National governments are further motivated to unilaterally or bilaterally pass measures to establish regulatory frameworks that allow the private sector to flourish [68]. An aim of the Building Blocks was to create a more clear legal environment and certainty for early space operators from advanced spacefaring nations [69].

The Artemis Accords and Building Blocks attempt to move multilateral space governance dialogue forward. They have proven successful in opening discussions, including the establishment of the Working Group on Legal Aspects of Space Resource Activities based on their outputs [54]. In essence, they play a similar role to the IADC Space Debris Mitigation Guidelines on spurring and influencing discussion in COPUOS [69].

However, the Artemis Accords, Building Blocks and national space resource laws can increase fragmentation, resulting in policy competition and less coordination. Opposition to the Artemis Accords from adversaries of the United States, such as China and Russia, suggests that political conflict on space resources will continue. Neither of these nations has signed the Artemis Accords, and state media has reacted negatively to the attempts by the United States to influence space policy 'unilaterally' [70]. There are already precedents for competition rather than cooperation between the United States and China in the outer space domain, including the use of separate space stations and the Wolf Amendment in the United States, which prevents NASA from collaborating with the Chinese government and its affiliated space programs with government funds [71].

The Artemis Accords and Building Blocks are well-intended, and their emphasis on keeping pace with emerging technologies and promoting adaptive regulation is a move in the right direction. However, such polycentric approaches are effective when the entire system is co-evolving. Otherwise, fragmentation and divisions are deepened by diverging regulatory approaches. The U.S. ODMSP effectively promoted polycentric governance because the standards were not disputed internationally, so similar guidelines and standards were adopted by the international community by consensus soon thereafter. On the contrary, the recent guidelines on space resources are not broadly supported by the international community, increasing the risk of less coordinated space exploration.

Weak enforcement capabilities

Since the 1980s, COPUOS discussions on outer space have only resulted in non-binding resolutions, principles, and guidelines, which can be effective when they are enforced through binding laws in national legislatures or voluntary commitments from commercial actors. For example, a nation can require that a space launch applicant demonstrate compliance with the Space Debris Mitigation Guidelines (2010) before obtaining a license [72]. A national regulatory authority can enforce this law by only approving applicants who comply with ISO standards. Because of their quantifiable nature, ISO standards can be required for licensing through national regulations with minimal friction. As a result, they serve a necessary bridge from international guidelines and norms to national policies. ISO standards can also become binding if they are enforced through other mechanisms, including commercial contracts or government procurement policies [50].

However, some nations can intentionally or unintentionally (i.e. lack of resources) not adopt the guidelines and standards into national law, resulting in more risk-friendly regulatory environments. Commercial space actors may be encouraged to apply for a license for space activities in these countries—a similar dynamic in maritime law led to the label 'flags of convenience' [68]. Hence, unless every state with space activities abides by the international guidelines, there is a risk of violations of rules and safety standards.

It is difficult for standard-setting organizations to create effective standards without international consensus on guidelines. If standards do not reflect agreed-upon guidelines, then they are less likely to be enforced universally, rendering them less effective [50]. Standards do develop before international guidelines, especially to support interoperability. Industry groups tend to form to develop standards for emerging space activities relatively early compared to intergovernmental or non-governmental organizations. For example, the Commercial SmallSat Spectrum Management Association (CSSMA) was established in 2016 to share regulatory knowledge and coordinate spectrum management among industry members, indicating early standard-setting for small satellite technologies [73]. However, standards are less likely to be put in widespread use through national regulations prior to UN guidelines, and their impact is likely limited to voluntary commitments by private actors. Furthermore, standards developed in industry forums may prioritize efficiency over risk and safety considerations. ISO 24113, which takes into account international guidelines, is yet to be updated to provide rules for small satellite constellations [50]. Such an update may not occur until new principles are established by COPUOS resolutions or IADC guidelines.

When private organizations voluntarily comply with safety standards, there is added security that they will be upheld. However, ensuring such a commitment from each private actor is unlikely, especially since private actors have no binding commitments from the OST. Since the OST requires states to authorize space activities, the standards can be more effectively enforced by national regulatory bodies. Voluntary compliance by the private sector may shorten feedback loops between learning and improved practice, but they do not provide sufficient security that safety protocols will be upheld.

Lack of inclusivity

COPUOS consists of governmental organizations with some non-governmental observers, and as such there is not a formal role for private actors or members of the scientific community [40]. Although private actors play an increasingly large role in space activities, the United Nations space governance framework was established before the commercial space sector emerged. Developing proper regulatory guidelines and safety standards requires broad coordination, operational experience, and technical expertise, so leaving out knowledgeable actors may reduce the effectiveness and evolutionary capacity of regulations.

Future space activities will include an even greater variety of actors and in-space interactions, and industry-led and non-governmental initiatives to develop frameworks and standards for these activities are occurring in other forums. The Global Expert Group on Sustainable Lunar Activities (GEGSLA) is a multi-stakeholder forum for discussions on coordinating lunar activities, with a goal of developing helpful reports for COPUOS [74]. As previously mentioned, developing industry guidelines has some benefits because they are voluntarily used by some private actors, they record the learned best practices of space operators, and they can inspire later standards and UN guidelines. However, such reports need to be supported by later UN resolutions and guidelines to become sufficiently effective and promote broad coordination. Proposals developed via non-governmental or industry initiatives are unlikely to reflect the concerns of all state actors who have voting power in COPUOS forums, rendering this process more difficult. Beyond disagreements on frameworks, a lack of inclusion of all state actors in early discussions for new proposals may also create difficulties for adoption by consensus. When the EU Code of Conduct for Outer Space Activities was proposed in 2010, it faced opposition from actors such as India and China, in part because they were not consulted during its formulation [75]. A more appropriate forum for early-stage discussions about emerging and future space activities would be both inclusive of all states and relevant non-state actors.

Adaptive Governance

Introduction

Adaptive governance involves mechanisms that allow for flexible rules, norms, and institutions which change over time as a result of policy experimentation and new learning about the environment [23]. Adaptive governance is increasingly being considered in space policy discussions [76]. The Building Blocks for the Development of an International Framework for the Governance of Space Resource Activities explicitly calls for a framework that will “Adhere to the principle of adaptive governance by incrementally regulating space resource activities at the appropriate time.” [64]. The Artemis Accords demonstrate compatibility with the Building Blocks and adaptive governance by encouraging Accords signatories to use their experiences and learning in space activities to contribute to multilateral discussions to review and revise future practices [52]. Emerging industry-led groups to manage space activities also demonstrate self-organization, a key characteristic of adaptive governance [77].

Although adaptive governance is being applied to outer space in forums outside the UN, these forums generally do not include all leading spacefaring nations and nonstate actors, and they are therefore less effective at encouraging widespread agreement and enforced adoption (see previous section). The current UN framework is not designed to support adaptive governance to address complex, global challenges, inhibiting our abilities to address global problems within its framework [24]. The lack of adaptability in UN forums is demonstrated by its governance mechanisms that facilitate slow regulatory processes and stalled discussions over conflicting beliefs. Introducing new, multilateral adaptive governance mechanisms would allow for a more coordinated and flexible space governance framework. In the fourth section (Section: Policy Proposals), specific mechanisms are proposed to improve adaptability within the current space governance framework.

Our current governance framework is still characterized by a poor responsiveness to anticipated future space activities. Our recent experience with space resources has demonstrated the lack of institutions that demonstrate both broad inclusivity and adaptability. As space activities increase and transformative technologies emerge, our current political trajectory is likely to lead to large regulatory oversights driven by a lack of adaptability, or a high degree of fragmentation driven by a lack of inclusivity and regulatory coordination. Fisher and Sandberg (2021) suggest that our current global governance system is insufficient for such challenges, calling for new governance frameworks. In the fifth section (Section: Structural Reforms), we consider several proposals which could lead to a new space governance framework built on the principles of adaptability, scalability, and coordination [78].

Space activities in the medium-term and long-term future will involve more human-environment interactions and create permanent changes to the outer space environment. As a result, larger risks associated with unsafe practices or a lack of coordination and cooperation will emerge. An adaptive and inclusive governance framework would be capable of effective regulation and coordination in the near-term through the long-term as the implications of space activities increase. We describe the evolution of space activities over time by describing three eras: the current New Space Era, the In-Space Economy Era (ISEE), and the Transformative Technologies Era (TTE).

Future Space Eras

New Space Era

Current priorities in space governance address emerging space activities from mainly private actors to navigate the New Space Era [79]. This era marks a transition from previous space exploration that was defined by national actors. While commercial operators are increasingly important, national governments are still the main source of demand [80]. For example, NASA's Commercial Orbital Transportation Services ran from 2005 to 2013 and utilized partnerships and contracts with private actors for operations in low-earth orbit, including commercial resupply for the International Space Station [81]. Government demand supported the growth of private space companies, including SpaceX.

Another challenge of the current era involves addressing the supervision of nascent in-space activities by private operators. These include on-orbit servicing, space resource utilization, active debris removal, planetary protection, and rendezvous and proximity operations [48].

In-Space Economy Era

Within a couple decades, private spaceflight and permanent operations on the lunar surface, lunar orbit, martian surface, and asteroid belt will give rise to a new era in space governance, which this paper labels the In-Space Economy Era (ISEE). Demand for goods and services in outer space will encompass many sectors, operators, activities, and links to the terrestrial economy. For example, private spaceflight (including tourism) and space entertainment will become more accessible, space-based manufacturing of health and medical products will increase, space-based energy services will become available, on-orbit servicing and refueling will be more widespread, and travel to the lunar surface will become more frequent [1].

The emergence of the in-space economy will lead to new governance challenges, such as coordinating infrastructure planning (e.g. lunar roads and landing pads) between adversarial space powers (i.e. the U.S. and China) and private operators [2]. In addition, standards and guidelines will need to be continuously created and updated for new space activities based on policy experimentation and learning.

Transformative Technologies Era

Transformative technological breakthroughs will mark the end of the ISEE and begin the next age of space exploration, labeled the Transformative Technologies Era (TTE). Advances in artificial intelligence, biotechnology, nanotechnology, geoengineering, and other technologies will vastly improve human capabilities for space activities [3]. Large-scale activities will emerge, including increases in space-based manufacturing, terraforming, and deeper exploration of the solar system and beyond with probes. Megastructures might include large-scale energy generation systems (e.g. Dyson spheres) [5]. or floating space settlements (e.g. O'Neill Cylinders) [6], but the wide array of possible engineering solutions and design choices of the TTE means the future of space infrastructure involves uncertainty [82]. The technological capabilities of actors in this era will bring significant governance challenges because of the increasing scale of activities, the asymmetrical power dynamics that such advanced technologies can create, and the catastrophic outcomes that can result from improper usage.

Complex Systems

As eras progress, the outer space domain is becoming more integral to humanity's sphere of influence, and thus activities in outer space are becoming more important to shaping our geopolitical environment and civilizational trajectory. We apply complex systems thinking to the space domain to understand the implications and risks associated with the growing importance of outer space. We then propose that adaptive governance frameworks applied to outer space can improve management of evolving risks and complexity..

Complex systems are characterized by a wide array of actors, broad connectivity and interactivity, and influence from external events and shocks [83]. Large-scale risks can emerge and propagate throughout a system from complex interactions between system elements [78]. When unanticipated events occur in complex systems, they demonstrate emergent properties, rapid changes, nonlinear dynamics, and widespread effects throughout the system [84].

Watson (2016) argues that the risks associated with complex systems can increase in severity over time [85]. Pegram and Kreienkamp (2019) outline three mechanisms of complex systems that track the increasing severity of breakdowns: tipping points and thresholds, fast feedback effects, and cascading failures [83].

We argue that the possibility of catastrophic events in the outer space domain increases across the three aforementioned eras of space activities. Tipping points can be identified which mark the introduction of new eras, each of which has risks involving significantly quicker feedback loops and further cascading failures than the prior era. The beginning of each space era marks a transition to a greater array of actors, a significantly higher degree of interconnectedness, and greater importance of outer space to our social, environmental, and technological systems. In the next section, we outline tipping points which might mark era transitions and the changing patterns in which breakdowns can propagate.

Outer Space Complexity

When humans first landed on the Moon, space was largely disconnected from terrestrial systems—the first moon landing had few implications on earth beyond the resulting societal reactions. In the New Space Era, earth's orbital environment is very interconnected with the economy, and thus breakdowns can arise from earth's orbital sphere. The orbital environment is relied upon for communication, navigation, earth observation, and other use cases across societal and technological systems [79]. Hence, all of these systems may be disrupted by collisions or other accidents in orbit. Failures would cascade to affect more aspects of society and the environment.

Several necessary conditions (i.e. tipping points) have to be in place for the transition to the ISEE. Reusable launch vehicles, such as the SpaceX Starship, were essential to lower launch costs to sufficient levels to support a hybrid (earth and space) economy [86]. The birth of the in-space economy still depends upon further coordinated actions between public and private stakeholders. A greater diversity in demand for in-space services must be supplied to support commercial businesses, and the public sector would need to provide the public goods that might not be commercially profitable, such as lunar roads and other infrastructure. Initial demand for the in-space economy might come from sources such as private spaceflight and more government contracts for lunar, martian, or asteroid missions. This initial spark in demand would have cascading effects that increase demand for other goods and services.

In the ISEE, the importance of outer space to our geopolitical and economic systems will significantly increase, exposing society to risks of further cascading failures. The ISEE will exhibit a high degree of interconnectedness among a wide array of actors beyond the orbital environment. These actors will primarily be in lunar orbit and on the lunar surface, but some activities may be conducted on the martian surface or the asteroid belt (e.g. asteroid mining) [87]. There will be deep connections across numerous sectors between the in-space economy and our terrestrial systems [1]. Cascading failures may be caused by failures of energy infrastructure in space that is relied upon by terrestrial users, or supply chain bottlenecks that lead to shortages on earth of goods manufactured in space (e.g. pharmaceuticals manufactured under microgravity conditions).

The tipping points which mark the transition to the TTE might be signaled by the development of technologies that would allow for space exploration and activities on much grander scales: geoengineering and terraforming, genetic enhancement, large closed-loop biological support systems, and transformative artificial intelligence (TAI) systems. TAI breakthroughs are particularly consequential to mark the TTE because of the vast capabilities that TAI would enable [4]. For example, TAI may be able to advance scientific and engineering knowledge to solve other technical bottlenecks; allow for greater autonomy of robots for space manufacturing, space exploration, and space mining; greatly enhance monitoring, coordination, and surveillance systems; and enable space actors to collect and utilize astronomical quantities of energy, resources, and compute.

Following the emergence of transformative technologies, the outer space environment will be integral and of utmost importance to our economy and geopolitical system. The space-based risks that will emerge in the TTE have much greater potential to cascade and cause catastrophic societal transformations than risks during the ISEE. Some of these risks have the potential to permanently alter the trajectory of humanity in significantly negative ways. The implications of trajectory-shaping tipping points such as the emergence of TAI, deployment of self-replicating probes, construction of space megastructures, directed panspermia, and other future technologies are not fully known. The worst-case outcomes associated with the TTE are further described in the next section (section: Scenario Analysis).

Global Governance Context

Criticisms of the polarization, bureaucracy, and lack of adaptivity of current multilateral institutions are widespread [88]. Furthermore, the view that existing global governance institutions are not fit for addressing modern global challenges is widely held [89]. Maintaining a global governance structure with widespread cooperative dynamics is necessary to ensure the probability of catastrophic outcomes is not increased by asymmetric power imbalances [90].

Our existing multilateral institutions were created in an era of less interconnectedness and a significantly smaller array of existential and catastrophic risks. In particular, our existing multilateral institutions are ill-equipped to manage existential risks because their designs, which were intended to create interdependence and cooperation in a post-war era, are a poor fit for managing high levels of complexity [83]. Green, Hale, and Colgan (2019) use the term 'existential politics' to describe a new set of implications for governance in the modern era, where questions are not about resources and benefits sharing but the survival of particular ways of life and values [91]. The fundamentally different nature of that which is at stake for global governance requires a rethinking of how our multilateral institutions should operate internally and with respect to the world [24].

Modern space governance institutions, like many other multilateral institutions, were designed to address complicated, not complex, issues. They were created when the main risks in space activities were rare accidents, and interactions between space operators and other actors did not demonstrate unpredictability and emergent consequences. Their regulatory systems involve bureaucratic mechanisms that are inhibited by stalled discussions and attempts to address issues in a top-down manner.

'Metagovernance' is the creation of frameworks that allow for governance to emerge and adapt, or overarching organizational structures that enhance the process of governance self-organization [92]. 'UN metagovernance' can be improved through systemwide mechanisms to improve the formation of valuable partnerships between various actors [93]. Allowing for governance to emerge and adapt to changes in the space domain can ensure that regulatory mechanisms keep pace with space activities.

Role of Space Governance

We are particularly interested in governance frameworks and mechanisms that are designed to avoid the worst case outcomes in outer space. To analyze the worst-case scenario outcomes in the space domain, we identify three categories of risks. Global catastrophic risks (GCRs) are threats that can cause significant damage to the overall well-being of humanity on a large scale [94]. If a GCR eliminates a large portion of the human population or our technological progress, human civilization may look vastly different when the population recovers [94]. Existential risks (x-risks) are a unique category of risks which are threats of adverse outcomes that would either eliminate earth-originating intelligent life or permanently limit its potential [95]. Another category of risks are known as suffering risks (s-risks), defined as risks "where an adverse outcome would bring about suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far." [96]. GCRs can emerge in the outer space domain throughout all three of the identified eras. Although there is a chance of x-risks or s-risks emerging in the space domain during the New Space Era or ISEE, these risks significantly increase when the TTE begins.

Outer space itself is a domain in which adverse outcomes associated with these risks can emerge, and thus space governance frameworks must be designed with a capacity for evolution and risk management. Developing the technical standards for individual risks in the space domain, such as those associated with artificial intelligence and biological contamination, should be managed by subject-matter experts integrated into the space governance framework. Properly addressing risks also depends upon inclusive coordination mechanisms that effectively manage interactions between environmental, technological, and human factors. Governance mechanisms must be sufficiently adaptable to adjust rules and regulations in significant ways as tipping points are reached, such as the introduction of the ISEE and TTE eras. Furthermore, mechanisms should be sufficiently adaptable in the short-term to mitigate propagating negative effects from rapid feedback loops during unexpected, catastrophic events.

To effectively respond to the risks and uncertainties, Ashby's Law of Requisite Variety implies that the governance system must be able to exhibit more variety than the system itself [97, 78]. Furthermore, the speed of coordinated action must exceed the speed of cascading failures and feedback loops of emergent risks. Hence, the principles of adaptive governance can guide the design of space governance mechanisms to manage such risks [78].

Analysis of the evolution of space activities and associated risks leads to two key conclusions. First, space governance should consider the changing relationship of the space domain with respect to the geopolitical and economic environment.

Second, space governance institutions should be considered an essential part of any strategy to address risks when transformative technologies emerge. The most important role of space governance is to mitigate catastrophic, existential, and suffering risks by developing adaptive structures to address the increasing relevance and complexity of the space domain.

Scenario Analysis

Methodology

Knowledge of future developments in outer space and their associated societal, technological, and environmental implications cannot be fully determined because of the complexity and unpredictability of our geopolitical system in the long-term future. As a result of our uncertainty, it is difficult to assign probabilities or establish a likelihood ranking between various possible scenarios [98]. However, we can still enumerate several possible scenarios and take active measures to mitigate the possibilities of the worst-case outcomes [99]. Identifying the effectiveness of possible mitigating actions against negative outcomes requires exploratory uncertainty analysis, where a wide range of possible scenarios and outcomes are explored [7]. Scenarios can be described as "projected futures that claim less confidence than probabilistic forecasts." [99]. Decision-making under deep uncertainty (DMDU) models can be used to identify responses that are effective across a wide range of possible scenarios, especially by introducing "monitor and adapt" strategies [7]. For example, Dynamic Adaptive Planning (DAP) involves the specification of initial actions, contingent actions, a monitoring system, and 'signposts' (i.e. tipping points) for conditions being monitored—when signposts are reached, contingent actions are put in place [98]. DMDU models can support adaptive governance by allowing for flexible responses based on how events proceed, but this first requires identifying a range of scenarios and negative outcomes to be avoided.

Scenarios analyzed in this paper were limited to the most negative outcomes to be avoided, primarily occurring after transformative technologies are developed. The maximally negative outcomes were identified as those which may cause permanent, irrecoverable damage to the trajectory of humanity, or astronomical amounts of suffering for other sentient beings. Recognition and understanding of these scenarios is necessary to assess the role of space governance in mitigating or propagating catastrophic, existential, and suffering risks. The methodology and risk mitigation strategy proposed by this paper follows from the 'maxipok rule' identified by Bostrom (2002): "Maximize the probability of an okay outcome, where an okay outcome is any outcome that avoids existential disaster." [95].

To select scenarios for the analysis, a survey of recent literature on GCRs, x-risks, and s-risks was first conducted to identify existing lists of risks, including academic research and policy reports [78, 100, 36, 32]. A classification system similar to that proposed by Avin (2018) was then used to classify the risks to determine which are most relevant to space activities and space governance [101]. Our system classified risks based on whether they damage critical systems relevant to future space activities, contain effects which cascade across astronomical distances (i.e. 'astronomical spreading'), and should be addressed by space governance frameworks. Risks are not considered within the domain of space governance if an alternative governing institution or coordination mechanism would be responsible for mitigating the risk rather than institutions specifically designed for space governance. For example, preventing a terrestrial great power conflict would be the responsibility of the UN Security Council, bilateral negotiations, or another body whose primary responsibility is not space governance. Risks meeting all three of the aforementioned criteria were considered in the scenario analysis.

Many risks that were not considered may still be relevant to space governance in shorter time frames. For example, near-term space colonies will likely not be self sufficient, so an environmental or social catastrophe on earth might lead to their collapse. Since our scenario analysis relates to space governance in the longterm future, these risks were not included. Other crises which are localized are not considered as demonstrating astronomical spreading. For example, weapons of mass destruction, state collapse, and terrorist attacks are treated as international, national or local issues without implications across astronomical distances. We argue that the GCRs associated with these events will be significantly reduced once humanity and its descendants are an advanced spacefaring civilization.

Based on the classification, eight risks are considered: resource depletion, transformative artificial intelligence (TAI), biological contamination, global governance collapse, totalitarianism, directed panspermia, wild animal suffering, and digital minds suffering.

Negative scenarios based on these risks are classified into three high-level categories: sustainability failures, biological spreading, and transformative artificial intelligence (TAI). In the following section (Policy Proposals), concrete institutional proposals are presented which are promising adaptive mechanisms to address the possible negative outcomes identified. In the final section (Structural Reforms), values, norms, and longtermist reforms to guide space expansion are proposed to further improve adaptability and encourage more ambitious thinking about longterm space governance and activities.

Global Catastrophic, Existential, and Suffering Risks			
	Relevant Critical System	**Astronomical Spreading**	**Space Governance Issue**
Environmental			
Asteroid Impact			
Ecological Collapse			
Extreme Climate Change			
Extreme Weather			
Natural Disaster			
Supervolcanic Eruption			
Societal			
Financial Collapse			
Food Crisis			
Mass Migration			
Pandemics	X		X
Biological Contamination	X	X	X
Resource Depletion	X	X	X
Urban Expansion			
Water Crisis			
Technological			
Transformative AI	X	X	X
Geopolitical			
Weapons of Mass Destruction	X		X
Global Governance Collapse	X	X	X
State Collapse	X		X
Terrorist Attack	X		X
Totalitarianism	X	X	X
Suffering			
Directed Panspermia	X	X	X
Spread of Wild Animals	X	X	X
Spread of Digital Minds	X	X	X

Table 1: List of GCRs, s-risks, and x-risks considered for relevance to longterm space governance

Sustainability Failure

Common-Pool Resources

Common-pool resources (CPR) are large but subtractable quantities of resources where it is possible, although difficult, to recognize and exclude resource users [102]. The outer space domain is not a single common-pool resource, but a system of many CPRs with different properties [103]. Geographical regions of relevance in the present and near future include low-earth orbit, geosynchronous orbit, the lunar surface, the Martian surface, the orbital environments of the moon and Mars, and the surfaces of asteroids in the asteroid belt. Each region has various resources and use cases. For example, earth's geosynchronous orbit has a limited number of orbital slots and radio frequencies, which are governed by the International Telecommunications Union. Although the volume of orbital space on a celestial body is constant, the usability of this space can be depleted by unsafe levels of space debris. The lunar surface has valuable locations such as the Peaks of Eternal Light and cold traps, and a wide array of useful materials such as helium-3, uranium, and rare earth elements [27]. Sustainability risks are associated with each common-pool resource in the space domain, and risk mitigation efforts need to be individualized for each resource. Negative outcomes associated with sustainability failures are rendering orbits unusable and resource depletion.

Orbits

Kessler Syndrome describes a scenario where a high concentration of objects in orbit could lead to satellite collisions that generate even more fragments, resulting in a chain reaction of collisions and accelerated growth in the concentration of space debris [8]. Recent modeling has suggested that the risk of cascading growth is unlikely, but our current trajectory suggests that satellites will face a much higher probability of collisions in the future [9]. Furthermore, there is significant uncertainty surrounding the number of satellites which will eventually be launched around earth or other celestial bodies. The development of active debris removal technologies should also be considered, and our ability to remove large quantities of debris may be able to prevent Kessler syndrome or reverse its effects after it has begun.

Resource Depletion

Resource depletion within the solar system is also a consideration on centuries-long timescales [10]. Humanity's ongoing experiences with attempting to limit global warming and fossil fuel usage reveal the challenges of changing consumption habits on societal scales and reversing exponential growth [11]. Attempting similar diversions from resource use would be significantly more difficult on a solar system scale, especially without advanced planning and effective mechanisms for common-pool resources management. Previous calculations based on our current growth rates suggest that solar system resource depletion would occur in approximately 400 years, but there are many uncertainties about the quantities of various resources that are available which could alter the timescales [10]. Depletion of solar system resources would be catastrophic for inhabitants of the solar system, posing a GCR.

Avoiding a depletion of solar system resources on centuries-long timescales is necessary to reflect the concerns of future generations, which receive attention in the United Nations 'Our Common Agenda' report from 2021 [104]. Resources can be treated as intergenerational public goods, and resource depletion would significantly reduce intergenerational equality [105].

Effective common pool resource management generally depends on nested—or tiered—levels of governance, collective choice decision making, understood boundaries of the CPR system, and monitoring and adaptation [106]. TAI might lead to accelerating automation, capital production, and resource usage [12]. Such an outcome would inhibit the ability of governance systems to reflect each of the aforementioned conditions for effective CPR management. Instead, there would be one or several TAI systems utilizing resources at unprecedented levels, resulting in a lack of tiered governance and collective choice decision making among resource users. TAI systems might be capable of understanding and operating within boundaries, or monitoring the environment and adaptively reducing output if resource pools are critically low. However, there is uncertainty with respect to whether these mechanisms will be instilled within TAI, and they would not occur in a system attempting to maximize its productivity and output.

The orthogonality thesis suggests that any goal of a TAI system can be paired with any level of intelligence [4]. An implication is that deployed and uncontrolled TAI systems in the space domain can pursue particular objectives, such as exploration or resource exploitation, without limitation. As a result, TAI systems will pose an unprecedented threat to space resource sustainability.

Sustainable Safety

Yang and Sandberg (2022) suggest treating safety as a common-pool resource which can be ensured through an adaptive governance framework [107]. Safety has similarities with other CPRs because it is shared by all involved actors, and all stakeholders can contribute to its depletion. The game-theoretic framework is different because one actor can have an outsized harmful impact, whereas the risk of CPRs involve many actors having a large combined impact through many individual depletions of the common pool, known as the 'tragedy of the commons' [108]. In the safety scenario, one well-intentioned actor who acts independently from the group can cause significant harm [109]. For example, one space operator who believes that landing a spacecraft on a particular region of the moon is safe might actually cause dust interference with critical operations for survival (e.g. oxygen production). In some cases, the risk of one actor's wrongdoing can be reduced through a distributed system of control of critical resources and failure points, such as water and oxygen supply [110]. The dynamics of safety as a resource suggest that some principles of effective common pool resources management, such as effective monitoring and the broad inclusion of all actors, can be applied [106].

Biological Spreading

Astrobiological Life

There is yet to be a proven discovery of extraterrestrial life. However, it is possible that advances in astronomy and astrobiology, such as more exoplanet data and research on potential signs of biological and technological signatures, will enable a discovery [111]. The discovery of astrobiological life would increase the probability of harmful biological contamination if humanity makes physical contact with the celestial body on which life is discovered. There are several regions within the solar system which are potentially habitable in the present day, including Saturn's moon Titan and Jupiter's moon Europa [112]. There is also a planet within the habitable zone of Proxima Centauri, known as Proxima Centauri b [113]. Furthermore, anthropogenic processes such as terraforming might provide the conditions for life to emerge where it does not currently exist [114]. Travel to a habitable planet might risk forward contamination if microbial biological life exists, or backward contamination if a launched probe returns to earth without proper planetary protection measures in place. Article IX of the OST includes a provision to avoid "harmful contamination" of celestial bodies [47]. However, developing further regulatory frameworks for preventing biological contamination that keep pace with scientific and technological breakthroughs is necessary to effectively mitigate these risks [115].

Safety Standards for Contamination

Harmful biological contamination may pose a GCR, mainly relating to backward contamination of the earth [13]. Missions such as the Mars Sample Return, or future return missions, could lead to such an outcome if active biological life is found and proper planetary protection and quarantine measures are not taken [116]. However, measures to avoid cross-contamination are well researched and standardized [14]. Since the events which could lead to backward contamination are well known and controlled (i.e. return missions), the complexity and emergence of such a GCR is suppressed.

GCRs from biological contamination should still be continuously addressed through adaptive standards setting based on learning. The risk of biological contamination of celestial bodies could be reduced through technical standards reflecting our knowledge of biological disaster management [14]. Involvement of the pandemic and biological disaster preparedness community in designing standards and procedures for risk mitigation, response, recovery, and adaptation is essential [117].

Spreading Biological Life

Directed panspermia involves the intentional spreading of biological material across interstellar distances to increase the probability and longevity of survival for the descendants from terrestrial origins of life [15]. Most proposals involve spreading microbes or other genetic material. The risk of introducing organic material to extraterrestrial environments would raise ethical considerations because of the possibility of harmful biological contamination and a lack of control over how the introduced organic material would react and evolve [16]. If directed panspermia occurs, prior discussions of scientific and technical standards would allow for the careful selection of pansperms to minimize the probability of harmful contamination [15]. However, we lack control over how evolution may unfold after directed panspermia, and life may evolve which experiences vast amounts of suffering.

Animals might also be introduced to other celestial bodies or artificial settlements to which humans expand [17]. This might occur to ensure a food source through establishing factory farming, or a similar method of ensuring sufficient food production in a physical space- and resource-constrained environment. This would only pose risks of forward contamination if the celestial body contained astrobiological life. However, any spreading of animal life would pose ethical concerns associated with animal well being and suffering.

The spreading of life does not pose a risk of widespread emergent phenomena that could cause cascading failures and existential catastrophes, so the reduced risk for humanity might lead to less political and technological restraint of these activities. However, the spreading of animals for farming may lead to greater animal suffering, and this outcome should be treated as a significant suffering risk (s-risk) [118]. Directed panspermia might result in the emergence of life elsewhere that contains large amounts of suffering, posing another s-risk.

The benefits of spreading life might exceed the costs if contamination risks and s-risks are mitigated, such as by ensuring the well being of animals and conducting deeper analysis on the implications of directed panspermia. Hence, these activities should not necessarily be avoided, but they should only be conducted if proper safety measures are taken, the implications are deeply analyzed, and the conclusions of research suggest there are effective ways to mitigate the associated s-risks.

The actors who launch missions to habitable environments or conduct directed panspermia may be state or non-state actors. It is plausible that these actors would be space-based in the future because many spacefaring actors are likely to be focused on further space expansion, and launches would not need to overcome earth's gravitational well. Authorization and supervision of such space-based activities, either from earth or within space, is essential to ensuring that safety protocols are followed, especially if space expansionists have motivation to locate themselves in space to avoid strict safety standards.

Transformative Artificial Intelligence (TAI)

Introduction

There is likely to be a heavy involvement of TAI and autonomous systems in future spacefaring because of their advanced capabilities and the adverse conditions of the space domain [119]. Karnofsky (2016) defines TAI as ”AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution.” [120]. If TAI is realized within the next several decades, timelines for early space activities will be greatly accelerated. TAI would shorten the time taken for large space constructions to emerge, such as energy production or large manufacturing facilities. TAI systems would also be capable of conducting entirely autonomous operations at long distances, expanding the geographical scope of space activities in a quicker time frame. Space activities in every industry would increase exponentially. Surveys of AI experts revealed the following predictions regarding the arrival year of TAI: twenty percent chance by 2036, fifty percent chance by 2060, and seventy percent chance by 2100 [121].

There are many perspectives on what TAI will look like: a superintelligent AI that is cognitively superior to humans, an AI ecology where multiple, specialized AI agents perform a wide variety of tasks (i.e. Comprehensive AI Services, or CAIS), and a general purpose technology (GPT) which can be used by humans for a wide variety of social, scientific, military, economic, and political applications that transform the structure of our institutions [122]. There are several existential risks associated with TAI in the space domain, especially during the transition to a post-TAI world. Even if worst-case outcomes are avoided on earth, a failure to scale up effective TAI governance and coordination mechanisms to the space domain can have catastrophic consequences. Four TAI-related scenarios are considered in more detail below.

Dependence

Outcome One (Misaligned TAI)

Misaligned TAI

“Misaligned” TAI refers to TAI systems which behave in ways that are unintended, given the objectives provided by humans [18]. Misaligned AI systems differ from AI that behaves as intended by a human operator who has malicious intent—these scenarios are covered in more detail in a scenario discussed later (section: Malicious Use of TAI). If usage of a misaligned system is allowed with few constraints in the space domain, there will be countless opportunities for a TAI system to choose an unintended pathway to accomplish its goal. Significant x-risks and s-risks can emerge from the usage of misaligned TAI in outer space.

Misaligned TAI can be created because when an agentic system is choosing its ‘policy’ to accomplish some assigned goal, it may decide to ‘game’ the system to maximize its objectives rather than pursue them as humans intended [123]. This concept is closely related to Goodhart’s law, which states that “any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.” [124]. With enough dependence on misaligned TAI systems, the civilizational trajectory pursued might differ significantly from humanity’s desires. Human agents might nevertheless become too dependent on TAI to maintain the knowledge, capabilities, and power to alter our path forward [123].

One example of misaligned TAI in the space domain could involve a space actor creating self-replicating von Neumann probes for space exploration, which then replicate at unprecedented and unexpected rates and use a sizable amount of matter within the solar systems and other star systems for replication [125]. Such an outcome resembles the ‘paperclip maximizer’ proposed by Bostrom (2012) [126]. In addition, advanced TAI systems may enable the creation of sentient digital minds, which would demonstrate greater survivability in the adverse space environment than biological humans. Outer space has sufficient quantities of energy and compute resources to create servers and run astronomical amounts of digital minds, assuming they are technologically feasible [22]. If a space actor attempts to create such digital minds in outer space using misaligned TAI systems, the experiences of the digital minds would be dictated by the TAI system’s metrics, potentially resulting in vast amounts of suffering.

Limiting Usage

If misaligned TAI systems are developed, regulatory measures such as limiting usage and strict oversight may be in place to prevent unintended actions with potentially vast consequences. If we are careful with deploying systems, these measures might be sufficient to prevent a loss of human autonomy on earth and prevent some worst-case outcomes associated with TAI.

For various reasons, a reliance on caution and oversight might be less effective in outer space than on earth, instead resulting in more ‘goodharting’ and less human autonomy maintained over decision-making. First, the adverse conditions in outer space and less institutional friction might result in a much greater proportion of tasks being entirely automated. Many powerful institutions on earth (e.g. governments, corporations, etc.) have deep rooted policies which may demonstrate friction to enhancement or replacement by TAI [127]. Second, space actors might have greater trust for TAI systems and less resistance to complete automation than earth-based actors. Trust might be a key factor that bottlenecks the pace of TAI deployment on earth, including within corporations and governance institutions [128]. Trust for TAI systems might be greater in space than on earth because many spacefaring actors are likely to prioritize further space expansion and development above other considerations, which requires TAI-driven technological progress—safety considerations might be less emphasized. Third, many space activities that occur will be unprecedented on earth because of the vast amounts of energy and resources uniquely available in space. These include self-replicating probes, large-scale energy systems such as Dyson spheres, and large sentient simulations that require astronomical quantities of energy. A lack of precedence suggests we would have less insight into the risks associated with these activities.

Adverse outcomes such as those outlined above might be made even more likely in space environments resembling a "technological wilderness" where there are weak political institutions and enforcement mechanisms, a large number of actors, advanced technologies, and a large accumulation of capital. Such a scenario would lead to strong competitive dynamics between commercial organizations and an inability of governance mechanisms to reduce the competitive pressures or enforce regulations and safety measures. The current trajectory of the in-space economy suggests that the lunar environment, and perhaps the martian environment, might demonstrate these characteristics. If we lack scalable space governance frameworks, then the lack of oversight will increase over time.

If space development across large distances occurs quickly after TAI is developed, implementing political restrictions on TAI usage in the space domain might depend upon strong space governance frameworks being developed before the technologies become feasible. Without proper frameworks and enforcement capabilities, our ability to regulate TAI in outer space would be curtailed. Furthermore, the relevance of AI in space development suggests that experts on AI systems should be consulted in the development of relevant space standards. For example, regulations for advanced in-space autonomous systems might include thresholds for human oversight, limitations on resource exploitation, and an understanding of the internal decision-making algorithms of deployed systems [129].

Outcome Two (Premature Values Lock-in)

In the previous subsection, a scenario was considered where TAI systems are not aligned with human interests. In this section, we consider scenarios where TAI systems are aligned with values selected by particular stakeholders (e.g. space-based actors or AI researchers), but those values are not necessarily representative of all of humanity or future generations. In particular, we consider the risk of the lock-in of these values, given they are not optimal. A possible pathway to such an outcome involves comprehensive AI services (CAIS).

The CAIS model suggests a world in which there are unique AI systems to autonomously conduct various tasks, including AI systems optimized for R&D (i.e. the creation of new AI systems for particular purposes) [130]. Hence, AI systems can collectively perform all tasks, but none of them are independently generally intelligent. Competitive pressures may cause CAIS to be quickly deployed by various actors for a variety of purposes [131]. As a result, there is a risk of high dependence on CAIS to maintain and progress society, including making trajectory-shaping decisions.

As was mentioned in the previous subsection, complete dependence on TAI systems such as CAIS may be more likely in outer space than on earth because of greater trust and less institutional friction to its widespread adoption. The "technological wilderness" scenario for outer space strengthens this argument. There might be near-unanimous consensus among powerful actors in space that the use of CAIS systems is a net benefit, resulting in an explosion of usage.

Through the widespread use of CAIS, the space domain might become locked into dependence on TAI because of the transfer of coordination and oversight responsibilities to CAIS systems, the lack of skills of humans to manage outer space affairs, and progressive improvements in CAIS systems [19]. The outcome would be a TAI-controlled lock-in that Christiano (2019) describes as "going out with a whimper". Even if we avoid such an outcome on earth, a loss of autonomy in outer space could cause humanity to permanently lose control over how spacefaring proceeds.

In such a scenario, humanity would lose the ability to influence society and change its longterm trajectory [123]. Decisions with respect to further space expansion would be made by TAI rather than humans, including the design and coordination mechanisms for settlements. A fully-aligned TAI might make decisions that are perfect matches to those reflecting the values of humanity, known as Friendly AI [132]. In this scenario, complete dependence on TAI is not necessarily negative. However, if TAI is only aligned with the values and interests of some humans, such as early spacefaring actors or TAI developers, then the longterm values which shape the trajectory of humanity will not be representative of current or future generations, and we would be unable to adjust these values at later dates. This outcome would permanently curtail the potential of humanity, and thus it could be considered an x-risk.

Earth-based actors might demonstrate a similar adoption of CAIS to space-based actors, resulting in a similar dependence on TAI. However, there are several reasons why such a dependence might not immediately emerge on earth: resistance to change from individuals and institutions, a greater diversity of actors with competing preferences, and greater autonomy for humans in the terrestrial environment. Eventually, earth-based actors may find it beneficial to transfer all responsibilities to TAI systems, including coordination and governance. However, the longer transition period would allow for inclusive dialogue and a 'long reflection', where humanity can spend as much time as necessary to consider the values which it prefers to lock-in via TAI systems [30]. It will be significantly more difficult for space actors to engage in a long reflection before further expansion and lock-in of values.

Assuming CAIS are not aggressively competing, we may end up in a world where longterm values and activities on earth look vastly different from longterm values and activities in outer space. However, most members of future generations are likely to live outside the terrestrial environment, and thus engaging in a long reflection strictly on earth while space development proceeds is insufficient to mitigate risks associated with lock-in.

The probability of such an outcome could be mitigated through stronger, scalable coordination mechanisms for space governance and stronger ties between earth-based and space-based actors. Ensuring proper supervision and oversight of space actors can prevent the premature lock-in of values that actors on earth may attempt to avoid.

Totalitarianism

Outcome Three (Influence-Seeking TAI)

Influence-Seeking TAI

To better analyze the potential implications of influence-seeking TAI systems on space governance and space exploration, we consider the implications of Advanced, Planning, Strategically aware (APS) systems [20]. Such systems can outperform humans on advanced tasks that imply power (e.g. scientific research), are capable of making and executing plans to accomplish goals, and the models of the world used by such systems are highly accurate (including power dynamics) [20]. An APS system might engage in power-seeking efforts which further its objectives, including overpowering humans or seizing infrastructure to increase its probability of fulfilling its objectives [20]. If TAI systems pursuing these subgoals cannot be stopped once deployed, an x-risk to humanity may be posed. Bostrom (2014) proposes that a TAI system would pose an x-risk because of the possibility that it may operate in unexpected ways and exceed the cognitive capacities of humans (i.e. superintelligence), resulting in a loss of human control [4].

Existing space governance structures are least likely to matter in a superintelligence scenario where the intelligent system would replace human coordination and governance. However, if political measures prevent the emergence of a superintelligence on earth, the space domain may become an integral component of the geographical space in which such a superintelligent AI can emerge. Hence, the ability to enforce stringent regulations in the space domain would be necessary to mitigate existential risks associated with influence-seeking AI.

Preventing Development

If earth-based actors become extremely concerned about the implications of a superintelligent TAI, they may attempt to enforce a ban on such a system from being trained altogether. This would lead to longer timelines for the creation of a superintelligent TAI. Training a TAI system will require vast amounts of compute and advanced hardware [133]. Such a project would thus require large-scale energy systems, servers, and fabrication facilities to produce advanced AI hardware (e.g. AI chips) that are only feasible for a small number of earth-based actors [134]. Developing this infrastructure entirely within the space domain might require a significant amount of time because of the necessary breakthroughs in space resources and large-scale manufacturing that need to occur beforehand. Hence, training a TAI system with space-based infrastructure is likely infeasible on short timescales. Such an operation could be made easier if there are not payload checks which prevent advanced AI chips from being transported to space, since this would eliminate the need to construct an off-earth fabrication facility.

Attempts by powerful or rogue actors to establish the necessary compute infrastructure entirely within space, for example on the lunar or martian surface, would be visible, allowing for a unified response from ethical actors on earth. Such a scenario could be compared to developing the large-scale infrastructure required to sufficiently enrich radioactive materials to produce nuclear weapons, but in an even more transparent environment [135].

Superintelligent TAI systems may be preceded by advanced, specialized AI systems that allow for rapid space development. The specialized systems would be AI systems that significantly improve the abilities of humans to solve various science and engineering challenges—similar to DeepMind's AlphaFold—which enable space development at much quicker rates [136]. These earlier stage TAI systems were referred to by Karnofsky (2021) as a technology Process for Automating Scientific and Technological Advancement (PASTA) [137]. As a result of this technology, rapid space development could allow for powerful, space-based actors to emerge. The space-based infrastructure necessary to train a superintelligent system could then be constructed more easily. TAI systems trained in outer space might also have less energy and compute limitations than systems on earth because of the vast quantities of resources in outer space, potentially reducing bottlenecks to the development of superintelligent systems. In these scenarios, we should ensure that space governance mechanisms exist to regulate activities in the space domain at least as effectively as on earth.

Outcome Four (Great Power)

Space and Totalitarianism

If any variation of TAI systems, including CAIS or PASTA, are disproportionately available and utilized by particular actors in outer space, the associated opportunities for rapid scientific and technological development would result in large power asymmetries. The power imbalance may quickly become large enough for complete space supremacy, including the ability to limit others' access to space and approach totalitarian lock-in outcomes on astronomical scales. In contrast to the prior subsection, this subsection considers a totalitarian lock-in achieved by an advanced public or private actor using advanced technologies for malicious ends—not a superintelligent TAI system. Limiting the probability of such an outcome depends upon proper monitoring of technological development in outer space, defending inclusive and open access to the space domain, and promoting in-space economic competition and decentralization.

Attempts to limit access to outer space are impermissible under Article I of the OST, which states that space "shall be free for exploration and use by all States without distrimination of any kind." [47]. Such limitations could also result in a dangerous power imbalance. Although it is possible that establishing complete totalitarian rule in outer space is difficult because of its scale, the adversarial extraterrestrial environment might also be particularly suitable for the rise of a totalitarian power in the long-run future [21]. The necessity to exercise caution to prevent operational safety breakdowns, which may expose settlers to high radiation or unpressurized environments, could provide justification for an assertive authority using strong measures to prevent misbehavior, such as widespread surveillance [21]. Furthermore, the scarcity of critical resources in space—such as food, water, and oxygen—increases the probability of powerful, centralized economic monopolies with excessive power [21]. If sentient beings in outer space are primarily digital minds, then totalitarian control could be achieved by controlling the server on which the digital minds are run. [22].

Specialized TAI

If TAI allows great powers or private actors to guide rapid scientific and technological development, a single actor may be able to strategically establish control over the space domain and limit access for other operators. Unlike generally intelligent TAI systems which can accomplish all tasks, a scenario with PASTA might see various actors developing specialized systems to perform unique subsets of tasks. Thus, different actors would be more effective at automating different cognitive tasks. For example, one actor may have highly-developed AI systems that are suitable for advancing biology while another may have AI systems that are suitable for understanding the fundamental laws of physics. Some specialized systems might be more suitable for technological development and control in the space domain, such as systems that can develop engineering solutions using space-based materials, determine the optimal method for terraforming a celestial body, or identify the most valuable asteroids for mining. Beyond allowing actors to develop sizable advantages in space development, AI systems for exploitation of space resources might enable actors to access significant quantities of energy and resources to support other activities, including the development of even more advanced TAI systems.

Developing specialized AI systems for spacefaring requires requisite knowledge about the task to be able to train the system, so having existing knowledge and experience within the space domain would be essential for being able to develop such systems. Hence, existing leadership in the space domain is relevant to determining which actor can develop better AI systems for future space development to establish an asymmetrical advantage. The United States is still widely considered to be the leading nation in outer space, but recent analysis suggests that China may become the leading space power by 2030 [43].

Space Geography

There are also inherent first-mover advantages to be gained in the space domain based on its geography. Lagrange points are locations in space where objects are able to remain in place because of the gravitational forces of two celestial objects. There are five Lagrange points (L1-L5) in each system, but two are considered stable (L4-L5) and three are unstable (L1-L3)—objects located at unstable points require regular corrections, making them less suitable for megastructures [138]. Presently, the Solar and Heliospheric Observatory (SOHO) is located at L1 of the earth-sun system, and the James Webb Space Telescope will be located at L2. The L4 and L5 points have been proposed as ideal locations for large space colonies, including large rotating cylindrical habitats known as O'Neill Cylinders [6]. These locations may also be ideal for large energy-generation systems or manufacturing facilities. Given the limited number of stable Lagrange points in the earth-sun system, control of the Lagrange points may allow a particular space actor to develop a large technological and economic advantage.

Shared guiding principles for space development and planning can limit a first-mover from establishing asymmetrical control. For example, the Lagrange points are fairly large regions of space, and proactive coordination and commitments to share their usage can be established. Freedom engineering, which proposes engineering solutions that solve the challenges of establishing space settlements while minimizing coercion, can be used to minimize the probability of a centralized authority with uncontrolled power [110]. For example, a distributed water and oxygen supply would prevent a single actor from controlling space settlers' necessities for life [110]. As we discover new information about the celestial environment, we can continue to engage in collaborative planning and develop applications of freedom engineering.

Malicious Use of TAI

The transformative capabilities of TAI imply deep uncertainty about the future of space activities, who will control them, to what extent control can be locked in, and the values those in control will have. A malicious actor may be able to use TAI systems to maintain power advantages and suppress competition.

The specific objectives given to a TAI system might be attempts to consolidate power by a single actor to achieve unilateral control [139]. For example, TAI systems could be designed to monopolize control of critical resources (e.g. water and oxygen), develop megaprojects that result in vastly unequal resource distribution (e.g. Dyson spheres) [5], develop scientific breakthroughs that specifically lead to asymmetric advantages in space exploration capabilities (e.g. self-replicating probes) [140], and strengthen widespread surveillance of closed biological support systems.

TAI systems may also strengthen the case for totalitarianism by making monopolized control of critical resources more efficient and realistic than without TAI. By enhancing capabilities for central planning, a single power using TAI could efficiently allocate necessary resources to a large space-based population with a low risk of miscalculations of supply and demand [127]. If one actor develops these power imbalances and prevents other space–based actors from sharing power, a single authority might be able to achieve totalitarian control in outer space.

Becoming a Singleton

Tailoring specialized TAI development to the outer space domain, including establishing a space-based foothold, might be an optimal strategy for enabling long-term control. A technologically advanced space-based actor would become significantly more powerful than any actor on earth because of the vast amounts of resources and energy available, as well as the freedom of movement that space supremacy would enable. Large-scale space projects realized by an advanced space-based actor—including megastructures for transportation, communication, manufacturing, or energy generation—would increase the power imbalance over time and reinforce control over space development. A totalitarian power might aim to eventually establish surveillance across a wide array of celestial bodies, floating settlements, and other star systems. Self-replicating probes are a commonly described way to quickly spread at interstellar, or intergalactic, distances [141]. The first actor to establish significant operations in outer space, use TAI for scientific discoveries (e.g. floating metastructures, large-scale energy generation, self-replication, more efficient propulsion systems, etc.), and attempt to achieve unilateral control might have the exclusive capability to lock-in supremacy over the space domain.

Risk Category	Worst-Case Scenarios	Risk Classification
Sustainability Failure	Kessler Syndrome, Resource Depletion	GCR
Biological Spreading	Spread of Wild Animals, Directed Panspermia, Astrobiological Contamination	S-Risk
Transformative Artificial Intelligence	Dependence (Misaligned TAI or Premature Values Lock-in), Totalitarianism (Influence-Seeking AI or Great Power)	S-Risk or X-Risk

Table 2: Classification of worst-case scenarios included in the scenario analysis

.

Policy Proposals

Introduction

Institutions involved in space governance have demonstrated a tendency to avoid addressing future issues. However, the uncertainties and rapid development of the space domain suggests the need for foresight and adaptive, scalable governance mechanisms. These mechanisms are appropriate for addressing emerging space activities both now and in the far future.

Improving processes for managing catastrophic risks and investing in risk mitigation research are effective ways to reduce the probabilities of catastrophic events [25]. Adaptive governance has been shown to be an effective framework for promoting the sustainable use of common-pool resources and mitigating global catastrophic risks [26]. Such frameworks strengthen our ability to maintain governance mechanisms that can co-evolve with emerging space activities and risks.

Adaptive governance principles have been previously considered in the context of outer space. Migaud (2021) devised criteria for the application of adaptive governance to outer space [76]. Multiple analyses have also been conducted to apply common pool resources governance approaches to the space domain [142]. Commons management has also been applied to the lunar environment, [87]. and polycentric approaches to lunar governance have been analyzed [103].

This section outlines concrete proposals for reforms to the existing international space governance framework. It takes into account the theoretical foundations of adaptive governance and common-pool resources, proposals from recent space governance literature related to these frameworks, and the shortcomings of the current space governance framework. These solutions would effectively shorten cycles between policy implementation and feedback, address regulatory gaps, and promote inclusivity in the governance process. These solutions meet the further constraint of being scalable across new technologies and regions of space accessible in the far future. In particular, this section proposes four new mechanisms for the multilateral space governance framework: shared infrastructure, horizon scanning, a conflict resolution mechanism, and a verification agency. In the next section (Structural Reforms), four more transformative, longtermist proposals are presented to adjust the norms, values, and institutional frameworks that guide space governance and activities.

Shared Infrastructure

Multilateral governance mechanisms should develop and manage shared infrastructure that is available to all space operators who comply with established norms, guidelines, and standards [27]. This infrastructure may include landing pads, energy, navigation, surface mapping, and space traffic management systems. The infrastructure should follow open international standards that are established for space activities to promote both safety and interoperability, which would enhance coordination.

Advanced Technology Controls

In the TTE, the shared infrastructure could include centralized control over the usage of advanced, high-risk technologies in the space domain to prevent malicious use, such as biological and chemical substances (e.g. for terraforming) or TAI. Establishing shared outer space infrastructure for the space economy prior to the development of these technologies would serve as an effective precedent for centralized control of such technologies when they are developed.

Central control of TAI access is a frequent proposal in the AI governance literature [139]. Access to agentic or general purpose AI in outer space could be limited to prevent cases where usage may promote monopolistic behaviors, lead to the development of dangerous technologies, or result in unintended harmful actions. Whether an authority to manage TAI centrally is necessary depends on the extent to which TAI is aligned, and whether it has built in mechanisms to not pursue malicious objectives which violate established norms. A fully aligned TAI might need to internally prevent itself from operating in ways that provide advantages uniquely available in the space domain, including the pathways to totalitarian power described in the previous section (Totalitarianism): surveillance (justified for safety reasons) or economic monopolies over critical resources. If advanced TAI systems cannot internally 'enforce' decentralization by preventing themselves from establishing monopolies that are not in common interest, it would be beneficial to have a central, shared infrastructure system in outer space to moderate usage.

Incentive Design

Shared infrastructure, where usage depends on following standards and guidelines, can encourage broad compliance if the costs of not complying exceed the cost of compliance [27]. Furthermore, shared infrastructure is effective because it is scalable to the regulation of new space activities, technologies, or regions of settlement across vast distances of space. If infrastructure and centralized control of high-risk technologies is continuously scaled to include management of activities at the forefront of space exploration, then catastrophic and existential risks could remain mitigated in the long-run future. Actors would be actively encouraged to follow rules because they would be able to conduct more space activities than without the shared infrastructure. While the extent to which civilization can remain coordinated decreases with distance, we argue in the next section (Communication Network) that a scalable, shared communication network for all space activities is preferable to a lack of any coordination mechanisms.

Shared infrastructure also includes several benefits that arise from the unique game-theoretical framework of the space domain. Governance of high-risk activities requires adherence to rules from all actors. One actor's risky behavior could lead to catastrophic consequences, even with compliance from the rest [109]. Shared infrastructure is more effective than other methods, such as binding norms, at limiting noncompliant actors because operations in space would be made more difficult without access to shared infrastructure. More binding rules might not minimize the number of noncompliant actors since some states might be inclined to not sign treaties to maintain 'flags of convenience', or powerful actors might ignore binding rules without fear of retribution. Since shared infrastructure could also be used by private actors, multilateral institutions could directly persuade the commercial sector to abide by its guidelines instead of relying on national legislation. This would shorten feedback loops between rules and practice as well as minimize the effects of 'flags of convenience'.

Shared infrastructure would not eliminate the possibility of space activities for non-compliant actors. For example, a powerful actor could develop alternative large-scale infrastructure to support its own space activities. However, separate infrastructure would only be available to great powers or advanced private actors because of the financial costs involved in independently developing infrastructure for space exploration. Furthermore, global coordination has previously occurred on large-scale projects in the space domain, including collaboration on the International Space Station (ISS). However, when China emerged as an advanced space power after the ISS was created, it developed its own Tiangong space station, emphasizing the importance of allowing new actors to be integrated in existing shared infrastructure to maintain coordination. Developing independent infrastructure would also be time-consuming, noticeable, and intentional—a lack of cooperation could result in political consequences given the risks involved.

Additional Benefits

There are additional benefits to creating shared infrastructure. First, it increases inclusivity in space exploration by providing services that are expensive to develop, such as launch and landing pads, to all states and private operators. Second, developing shared infrastructure in the near future might kickstart the space economy, opening up beneficial commercial opportunities. Third, shared infrastructure could improve coordination between actors within the space economy. Shared launch pads could be located in areas where they would not interfere with other space activities. In addition, a shared framework for space situational awareness (SSA), space traffic management (STM), or conjunction warnings would reduce risk to spaceflight and satellites. Currently, a global system for managing space traffic does not exist. Fourth, shared infrastructure would reduce the probability of one actor gaining autocratic power, even without transformative technologies [21]. In the far future, reliable networks for transportation and trade throughout the solar system, or beyond, could be shared infrastructure that would lower the barriers to entry for space resource utilization and manufacturing activities, resulting in greater economic competition. Finally, shared infrastructure based on standards would enhance interoperability, enhancing the ease of coordination on a technical level.

Graduated penalties are a component of effective common pool resource management [106]. Hence, punishments for non-compliance with guidelines should not necessarily extend to all shared infrastructure immediately. For example, removing a state's access to an STM system would make spacefaring more dangerous for all actors. Some needed services for space exploration, such as access to a launch pad or refueling services, could be revoked earlier without large negative consequences. Meanwhile, access to advanced technologies such as TAI or terraforming via the centralized sharing mechanism could be revoked immediately. Access to these technologies could also be earned through a history of compliance.

Horizon Scanning

Primary Role and Responsibilities

There are several contributing factors to the systemic neglect for the long-term future by most governance mechanisms: future generations generally lack political representation, most constituents prioritize short-term issues, government officials hold office for short periods of time, and many politicians discount the future because of uncertainty about how it will unfold [29].

Horizon scanning is a strategy used to consider possible future scenarios and determine risks and regulatory gaps in the current framework [28]. It can bring more clarity to how the future may unfold, provide concrete information to lawmakers, and increase public awareness about future issues. Thus, it can increase the extent to which the long-term future is considered within the governance system.

A horizon scanning agency should be established within the space governance regulatory framework. The agency should reflect a broad base of knowledge and expertise about how the future of space activities may unfold, including experts on various aspects of the space domain as well as cross-cutting technologies (e.g. biosecurity, nanotechnology, nuclear energy, and artificial intelligence). Perspectives from academia, industry, governments, and NGOs should all be included. The diversity of perspectives prevents the development of conformity or narrow-mindedness [25].

The horizon scanning agency should pay particular attention to identifying aspects of the current standards and guidelines that may open loopholes or security risks. In addition, the agency should analyze the implications of emerging and future activities and technologies to adapt governance frameworks appropriately. Focusing on these challenges will allow the agency to develop concrete research outputs upon which policymakers can act, strengthening the adaptivity of the governance mechanisms.

By identifying specific failure modes, the horizon scanning agency goes beyond the traditional scope of horizon scanning research and engages in 'red teaming'. A red team is a group of experts to conduct scenario analyses and identify possible failures [25]. Red teaming is usually used to identify possible ways in which an adversary could damage or exploit a system [143]. By proactively identifying the gaps, solutions can be implemented before breakdowns of the system occur. Such a process should scrutinize all sets of guidelines and standards, taking into account the extent to which they are followed and implemented via national legislation.

Future Technologies

For entirely new technologies or space activities, the horizon scanning agency should have the authority to place topics on agendas of multilateral fora. They may even have the authority to establish new working groups to address urgent topics, provided there is sufficient funding to support these groups. By allowing future-oriented researchers to establish working groups, polarization and competing interests would not be able to inhibit the discussion of relevant topics for space governance. If these working groups exist under the umbrella of a legitimate organization such as COPUOS, they may maintain more credibility.

For urgent developments that may pose risks which cannot be effectively managed by current rules, the horizon scanning agency should be able to propose concrete changes based on its analysis. These changes may be updates to existing standards and guidelines, or the establishment of inclusive discussions outside the UN to develop proposals for new frameworks and rules. A stronger version of the horizon scanning agency would have the authority to update standards and guidelines independently. Furthermore, it would have the funding available to establish working groups and multilateral discussions at its own discretion. The horizon scanning agency would effectively be delegated authority by all space actors to represent the interests of future generations in space governance. Although it may only occasionally make direct changes to rules, this would allow for sufficient adaptivity in instances where concerns related to the speed of feedback loops and cascading failures render negotiations by a multilateral body insufficiently adaptable.

Additional Responsibilities

The horizon scanning agency could have additional responsibilities beyond conducting its own research and analysis. For example, it could summarize external research on long-term outer space possibilities, including think tank and national government analysis, in concise reports for policymakers. Furthermore, it could analyze long-term trends to identify changes to space activities and the environment which might not be immediately recognizable. Slow variables are conditions that change over long periods of time, rendering changes more difficult to detect [83]. Horizon scanning efforts can involve tracking slow variables and identifying tipping points for environmental or technological conditions at which regulatory or further risk mitigation actions might be necessary [144].

The inclusion of a horizon scanning agency within the multilateral space governance framework is aligned with best practices for risk management, especially catastrophic risks. The 'three lines of defense' model for addressing risk is common in industry to address significant threats [25]. The three levels of risk management include: (1) units to own risk management within existing departments, (2) an office specifically for managing risks, led by a Chief Risk Officer, and (3) an independent organization to conduct audits on the risk management process [145]. By identifying risks across the entire domain of space governance, the horizon scanning agency fulfills the second level of the framework. The first level would be fulfilled by localized regulatory units and frameworks in space, and the third level could be fulfilled by a multilateral forum which is responsible for ensuring that the horizon scanning agency has sufficient expertise and diversity to eliminate groupthink. Since the forum would not be independent of the governance process, additional funding could allow for external auditors to assist with the assessment and improvement of the horizon scanning process.

Conflict Resolution Mechanism

Existing Mechanisms

Presently existing conflict resolution mechanisms for space disputes are infrequently used [40]. The Liability Convention allows for claims to be brought against launching states via diplomatic channels, and it allows for claims to be settled by a Claims Commission if no diplomatic solution can be found [146]. However, using this mechanism necessitates the use of formal diplomatic channels between states. The mechanism is less accessible to commercial, academic, or civil actors because only states can bring claims [40]. Another limitation of the Liability Convention as a settlement mechanism is that its decisions are non-binding [147]. The Liability Convention has very rarely been utilized for settling disputes. One instance in which it was invoked involved a Soviet satellite that crashed in Canada in 1978, but a settlement was reached without a need for the Claims Commission [148].

Another existing venue for addressing space disputes is the Optional Rules for Arbitration of Disputes Relating to Outer Space Activities, developed by the Permanent Court of Arbitration (PCA) in 2012. The PCA rules are inclusive of all actors, including recognition for non-governmental organizations or commercial actors. They are also binding, offer flexibility for possible outcomes, and recognize the need for arbitrators with scientific and legal expertise in the space domain [147]. Using the rules is voluntary, and their purpose is to offer a framework for how two parties could settle a dispute through arbitration. The rules are yet to be used to solve a known space-related arbitration dispute, perhaps due to unawareness or a preference for using more general, well-established rules for arbitration for space-related disputes [149]. For example, standard arbitration procedures such as the London Court of International Arbitration and the International Centre for Dispute Resolution have been used in space-related conflicts [149].

Another recent initiative to improve space-based conflict resolution is the Courts of Space in Dubai, which established a working group to conduct scenario analysis and develop a Space Dispute Guide [150]. The effectiveness of the Courts of Space will depend on how broadly it is accepted and used, as well as how effectively it can reach and enforce solutions and adaptive measures to limit interference and improve coordination in space activities.

New Mechanism

A robust conflict resolution mechanism is necessary for effective governance of a complex, evolving system such as outer space [106]. An improved conflict resolution mechanism could support adaptive capacity by operating more in unison with the other components of the framework proposed by this paper. Such a mechanism would have two additional avenues to address a conflict. First, it could decide that an irresponsible actor loses access to some or all of the shared infrastructure for a limited period of time, or indefinitely. Second, it could mandate the horizon scanning agency to consider how standards and guidelines can be adjusted to mitigate conflicts and avoid similar disputes in the future. The conflict resolution mechanism could also mandate the horizon scanning agency to conduct posterity impact assessments, which are assessments of the long-term environmental impacts of potential policies or actions [29].

There have historically not been many disputes in the outer space domain, but the increase in the quantity and interconnectedness of actors is likely to lead to many more disputes. Because there is uncertainty regarding the nature of many of these disputes, a conflict resolution mechanism that supports adaptive capacity would be most effective.

Example

A concrete example of a space-based dispute between increasingly interconnected actors is the risk of satellite constellations, such as SpaceX's Starlink, interfering with astronomical observations [151]. SpaceX has collaborated with astronomers and tested solutions to reduce the reflectivity of their satellites, but the solutions might not be sufficient, or other satellite constellation operators might not voluntarily cooperate with astronomers [152]. Presently, astronomers would not be able to raise a claim via the Liability Convention without bringing it through a state actor, causing the dispute to be a diplomatic issue when it primarily involves commercial and scientific interests. An adaptive dispute resolution mechanism could allow astronomers to raise a concern about satellite constellation interference with their observations, and adaptive solutions could be reached. The adaptive dispute resolution mechanism might mandate the horizon scanning agency to develop new standards to ensure astronomical observation can continue with minimal disruption. Quantitative restrictions on the reflectivity and luminosity of small satellites could then be embedded in the standards, especially in regions where they may conflict with astronomy. Tipping points could also be established, beyond which a party is expected to take a further action [29]. For example, the reflectivity restrictions for future small satellites could become more stringent if thresholds of small satellite density in low-earth orbit are met. Through the identification of tipping points, adaptive resolutions could be reached despite the lack of information about effective technical solutions or future small satellite trends [153]. Requiring policy actions after tipping points are features of many long-term planning and risk assessment strategies. This includes decision making under deep uncertainty (DMDU) models, which can be effective ways to deal with technological and environmental uncertainties [25].

The new standards enacted would not only apply to the actor against which the dispute was raised, but all operators in the space domain. Hence, the decisions reached by the conflict resolution mechanism would be able to directly update guidelines and standards. In addition, actors could raise concerns prior to any harm or wrongdoing so that precautionary measures could be taken. Thus, the conflict resolution mechanism would have an impact on positively shaping adaptive space governance that exceeds the traditional role of a dispute settlement mechanism.

Verification Agency

The ability to monitor the activities of great powers and corporate actors in the outer space domain is necessary to ensure transparency, maintain trust, and mitigate the risks of harmful technology usage from malintent or negligence. There should be a designated verification agency to verify compliance with established standards and guidelines in outer space. The verification agency would verify compliance with technical standards as well as guidelines for usage of other advanced technologies such as TAI, biotechnology, or nanotechnology.

Legal Precedents

Through the Treaty on the Non-Proliferation of Nuclear Weapons (NPT), the International Atomic Energy Agency (IAEA) is legally authorized as the verification agency to ensure that states have safeguards and do not violate the provisions within the NPT on developing nuclear technologies [154]. The IAEA has the legally binding authority to verify safeguards at nuclear installation, including both the accuracy and completeness of reports. It does not have a similar authority to verify compliance with nuclear safety standards, but many actors voluntarily allow the IAEA to conduct safety reviews [154].

Article IV of the OST bans the placement of nuclear weapons or other weapons of mass destruction (WMD) in outer space [47]. However, there is no legal precedent that would require space operators with facilities and installations in outer space to allow compliance checks or monitoring of the usage of advanced technologies. A binding agreement that allows a designated agency to verify compliance with safety standards, weapons bans, and other technology restrictions would grant the necessary authority to the agency. UNOOSA might be a natural organization under which the verification agency would fall because of its inclusivity and centrality to space governance. Voluntary commitments from state and private actors to allow inspections and verification checks might suffice, as long as all actors in outer space consent to the inspections. Compliance with verification checks can be required for use of the shared infrastructure mechanisms outlined above, which may provide the necessary incentives.

Role and Responsibilities

The verification agency should have the diversity of knowledge and expertise required to ensure it can adequately perform its work. Experts with an industry or government background in space activities would be effective at checking compliance with space safety standards, but experts in advanced technologies (e.g. nuclear, biotechnology, AI) should be involved to monitor potentially risky behaviors. The verification agency could also serve an expanded role of maintaining a register of space activities, which would increase trust and transparency between actors [155].

Incidents of non-compliance identified by the verification agency could result in a loss of access to shared infrastructure. Furthermore, there would be negative political ramifications to not following regulations and guidelines, especially if the violation arose from malintent or posed a catastrophic or existential risk. The increased probability of the verification agency discovering negligent practices may disincentivize misbehavior and limit the extent to which powerful actors conduct impermissible, dangerous, or otherwise secretive research or activities in the space domain.

Structural Reforms

Introduction

We previously described how effective metagovernance involves frameworks that allow for governance to emerge through self-organization (Section: Global Governance and GCRs) [92]. Thus, governance frameworks provide limitations on the extent to which self-organization is possible, how quickly it can occur, and the forms which governance institutions and mechanisms can take. Space governance institutions are not robust and scalable across vast distances of outer space because of physical limits on communication and travel speeds. Hence, considering how existing space governance institutions—such as national governments—can supervise and govern outer space in the longterm future is erroneous. Instead, frameworks for space governance can be applied across long distances and timescales which provide an umbrella under which institutions can form and evolve. Through appropriate structural reforms to the space governance framework, we can increase the likelihood that space governance institutions in the longterm future emerge and remain aligned with our goals.

Governance institutions, of which the United Nations and national governments are terrestrial examples, are responsible for enacting and enforcing policies, regulations, standards, and risk mitigation efforts. The vast distances of space suggest that the relationship between these institutions will evolve into multiple, decentralized units of decision-making—a property of polycentric governance [26]. Effective space governance frameworks would allow for new governance institutions to form as new space activities and regions of exploration and settlement emerge. This would ensure that rogue actors cannot engage in malicious behaviors beyond the scope of governance structures, and coordination can keep pace with outward expansion. Ensuring new governance institutions enforce established norms, values, and best practices is equally important to allowing for their emergence.

Humanity and its descendants will have more time and capacity to think about moral values in the future because of the vast amounts of time and compute resources that will be available to support moral reflection. Thus, we should also consider that our concept of morality may evolve, impacting the norms, values, and best practices that we wish to promote and spread. For example, if we learn that advanced artificial intelligence systems in the far future are sentient beings that can experience suffering, such as digital minds or whole brain emulations, we may expand our moral circle to allow for their rights to be protected to mitigate s-risks [31].

The ‘long reflection’ is a concept of a period of time, after our cognitive abilities are greatly enhanced by artificial intelligence systems, that humanity can reflect upon and refine the values it wishes to pursue in the longterm future [30]. The long reflection would be significantly more difficult to enforce if humanity and its descendants were already spacefaring since mechanisms for inclusive discussion, coordination and updating values would be significantly weaker. Reflection with an advanced spacefaring civilization may cause our concept of morality to update in particular regions of the universe differently from others. For a long reflection to be feasible across astronomical distances, it is necessary for certain values of cooperation and inclusivity to be widespread before expansion begins, and for large-scale coordination mechanisms across humanity's sphere of influence to exist and allow for co-evolution of values across long distances. Under these conditions, the norms and values that guide humanity and institutional decision-making can be coordinated and updated during a long reflection into the far future.

It is also possible that our moral values are convergent, which may suggest that all institutions can independently undergo evolution based on experience and learning and reach similar norms and values in the longterm future. However, it is unclear whether morality would converge over long timescales within distinct groups, and it is also possible that some settlements do not reflect on values or lock in suboptimal values. The convergence of moral values is an unsolved philosophical question within the domain of metaethics [156]. If we assume that moral values will converge, or otherwise neglect the need for coordination of institutions across vast distances into the longterm future, we may allow for expansive space exploration and settlement to occur without frameworks in place to ensure coordination. Such uncoordinated and fragmented space expansion may be extremely difficult to reverse at later times, suggesting that the present is an ideal time to consider and establish optimal space governance frameworks.

This section considers four ideas that may guide the design of improved space governance frameworks for improved coordination and create the necessary conditions for such a framework through shifts in our present norms and values: adaptive forums, a communication network, moral circle expansion, and values handshakes. These proposals are more transformative and longtermist than the proposals in the previous section, and they may be more difficult to implement within our current society as a result. However, they would have a greater impact on ensuring longterm coordination and risk mitigation if implemented successfully. We introduce them not only to provide ideas of transformational changes for more longtermist governance, but also to encourage the space governance community to think about more ambitious changes to the current framework to ensure its adaptability and scalability.

Adaptive Forums

Uncertainty and Local Governance

Successful coordination within a space governance framework requires institutions to emerge and govern activities in all regions of outer space which are explored and settled. Our space governance framework should take into account the vast uncertainties about institutional design and fit within the space environment. Our uncertainty is deepened by a lack of complete understanding of the geographical features of outer space, and a lack of precedents for governing a system with the unique scale and geography of outer space.

Existing knowledge of the space domain has allowed for initial technical analysis into how some resource systems can be regulated, particularly the lunar ecosystem and asteroids [27]. We still lack complete knowledge about all regions of the outer space domain, and space exploration will begin before we fully understand the environment. Thus, guidelines and standards will need to further evolve based on learning after space activities take place. There is even more geopolitical, economic, scientific, and technical uncertainty about floating space settlements, terraformed planets, habitable exoplanets, or other regions of space which may eventually include settlements. Establishing effective rules for technologies that are yet to exist or unexplored regions is difficult because of our uncertainties.

Cockell (2010) emphasized the importance of the realities of local space environments in determining the proper governance rules and regulations, rendering the actors involved (e.g. local space operators) as the proper individuals to determine the rules [21]. Hence, it is essential to have an overarching governance framework that is sufficiently adaptable and suitable for self-organization and emergence of new rules and authorities. Adaptive governance frameworks are broadly applicable to the space domain, even to environments and technologies where we lack knowledge, because they allow for both universal and local regulations to update accordingly. Adaptability will translate into resilience as new space frontiers and technologies emerge.

Concept

We introduce the idea of adaptive forums as flexible institutions within an adaptive governance framework that can quickly emerge and evolve in a local environment where the geography and scope of activities are expanding. Adaptive forums are governing institutions that provide representation through some mechanism for all stakeholders involved in a particular region, activity, or resource system. The term 'adaptive' is used since membership in adaptive forums evolves over time as relevant stakeholders change, the mechanisms guiding how institutions make decisions can change, and new adaptive forums can emerge as new regions are explored and new activities are conducted. Adaptive forums emerge to allow for dialogue on the governance of a particular region, activity, or resource. Furthermore, they have the capacity to adopt and enforce rules and regulations within that defined jurisdiction.

Adaptive forums include representation of all actors who are relevant to that which they govern. The relation between the adaptive forum and the represented actor may vary across implementations. For example, all relevant stakeholders may have direct membership in a forum if a particular activity or resource is only relevant to a small group of stakeholders, such as a small colony in deep space. On the contrary, governance models such as representative democracy may be employed for governance of large space settlements, celestial bodies, or resource systems. In systems with a large number of non-human sentient beings, such as a server running digital minds, an adaptive forum which includes the entities and provides a mechanism for their representation can mitigate s-risks by enacting policies aligned with their preferences.

New adaptive forums should emerge when a particular system reaches a critical point such that discussion of rules and regulations becomes necessary to ensure norms are followed and safety measures are taken. This process is demonstrated today by industry-led groups that emerge to discuss standards and best practices for new technologies . Because these forums involve the stakeholders who have a significant level of learning and experience with respect to a particular topic, their outputs approximately reflect the most updated knowledge available . However, since they fall outside the traditional governance framework composed of international, national, and local governments, they lack mechanisms for enforcement and implementation. Furthermore, the informal process through which these forums arise results in a lack of broad representation of stakeholders across society, including traditional regulatory authorities and the academic community. Adaptive forums would capture more information, demonstrate greater inclusivity, and have more effective enforcement mechanisms than any current space governance institutions.

Implementation

The creation process and enforcement mechanisms of adaptive forums can be strengthened through guidelines for when an adaptive forum should emerge and what its initial composition should be. Guidelines should aim at ensuring inclusion of all stakeholders in adaptive forums, including commercial organizations, local regulatory authorities, and academic research organizations. Furthermore, they should provide lower and upper bounds for when a region, activity, or resource system is sufficiently relevant to merit an adaptive forum. The purpose of guidelines is to ensure that governance mechanisms are created when necessary, and that they are seen as legitimate such that they have enforcement power. Because of the unique nature of each governed system and the unpredictability of emergent phenomena, guidelines for adaptive forum formation should provide boundaries and norms rather than attempt top-down enforcement through complicated rules dictating the forum creation process.

Guidelines should require that quantified thresholds are developed that determine which stakeholders meet the requirements to be included in the initial composition of adaptive forums. Being included in adaptive forums may resemble the concept of citizenship, or in other cases, it may resemble membership in a regulatory organization.

Third-party organizations could also exist to determine when adaptive forums become necessary in particular regions and the thresholds for inclusion in new adaptive forums. These organizations may utilize horizon scanning techniques and arbitration to make their decisions, and thus they are an extension of the horizon scanning agency and conflict resolution mechanism for space governance discussed in the previous section.

Once thresholds for adaptive forums are created, other actors who meet these thresholds at future dates will automatically gain membership in the forum. Furthermore, members who fall below all thresholds could be removed from adaptive forums. As a result of the threshold requirements, the membership of adaptive forums is always evolving. However, it always includes all stakeholders relevant to a particular system in some capacity, and thus it ensures representation.

Over time, the optimal criteria to determine relevance to a particular activity or resource system may change. Thus, the forum should maintain the ability to update thresholds to determine entry requirements. Furthermore, the number of actors relevant to a particular adaptive forum may change by orders of magnitude over time. For example, a distant region of outer space may initially be utilized by a small number of actors, but it may eventually become a hub for widespread space activities. As a result, the organizational structure of an adaptive forum should be flexible over time. For example, mechanisms should exist for a forum that is initially structured as a small council to transition into a larger representative democracy by incorporating stakeholders in different fashions.

Because the adaptive forum has the ability to update its own structure and rules for membership, there should be external checks and balances imposed to curtail its power. A third-party conflict resolution mechanism should have the power to consider complaints, identify flaws in adaptive forums, introduce solutions, or override decisions by the adaptive forum under certain conditions.

Adaptive forums have the authority and legitimacy to govern their respective systems. This means they can create and enforce necessary regulations more effectively than legacy structures, whose capacities and institutional fit will be challenged if structural reforms do not occur. Adaptive forums would also be more effective than legacy structures at proactively addressing emerging and future trends through horizon scanning, early identification of key stakeholders, and co-evolution with the environment.

Communication Network

We previously proposed shared large-scale infrastructure projects as a mechanism to improve coordination between actors involved in spacefaring (Section: Shared Infrastructure). We argued that shared infrastructure is scalable and able to promote coordination across vast distances and timescales. We further propose a shared network of probes for communication, coordination, and monitoring across large distances of space. Such probes could be spread across interplanetary and interstellar distances, allowing for communication across humanity's sphere of influence as long as the expansion of the network outpaces space exploration. Beyond enabling communication, the network would provide more information about the space domain, monitor resource usage and levels, and supervise space activities as they continue to emerge in the near future.

Previous work has suggested that advanced technological civilizations may attempt to devise large communication networks of probes across star systems [157]. Furthermore, probes may be launched to star systems with inhabited planets, inside of which they could become trapped [158]. A network of 'node probes', with one or several probes in each star system, would allow for two way communication that cannot be achieved without a sufficiently dense network of probes [159].

Developing such a communication network would be an intergenerational project, and the present is an appropriate time to begin developing the technical capacity—ensuring the project can be launched approximately when transformative technologies emerge is essential. Some actors are already researching the technologies that pose bottlenecks to interstellar probes [160]. Breakthrough Starshot is developing a light sail to travel to Alpha Centauri at relativistic speeds [161]. Although these projects intend to launch one-way probes, similar technologies are necessary for a communication network. The network may initially only include probes throughout the solar system, but it could quickly be expanded to include probes at greater distances when this becomes necessary to monitor space activities.

A stable, star-spanning network of probes would still have limitations on its ability to ensure coordination because of the time required for information to travel. Hence, such a system could not be relied upon to ensure that norms and values between interstellar civilizations can co-evolve with short feedback loops. However, information would be able to slowly diffuse throughout the galaxy, resulting in frontiers of information which spread from the source at quicker rates than events unfold or civilizations expand. Recipients of information at far distances would be able to update their plans accordingly, including changing policies or planning for a conflict with a misaligned actor. The system would be a significant improvement over a lack of coordination, especially for creating stability.

If the communication network is redundant and robust, its maintained existence would be independent of any natural or artificial catastrophes. For example, if a supernova were to occur in one region of the galaxy, undamaged probes that detect the event would be able to transmit the information throughout the network, and replacement probes could be sent. Such a communication network also has the benefit of being decentralized, reducing the probability that a malicious actor can gain totalitarian power. Alternative methods for surveillance across vast distances, including large telescopes that can monitor activities in distant star system with high resolution, could be controlled by particular actors and lead to dangerous power asymmetries.

In the next two subsections, we explain how ensuring most spacefaring actors are initially value-aligned can limit the extent to which malicious actors can harm coordination across the network.

Moral Circle Expansion

When humanity or its descendants first start expanding at vast distances from earth (i.e. beyond the solar system), the values of actors engaging in spacefaring may be vastly different from each other. Furthermore, they are likely to reflect the values of competing great powers or private organizations—artificial intelligence or space industry corporations—that are driving space expansion. If this era of accelerated space exploration begins with fragmented institutions, a lack of coordination, and unclear values, it will become significantly more difficult in the future to establish shared norms across humanity's entire sphere of influence and coordinate between actors. If reflection on our values in the present day can lead us to agree upon some robustly beneficial norms and values for the longterm future before space exploration begins, early-stage coordination and value alignment are more likely to follow.

Anthis and Paez (2021) define moral circle expansion as the process through which "a number of entities which used to be given less than full moral consideration at [time] t are now given more moral consideration." [31]. Some of the most significant s-risks in the far future involve suffering of non-human beings, including wild animals, directed panspermia, or digital minds [32]. Furthermore, hypothetical discussions of a 'galactic club' of extraterrestrial civilizations lead to questions over whether human rights would be protected by more powerful civilizations, suggesting that moral circle expansion to include all sentient beings could also serve to safeguard humanity in extraterrestrial contact scenarios [162]. Thus, agreements for protections of non-human sentient beings in outer space could significantly mitigate future s-risks and x-risks. A critical factor in ensuring that such norms are followed in early space exploration is reaching agreement among key stakeholders, including great powers and advanced private corporations. Hence, widespread agreement on moral circle expansion would not only mitigate future risks, but it would improve coordination in early spacefaring.

A United Nations resolution to protect the interests of non-human sentient beings in outer space would be a positive step, but successful implementation would require national regulations within great powers and buy-in from corporations, including future corporations that are space-based or beyond the supervision of national governments in outer space. Furthermore, the willingness of state actors to support such a resolution is unclear. Ensuring that moral circle expansion guides key stakeholders in national governments and private corporations could be a more effective way to influence future values and norms among space actors. Ultimately, bilateral agreements between spacefaring nations or corporate governance measures to commit to such norms could help ensure that the values of space actors are aligned with future generations.

Values Handshakes

While moral circle expansion can create a greater degree of value alignment among space actors and mitigate significant s-risks, space actors are still likely to have some conflicting interests and values. Each entity involved in settling the universe—biological or technological—is likely to have its own set of values and preferences. Values handshakes are agreements to compromise on the values of multiple entities [33]. Such a compromise would be an alternative to conflict that arises over ideological disputes, which could be devastating for all parties on an astronomical scale.

The concept of values handshakes is closely related to acausal trade, which involves the ability to coordinate with other actors for mutual benefit without any communication or effect on each other [163]. Agents can hypothetically agree on values handshakes without ever coming in contact. This mechanism seems to work in a multiverse setting where superintelligences can approximate the distribution of preferences of superintelligences across universes, reaching a values handshake with all possible actors [163]. Although perfect acausal trade within our galaxy is not possible because of the plausibility of future contact, a weaker version of acausal trade can be applied which involves coordination with future entities by making certain pre-commitments to robustly beneficial norms and values [164].

Although various entities involved in space exploration may have different values, most entities are likely to have values that are considered reasonable . These may include promoting commercial interests, contributing to scientific advancement, spreading aesthetics and beauty, or maximizing happiness. Actors who recognize the major risks and uncertainties in spacefaring with transformative technologies can pre-commit to respecting the values of other reasonable agents at a smaller cost to themselves, resulting in large benefits from improved coordination and less conflict. Furthermore, agents can pre-commit to the peaceful use of outer space as long as others respect their pre-commitments. Such agreements could be considered values handshakes.

If early spacefaring actors agree to moral circle expansion and values handshakes, there is more likely to be broad coordination and collaboration among them. Furthermore, all future agents will have strong incentives to agree to these norms and values to avoid disputes with more advanced spacefaring actors.

Making pre-commitments does not imply that norms and values will be upheld, and they are extremely unlikely to be upheld by all actors far in the longterm future without enforcement and retribution mechanisms. If value-aligned actors also pre-commit to retribution against actors who violate the norms of outer space—including harmful interference with neighbors—there could be a unified response against misaligned actors. If actors are initially aligned with shared values and norms, the coalition of value-aligned actors is likely to exceed the strength of unaligned actors. While unaligned actors would be able to cause massive damage and suffering in their vicinity, such as seizing nearby infrastructure or running digital minds on servers that experience vast amounts of suffering, actors at greater distances in all directions would have time to recognize and respond to the threat. The speed at which information spreads would greatly exceed the speed at which a civilization can spread [165].

Warfare is likely to be extremely costly and difficult on large scales because of the vast distances involved, so actors are likely to have a preference for refraining from warfare with more advanced or similarly advanced powers. Because of the long travel times in outer space which reduce coordination, humanity's sphere of influence is likely to be very decentralized without central points of failure. As a result, threats would be able to be mitigated before they can propagate throughout the entire network or cause widespread damage beyond local regions.

As humanity spreads across further distances in outer space, the network of actors will become much larger and distributed. The duration of conflicts may increase since the time taken for information to reach other actors and a response to be initiated would increase. However, the strength of the coalition of actors with aligned values would likely become stronger relative to misaligned actors, and the necessary time for a misaligned actor to seize the entire network would be much longer. Hence, as more time passes, threats on existential scales would decrease. On the order of billions of years, our galaxy may reach a stable state where norms and values are largely locked in, and threats to the network are negligible [35].

Future Research Directions

This paper provides an introductory analysis of various long-term risks that may emerge in the outer space domain, the shortcomings of our current space governance framework to manage them, and proposals for an improved space governance framework. More research to extend several elements of the analysis in this paper are strongly encouraged.

The Space Futures Initiative was recently established to conduct research to improve the long-term prospects of humanity and its descendants in outer space, and its research agenda includes many of the topics addressed in this paper. Below, we propose several ideas for further research beyond this paper.

First, more research should be done on the nature, quantity, and timelines of space activities during the In-Space Economy Era. The extent to which these activities emerge greatly affects the importance of the space domain in shaping humanity's response to transformative technologies. Furthermore, activities within an interconnected in-space economy can amplify and mitigate existing terrestrial catastrophic and existential risks. For example, satellites can provide resilient backup infrastructure for communications and utilities, and biological research in microgravity conditions can lead to the development and manufacturing of medicines to treat a variety of diseases [1].

Second, further work should be conducted on the additional risks that may arise in the longterm future when there are more sentient beings in outer space. Cyberattacks on space-based infrastructure might become a greater threat, especially in very longterm futures when space-based servers are running digital minds. If weapons of mass destruction (biological, nuclear, chemical, or technological) become relatively easy to manufacture, it will become extremely important to have levels of surveillance, monitoring, policing, and enforcement in space which match those of earth.

Third, the extent to which a totalitarian lock-in is likely in outer space could be further explored. Local regions of space may be naturally suitable for a totalitarian ruler because of the adverse conditions, or freedom engineering may be able to prevent such an outcome [21]. On large scales, totalitarian control may be feasible for solar systems through megastructures at critical locations for energy and surveillance, but this may be more difficult than local control of a space colony. Furthermore, the relationship between civilizations across star systems could be further explored. A totalitarian lock-in may be less of a concern across star systems, but the possibility of values lock-in that persists across interstellar settlement could be further analyzed. In addition, governance of servers running digital minds is a potentially important question related to totalitarianism and mitigating s-risks.

Finally, more research should be conducted on possible worlds in which specialized AI systems could be used by humans to significantly accelerate scientific and technological progress necessary for space development prior to the existence of an artificial general intelligence. In these worlds, space governance plays a significant role in determining the initial conditions in which a superintelligent system may emerge. Furthermore, if a superintelligent system proves to be impossible or requires much longer timescales than many experts currently anticipate, the utilization of narrow AI systems for space development might lead to extremely powerful influence over the long-term future.

Conclusion

The first section of this paper (section: Current Framework) highlighted the shortcomings of our current multilateral space governance framework, including the increase in fragmentation, lack of coordination, and weak enforcement mechanisms. The second section (Section: Adaptive Governance) discussed how space activities may quickly evolve with transformative technologies, and it introduced theoretical frameworks that might be more effective than the current framework at managing the complexity and interconnectedness of outer space. This section also placed space governance in the context of global governance and UN metagovernance, which display similar shortcomings.

The third section (Section: Scenario Analysis) analyzed catastrophic, existential, and suffering risks that might become more severe unless space governance mechanisms are improved. Three categories of risks were analyzed: sustainability failures, biological spreading, and transformative artificial intelligence. In the fourth section (Section: Policy Proposals), four mechanisms were introduced that can be implemented within the existing space governance framework to enhance its adaptive capacity: shared infrastructure, horizon scanning, a conflict resolution mechanism, and a verification agency. In the fifth section (Section: Structural Reforms), we outlined four longtermist structural and ideological shifts to transform current space governance and activities: adaptive forums, a communication network, moral circle expansion, and values handshakes. We expect these proposals to be less tractable than the policy proposals, but we introduce them to encourage further discussion and more ambitious thinking about space governance solutions for longterm futures.

Large transformations to our space governance framework and space exploration norms are likely necessary to ensure a prosperous long-term future, but acceptance of the critical importance of longtermism and future generations by modern key stakeholders are necessary conditions to enable such reforms. The United Nations 'Our Common Agenda' report, issued by the Secretary-General, calls for a 'Summit of the Future' to take place in 2023 [104]. The Summit would include a dialogue on outer space and pay special attention to future generations. Such an occasion might be an opportunity to initiate the process of implementing transformative space governance reforms. Inclusivity of both state and nonstate actors in the ongoing dialogue is also essential to success, including both commercial and scientific interests.

This paper intends to be a step towards thinking about long-term space futures within the space policy and space governance communities. In particular, we hope to demonstrate that long-term space futures have some path dependencies on our existing space governance frameworks, and scalable, coordinated space governance frameworks can improve the prospects of long-term space futures. By concretely describing the risks within our space futures, we hope to have properly addressed the urgency of considerations of future generations and the long-term future in modern space governance dialogue.

Acronyms

ADR - Active Debris Removal

CCSDS - Consultative Committee for Space Data Standards

CNSA - China National Space Administration

CONFERS - Consortium for Execution of Rendezvous and Servicing Operations

COPUOS - Committee on the Peaceful Uses of Outer Space

DARPA - Defense Advanced Research Projects Agency

DMDU - Decision Making under Deep Uncertainty

GCR - Global Catastrophic Risk

GEGSLA - Global Expert Group on Sustainable Lunar Activities

GPT - General Purpose Technology

IADC - Inter-Agency Space Debris Coordination Committee

IAEA - International Atomic Energy Agency

ISEE - In-Space Economy Era

ISS - International Space Station

ISO - International Organization for Standardization

LSC - Legal Subcommittee (COPUOS)

NASA - National Aeronautics and Space Administration

NPT - Non-proliferation Treaty

ODMSP - Orbital Debris Mitigation Standard Practices (United States)

PCA - Permanent Court of Arbitration

RPO - Rendezvous and Proximity Operations

TAI - Transformative Artificial Intelligence

TTE - Transformative Technologies Era

UN - United Nations

UNOOSA - United Nations Office for Outer Space Affairs

S-risk - Suffering Risk

SSA - Space Situational Awareness

STM - Space Traffic Management

STSC - Scientific and Technical Subcommittee (COPUOS)

WMD - Weapons of Mass Destruction

X-Risk - Existential Risk

Acknowledgements

I wish to thank Giuliana Rotola and Anders Sandberg for invaluable mentorship, as well as Will Aldred, Jacob Arbeid, Som Bagchi, Nathan Barnard, Catherine Brewer, Madeleine Chang, Martin Elvis, Marcel Ernst, Lukas Finnveden, Zach Freitas-Groff, Kirke Joamets, Christopher Lankhof, Fin Moorhouse, Morgan Livingston, Yannick Mühlhäuser, Richard Ngo, Cullen O'Keefe, Rory Svarc, Stefan Torges, and Olaf Willner for useful comments and feedback. This research was made possible through funding from the Cambridge Existential Risk Initiative. The opinions expressed in this publication are mine, and all mistakes are my own.

References

[1] James Black, Linda Slapakova, and Kevin Martin. Future Uses of Space Out to 2050: Emerging threats and opportunities for the UK National Space Strategy. Technical report, RAND Corporation, March 2022. URL: https://www.rand.org/pubs/research_reports/RRA609-1.html.

[2] Jeffrey Montes, Jessy Kate Schingler, and Phillip Metzger. Asce paper: Pad for humanity: Lunar spaceports as critical shared infrastructure. *Open Lunar Foundation*, 2021. URL: https://www.openlunar.org/library/pad-for-humanity.

[3] Michio Kaku. *The Future of Humanity: Terraforming Mars, Interstellar Travel, Immortality, and Our Destiny Beyond*. Penguin Books, 2019. URL: https://www.penguin.com.au/books/the-future-of-humanity-9780141986067.

[4] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. Google-Books-ID: 7_H8AwAAQBAJ.

[5] Freeman J. Dyson. Search for Artificial Stellar Sources of Infrared Radiation. *Science*, 131:1667–1668, June 1960. ADS Bibcode: 1960Sci...131.1667D. URL: https://ui.adsabs.harvard.edu/abs/1960Sci...131.1667D, doi:10.1126/science.131.3414.1667.

[6] G. K. O'Neill. The colonization of space. *Physics Today*, 27:32–40, January 1974. ADS Bibcode: 1974PhT....27i..32O. URL: https://ui.adsabs.harvard.edu/abs/1974PhT....27i..32O, doi:10.1063/1.3128863.

[7] Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper. Introduction. In Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, editors, *Decision Making under Deep Uncertainty: From Theory to Practice*, pages 1–20. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-05252-2_1.

[8] Donald J. Kessler and Burton G. Cour-Palais. Collision frequency of artificial satellites: The creation of a debris belt. *Journal of Geophysical Research: Space Physics*, 83(A6):2637–2646, 1978. URL: https://onlinelibrary.wiley.com/doi/abs/10.1029/JA083iA06p02637, doi:10.1029/JA083iA06p02637.

[9] Jakub Drmola and Tomas Hubik. Kessler Syndrome: System Dynamics Model. *Space Policy*, 44-45:29–39, August 2018. URL: https://www.sciencedirect.com/science/article/pii/S0265964617300966, doi:10.1016/j.spacepol.2018.03.003.

[10] Martin Elvis and Tony Milligan. How much of the Solar System should we leave as wilderness? *Acta Astronautica*, 162:574–580, September 2019. URL: https://www.sciencedirect.com/science/article/pii/S0094576517318507, doi:10.1016/j.actaastro.2019.03.014.

[11] W. Neil Adger, Suraje Dessai, Marisa Goulden, Mike Hulme, Irene Lorenzoni, Donald R. Nelson, Lars Otto Naess, Johanna Wolf, and Anita Wreford. Are there social limits to adaptation to climate change? *Climatic Change*, 93(3):335–354, April 2009. doi:10.1007/s10584-008-9520-z.

[12] Philip Trammell and Anton Korinek. Economic growth under transformative AI. *Global Priorities Institute*, September 2020. URL: https://globalprioritiesinstitute.org/philip-trammell-and-anton-korinek-economic-growth-under-transformative-ai/.

[13] J. D Rummel and L Billings. Issues in planetary protection: policy, protocol and implementation. *Space Policy*, 20(1):49–54, February 2004. URL: https://www.sciencedirect.com/science/article/pii/S0265964603000845, doi:10.1016/j.spacepol.2003.11.005.

[14] *Assessment of the Report of NASA's Planetary Protection Independent Review Board*. 2020. URL: https://www.nap.edu/read/25773/chapter/1, doi:10.17226/25773.

[15] M. Meot-Ner and G. L. Matloff. Directed panspermia - A technical and ethical evaluation of seeding nearby solar systems. *Journal of the British Interplanetary Society*, 32:419–423, November 1979. ADS Bibcode: 1979JBIS...32..419M. URL: https://ui.adsabs.harvard.edu/abs/1979JBIS...32..419M.

[16] Gary David O'Brien. Directed Panspermia, Wild Animal Suffering, and the Ethics of World-Creation. *Journal of Applied Philosophy*, 39(1):87–102, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/japp.12538. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/japp.12538, doi:10.1111/japp.12538.

[17] Wild animal welfare in the far future. *EA Forum*. URL: https://forum.effectivealtruism.org/posts/MKmowJNCeJCaitK3x/wild-animal-welfare-in-the-far-future.

[18] Simon Zhuang and Dylan Hadfield-Menell. Consequences of Misaligned AI. February 2021. arXiv:2102.03896 [cs]. URL: http://arxiv.org/abs/2102.03896.

[19] Sam Clarke. Clarifying "What failure looks like". *AI Alignment Forum*. URL: https://www.alignmentforum.org/posts/v6Q7T335KCMxujhZu/clarifying-what-failure-looks-like.

[20] Joseph Carlsmith. Is Power-Seeking AI an Existential Risk? June 2022. arXiv:2206.13353 [cs]. URL: http://arxiv.org/abs/2206.13353.

[21] Charles Cockell. Essay on the Causes and Consequences of Extraterrestrial Tyranny. *Journal of the British Interplanetary Society*, 63:15–37, January 2010.

[22] Holden Karnofsky. Digital People Would Be An Even Bigger Deal. *Cold Takes*, July 2021. URL: https://www.cold-takes.com/how-digital-people-could-change-the-world/.

[23] Thomas Dietz, Elinor Ostrom, and Paul Stern. The Struggle to Govern the Commons. 302:7, 2003.

[24] Mark Beeson. *Rethinking Global Governance*. Bloomsbury Publishing, February 2019. Google-Books-ID: VxxHEAAAQBAJ.

[25] Toby Ord, Angus Mercer, and Sophie Dannreuther. Future Proof. *The Centre for Long-Term Resilience*, 2021. URL: https://www.longtermresilience.org/futureproof.

[26] Elinor Ostrom. Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change*, 20(4):550–557, October 2010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959378010000634, doi:10.1016/j.gloenvcha.2010.07.004.

[27] Martin Elvis, Alanna Krolikowski, and Tony Milligan. Concentrated Lunar Resources: Imminent Implications for Governance and Justice. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2188):20190563, January 2021. arXiv:2103.09045 [astro-ph, physics:physics]. URL: http://arxiv.org/abs/2103.09045, doi:10.1098/rsta.2019.0563.

[28] Effie Amanatidou, Maurits Butter, Vicente Carabias, Totti Könnölä, Miriam Leis, Ozcan Saritas, Petra Schaper-Rinkel, and Victor van Rij. On concepts and methods in horizon scanning: Lessons from initiating policy dialogues on emerging issues. *Science and Public Policy*, 39(2):208–221, March 2012. doi:10.1093/scipol/scs017.

[29] Tyler M John and William MacAskill. Longtermist Institutional Reform. *Legal Priorities Project*, 2021. URL: https://globalprioritiesinstitute.org/wp-content/uploads/Tyler-M-John-and-William-MacAskill_Longtermist-institutional-reform.pdf.

[30] Research Agenda. *Global Priorities Institute*, 2020. URL: https://globalprioritiesinstitute.org/research-agenda-web-version/.

[31] Jacy Reese Anthis and Eze Paez. Moral circle expansion: A promising strategy to impact the far future. *Futures*, 130:102756, June 2021. URL: https://www.sciencedirect.com/science/article/pii/S0016328721000641, doi:10.1016/j.futures.2021.102756.

[32] Brian Tomasik. Risks of Astronomical Future Suffering. July 2019. URL: https://longtermrisk.org/risks-of-astronomical-future-suffering/.

[33] Scott Alexander. The Hour I First Believed. *Slate Star Codex*, April 2018. URL: https://slatestarcodex.com/2018/04/01/the-hour-i-first-believed/.

[34] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308, February 2006. URL: https://www.sciencedirect.com/science/article/pii/S037015730500462X, doi:10.1016/j.physrep.2005.10.009.

[35] Holden Karnofsky. All Possible Views About Humanity's Future Are Wild. *Cold Takes*, 2021. URL: https://www.cold-takes.com/all-possible-views-about-humanitys-future-are-wild/.

[36] Toby Ord. *The Precipice*. 2020. URL: https://theprecipice.com.

[37] Working Groups of the Committee and its Subcommittees. *United Nations Office for Outer Space Affairs*. URL: https://www.unoosa.org/oosa/en/ourwork/copuos/working-groups.html.

[38] Report of the Legal Subcommittee on its sixtieth session. *Committee on the Peaceful Uses of Outer Space*, 2021. URL: https://www.unoosa.org/oosa/en/ourwork/copuos/lsc/2021/index.html.

[39] Space Law Treaties and Principles. *United Nations Office for Outer Space Affairs*. URL: https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties.html.

[40] Fabio Tronchetti. *Fundamentals of Space Law and Policy*. Briefs in Space Development. Springer, 2013.

[41] Daniel L. Oltrogge and Ian A. Christensen. Space governance in the new space era. *Journal of Space Safety Engineering*, 7(3):432–438, September 2020. URL: https://www.sciencedirect.com/science/article/pii/S2468896720300550, doi:10.1016/j.jsse.2020.06.003.

[42] Start-Up Space Report 2022. *BryceTech*, 2022. URL: https://brycetech.com/reports.

[43] Nicholas Eftimiades. Small satellites: The implications for national security. *Atlantic Council*, May 2022. URL: https://www.atlanticcouncil.org/in-depth-research-reports/report/small-satellites-the-implications-for-national-security/.

[44] NASA: Artemis. *NASA*, 2022. URL: https://www.nasa.gov/specials/artemis/index.html.

[45] Tereza Pultarova. Russia, China reveal moon base roadmap but no plans for astronaut trips yet. *Space.com*, June 2021. URL: https://www.space.com/china-russia-international-lunar-research-station.

[46] Lucien Rapp, Maria Topka, and Lucas Mallowan. Which Jurisdiction for Private In-space Assembled Autonomous Platforms? *Space Policy*, 56:101413, May 2021. URL: https://www.sciencedirect.com/science/article/pii/S0265964621000059, doi:10.1016/j.spacepol.2021.101413.

[47] Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies. *United Nations General Assembly*, 1967. URL: https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/introouterspacetreaty.html.

[48] Kaitlyn Johnson. Key Governance Issues in Space. September 2020. URL: https://www.csis.org/analysis/key-governance-issues-space.

[49] CONFERS. URL: https://www.satelliteconfers.org/.

[50] H. Stokes, Y. Akahoshi, C. Bonnal, R. Destefanis, Y. Gu, A. Kato, A. Kutomanov, A. LaCroix, S. Lemmens, A. Lohvynenko, D. Oltrogge, P. Omaly, J. Opiela, H. Quan, K. Sato, M. Sorge, and M. Tang. Evolution of ISO's space debris mitigation standards. *Journal of Space Safety Engineering*, 7(3):325–331, September 2020. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468896720300689, doi:10.1016/j.jsse.2020.07.004.

[51] Report of the Scientific and Technical Subcommittee on its fifty-ninth session. *Committee on the Peaceful Uses of Outer Space*, 2022. URL: https://www.unoosa.org/oosa/en/ourwork/copuos/stsc/2022/index.html.

[52] Rossana Deplano. The Artemis Accords: Evolution or Revolution in International Space Law? *International & Comparative Law Quarterly*, 70(3):799–819, July 2021. Publisher: Cambridge University Press. URL: https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/artemis-accords-evolution-or-revolution-in-international-space-law/DC08E6D42F7D5A971067E6A1BA442DF1, doi:10.1017/S0020589321000142.

[53] Report of the Legal Subcommittee on its sixty-first session. *Committee on the Peaceful Uses of Outer Space*, 2022. URL: https://www.unoosa.org/oosa/en/ourwork/copuos/lsc/2022/index.html.

[54] Working Paper on the Establishment of a Working Group on Space Resources. *Committee on the Peaceful Uses of Outer Space*, 2021. URL: https://www.unoosa.org/documents/pdf/copuos/lsc/space-resources/Non-paper-on-the-Establishment-of-a-Working-Group-on-Space_Resources-at-COPUOS_LSC-27-05-2021.pdf.

[55] Victor Galaz, Beatrice Crona, Henrik Österblom, Per Olsson, and Carl Folke. Polycentric systems and interacting planetary boundaries — Emerging governance of climate change–ocean acidification–marine biodiversity. *Ecological Economics*, 81:21–32, September 2012. URL: https://www.sciencedirect.com/science/article/pii/S0921800911004964, doi:10.1016/j.ecolecon.2011.11.012.

[56] Elinor Ostrom. Coping with Tragedies of the Commons. *Annual Review of Political Science*, 2(1):493–535, June 1999. URL: https://www.annualreviews.org/doi/10.1146/annurev.polisci.2.1.493, doi:10.1146/annurev.polisci.2.1.493.

[57] Rakhyun E. Kim. The emergent network structure of the multilateral environmental agreement system. *Global Environmental Change*, 23(5):980–991, October 2013. URL: https://www.sciencedirect.com/science/article/pii/S095937801300112X, `doi:10.1016/j.gloenvcha.2013.07.006`.

[58] ISO and the UN - Working together for international standardization. *International Organization for Standardization*. URL: https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100432.pdf.

[59] U.S. Government Orbital Debris Mitigation Standard Practices. *United States*, November 2019. URL: https://orbitaldebris.jsc.nasa.gov/library/usg_orbital_debris_mitigation_standard_practices_november_2019.pdf.

[60] Brian Greenhill and Yonatan Lupu. Clubs of Clubs: Fragmentation in the Network of Intergovernmental Organizations. *International Studies Quarterly*, 61(1):181–195, March 2017. `doi:10.1093/isq/sqx001`.

[61] Jeff Foust. Japan passes space resources law. *SpaceNews*, June 2021. URL: https://spacenews.com/japan-passes-space-resources-law/.

[62] Sean Potter. Saudi Arabia Signs Artemis Accords. *NASA*, July 2022. URL: http://www.nasa.gov/feature/saudi-arabia-signs-artemis-accords.

[63] The Artemis Accords: Principles for Cooperation in the Civil Exploration and Use of the Moon, Mars, Comets, and Asteroids for Peaceful Purposes. *NASA*, 2020. URL: https://www.nasa.gov/specials/artemis-accords/img/Artemis-Accords-signed-13Oct2020.pdf.

[64] Building Blocks for the Development of an International Framework for the Governance of Space Resource Activities. *The Hague International Space Resources Governance Working Group*, November 2019. URL: https://boeken.rechtsgebieden.boomportaal.nl/publicaties/9789462361218#152.

[65] The Hague International Space Resources Governance Working Group. *Leiden University*. URL: https://www.universiteitleiden.nl/en/law/institute-of-public-law/institute-of-air-space-law/the-hague-space-resources-governance-working-group.

[66] Martin Švec, Petr Boháček, and Nikola Schmidt. Utilization of Natural Resources in Outer Space: Social License to Operate as an Alternative Source of Both Legality and Legitimacy. URL: https://www.ogel.org/article.asp?key=3872.

[67] Agreement Governing the Activities of States on the Moon and Other Celestial Bodies. *United Nations General Assembly*, 1979. URL: https://www.unoosa.org/pdf/gares/ARES_34_68E.pdf.

[68] Maria Lucas-Rhimbassen. The COST of Joining Legal Forces on a Celestial Body of Law and Beyond: Anticipating Future Clashes between Corpus Juris Spatialis, Lex Mercatoria, Antitrust and Ethics. *Space Policy*, 59:101445,

February 2022. URL: https://www.sciencedirect.com/science/article/pii/S0265964621000370, doi:10.1016/j.spacepol.2021.101445.

[69] Fengna Xu and Jinyuan Su. New Elements in the Hague Space Resources Governance Working Group's Building Blocks. *Space Policy*, 53, August 2020. URL: https://www.sciencedirect.com/science/article/pii/S026596462030028X.

[70] Elliot Ji, Michael B. Cerny, and Raphael J. Piliero. What Does China Think About NASA's Artemis Accords? September 2020. URL: https://thediplomat.com/2020/09/what-does-china-think-about-nasas-artemis-accords/.

[71] Department of Defense and Full-Year Continuing Appropriations Act, 2011. *United States Congress*, 2011. URL: https://www.congress.gov/112/plaws/publ10/PLAW-112publ10.htm.

[72] Space Debris Mitigation Guidelines of the Committee on the Peaceful Uses of Outer Space. *United Nations Office for Outer Space Affairs*, 2010. URL: https://www.unoosa.org/pdf/publications/st_space_49E.pdf.

[73] CSSMA. URL: https://cssma.space/.

[74] About GEGSLA – Moon Village Association. URL: https://moonvillageassociation.org/gegsla/about/.

[75] Rajeswari Pillai Rajagopalan. Debate on Space Code of Conduct: An Indian Perspective. *ORF*. URL: https://www.orfonline.org/research/debate-on-space-code-of-conduct-an-indian-perspective/.

[76] Michael R. Migaud, Robert A. Greer, and Justin B. Bullock. Developing an Adaptive Space Governance Framework. *Space Policy*, 55:101400, February 2021. URL: https://www.sciencedirect.com/science/article/pii/S0265964620300424, doi:10.1016/j.spacepol.2020.101400.

[77] Carl Folke, Thomas Hahn, Per Olsson, and Jon Norberg. Adaptive Governance Of Social-Ecological Systems. *Annual Review of Environment and Resources*, 30(1):441–473, November 2005. URL: https://www.annualreviews.org/doi/10.1146/annurev.energy.30.050504.144511, doi:10.1146/annurev.energy.30.050504.144511.

[78] Len Fisher and Anders Sandberg. A Safe Governance Space for Humanity: Necessary Conditions for the Governance of Global Catastrophic Risks. *Global Policy*, n/a(n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.13030. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.13030, doi:10.1111/1758-5899.13030.

[79] Oltjon Kodheli, Eva Lagunas, Nicola Maturo, Shree Krishna Sharma, Bhavani Shankar, Jesus Fabian Mendoza Montoya, Juan Carlos Merlano Duncan, Danilo Spano, Symeon Chatzinotas, Steven Kisseleff, Jorge Querol, Lei

Lei, Thang X. Vu, and George Goussetis. Satellite Communications in the New Space Era: A Survey and Future Challenges. *IEEE Communications Surveys & Tutorials*, 23(1):70–109, 2021. Conference Name: IEEE Communications Surveys & Tutorials. `doi:10.1109/COMST.2020.3028247`.

[80] NASA and Commercial Space. Publisher: Brian Dunbar. URL: https://www.nasa.gov/offices/oct/partnership/comm_space/.

[81] Jerry Wright. NASA Releases COTS Final Report. *NASA*, April 2015. URL: http://www.nasa.gov/content/nasa-releases-cots-final-report.

[82] Daniel Deudney. *Dark Skies: Space Expansionism, Planetary Geopolitics, and the Ends of Humanity*. Oxford University Press, Oxford, New York, 2020.

[83] Tom Pegram and Julia Kreienkamp. Governing Complexity Design Principles for Improving the Governance of Global Catastrophic Risks. *The Global Challenges Foundation*, January 2020. URL: https://globalchallenges.org/a-knowledge-overview-on-global-catastrophic-risks-and-the-global-governance-gap/.

[84] Jessica F. Green. Governing Complex Systems: Social Capital for the Anthropocene. By Oran R. Young. Cambridge, MA: The MIT Press, 2017. 296p. 30.00 paper. *Perspectives on Politics*, 16(1):278–280, March 2018. Publisher: Cambridge University Press. URL: https://www.cambridge.org/core/journals/perspectives-on-politics/article/governing-complex-systems-social-capital-for-the-anthropocene-by-oran-r-young-cambridge-ma-the-mit-press-2017-296p-9000-cloth-3000-paper/60D2A2635058B29E1C7F54D1C9C18847, `doi:10.1017/S1537592717003383`.

[85] Matthew Watson. Time to stage trials of engineering the atmosphere to cool Earth. *New Scientist*, 2016. URL: https://www.newscientist.com/article/2113880-time-to-stage-trials-of-engineering-the-atmosphere-to-cool-earth/.

[86] Nicholas Crisp, Katharine Smith, and Peter Hollingsworth. Small Satellite Launch to LEO: A Review of Current and Future Launch Systems. *Transactions of the Japan Society for Aeronautical and Space Sciences, Aerospace Technology Japan*, 12(ists29):Tf_39–Tf_47, 2014. `doi:10.2322/tastj.12.Tf_39`.

[87] Alanna Krolikowski and Martin Elvis. Marking Policy for New Asteroid Activities: In Pursuit of Science, Settlement, Security, or Sales? *Space Policy*, 47:7–17, February 2019. URL: https://www.sciencedirect.com/science/article/pii/S0265964618300262.

[88] David Kennedy. *A World of Struggle*. Princeton University Press, February 2016. URL: https://press.princeton.edu/books/hardcover/9780691146782/a-world-of-struggle.

[89] Ian Goldin. *Divided Nations: Why global governance is failing, and what we can do about it*. OUP Oxford, March 2013. Google-Books-ID: uJkxFIWE-QBEC.

[90] Christopher Nathan and Keith Hyams. Global policymakers and catastrophic risk. *Policy Sciences*, 55(1):3–21, March 2022. `doi:10.1007/s11077-021-09444-0`.

[91] Jessica Green, Thomas Hale, and Jeff Colgan. The Existential Politics of Climate Change. *Global Policy Journal*. URL: https://www.globalpolicyjournal.com/blog/21/02/2019/existential-politics-climate-change.

[92] Stamatios Christopoulos, Balazs Horvath, and Michael Kull. Advancing the Governance of Cross-Sectoral Policies for Sustainable Development: A Metagovernance Perspective. June 2012. URL: https://onlinelibrary.wiley.com/doi/10.1002/pad.1629.

[93] Marianne Beisheim and Nils Simon. Multistakeholder Partnerships for the SDGs: Actors' Views on UN Metagovernance. *Global Governance: A Review of Multilateralism and International Organizations*, 24(4):497–515, December 2018. Publisher: Brill Nijhoff. URL: https://brill.com/view/journals/gg/24/4/article-p497_3.xml, `doi:10.1163/19426720-02404003`.

[94] Nick Bostrom and Milan M. Cirkovic. *Global Catastrophic Risks*. OUP Oxford, September 2011. Google-Books-ID: sTkfAQAAQBAJ.

[95] Nick Bostrom. Existential Risks: Analyzing Human Extinction Scenarios. 2002. URL: https://nickbostrom.com/existential/risks.

[96] Max Daniel. S-risks: Why they are the worst existential risks, and how to prevent them (EAG Boston 2017). *Center on Long-Term Risk*, June 2017. URL: https://longtermrisk.org/s-risks-talk-eag-boston-2017/.

[97] W Ross Ashby. Requisite variety and its implications for the control of complex systems. cybernetica 1 (2): 83-99. 1958.

[98] Warren E. Walker, Vincent A. W. J. Marchau, and Jan H. Kwakkel. Dynamic Adaptive Planning (DAP). In Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, editors, *Decision Making under Deep Uncertainty: From Theory to Practice*, pages 53–69. Springer International Publishing, Cham, 2019. `doi:10.1007/978-3-030-05252-2_3`.

[99] R. J. Lempert. Robust Decision Making (RDM). In Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, editors, *Decision Making under Deep Uncertainty: From Theory to Practice*, pages 23–51. Springer International Publishing, Cham, 2019. `doi:10.1007/978-3-030-05252-2_2`.

[100] Global Catastrophic Risks 2020. *Global Challenges Foundation*. URL: https://globalchallenges.org/wp-content/uploads/Global-Catastrophic-Risks-2020-Annual-Report.pdf.

[101] Shahar Avin, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. Ó hÉigeartaigh, William J. Sutherland, and Martin J. Rees. Classifying global catastrophic risks. *Futures*, 102:20–26, September 2018. URL: https://www.sciencedirect.com/science/article/pii/S0016328717301957, doi:10.1016/j.futures.2018.02.001.

[102] Elinor Ostrom. The Challenge of Common-Pool Resources. *Environment: Science and Policy for Sustainable Development*, 50(4):8–21, July 2008. URL: http://www.tandfonline.com/doi/abs/10.3200/ENVT.50.4.8-21, doi:10.3200/ENVT.50.4.8-21.

[103] Lukas Kuhn. Polycentricity for Governance of the Moon as a Commons - Open Lunar Foundation. 2021. URL: https://www.openlunar.org/library/polycentricity-for-governance-of-the-moon-as-a-commons.

[104] United Nations. *Our Common Agenda - Report of the Secretary-General*. United Nations, S.l., 2021. OCLC: 1273675504.

[105] Marie-Claire Cordonier Segger, Marcel Szabó, and Alexandra R. Harrington, editors. *Intergenerational Justice in Sustainable Development Treaty Implementation: Advancing Future Generations Rights through National Institutions*. Treaty Implementation for Sustainable Development. Cambridge University Press, Cambridge, 2021. URL: https://www.cambridge.org/core/books/intergenerational-justice-in-sustainable-development-treaty-implementation/8FEAC2DAA000B10B0F3C01F395483C0C, doi:10.1017/9781108768511.

[106] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, November 1990. Google-Books-ID: 4xg6oUobMz4C.

[107] Vicky Chuqiao Yang and Anders Sandberg. Collective Intelligence as Infrastructure for Reducing Broad Global Catastrophic Risks. page 13.

[108] Garrett Hardin. The Tragedy of the Commons. *Science*, 1968. URL: http://cgm.cs.mcgill.ca/~avis/courses/566/commons.html.

[109] Nick Bostrom, Thomas Douglas, and Anders Sandberg. The Unilateralist's Curse and the Case for a Principle of Conformity. *Social Epistemology*, 30(4):350–371, July 2016. URL: http://www.tandfonline.com/doi/full/10.1080/02691728.2015.1108373, doi:10.1080/02691728.2015.1108373.

[110] Charles S. Cockell. Freedom Engineering – Using Engineering to Mitigate Tyranny in Space. *Space Policy*, 49:101328, August 2019. URL: https://www.sciencedirect.com/science/article/pii/S0265964618301036, doi:10.1016/j.spacepol.2019.07.002.

[111] Exoplanet Biosignatures: Observational Prospects | Astrobiology. URL: https://www.liebertpub.com/doi/full/10.1089/ast.2017.1733.

[112] K. P. Hand, C. Sotin, A. Hayes, and A. Coustenis. On the Habitability and Future Exploration of Ocean Worlds. *Space Science Reviews*, 216:95, July 2020. ADS Bibcode: 2020SSRv..216...95H. URL: https://ui.adsabs.harvard.edu/abs/2020SSRv..216...95H, doi:10.1007/s11214-020-00713-7.

[113] Victoria S. Meadows, Giada N. Arney, Edward W. Schwieterman, Jacob Lustig-Yaeger, Andrew P. Lincowski, Tyler Robinson, Shawn D. Domagal-Goldman, Russell Deitrick, Rory K. Barnes, David P. Fleming, Rodrigo Luger, Peter E. Driscoll, Thomas R. Quinn, and David Crisp. The Habitability of Proxima Centauri b: Environmental States and Observational Discriminants. *Astrobiology*, 18(2):133–189, February 2018. Publisher: Mary Ann Liebert, Inc., publishers. URL: https://www.liebertpub.com/doi/full/10.1089/ast.2016.1589, doi:10.1089/ast.2016.1589.

[114] Martin Beech. *Terraforming: The Creating of Habitable Worlds*. Springer Science & Business Media, April 2009. Google-Books-ID: dwm72BO5zoMC.

[115] George Profitiliotis and Maria Loizidou. Planetary protection issues of private endeavours in research, exploration, and human access to space: An environmental economics approach to forward contamination. *Advances in Space Research*, 63(1):598–605, January 2019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0273117718307889, doi:10.1016/j.asr.2018.10.019.

[116] Ker Than. How do we protect planets from biological cross-contamination? | Stanford University School of Engineering. May 2020. URL: https://engineering.stanford.edu/magazine/article/how-do-we-protect-planets-biological-cross-contamination.

[117] Owen Cotton-Barratt, Max Daniel, and Anders Sandberg. Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter. *Global Policy*, 11(3):271–282, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1758-5899.12786. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1758-5899.12786, doi:10.1111/1758-5899.12786.

[118] Tobias Baumann. S-risks: An introduction. *Center for Reducing Suffering*, 2017. URL: https://centerforreducingsuffering.org/research/intro/.

[119] Cuebong Wong, Erfu Yang, Xiu-Tian Yan, and Dongbing Gu. An overview of robotics and autonomous systems for harsh environments. In *2017 23rd International Conference on Automation and Computing (ICAC)*, pages 1–6, September 2017. doi:10.23919/IConAC.2017.8082020.

[120] Holden Karnofsky. Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity. *Open Philanthropy*, May 2016. URL:

https://www.openphilanthropy.org/research/potential-risks-from-advanced-artificial-intelligence-the-philanthropic-opportunity/.

[121] Holden Karnofsky. AI Timelines: Where the Arguments, and the "Experts," Stand. *Cold Takes*, September 2021. URL: https://www.cold-takes.com/where-ai-forecasting-stands-today/.

[122] Allan Dafoe. AI Governance: Opportunity and Theory of Impact. September 2020. URL: https://forum.effectivealtruism.org/posts/42reWndoTEhFqu6T8/ai-governance-opportunity-and-theory-of-impact.

[123] Paul Christiano. What failure looks like. *AI Alignment Forum*, March 2019. URL: https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

[124] Charles Goodhart. Problems of monetary management : the U.K. experience. *Papers in monetary economics 1975 ; 1*, 1, 1975.

[125] C. Sagan and W. I. Newman. The Solipsist Approach to Extraterrestrial Intelligence. *Quarterly Journal of the Royal Astronomical Society*, 24:113, June 1983. ADS Bibcode: 1983QJRAS..24..113S. URL: https://ui.adsabs.harvard.edu/abs/1983QJRAS..24..113S.

[126] Nick Bostrom. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2):71–85, 2012. Publisher: Springer Verlag. doi:10.1007/s11023-012-9281-3.

[127] Center for Security and Emerging Technology, Matthew Daniels, and Ben Chang. National Power After AI. Technical report, Center for Security and Emerging Technology, July 2021. URL: https://cset.georgetown.edu/publication/national-power-after-ai/, doi:10.51593/20210016.

[128] Holden Karnofsky. Weak point in “most important century”: full automation. *Cold Takes*. URL: https://www.cold-takes.com/weak-point-in-most-important-century-full-automation/.

[129] Model Artificial Intelligence Governance Framework: Second Edition. *Personal Data Protection Commission*, January 2020. URL: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.ashx.

[130] K Eric Drexler. Reframing Superintelligence. *Future of Humanity Institute*, 2019. URL: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf.

[131] Andrew Critch. What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs). *LessWrong*. URL: https://www.lesswrong.com/posts/LpM3EAakwYdS6aRKf/what-multipolar-failure-looks-like-and-robust-agent-agnostic.

[132] Eliezer Yudkowsky. Complex Value Systems in Friendly AI. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, editors, *Artificial General Intelligence*, volume 6830, pages 388–393. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. Series Title: Lecture Notes in Computer Science. URL: http://link.springer.com/10.1007/978-3-642-22887-2_48, doi:10.1007/978-3-642-22887-2_48.

[133] Andrew Lohn and Micah Musser. AI and Compute. *Center for Security and Emerging Technology*. URL: https://cset.georgetown.edu/publication/ai-and-compute/.

[134] Saif Khan and Mann. AI Chips: What They Are and Why They Matter. Technical report, Center for Security and Emerging Technology, April 2020. URL: https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/, doi:10.51593/20190014.

[135] Weapon Materials Basics. *Union of Concerned Scientists*. URL: https://www.ucsusa.org/resources/weapon-materials-basics.

[136] AlphaFold. URL: https://www.deepmind.com/research/highlighted-research/alphafold.

[137] Holden Karnofsky. Forecasting Transformative AI, Part 1: What Kind of AI? URL: https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/.

[138] What is a Lagrange Point? *NASA Solar System Exploration*, March 2018. URL: https://solarsystem.nasa.gov/resources/754/what-is-a-lagrange-point.

[139] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, and Allan Dafoe. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. 2018. URL: https://docs.google.com/document/d/e/2PACX-1vQzbSybtXtYzORLqGhdRYXUqiFsaEOvftMSnhVgJ-jRh6plwkzzJXoQ-sKtej3HW_0pzWTFY7-1eoGf/pub.

[140] R. A. Freitas, Jr. A self-reproducing interstellar probe. *Journal of the British Interplanetary Society*, 33:251–264, January 1980. ADS Bibcode: 1980JBIS...33..251F. URL: https://ui.adsabs.harvard.edu/abs/1980JBIS...33..251F.

[141] Stuart Armstrong and Anders Sandberg. Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica*, 89:1–13, August 2013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0094576513001148, doi:10.1016/j.actaastro.2013.04.002.

[142] Joan Johnson-Freese and Brian Weeden. Application of Ostrom's Principles for Sustainable Governance of Common-Pool Resources to Near-Earth Orbit. *Global Policy*, 3(1):72–81, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1758-5899.2011.00109.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1758-5899.2011.00109.x, doi:10.1111/j.1758-5899.2011.00109.x.

[143] Ang Yang, H.A. Abbass, and R. Sarker. Characterizing warfare in red teaming. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2):268–285, April 2006. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). doi:10.1109/TSMCB.2005.855569.

[144] Steven W. Popper. Reflections: DMDU and Public Policy for Uncertain Times. In Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, editors, *Decision Making under Deep Uncertainty: From Theory to Practice*, pages 375–392. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-05252-2_16.

[145] Toby Ord. Proposal for a New 'Three Lines of Defence' Approach to UK Risk Management. *Future of Humanity Institute*. URL: https://www.fhi.ox.ac.uk/wp-content/uploads/three_lines_defence.pdf.

[146] Convention on International Liability for Damage Caused by Space Objects. *United Nations General Assembly*, 1972. URL: https://www.unoosa.org/pdf/gares/ARES_26_2777E.pdf.

[147] Fabio Tronchetti. The PCA Rules for dispute settlement in outer space: A significant step forward. *Space Policy*, 29(3):181–189, August 2013. URL: https://www.sciencedirect.com/science/article/pii/S026596461300057X, doi:10.1016/j.spacepol.2013.06.007.

[148] Brian C. Weeden and Tiffany Chow. Taking a common-pool resources approach to space sustainability: A framework and potential policies. *Space Policy*, 28(3):166–172, August 2012. URL: https://www.sciencedirect.com/science/article/pii/S0265964612000604, doi:10.1016/j.spacepol.2012.06.004.

[149] Charles Rosenberg and Vivasvat Dadwal. The 10 Year Anniversary of the PCA Outer Space Rules: A Failed Mission or The Next Generation? *Kluwer Arbitration Blog*, February 2021. URL: http://arbitrationblog.kluwerarbitration.com/2021/02/16/the-10-year-anniversary-of-the-pca-outer-space-rules-a-failed-mission-or-the-next-generation/.

[150] Dubai's Courts of Space launches international Working Group to explore space-related legal innovations. *DIFC Courts*. URL: https://www.difccourts.ae/media-centre/newsroom/dubais-courts-space-launches-international-working-group-explore-space-related-legal-innovations.

[151] Aslan Abashidze, Irina Chernykh, and Maria Mednikova. Satellite constellations: International legal and technical aspects. *Acta Astronautica*, 196:176–185, July 2022. ADS Bibcode: 2022AcAau.196..176A. URL: https://ui.adsabs.harvard.edu/abs/2022AcAau.196..176A, doi:10.1016/j.actaastro.2022.04.019.

[152] Meghan Bartels. Astronomers and SpaceX coming together to make Starlink megaconstellation less disruptive to science. *Space.com*, June 2020. URL: https://www.space.com/spacex-starlink-satellites-astronomers-visibility-response.html.

[153] Judy Lawrence, Marjolijn Haasnoot, Laura McKim, Dayasiri Atapattu, Graeme Campbell, and Adolf Stroombergen. Dynamic Adaptive Policy Pathways (DAPP): From Theory to Practice. In Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, editors, *Decision Making under Deep Uncertainty: From Theory to Practice*, pages 187–199. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-05252-2_9.

[154] Mark Hibbs. Why Does the IAEA Do What It Does? *Carnegie Endowment for International Peace*, June 2017. URL: https://carnegieendowment.org/2017/11/06/why-does-iaea-do-what-it-does-pub-74689.

[155] Nivedita Raju. Transparency and Confidence Building for the Moon - A Primer - Open Lunar Foundation. URL: https://www.openlunar.org/library/transparency-and-confidence-building-for-the-moon.

[156] David O. Brink and David Owen Brink. *Moral Realism and the Foundations of Ethics*. Cambridge University Press, February 1989. Google-Books-ID: viUm7tVhAnIC.

[157] R. N. Bracewell. Communications from Superior Galactic Communities. *Nature*, 186:670–671, May 1960. ADS Bibcode: 1960Natur.186..670B. URL: https://ui.adsabs.harvard.edu/abs/1960Natur.186..670B, doi:10.1038/186670a0.

[158] John Gertz. Oumuamua and Scout ET Probes. Technical report, April 2019. Publication Title: arXiv e-prints ADS Bibcode: 2019arXiv190404914G Type: article. URL: https://ui.adsabs.harvard.edu/abs/2019arXiv190404914G.

[159] John Gertz. Strategies for the Detection of ET Probes Within Our Own Solar System. Technical report, November 2020. Publication Title: arXiv e-prints ADS Bibcode: 2020arXiv201112446G Type: article. URL: https://ui.adsabs.harvard.edu/abs/2020arXiv201112446G.

[160] Philip Lubin. A Roadmap to Interstellar Flight. January 2022. arXiv:1604.01356 [astro-ph, physics:physics]. URL: http://arxiv.org/abs/1604.01356, doi:10.48550/arXiv.1604.01356.

[161] Breakthrough Initiatives. URL: https://breakthroughinitiatives.org/initiative/3.

[162] Michael Bohlander. Joining the "Galactic Club": What Price Admission? – A hypothetical case study of the impact of human rights on a future accession of humanity to interstellar civilisation networks. *Futures*, 132:102801, September 2021. URL: https://www.sciencedirect.com/science/article/pii/S0016328721001105, doi:10.1016/j.futures.2021.102801.

[163] Caspar Oesterheld. Multiverse-wide Cooperation via Correlated Decision Making. *Center on Long-Term Risk*, August 2017. URL: https://longtermrisk.org/multiverse-wide-cooperation-via-correlated-decision-making/.

[164] Stuart Armstrong. Acausal trade: conclusion: theory vs practice. *LessWrong*. URL: https://www.lesswrong.com/posts/5bd75cc58225bf0670375427/acausal-trade-conclusion-theory-vs-practice.

[165] Jonathan Carroll-Nellenback, Adam Frank, Jason Wright, and Caleb Scharf. The Fermi Paradox and the Aurora Effect: Exo-civilization Settlement, Expansion, and Steady States. *The Astronomical Journal*, 158:117, September 2019. ADS Bibcode: 2019AJ....158..117C. URL: https://ui.adsabs.harvard.edu/abs/2019AJ....158..117C, doi:10.3847/1538-3881/ab31a3.