Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence in Safety-Critical Aerospace Systems: Applications, Challenges, and Implications

Marcos Fernandez-Tous

marcos.fernandeztous@und.edu

University of North Dakota

Ian Werkheiser

The University of Texas Rio Grande Valley

Sreejith Vidhyadharan Nair

University of North Dakota

Michael Dodge

University of North Dakota

Sai Susmitha Guddanti

University of North Dakota

Research Article

Keywords: Artificial Intelligence, space technology, accountability, ethics, space governance

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11271910/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Artificial Intelligence in Safety-Critical Aerospace Systems: Applications, Challenges, and Implications

Dr. Marcos Fernandez-Tous. Assistant Professor in the Department of Space Studies at the University of North Dakota. Grand Forks, ND, U.S. Marcos.fernandeztous@UND.edu ORCID: 0000-0003-1898-4872
Dr. Michael Dodge. Associate Professor in the Department of Space Studies at the University of North Dakota. Grand Forks, ND, U.S. michael.s.dodge@UND.edu
Dr. Sreejith Vidhyadharan Nair. Research Assistant Professor in the Department of Aviation at the University of North Dakota. Grand Forks, ND, U.S. S.vidhyadharan@UND.edu
Dr. Ian Werkheiser. Professor in the Department of Philosophy at the University of Texas Rio Grande Valley. Edinburg, TX, U.S. ian.werkheiser@utrgv.edu ORCID: 0009-0002-9205-1608.
Sai Susmitha Guddanti. GRA in the Department of Space Studies at the University of North Dakota. Grand Forks, ND, U.S. saisusmitha.guddanti@und.edu ORCID: 0000-0003-4660-6430

Abstract
This paper investigates the role of artificial intelligence (AI) in safety-critical aerospace systems. It emphasizes its function not merely as an automatic aid but as an increasingly autonomous decision-making agent embedded within complex engineering environments. The investigation of AI applications is centered around Reinforcement Learning techniques and other approaches that are considered supporting this concept. An interdisciplinary framework is adopted, integrating technical analysis of AI applications with system-level considerations in design and operations. In order to provide a broad picture of the implications of the use of AI-based tools into aerospace systems, and as a novelty in the field, it includes issues of governance, legal responsibility, accountability, and ethics. The study is grounded in three application domains —hypersonic atmospheric reentry, autonomous satellite systems, and nuclear space propulsion— selected for their high-risk operational contexts and minimal tolerance for failure. Across these domains, the paper examines how AI-driven decisions directly influence system performance in all phases of a system's life, from design through deployment and operation. It highlights the need for robust frameworks to ensure reliability, transparency, and responsible deployment. A set of thematic constructs are developed which demonstrates how they manifest across the three application domains. The aim of this paper is to inform current and future system designs and policy frameworks, through an interdisciplinary approach grounded in current practices.

Keywords: Artificial Intelligence, space technology, accountability, ethics, space governance.

Statements and declarations
The present work has not received any financial support or funding from any organization.

1. Introduction

On July 23rd, Elon Musk, the CEO of SpaceX, was interviewed by *The Economist* [1]. In an interview largely focused on the implications of AI for politics and sociology, Musk interestingly mentioned several space-supported systems, such as Starlink, but only in the context of how they will make AI technology available to increasingly larger parts of the globe. In our opinion, an important missing piece lies precisely in the counterbalance: how AI will support space missions and the concomitant space infrastructure. The present article aims to address this gap.

A research project of this magnitude cannot be accomplished in a single article. For this reason, our approach aims to lay the foundations for further work in the area. We argue that the AI concept *avant la lettre*, can trace its origins back to the optimization of software loops for space missions in the 1960s. Since then, confidence in these systems, owing to the extraordinary results achieved, has led to AI being used in an increasing number of aspects of the space industry. In the present day, AI is used to solve technological problems that possess at least one of the following three characteristics: they are complex, multidimensional, and multi-objective. We discuss what we consider to be some representative examples.

Beyond a taxonomic description of AI applications, which we believe is necessary, this paper also analyzes the transformation of AI into a decision-making agent. This is where aspects of law, accountability, and ethics come to the forefront, although at different levels. Legal aspects are considered because they provide a general perspective,

one that reveals the lack of specific work on AI for space applications in the industry today. If, as we contend, there is a need to develop AI-specific legal material in the form of regulations or non-binding standards, then accountability is the aspect that will define the boundaries of the problem. Accountability comes with responsibility, which introduces the third element into the discussion: the ethical component. We believe that, if anything defines the beginning of the second quarter of the 21st century, it is the multidimensional research effort required to solve today's problems. It is to this endeavor that researchers such as ourselves are committed.

The rest of the paper is organized as follows. Section 2 establishes the limits of the discussion on AI technology adopted in this article, which is driven by a particular application: Reinforcement Learning. Section 3 describes three paradigmatic AI applications in space, namely hypersonic atmospheric entry, autonomous satellite systems, and nuclear space propulsion. Different AI applications within each of these cases are categorized for the first time. The legal implications of AI are discussed in Section 4, and accountability and ethical aspects are considered in Sections 5 and 6, respectively. Conclusions and future work are presented in the final section of the paper, Section 7.

2. AI in Safety-Critical Aerospace Systems: Themes and Paradigm

One of the earliest instances in which an autonomous system was incorporated into a space propulsion system occurred during the SMART1 mission (2003–2006). This technology demonstrator from the European Space Agency was designed to validate several advanced subsystems, including a Solar Electric Propulsion system that employed a Hall effect thruster manufactured by Snecma [2]. The propulsion system enabled orbit raising from geostationary transfer orbit to cislunar space. Two based optimization loops were implemented for propulsion control. The first operated under steady state conditions, with xenon propellant injected at a constant mass flow rate. The second loop functioned during throttling, adjusting the propellant flow in response to changes to the magnetic field strength and discharge voltage. In addition, an autonomous software module was applied to the spacecraft's Failure Detection, Isolation, and Recovery (FDIR) system, a capability that was limited to the management of single failure scenarios. While these early applications lacked the complexity associated with current AI applications, they proved highly effective. The successful implementation of the optimization and failure detection schemes contributed to extending the SMART1 mission by approximately one year. Since then, the role of AI autonomy in propulsion and in spacecraft operations has expanded significantly into other applications such as autonomous orbit station keeping and coordinated maneuver planning for satellite constellations [3].

This work focus is limited to the class of machine learning methods known as Reinforcement Learning (RL). Other AI approaches will be the objective of future research efforts. To establish a common ground of understanding, we consider it important to go through some definitions of RL elements that we will be considering in the paper. They are summarized as follows:

- **State definition**: State refers to a representation of the environment at a given time step that encapsulates all relevant information necessary for decision-making by the agent. It serves as the input to the policy, ideally satisfying the Markov property, whereby the current state sufficiently summarizes past information to predict future system evolution. Typically —but not always— the current state is defined through position, velocity (or momentum), and orientation (attitude), together with their change over time.
- **Policy-based action selection**: Given the current state, an RL algorithm infers the optimal next action according to a predefined policy. This policy encodes in a particular manner those performance objectives that will be optimized. The inference process combines accumulated experience from past interactions with exploratory actions introduced through stochastic sampling, thereby emulating a learning process in which the agent operates with partial knowledge of the environment and limited control over all influencing variables. An effective policy thus balances exploration efficiency, robustness, and solution accuracy.
- **State evaluation**: The resulting next state is evaluated using sensor inputs. Because of their nature, they are at the interface between the physical system (hardware) and the control or decision-making software. While inputs are considered to make inferences, the intrinsic sensor operation —how data is obtained from the surrounding environment— is not part of RL.
- **Reward assignment**: Based on the outcome of the state evaluation, a reward signal is assigned. These rewards act as feedback stimuli that guide the learning process, encouraging behaviors that move the system toward an optimal solution.

- **Iteration**: The process is repeated iteratively, allowing the policy to improve progressively as the agent learns from successive interactions with the environment: RL is a process that deploys in time.

It is important to disambiguate this approach to Artificial Intelligence from similar terms and concepts given the growth in popularity of labelling things "AI" in recent years and the concomitant confusion in what the term AI really means. AI in our sense is different from classical closed-loop control systems, which operate through predefined feedback relationships that continuously regulate system behavior in response to deviations from desired states using fixed control laws validated in advance through conventional engineering methods. Feedback control theory defines such systems in terms of deterministic feedback mechanisms that compare system outputs against desired reference states and adjust behavior accordingly [4]. AI systems are characterized not primarily by this feedback regulation but by computational capacities associated with perception, inference, learning, prediction, and adaptive decision-making. AI systems based on RL can be seen as intelligent agents that perceive environments and select actions to achieve specified goals, emphasizing adaptive inference rather than predetermined response structures. Autonomous system is also a term related —but not identical— to AI. Autonomous systems are defined by operational self-governance: they are systems capable of perceiving environmental conditions, selecting among possible actions, and executing tasks with limited or absent real-time human intervention. Autonomous systems are defined by their ability to achieve goals under uncertainty regardless of the specific mechanisms used internally, whether conventional control architectures, AI methods, or combinations of both [5]. In this paper, we discuss issues that arise in AI both when they are operating autonomously and when they are replacing or complementing human decisions.

Another concept closely related to artificial intelligence is Digital Twinning (DT)**.** As the term suggests, a digital twin consists of a virtual representation of a physical system that runs in parallel with its real-world counterpart, enabling the testing and evaluation of specific behaviors during experiments. This approach reduces the need for physical modifications when exploring alternative configurations and allows system limits to be investigated without incurring operational risk. Although DT does not imply AI —neither AI implies DT— these two concepts are often seen working together to increase their individual potential.

Conceptually, the roots of digital twinning can be traced back to the Apollo 13 accident, when an explosion in an oxygen tank of the Service Module required ground teams to reproduce, as accurately as possible, the conditions inside the spacecraft in order to devise recovery procedures. That improvised situation shared many similarities with modern DT applications, in which automated data exchange enables the continuous synchronization of virtual and physical models. In modern applications, however, this duplication is primarily motivated by safety and design simplification rather than by contingency needs: Contrary to the Apollo 13 case, today DT applications are not improvised. For instance, digital twin approaches can be found during testing of nuclear reactor configurations on the ground. This example is interesting because of its natural extension to space propulsion applications. Specifically, DTs offer a powerful framework for investigating dynamic phenomena such as time delays in reactor responses to reactivity insertion, which are difficult or unsafe to test experimentally.

A persistent challenge in DT development today is the increasing need of model fidelity. Lin et al. propose a proactive strategy in which key parameter values of the physical system are forecast within the virtual model, allowing discrepancies to be identified and mitigated in advance, thereby reducing the likelihood of unexpected or unreliable outcomes [6]. Sometimes, the existing literature refers to this approach as Simulation to Reality Domain Adaptation (SRDA); SRDA is achieved through a physics-informed model, or a set of often complex relations between physics parameters**.** In recent years, solving model complexities has been eased thanks to advances in computational power and data availability, allowing trustability levels unthinkable just a few years back. One of the leading U.S. centers for the application of digital twin technologies in nuclear reactor modeling is the Idaho National Laboratory, which has achieved several first-of-a-kind development in this area [7].

In regard to AI in-situ applications, the onboard power required to support computations must be carefully evaluated. This consideration is more critical in spaceborne AI applications than in their terrestrial counterparts, where access to energy is far less constrained. The associated power and memory demands arise from several sources: physics-based models used to inform machine-learning algorithms (a major area of progress in recent years, as indicated above), telemetry streams including event logs, real-time state estimation buffers interfacing with sensors (e.g., Kalman filters), lookup tables or deep neural networks depending on the selected ML architecture, or the dedicated ML

processing hardware itself. How much energy these applications require is an analysis on its own, and out of the scope of the present discussion.

To anchor the discussion in real world or potentially realistic examples, the following section examines applications of artificial intelligence across three representative space use cases —hypersonic re-entry, autonomous satellite systems, and nuclear Space propulsion. These applications are grouped into two primary project phases: design and development, and operations. Each phase comprises several subcategories. To avoid redundancy, those capabilities that span both phases are grouped under a third category, referred to as transversal applications. The coding following the label is used to simplify the ensuing discussion in the use cases —see below.

- Design and development phase: AI applications supporting system conception, modeling, and validation include:
    - **Digital Twin (DT) technologies (D1).**
    - **Replacement or augmentation of scarce experimental data through hybrid or surrogate models (D2)**: Besides DT applications, AI is also a useful tool to reproduce data that can be analyzed through models, particularly in cases where experiments are too expensive or complex to perform.
    - **Alternative design strategies, including dual-system architectures (D3)**: We refer to dual systems as spacecraft architectures employing two independent propulsion systems, each of which can be activated as needed depending on the required thrust level and impulse for a given maneuver.
    - **Sensor fusion in the design phase (D4)**: Sensor fusion refers to technologies used in space programs to acquire inputs describing the environment from multiple sensors. It also involves combining this information into a coherent representation by applying defined rules that assign different weights to the various data sources and filter out what is considered noise.

- Operational phase: During mission execution, AI enables enhanced autonomy and robustness through:
    - **Real-time decision making (O1)** Real-time decision making refers to the capability of a system or agent to process incoming data and execute decisions within tight time constraints such that the outcome remains valid and effective in a dynamically evolving environment. It emphasizes timely responsiveness, where computational latency is bounded to meet operational deadlines inherent to the system's context.
    - **Sensor fusion during operations, including hardware–software interaction (O2)**: Analogous to D4, with applications to the operational phases instead.
    - **Hazard avoidance (O3)**: It specifically refers to the maneuvers that a spacecraft performs to avoid any collisional hazard along its nominal trajectory.
    - **Station keeping (O4)**: Typically applied in environments where atmospheric effects are significant, station-keeping maneuvers restore a satellite to its nominal altitude when it loses orbital energy due to aerodynamic drag. More broadly, the term also applies to maintaining a spacecraft at a desired position, such as near Lagrange points. In these cases, atmospheric drag is absent, but other perturbations —such as solar radiation pressure and gravitational disturbances —may still displace the spacecraft from its nominal position, requiring corrective maneuvers.
    - **Constellation management and synchronization (O5)**: It refers to the coordinated management of multiple satellites operating in synchronized trajectories —for example, within the same orbital plane —where inadequate coordination could otherwise lead to close approaches or collisions between constellation elements.
    - **Alternative operational modes and dual-system concepts (O6)**: Analogous to D3, with applications to the operational phases instead.

- Transversal applications. These capabilities are applicable across both the design and operational phases:
    - **Autonomy and prediction (T1)**: it refers to a spacecraft's ability to determine and execute the most optimal maneuver autonomously. This capability is particularly important in situations where communications are degraded or unavailable, such as during blackout effects (e.g., in hypersonic flight) or when significant signal delays arise due to the spacecraft's distance from Earth.

- **Multi-dimensional approaches (T2)**: This term refers to a key characteristic of problems for which artificial intelligence has proven to be very effective. Such problems involve multiple interacting dimensions and variables, making them too complex to be solved efficiently using traditional methods within a reasonable timeframe.
- **Optimization (T3)**: Another key characteristic of the problems considered here is that they typically do not admit a single unique solution. Instead, identifying the "best" solution requires finding an optimal compromise among multiple, often conflicting, objectives or constraints.
- **Fault detection and health monitoring (T4)**: A longstanding capability in space missions, whose performance has been significantly enhanced with the introduction of artificial intelligence techniques.
- **Transition from tactical to strategic military missions (T5)**: In recent years, this transition has been driven by the need for improved resource optimization and faster responsiveness to rapidly evolving geopolitical scenarios. As a result, satellites increasingly need to be repositioned or reassigned to missions that were not anticipated during the original planning phase.
- **Servicing (T6)**: This term encompasses a range of services applied to spacecraft, from maintainability actions to performance enhancements. Examples of on-orbit servicing include refueling, repairing damaged components, and replacing or upgrading subsystems, such as the servicing missions performed on the Hubble Space Telescope.

3. Application Domains and Use Cases

This section applies the conceptual framework to three representative domains. Each use case is analyzed through the themes introduced in Section 2. This analysis ensures a consistent and comparative perspective across these domains.

3.1 Hypersonic Atmospheric Reentry

Although hypersonic aerodynamics problems arise in several areas of the space domain —such as scramjet combustors, high-speed nozzle expansions, and weapon systems like DARPA's HAWC —this section focuses on the specific case of Earth atmospheric reentry. This case was selected because its complexity encompasses multiple opportunities for AI applications that align with the categories introduced in Section 2. A notable example of hypersonic reentry occurred during the Apollo missions, where entry velocities reached approximately Mach 33 [8]. Moreover, the physical phenomena associated with Earth reentry are largely transferable to Mars entry scenarios, making the problem both relevant and timely. At such extreme velocities, strongly coupled phenomena are triggered around the capsule, such as chemistry kinetics and plasma fluid mechanics. This coupling naturally creates complex conditions that motivate the use of AI-based solvers.

D2 Augmentation of scarce experimental data through modeling: Hypersonic flight data are extremely limited due to both the relative novelty of many applications and the inherent difficulty of acquiring reliable measurements during flight. As a result, most research in this area relies heavily on models based on computational fluid dynamics (CFD) simulations. This modeling-driven approach makes hypersonic aerodynamics a particularly natural and well-established application domain for AI-based methods. One aspect to note, the universe of solutions possible and acceptable will be much influenced by the data obtained, which in turn depends on the model used. Thus, depending on this model, there is potential for an inadvertent bias in future decision-making processes.

T1 Autonomy and prediction:

- Many applications of AI to hypersonic reentry problems focus on the determination of optimal trajectories —that is, identifying the most efficient path from space to the ground. While this problem admits an immediate solution if atmospheric effects are neglected, the presence of the atmosphere introduces multiple, tightly coupled constraints. For example, the heat absorbed by the vehicle must remain below a prescribed maximum. Carbon-carbon composites, used in thermal protection systems (TPS), more commonly known as heat shields, begin to decompose at temperatures above approximately 3,000 °C. In addition, mechanical properties degrade significantly with increasing temperature. Thus, the aerodynamic loads imposed by the surrounding flow must be kept within structural limits. Finally, mission considerations may also require minimizing the duration of the communications blackout period.

- The optimization of hypersonic trajectories is often structured using deep neural networks (DNNs), which emulate multipolar decision-making processes in complex systems. Over the years, however, several inefficiencies have been identified in the exclusive use of DNNs —most notably their limited ability to avoid redundancies in input-output evaluations. These redundancies lead to unnecessary computational cost and increased solution time. To address these issues, researchers have increasingly turned to physics-informed networks (PINs) as a reinforcement mechanism. In such approaches, only the most critical variables —or those expected to exhibit the largest changes over a given time step— are explicitly resolved at each iteration. Criteria for selecting these variables have been proposed by Gong et al., who eliminate variables that are highly correlated with past information, since their values can be inferred from historical data [9]. This strategy significantly reduces computation time and accelerates convergence. Another key challenge consists of avoiding the erroneous inference of correlations, which can lead the optimizer toward local minima rather than global solutions. Alternative approaches address this issue by constraining the hyperspace of admissible solutions using physical boundaries —for example, upper temperature limits dictated by vehicle materials— thereby preventing the exploration of physically unrealistic solutions. Pate and Subbarao (2024) refer to these constrained spaces as zonotopes [10]. When constructed appropriately, zonotopes form mathematically closed systems: a trajectory starting within the admissible hypervolume cannot evolve toward solutions outside it, ensuring self-consistency. Additionally, the incorporation of digital twinning has been proposed by Zhang et al., wherein the physical system is evaluated through a high-fidelity digital replica while the AI component determines the relevant set of variables and their admissible values [11]. This division of labor is effective as long as latency between system components is sufficiently low, so that simulations do not lag behind the process of AI decision-making. Finally, and most importantly, for these frameworks to be auditable, the structure of the research construct must remain fully traceable in order to ensure transparency and accountability.
- With respect to communication and autonomy, blackout events are an undesirable phenomenon that occurs when radio signals carrying information from the vehicle interact with the plasma layer formed around it. The time interval during a blackout phenomenon depends to a high degree on the signal frequency [12]. This plasma sheath is generated by extremely high temperatures and associated chemical reactions that arise during hypersonic flights. Models to analyze the formation and behavior of this plasma layer have been developed in the past [13–15]. These models typically combine classical electromagnetic formulations with fluid mechanics, within the framework known as magnetofluid dynamics. We have not identified works that explicitly incorporate AI algorithms to model or predict plasma behavior, although such applications are likely to emerge in the near future. Instead, current mitigation approaches focus on adaptive frequency management and dynamic link selection, which can partially exploit intermittent communication windows. These windows are important to keep meaningful human control from the ground, at the very least in a monitoring role. Added to this, onboard sensors may fail to provide information beyond the plasma sheath. The absence of a communication link between the vehicle and the exterior of the capsule shifts the burden of state estimation toward onboard inertial navigation systems for position and attitude determination. The execution of AI algorithms onboard —a paradigm known as Edge AI— may help reduce the uncertainties introduced in the system while these conditions last.

T3 Optimization aspects: Aerothermal prediction is critical for controlling vehicle temperatures and heat transfer rates during hypersonic flight. Thermal Protection Systems (TPS) are typically designed as conical structures with blunted tips, which act as shields for the vehicle. The selection of TPS materials is governed by trade-offs among weight, cost, and reliability. Common materials reported in the literature include ablative phenolics, ceramics, and carbon–carbon composites. Their properties must be carefully balanced against thermal protection performance, since overestimating TPS thickness may result in unnecessary mass penalties. This naturally leads to another optimization problem, where AI-based methods can play a significant role. One active thermal control strategy involves transpiration cooling, in which a coolant permeates through a porous TPS surface. In this case, the optimization problem focuses on minimizing coolant usage —since it represents inert payload mass— through careful design of pore geometry and internal flow channels. This approach was demonstrated by Wu et al. in the design of a nose cone [16]. Accurately resolving the Navier–Stokes equations for viscous, compressible, chemically reacting flows with high Reynolds and Prandtl numbers —and rapidly varying conditions around the vehicle— remains computationally challenging. As a result, recent approaches increasingly adopt physics-informed neural networks (PINNs), as discussed earlier, to augment or replace traditional solvers. A recent example of such an approach is provided by He et al. [17]. Hybrid methods incorporating PINNs currently represent one of the most promising research directions in aerothermal prediction and TPS design.

T4 Fault detection and health monitoring: Due to the multiplicity of system states that may arise when sensors or actuators fail —either by providing incorrect measurements or by implementing erroneous trajectory corrections— such elements are particularly well suited for reinforcement with AI-based tools. Zhou et al. propose the use of deep neural networks (DNNs) to enable the simultaneous and accurate identification of fault type, location, and time of occurrence in a hypersonic vehicle [18]. When labeled data are scarce and insufficient to adequately train DNN-based fault detectors, some researchers compare known fault signatures with model-generated data using convolutional neural networks [19]. In the context of multisensor failures, fuzzy logic approaches have also been employed to trace and identify potential faults while estimating overall system health, and at least some of these methods have demonstrated long-term stability and robustness [20]. In summary, AI-based techniques constitute an effective barrier against the continued use of faulty sensors, thereby enhancing system reliability and operational safety.

O1 Real-time decision making: Because of time-lag constraints in communication and control, autonomous decision-making is essential during the operation of hypersonic vehicles. These vehicles are highly sensitive to perturbations —primarily, though not exclusively, atmospheric disturbances— that can evolve over extremely short time scales. One viable approach to ensuring safe operation is to shift the learning and optimization process to the offline design phase. Reinforcement learning (RL) is particularly well suited for this purpose, as it enables the training of deep neural networks (DNNs) to identify optimal solutions through policy optimization methods, such as gradient-based approaches, using feasible trajectories that have been extensively pre-tested. Once trained, these policies can be deployed during the operational phase to support real-time decision-making. A key consideration in this context is that traceability need not be applied to individual control decisions or outcomes, but rather to the structure and logic of the AI framework itself. This aspect has important accountability consequences that are analyzed in section # of the present paper. The application of gradient-based policies within a reinforcement learning framework to determine optimal hypersonic trajectories was demonstrated by Bao et al. [21]. Variations and extensions of this approach have since been explored by Su et al. and Xu et al., addressing different mission profiles and optimization criteria [22, 23]. More recently, Sparrow Search Algorithms (SSA) have been introduced to mitigate the risk of convergence to local minima when seeking globally optimal solutions, further improving the robustness of autonomous trajectory optimization methods.

O2 Sensor fusion and state estimation: The application of AI to the operational phase typically requires the fusion of data from multiple sensor sources. These sensors provide measurements that indicate how closely the vehicle is following the optimal trajectory defined during the design phase. Deep neural network (DNN)–enhanced reinforcement learning (RL) algorithms have been employed to simplify both the policy and value functions, as in Das et al. [24]. However, real-time trajectory monitoring introduces the challenge of filtering sensor noise to achieve optimal performance. Addressing this challenge often involves combining RL approaches with Kalman filtering techniques, or derivatives thereof, such as the Extended Kalman Filter (EKF), which are better suited for nonlinear systems. An example of this application during hypersonic phases of flight is when vehicles are subject to significant atmospheric pressure and high thermal loads. This condition leads to nonlinear vibrational and torsional dynamics in the hypersonic structure. Zhang et al. propose the application of EKF-based processing of accelerometer data to characterize and estimate these effects [25]. Sensitivities in state estimation arising from uncertainty in initial conditions can be further mitigated through Monte Carlo-based methods. An example of this approach is provided by Singh et al., who applied such methods to ground-based radar tracking of ballistic targets [26]. From an auditability and systems assurance perspective, it is important to distinguish between three key components. Each of them may involve different responsible parties: the sensor hardware and data interfaces, the sensor-processing logic (e.g., KF, EKF, or UKF implementations), and the data fusion and decision-making layer. Treating these elements as separate but interacting subsystems enhances traceability, accountability, and system robustness.

3.2 Autonomous Satellite Systems

The number of satellite-based scientific and commercial missions has evolved in parallel with technological advances. This trend is particularly evident in the development of propulsion systems. Early space missions relied primarily on chemical propulsion. However, the adoption of simpler and less expensive propellants, such as cold gases (e.g., nitrogen), enabled more efficient handling and improved precision in orbital maneuvers. Further advancements came with the introduction of electric propulsion systems, which make far more efficient use of propellant —typically noble

gases such as xenon, and more recently alternatives including iodine, krypton, and argon. These technologies enabled lighter spacecraft designs, ultimately contributing to reduced launch costs.

NASA's Space Electric Rocket Test (SERT) program demonstrated the feasibility of Kaufman ion thruster technology in 1964, followed by SERT II in 1970. These early systems employed mercury as a propellant, while later ion propulsion systems transitioned to noble gases such as xenon [27, 28]. The impact of these early demonstrations extended into the 1990s, when NASA's Deep Space 1 mission used the xenon-fueled NSTAR ion propulsion system and subsequently encountered asteroid 9969 Braille and comet 19P/Borrelly [29]. Electric propulsion technologies have since been incorporated into missions such as DART, which carried the NEXT-C ion thruster as a technology demonstration [30], and BepiColombo, an ESA/JAXA mission whose xenon-fueled solar electric propulsion system supported its eight-year cruise to Mercury [31]. BepiColombo is scheduled to enter Mercury orbit in November 2026 [32].

Together, these developments have significantly expanded the range of applications and mission profiles achievable by spacecraft.

D3 Alternative design strategies, including dual-system architectures: Chemical and electric propulsion systems serve different mission needs, primarily —though not exclusively— based on the thrust and impulse levels required. This distinction suggests the potential benefits of dual-propulsion architectures for complex missions, in which a spacecraft can switch between chemical and electric propulsion depending on operational circumstances. An early example of such an approach is the Boeing 702 satellite platform developed in the late 1990s. This platform employed chemical propulsion during initial phases such as orbit raising and early station-keeping maneuvers, before transitioning to electric propulsion to reduce propellant consumption and operational costs. This strategy was feasible because station-keeping maneuvers —particularly in low Earth orbit— can be performed efficiently with lower-thrust systems. In contrast, when these platforms operated in geostationary orbit, they relied more heavily on chemical propulsion due to higher thrust requirements [33]. Dual-propulsion systems are also of interest for military applications, where station keeping and tactical re-tasking demand rapid changes in thrust and maneuver profiles, often under time-critical constraints. In such scenarios, optimal analyses can be used to select between chemical and electric propulsion modes, as well as to determine the optimal timing for switching between them based on mission objectives and fuel expenditure. The integration of artificial intelligence can further enhance this decision-making process by predicting maneuver requirements and optimizing propulsion mode selection in real time. Examples of the application of dual-propulsion concepts to collision-avoidance maneuvers can be found in the literature [34].

O3 Hazard avoidance: In 2007, China conducted an anti-satellite test in which the weather satellite Fengyun-1C, launched in 1999 into a near-circular polar orbit from the Taiyuan Space Center, was deliberately destroyed using a kinetic kill vehicle. In the months that followed, more than 2,200 debris fragments were identified and catalogued, of which only about 13% decayed and burned up in the atmosphere, according to contemporary estimates [35]. Two years later, the accidental collision between Iridium 33 and Cosmos 2251 in 2009 generated nearly 2,000 additional debris fragments in low Earth orbit [36]. While these are not the only collisions that have occurred, they remain among the most significant contributors to the space debris environment. Avoiding orbital debris is therefore critical for mission survivability. Although the debris population is mostly concentrated in low Earth orbit, this is not exclusively a LEO problem. Debris density decreases with altitude as we approach geostationary orbits, but the residence time of debris increases significantly at higher altitudes, where atmospheric drag is negligible and objects may persist for decades or longer. Collision avoidance typically requires fast, high-impulse maneuvers, which are most effectively performed using chemical propulsion. These maneuvers cannot be fully planned in advance, as they depend strongly on the specific geometry and timing of a potential conjunction event. Artificial intelligence applications in this domain focus on detection and avoidance systems for satellites in space, with methods centered on trajectory optimization for minimum fuel consumption [37, 38].

O4 Station keeping: The decay of a satellite's nominal orbit around the Earth results from multiple perturbing effects. In geostationary orbit (GEO), where the atmosphere is essentially negligible, orbital evolution is primarily driven by gravitational perturbations from the Sun and the Moon, as well as by the Earth's oblateness, which produces an anisotropic gravitational field commonly referred to as the J_2 effect. These perturbations necessitate specific corrective maneuvers, notably North-South and East-West station-keeping. North-South maneuvers are generally more complex and typically require two to three times more propellant than East-West maneuvers, which mainly involve longitudinal

repositioning along the orbit. Artificial intelligence techniques can assist in anticipating orbital deviations and planning station-keeping maneuvers more gradually and efficiently, thereby reducing propellant expenditure. In low Earth orbit (LEO), station keeping is dominated by atmospheric drag and is particularly well suited to autonomous control strategies. For example, Tafanidis et al. propose a RL-based approach in which the analysis of station-keeping alternative maneuvers is encouraged by maximizing policy entropy, and the control policy is represented by a flexible neural network, enabling adaptive and efficient station-keeping behavior [39].

O5 Constellation management and synchronization: Some satellite-based commercial applications require not individual spacecraft, but constellations or swarms of satellites operating in carefully designed phasing geometries. This is particularly true for wide-area communications coverage, such as in the case of Iridium and, more recently, Starlink constellation, as well as for certain remote-sensing missions where precise orbital synchrony is essential. Over time, however, orbital perturbations lead to a gradual decay of this synchronized motion, altering the relative distances and positions of the satellites within the formation. Restoring the desired configuration typically requires short bursts of energy to reposition individual satellites within the swarm. Neural-network-based approaches have been explored to address this problem (Morgan Choi & Lee, 2025), enabling adaptive control of relative motion. In addition, optimal formation & control strategies based on reinforcement learning have been reported by Yu et al. [40]. In that work, historical flight data combined with reinforcement learning were incorporated within an Adaptive Dynamic Programming (ADP) framework, allowing the controller to enlarge the universe of optimal actions.

T5 Transition from tactical to strategic military missions: Programs such as the Geosynchronous Space Situational Awareness Program (GSSAP) in geostationary orbit and the Tactically Responsive Space (TacRS) program operating in MEO and LEO represent significant initiatives within the military domain [41]. In this context, an additional factor makes these missions particularly attractive candidates for in-orbit refueling. Military space programs tend to prioritize reliability over cutting-edge technological advancement, as mission failure carries far more severe consequences than in most commercial applications. As a result, software obsolescence —which often limits the operational lifetime of commercial satellites before propellant depletion— tends to be less critical in military systems. Today, multi-objective missions are increasingly leveraging reinforcement learning techniques. In the near future, reinforcement learning is expected to play an important role in space applications involving cooperative multi-agent strategies. Related approaches are already being explored for coordinated surface deployments, both on land and at sea [42].

T6 Servicing: One of the earliest and most prominent examples of in-space servicing was the Hubble Space Telescope mission, designed as a modular platform that could be accessed and manipulated during extravehicular activities for repair, maintenance, and component replacement, including batteries. Today, renewed interest in in-space servicing is driven by multiple converging trends: the reusability of launch vehicles —such as SpaceX's Falcon 9, Blue Origin's New Shepard, and China's Long March family— following the precedent established by the Space Shuttle program, as well as the emergence of new mission concepts targeting the Moon and Mars. These developments have brought new servicing requirements to the forefront, including on-orbit repair, refurbishment, and refueling. Such operations could greatly benefit from the incorporation of artificial intelligence, enabling semi-autonomous rendezvous, approach, and docking maneuvers, as well as intelligent assessment of refueling needs and execution of refueling procedures. Companies such as Orbit Fab have proposed hardware solutions, including standardized refueling valves for use in space; however, no industry-wide standard has yet been universally adopted. This lack of uniformity risks increasing the market of solutions in excess, each of them adapted to specific manufacturers [43]. However, while standards typically provide assurance of market compatibility and support incremental innovation, premature standardization can also hinder radical innovation by locking in solutions that may later prove suboptimal [44]. Beyond refueling, additional services such as health monitoring, self-diagnosis, and decision-support systems represent further areas where artificial intelligence could play a transformative role in future in-space servicing architectures.

3.3 Nuclear Space Propulsion

3.3.1 System Overview

Nuclear propulsion is a high-efficiency method of spacecraft propulsion that uses energy released from nuclear reactions to generate thrust. It can be implemented through two architectures: Nuclear Thermal Propulsion (NTP) and Nuclear Electric Propulsion (NEP). In NTP, thermal energy produced by a fission reactor is directly transferred to a propellant, typically but not only hydrogen, which is then expanded through a nozzle to produce thrust. In contrast,

NEP converts reactor heat into electrical power through a liquid metal such as eutectic NaK or Li, which is used to drive electric propulsion systems. Important to note, no nuclear reactor systems for propulsion have been tested in space yet.

Both architectures rely on the precise control of a nuclear reactor, where system behavior is governed by tightly coupled relationships between neutron flux, reactivity, and core temperature [45]. Small perturbations of these variables can lead to rapid changes in reactor power, making stability and control central to safe operation. Reactivity is commonly regulated through rotating drums, which adjust the balance between neutron reflection and absorption within the core [46]. These characteristics define nuclear propulsion as a highly sensitive, tightly coupled, and safety-critical system, where accurate state estimation and timely control actions are essential. This operational context motivates the integration of artificial intelligence methods to enhance modeling, monitoring, and decision-making capabilities across both design and operational phases.

D2 + D1 Data Scarcity, Surrogate Models, and Digital Twins: The development of nuclear space propulsion systems is constrained by the limited availability of experimental data, as testing is often expensive, complex, and restricted by safety considerations [47]. Historically, only a limited number of full-scale programs, most notably the NERVA tests, have generated usable data, much of which is several decades old [48]. In addition to safety constraints, the cost and logistical complexity of building and operating physical test systems further limit experimental campaigns. This domain is particularly well suited for AI-driven approaches because much of the current understanding relies on computational models and simulations.

Artificial intelligence techniques are increasingly used to augment or replace traditional high-fidelity simulations through surrogate and hybrid models (D2). For instance, Sahadath et al. developed neural-operator–based surrogate models, including DeepONet and Fourier Neural Operators, to learn mappings from neutron source distributions to angular flux solutions [49]. These models achieve substantial computational speedups compared to conventional transport solvers while maintaining acceptable accuracy, enabling repeated evaluations and near real-time analysis for reactor systems. Similarly, Raissi et al. show how physics-informed neural networks (PINNs) embed governing physical laws directly into the learning process, allowing neural networks to solve nonlinear partial differential equations with limited data while preserving physical consistency; this can be translated to nuclear propulsion as well [50]. These approaches demonstrate how AI can approximate complex multiphysics behavior— such as neutron transport and coupled thermal phenomena— while significantly reducing computational cost and improving data efficiency.

The outputs of these surrogate and physics-informed models can further support the development of digital twin frameworks (D1), in which a virtual representation of the reactor system operates in parallel with the physical system. As discussed in Guddanti & Fernandez-Tous, digital twins can be integrated with advanced control strategies, including reinforcement learning-based controllers, to simulate and evaluate reactor behavior under varying operational conditions before deployment [51]. In this context, AI-enhanced models provide continuous state estimation and predictive capability, enabling the digital twin to act as both a validation and decision-support tool. Given the scarcity of real experimental data, such model-driven digital twin frameworks form a critical foundation for system design, testing, and validation.

T3 + T2 Optimization and Design Complexity: The design of nuclear propulsion systems inherently involves multi-dimensional and optimization problems, and competing constraints must be balanced across performance, safety, and mission requirements. Key design variables —such as fuel density, reactor geometry, and propellant conditions— interact in highly nonlinear ways, making traditional optimization approaches computationally intensive and often limiting the design space.

Artificial intelligence methods provide a powerful framework for addressing these challenges (T3). For instance, Schwarcz et al. employ a hybrid approach in which traditional Monte Carlo simulations are enhanced with machine learning to mitigate convergence to local minima in ground reactors [52]. This issue arises when the reactor is discretized into unit cells defined by parameters such as fuel density and moderator spacing; for space applications, the coolant role is replaced by the propellant's. Building on the earlier discussion, section 3.3 – D2+D1, of physics-informed neural networks (PINNs), Raissi et al. provide a framework for reducing the computational burden associated

with neutron transport, flux estimation, and related reactor physics calculations [50]. These methods embed governing physical laws directly into the learning process.

While conventional non-AI models continue to yield accurate results, they often lack the computational efficiency, particularly in terms of time-to-solution, required for practical or time-constrained applications [53]. Although they remain valuable for high-fidelity analysis and CFD augmentation, such approaches may not perform effectively in real-time or time-sensitive operational settings. In contrast, AI-driven methods emphasize computational efficiency and adaptability, particularly during the design phase, where greater traceability can be achieved and system behavior can be extensively recorded avoiding the risks associated with operational failures.

These optimization challenges are further characterized by their multi-dimensional nature (T2). This is particularly evident in mission-level applications, such as comparing crewed versus cargo trajectories, evaluating continuous-thrust profiles for NEP systems, or integrating nuclear propulsion within broader mission architectures. Maddock and Parsonage identify key drivers for crewed interplanetary mission design as a multi-objective problem, providing a foundation for future AI-based formulations [54]. To date, however, the full application of AI to these mission-level optimization problems remains limited.

T4 Fault Detection and Health Monitoring: Fault detection and health monitoring constitute a critical component of nuclear propulsion systems, given the sensitivity of reactor dynamics and the potential consequences of system failure.

Artificial intelligence techniques have significantly enhanced traditional fault detection approaches (T4). Similar to the hypersonic case discussed earlier, these methods can be broadly categorized into knowledge-driven and data-driven frameworks. Qi et al. provides a recent classification of these approaches in nuclear systems [55]. Knowledge-driven methods rely on model-based representations, often incorporating surrogate models where AI is embedded at the core of the decision-making process. In contrast, data-driven methods focus on sensor-level information, including both hardware reliability and software-based processing, and may employ techniques such as Kalman filtering, physics-informed models (e.g., PINNs), and machine learning approaches.

Within the data-driven category, a range of techniques have been explored, including neural networks, support vector machines, and unsupervised clustering methods [20, 55]. These approaches enable the identification of fault patterns from sensor data, improving the ability to detect anomalies under uncertain and noisy conditions. These capabilities are particularly important in scenarios involving sensor or actuator failures, where incorrect measurements or control actions may propagate through the system and lead to unstable reactor behavior. AI-based methods support the classification of fault type, location, and timing, thereby improving diagnostic accuracy and response time.

In addition, recent studies have explored AI-driven modeling of internal reactor phenomena, such as turbulent eddy viscosity in coolant flows [56]. Turbulence plays a critical role in enhancing heat transfer by increasing mixing and energy transport within the flow. While developed for terrestrial nuclear reactors, such approaches can be extended to space propulsion contexts, for example by improving the modeling of heat transfer to the propellant.

T1 + O1 Autonomous Control and Real-Time Decision Making: The operation of nuclear propulsion systems requires precise regulation of reactor dynamics, as small deviations in reactivity or temperature can quickly lead to unsafe conditions. This makes autonomous control and real-time decision-making essential, particularly when communication delays or mission constraints limit human intervention. Artificial intelligence enables the transition from traditional control strategies to more adaptive and responsive frameworks (T1, O1). Several of the methods discussed in earlier sections can be extended to operational contexts. For instance, Zhang et al. provides a comprehensive review of AI/ML techniques applied to nuclear reactors, including hybrid approaches that combine physics-informed neural networks (PINNs) with deep learning [57]. These methods allow physical governing equations to be solved in a manner applicable to both design and real-time operation, supporting more flexible and adaptive control strategies.

In parallel, Zohuri outlines a triad of AI applications in nuclear reactor systems: signal processing, fuzzy logic, and dynamic control [58]. In this framework, signal processing and fuzzy logic act as supporting layers that enable effective control decisions. Fuzzy logic, in particular, is well suited for nuclear systems, where highly complex and

nonlinear relationships are difficult to capture analytically. These relationships can instead be represented through pattern-based reasoning. In this approach, system behavior is inferred from observed input-output patterns rather than derived from explicit analytical models. Similarly, Fu et al. demonstrate the application of model predictive control (MPC) in space nuclear reactors, where a physics-based predictive model is used to optimize control actions under dynamic conditions, achieving improved tracking performance and disturbance minimization compared to traditional control methods [59].

Emerging approaches involving reinforcement learning (RL) presents a particularly promising framework for controlling a nuclear reactor. RL-based methods enable an agent to learn optimal control policies through interaction with a simulated environment, allowing it to adapt to system dynamics and disturbances without requiring explicit analytical models [60]. Though this work centered on ground reactors, it can be applied to space reactor seamlessly. RL approaches are particularly relevant in tightly coupled systems such as nuclear reactors, where traditional controllers may struggle to account for nonlinearities and uncertainties in real time. At the same time, the integration of AI into control loops introduces additional considerations, including system-level security risks, particularly in cloud-based implementations [61]. These aspects will be further discussed in the legal and governance sections of this paper.

O2 Sensor Fusion and State Estimation: Effective operation of nuclear propulsion systems relies on accurate estimation of internal reactor states from sensor data, which is often affected by noise, latency, and partial observability, making reliable state estimation a key challenge in safety-critical environments. Artificial intelligence methods support sensor fusion and state estimation (O2) by combining heterogeneous data sources into a coherent representation of system behavior. AI-enhanced frameworks can integrate traditional estimation techniques, such as Kalman filtering and its nonlinear extensions, with learning-based components that improve robustness to noise and model uncertainties [62]. This integration enables more reliable tracking of system states and enhances the quality of information available for control decisions.

In nuclear propulsion, sensor fusion is closely coupled with autonomous control. Reinforcement learning (RL)-based controllers, for instance, rely on accurate state representations to determine optimal actions, making the quality of fused sensor data critical to overall system performance [60]. Errors in state estimation can propagate through the control loop, potentially degrading performance or unsafe operating conditions [63].

From a systems perspective, it is important to distinguish between sensor hardware, signal processing algorithms, and the higher-level decision-making layer. Treating these components as interacting but separable subsystems improves traceability and supports clearer attribution of responsibility, particularly in safety-critical applications [64, 65].

4. Governance, Legal implications

The use of AI systems for space technologies will generate novel challenges related to legal management. The deployment of space technologies often carries with it considerations of both national and international legal systems, and AI-empowered technologies will be no exception to this norm. The challenge for extant space law will be how to categorize these systems, and whether responsibility for their use falls readily into traditional sovereign State analysis.

At the time of this writing, AI technologies should be seen, arguendo, as advanced tools utilized for a variety of technical reasons and efficiencies. It is under this understanding of the technology that this section of the essay examines the legal status of AI-empowered space systems. AI-empowered systems have not obtained the lofty status associated with sentient beings, as often seen in speculative fiction. Should they ever do so, unique issues of international and domestic law will concomitantly emerge, requiring deeper analysis as to how they may be ethically and legally deployed.

International space law posits the generally accepted parameters under which any space technologies may be used. The fundamental underpinning of international space law is the Outer Space Treaty (OST) of 1967, which created the structure for acceptable space activities by States-Party to the agreement [66]. Any AI technologies deployed into outer space, or operated on celestial objects outside of the Earth, will be subjected to the Outer Space Treaty regime. This treaty regime consists of five treaties generated by the United Nations, ranging from the Outer Space Treaty in

1967, to the Moon Agreement in the late 1970s [67]. Each of these treaties serves to give structure to myriad current and planned activities in space, with an eye towards permissibility rather than restriction. While there are some forbidden activities in space, such as disallowing the placement of nuclear weapons in orbit around the Earth or other celestial bodies, taken collectively the articles of these treaties stress cooperation, peaceful activity, and consideration of the needs of other States in space [66]. As such, AI-operated space objects generally fall under Article I of the OST, which permits the exploration and use of outer space by all States.

The extent to which the States operating space-based AI technologies are bound to the space treaties depends upon their ratification status; i.e., whether or not the State has become willingly bound to the treaty by ratifying it under the norms established for doing so, which typically are dictated by the legal instrument itself. If a State has ratified the Outer Space Treaty, or any of its progeny, it has thereby informed the world that it is willing to be bound by its terms. Unless a norm found within a treaty has risen to the level of a customary international law —a status tricky to determine, as it requires analysis of State practice (generally over long periods of time) and evidence of opinio juris — a State will not be bound to the particular rules stated in any given treaty [68, 69].

States operating extensively in space at the time of this writing have ratified, at the very least, the Outer Space Treaty [70]. States launching AI-equipped objects into space will therefore be bound to a system of international law that has governed space activity since nearly the first objects orbited the Earth. Given the age of the treaty system, and the fact that it was created in a time far in advance of AI technologies, a point of curiosity concerns how AI ought to be considered within the scaffolding of the OST regime.

Taking autonomous satellite systems as a case in point, a question arises as to what would happen if the AI onboard chose an action to reposition the object, and in the process of doing so it collides with another State's space object. The situation is similar to self-driving cars, which have already started service in certain parts of China and the U.S. In the former country, no strong legal constraints have apparently prevented a significant deployment in Shenzen by certain companies such as Baidu or Pony.ai. Nevertheless, the early state of the technology development makes the need of a regulatory package necessary. This situation can easily be extrapolated to the space satellite industry. Without appropriate regulations, collision events would be not only unfortunate but also legally simple to resolve. It is understood law that when one State causes damage to another State, it will generally be liable for such damages [71–73]. Historically, the actions one State took towards another that have caused damage are the result of human beings making consequential decisions. In the case of autonomous systems in space, determining the decisionmaker, and therefore the responsible party, is a murkier proposition. However, the OST regime is flexible enough to cover problematic situations involving AI-enabled space systems in the near to mid-term.

This is because the State Party that launched or controlled the object is bound to Article VI of the Outer Space Treaty, which mandates that States will be internationally responsible for their national activities in outer space, whether or not they are carried out by the State itself, or by its non-governmental entities. As currently constructed, AI-equipped objects in space will be thought of as mere tools by the law, and the State which launches or controls the object is the legally responsible body for any of its actions whilst in space. For the sake of speculation, even if AI were to rise to the level of sentience or true independence, the object's State of registry would still be responsible for the actions it conducts in space, just as the State would be responsible for the actions of an astronaut or cosmonaut making choices about where to go, or what to do.

Referring to the above case in point of autonomous satellite systems, should the choices made by its AI lead to damage to another State's space objects or personnel, then not only will the State be held responsible by Article VI of the OST, but will also be held liable under Article VII of the same treaty. That article requires States that launch or procure the launching of a space object that cause damage to another State to be internationally liable for such damages [66]. If the State in question has also ratified the Liability Convention, then it will be held liable for the damages caused in space by its AI-controlled object if it is deemed to be at fault for the accident. The Liability Convention Article III would therefore require some investigation into whether or not there is fault involved, which itself could be a complex and lengthy process, necessitating a deeper understanding of the nature of AI and its ability to make decisions independently of its maker [74]. On the other hand, a simpler scenario involves damage to the Earth caused by an AI-controlled object. For instance, if an atmospheric re-entry vehicle is returning to the Earth, and ends up colliding with an aircraft in flight, or causing damage to the surface of the Earth due to the AI's decisions, then the Liability Convention Article II would hold the launching State absolutely liable for any such damage, no matter how cautious or reasonable the State had been in designing, deploying, and utilizing the AI.

Because the novelty of using AI technologies in space could place States in jeopardy of liability in ways not predicted at the start of the Space Age, it would behoove the international community to consider identifying and adopting norms and behaviors that could prevent future damage scenarios. Such efforts should specify what is meant by AI-controlled space systems, identify clear lines of responsibility for their usage by States or non-governmental entities supervised by States, and detail how such technologies should intersect with and be guided by the extant liability provisions of the Outer Space Treaty and the Liability Convention. There is precedent for this methodology in the field of international space law. For instance, many of the key rules and behavioral norms codified in the Outer Space Treaty were first floated to the international community in the Declaration of Legal Principles Governing the Activities of States in the Exploration and Use of Outer Space, a document propounded in 1963 by the United Nations General Assembly XVIII [75]. Such 'declarations' or 'principles' documents are a form of soft law that allow States to experiment with norms to be solidified by later treaty documents, should sufficient international consensus lead to the adoption of such agreements. A Declaration on the Peaceful Uses of AI-Technology in Outer Space would stimulate international discourse on the ethical, responsible, and legal uses of an emergent field in technology that is certain to impact space activities for decades to come. If the discourse surrounding this area of soft law is fruitful, States might even be willing to move towards a treaty that helps codify international obligations surrounding AI usage.

Domestic laws must also be considered. Most States have not yet opted to write extensive laws specific to the realm of space activities, but some have legislated expansively. As with international law, integrating AI into the extant domestic laws will take time, but for now a space activity, such as a launch, must conform to the laws of the place in which it takes place. One major space actor with extensive legislation for such activities is the United States. If, for instance, the USA opted to launch an AI-equipped nuclear propulsion unit into space, it would need to comply not only with the international legal regime, but also with domestic laws such as the Commercial Space Launch Act as Amended, which requires a license to be procured through the Federal Aviation Administration (FAA) under 51 U.S.C. § 50905 [76, 77]. One condition for a license is that the Secretary of Transportation may prevent a launch if the technology is deemed unsafe under 51 U.S.C. § 50908(2)(c)(2), which will eventually require an analysis of how AI operates onboard spacecraft launched by the United States, in order to ensure the U.S. is not unreasonably risking the safety of others on Earth, or in the space environment [78].

Beyond the law, international and national policies will also guide and constrain the use of AI space technologies. Policies rely on State willingness to self-control activity outside of the strictures of international law. One benefit to this is that sovereignty, and the creative State activities that can extend therefrom, are maximized. A downside is that policies, including any developed to guide AI usage in space, remain effective only so long as the States utilizing them choose to do so.

A relevant example is compliance with the COSPAR Planetary Protection Guidelines as of 2026, which aim to protect celestial bodies of interest to the study of the formation of life from exploratory space activities [79]. If an AI-empowered object makes a decision that puts it too close to the celestial body under study, and as a consequence collides with that body, the ramifications for the search for life are unknown, and the consequences to the State ultimately responsible for the object remain to be hashed out in the international discourse. Similarly, AI-equipped atmospheric re-entry vehicles exploring certain celestial bodies will need to carefully land on their object of study, so as not to disturb biological and chemical surveys.

Ultimately, current legal instruments at both the international and, in some instances, national levels cover basic permissibility of use of and liability standards for AI-equipped spacecraft. As the use of these technologies expands and diversifies, and as the technology becomes more sophisticated, States will need to consider whether extant legal systems will be sufficient to govern their use, or if additional national and international efforts at clarifying legal obligations are warranted.

5. Accountability implications

In safety-critical aerospace systems, accountability must be distinguished from both responsibility and liability. Liability refers to the backward-looking legal obligation to remedy damage once harm has occurred and legal standards have been applied. Responsibility, by contrast, is forward-looking: refers to the prior assignment of duty namely the party, office, contractor, operator, or certifying authority that was expected to ensure a function was designed, integrated, operated, or supervised correctly [80–82]. Accountability connect these two terms [83]. It is the evidentiary and structural infrastructure that makes it possible to determine whether a responsibility was discharged and, if necessary, whether liability can later be adjudicated. For that reason, this section treats accountability not as a general ethical aspiration, but as a practical engineering condition for reconstructing what an AI-enabled aerospace system was qualified to do, what it actually did, and whether its behavior remained within validated operating envelope.

Classical aerospace systems have a traceable accountability chain. A component has a specification; a control law has a documented logic; a certification artifact records the assumptions, limits, tests, and failure modes; and telemetry can often be interpreted against that baseline [84]. AI-enabled systems disrupt that chain at several points. Training data may be incomplete or only partially representative of rare operational regimes. Neural architectures and initialization procedures can influence behavior without producing a conventional design rationale. Optimization can depend on stochastic exploration, reward shaping, and convergence criteria rather than deterministic derivation. In autonomous systems higher autonomy adds layers of perception, planning, adaptation, and decision-making that are not always visible in a traditional component-to-control-law audit trail [5]. The accountability problem therefore precedes the liability problem. Before one can ask who bears legal responsibility, the system must preserve enough evidence to show what was designed, what was validated, what was delegated to the AI, and what was recoverable after the event.

5.1 Why Space Operations Reframe the Accountability Problem

Space operations reframe AI accountability because the operational environment weakens many of the assumptions that commonly support terrestrial AI governance. In space-based and high-speed aerospace operations:

- communication with ground systems may be delayed, intermittent, degraded, or temporarily unavailable;
- onboard systems may need to make safety-critical decisions without immediate human supervision;
- sensor information may be incomplete, uncertain, or affected by harsh environmental conditions;
- opportunities for intervention may be limited once a decision has been executed;
- computational capacity, energy, communication bandwidth, and other system resources may be constrained;
- actions may have irreversible or long-lasting physical and operational consequences; and
- post-incident investigation may depend primarily on the evidence preserved onboard the system.

Accountability in such environments therefore depends on the ability to reconstruct the information available to the AI, the decisions it made, the constraints under which it operated, and whether its behavior remained within the validated operational envelope. Post-event explanations alone are insufficient when the system has not preserved adequate evidence of its internal state, decision process, and operational context. These operational constraints create three central accountability pressures:

- Limited opportunity for human intervention: Communication delays and blackout periods may require autonomous systems to make and execute decisions before ground operators can review or approve them.
- Potentially irreversible consequences: A single incorrect or poorly timed decision may permanently alter the condition, trajectory, safety, or viability of a mission.
- Constraints on evidence collection: The amount and type of evidence that can be recorded are limited by onboard memory, power consumption, mass, radiation exposure, processing capacity, and communication bandwidth.

For this reason, simply increasing logging or collecting more data cannot be treated as a universal solution. The process of generating and preserving evidence must itself be designed as a constrained subsystem. Accountability in space-based AI therefore requires advance decisions about which information should be recorded, how it should be protected and compressed, which evidence should receive priority, and how it can later be interpreted when continuous supervision and unlimited storage are unavailable.

5.2 Structural Traceability in Safety-Critical AI

In safety-critical aerospace AI, traceability should focus primarily on the structure of the AI framework rather than on producing a complete explanation for every individual runtime decision. Telemetry and event logs remain important, but outcome-level explanations alone are often insufficient because they may be:

- limited to a single decision or moment;
- sensitive to small changes in system state or input data;
- difficult to interpret consistently;
- computationally expensive to generate; and
- demanding in terms of onboard storage and communication resources.

These limitations become particularly important when AI systems must make rapid decisions using complex neural-network or reinforcement-learning models. In such settings, generating a detailed explanation for every action may compete with the processing, memory, and power resources required for safety-critical operation.

Traceability should therefore be established through the documented structure of the AI system, including its architecture, training process, data sources, operational constraints, validation procedures, decision boundaries, and monitoring mechanisms [85, 86]. Runtime evidence should support this structural record by showing the system state, relevant inputs, selected actions, confidence or uncertainty measures, and any departures from the validated operating envelope. The investigative burden should not depend on reconstructing an intuitive rationale for every action, but on determining whether the system operated according to its approved design and assurance framework.

The epistemic challenge is equally important. Explanations of individual decisions produced by stochastic or high-dimensional AI systems may be unstable [87]. Small changes in sensor data, initial conditions, environmental disturbances, or model parameters can produce different local explanations even when the system is operating under the same underlying policy.

For accountability purposes, the more reliable question is whether the deployed AI system remained within the structural and operational envelope that had been documented, tested, and approved. Establishing this requires evidence concerning the provenance, quality, and limitations of the training data; the model or policy architecture; the objectives, reward functions, and safety constraints; the training and optimization procedures; the assumptions applied during development; the scope of simulation and testing; the validated operating conditions; and the criteria used to authorize deployment.

Mathematical methods that represent reachable states, bounded uncertainty, and safety constraints provide useful examples of this structural approach. They incorporate safety expectations directly into the system's design and validation framework rather than relying only on an after-the-fact explanation of a single decision.

Post-incident investigation should therefore extend beyond asking why the AI selected a particular action at a particular moment. It should also determine:

- whether the system remained within its validated operating envelope;
- whether the deployed configuration matched the approved configuration;
- whether sensors and supporting systems operated correctly;
- whether environmental conditions exceeded modeled assumptions; and
- whether the system was used within its authorized operational limits.

When the system remains within its validated envelope, attention may shift toward the adequacy of the design assumptions, validation coverage, and acceptance criteria. When it operates outside that envelope, attention may shift toward configuration control, sensor integrity, unmodeled conditions, system degradation, or unauthorized use.

This approach does not reduce the importance of telemetry. Runtime evidence remains essential for reconstructing system states, mode transitions, inputs, commands, and environmental conditions. However, that evidence must be interpreted against the documented structure of the AI system. In aerospace AI, the system framework provides the accountability baseline, while the individual outcome is the event evaluated against that baseline.

5.3 Subsystem-Level Accountability

Accountability in autonomous aerospace systems should be decomposed across the subsystems involved in producing a decision. The sensing and data-interface layer generates the measurements on which the system depends. The sensor-processing layer transforms those measurements into estimates of the vehicle, mission, or environmental state using filtering and state-estimation methods. The data-fusion and decision-making layer then combines the estimated state with mission objectives, operational constraints, and learned or rule-based logic to select an action. This layered structure reflects the practical reality that different organizations may contribute to different parts of the system, including hardware manufacturers, software suppliers, AI developers, systems integrators, mission operators, and certification or safety authorities.

Subsystem decomposition supports accountability in three important ways. First, it enables the source of a failure to be identified more precisely. An unsafe action may result from a degraded sensor, an interface error, an inaccurate state estimate, an inappropriate fusion process, or a decision policy operating outside the conditions for which it was validated. Without this decomposition, these distinct possibilities may be reduced to the overly broad conclusion that "the AI failed."

Second, decomposition clarifies how responsibility is distributed across organizations and technical functions. A hardware-interface defect, a state-estimation mismatch, and a flaw in the design or training of a decision policy represent different accountability concerns and may involve different responsible parties.

Third, decomposition supports modular auditability. Independent verification and validation can assess whether each subsystem satisfies its own evidence and performance requirements and whether the interfaces between subsystems preserve sufficient information for later reconstruction. This is particularly important when software patches, model updates, or configuration changes affect one subsystem without replacing or revalidating the entire system. Accountability must therefore address not only the performance of individual layers, but also the integrity of the information transferred between them.

This accountability structure can be applied across different autonomous aerospace scenarios. In constellation management, measurements obtained through onboard navigation systems, inter-satellite ranging, crosslinks, and ground-based tracking may support relative-state estimation. The resulting estimates can then inform formation control, station-keeping, conjunction assessment, and collision-avoidance decisions, depending on the mission and available information. In nuclear thermal propulsion, reactor instrumentation, including measurements related to neutron flux, temperature, pressure, and actuator position may support reconstruction of the reactor and propulsion state, which in turn informs control-drum, power-management, and safety-related actions. During atmospheric reentry, inertial and navigation measurements, potentially supplemented by pressure, thermal, and other vehicle-state information, support state estimation and guidance under aerodynamic, structural, and thermal constraints.

Although the specific sensors, estimation methods, and control actions vary from one scenario to another, the basic accountability process remains the same. To make an AI-enabled aerospace system auditable, it must be possible to trace how raw observations were collected and processed, how they were converted into an estimate of the system state, and how that estimate led to a particular decision or control action. The evidence should also make clear which person, organization, requirement, or technical artifact governed each step in that process.

This traceability requirement guards against two common accountability failures. The first is the "problem of many hands," in which so many organizations and subsystems contribute to a decision that responsibility becomes diffused and no single party can meaningfully be held to account [88–90]. The second is the "accountability sink," in which a human is nominally placed in the decision loop but has little real ability to affect the outcome, serving mainly as the designated bearer of blame [91]. Subsystem decomposition addresses the first by keeping each contribution attributable to a specific layer and party, while the authority-boundary requirements discussed in Section 5.5 address the second by ensuring that assigned responsibility corresponds to actual decision authority.

5.4 Inference Methods and the Limits of Outcome Explanation

The choice of AI inference method has direct consequences for the kind of accountability evidence that can be produced. That choice should therefore be treated as a design-phase accountability requirement, not deferred until certification or post-incident analysis. Deep neural networks and reinforcement-learning policies can be powerful in hypersonic trajectory generation and attitude control because they can approximate complex mappings across high-dimensional, nonlinear flight regimes [21, 92]. Yet post-hoc attribution methods such as saliency, feature importance, and counterfactual probing should be treated as supplementary evidence rather than primary explanation [87]. They may help investigators explore sensitivity, but they rarely provide the kind of stable causal account expected in a safety-critical control investigation.

5.4.1 Edge AI and the Onboard Accountability Burden

Edge AI – the execution of AI algorithms onboard – increases the accountability burden because important evidence must be created, stored, and protected onboard rather than reconstructed later from ground-based archives. Unlike ground-controlled systems, onboard autonomous systems have limited memory, power, processing capacity, and communication bandwidth. Accountability records must therefore compete with navigation data, mission information, fault-management records, and other safety-critical functions. These records may also be exposed to radiation, thermal variation, and hardware faults that can damage or alter stored information.

For this reason, data integrity is not simply a technical issue. A corrupted log, missing model version, undocumented compression step, or unrecorded change in autonomy mode can make it difficult to reconstruct what the system did and why. When onboard records provide the main connection between the validated system design and the event being investigated, they must be treated as part of the safety and assurance case rather than as routine administrative data.

Three design implications follow. First, onboard accountability records should be treated as safety-relevant system data and protected through appropriate redundancy, error correction, access control, and fault detection. The level of protection should reflect the importance of those records to system assurance and post-event investigation. Second, decisions about which evidence should be preserved must be made during system design. An autonomous system should not be allowed to change or discard important records without predefined and documented rules. Third, the logging system should be able to monitor its own health. If the system detects that its ability to generate or preserve evidence has been degraded, that condition should be recorded and may trigger predefined limits on autonomous operation, additional safeguards, or a transition to a safer operating mode

5.5 Authority Boundaries and Human-AI Teaming

Accountability also depends on where decision authority sits along the autonomy spectrum. In supervised systems, the human operator retains formal authority, so the record should show what the AI recommended, what uncertainty information was available, whether the recommendation fell within the validated operating envelope, and how much time the operator had to respond. This is important because automation bias can make an AI recommendation operationally decisive even when the system is officially described as human supervised [93]. In mixed-initiative systems, authority is shared, and the boundary between human and machine decision-making must be clearly defined. The operational concept should state which actions may be taken autonomously, which require human approval, what happens when approval cannot be obtained in time, and how authority changes during communication loss or urgent conditions [85]. Any shift in effective decision authority should be recorded rather than hidden behind a general label such as *human supervised* [94].

In fully autonomous systems, accountability cannot depend on reviewing a human decision made at the time of the event. Instead, it rests on whether the system was properly qualified for its autonomous role, whether it remained within its approved operating envelope, and whether reliable evidence of its behavior was preserved. Autonomy mode should therefore be treated as a controlled system configuration. A vehicle may move between supervised, mixed-initiative, and fully autonomous operation during different mission phases, and each transition changes how accountability is distributed. These transitions should be logged with the time, triggering condition, basis of authority, active policy or software version, and the validated envelope that applied at the time.

6 Ethical Implications

The ethical implications of deploying AI in safety-critical aerospace systems are vital to any full discussion of these technologies. The high consequence, limited reversibility, and constrained opportunities for human intervention create conditions in which ethical issues like trust, risk acceptance, and human roles must be reinterpreted. These implications are complex and can be quite domain specific, as issues emerge from specific operational configurations as well as more abstract concerns. As a result, properties discussed in AI ethics such as trustworthiness, accountability, and safety must be embedded within system design and lifecycle practices rather than imposed externally [95, 96]. In this section, we will look at some of the ethical concerns that arise in the use cases in section 3 as an example of thinking through ethics in emerging technologies.

6.1 Human–Machine Interaction: From Operator to Collaborator

The integration of AI across both design and operational phases transforms the role of human agents. In design contexts (D1–D4), AI functions as an advanced analytical tool, with engineers retaining interpretive responsibility and accountability —see Section 5. In operational contexts, particularly those involving real-time decision-making (O1), autonomy under communication constraints (T1), and distributed coordination (O5), AI systems increasingly function as collaborators or proxies. This transition reflects the shift from human-in-the-loop to human-on-the-loop or human-out-of-the-loop systems [97]. As autonomy increases, several ethical challenges arise, such as reduced situational awareness, particularly in latency-constrained environments such as hypersonic reentry or deep-space operations; skill degradation as operators rely more heavily on automated systems; and responsibility gaps, especially in multi-agent systems where outcomes emerge from distributed interactions.

The concept of "meaningful human control" provides a framework for addressing these challenges. As articulated by Filippo Santoni de Sio and Jeroen van den Hoven, meaningful human control requires that systems be designed such that human agents retain the ability to understand, guide, and, where feasible, intervene in system behavior [94]. In the use cases under discussion, however, continuous real-time control is often infeasible due to communication latency or blackout conditions. As a result, meaningful human control must be interpreted *ex ante*, through the specification of system constraints, objectives, and validation procedures during design and testing. Given the diminished level of meaningful human control during the use of the systems, trust in the system in these critical moments becomes particularly important.

6.2 Trust, Transparency, and Participation

Across the use cases analyzed in Section 3, trust is not a binary attribute but a calibrated relationship of "appropriate trust" between human operators and AI system performance [98]. Trust is necessary both for the functioning of the system and the decision to use it in the first place. When trust is justified, it is also a stand-in for acceptance of the risks in the system or process. Risk acceptance is domain-specific. Hypersonic systems require acceptance of autonomous decision-making under communication constraints; satellite systems distribute risk across interacting agents; nuclear propulsion systems impose extremely low tolerance for failure. Acceptable risk of AI technology is context-dependent and must be evaluated relative to domain-specific consequences [99].

That said, in all the use cases we are considering, risk acceptance and trust must be accomplished early. Once systems transition to operational autonomy (O1–O6), particularly in time-critical environments, the ethical structure of trust shifts toward reliance under uncertainty. In hypersonic reentry, communication blackout conditions require onboard AI (T1, O1) to perform trajectory adjustments and state estimation without real-time human oversight. Similarly, in satellite constellation management (O5), distributed AI agents coordinate maneuvers in a multi-agent environment where local decisions produce system-level effects. In nuclear propulsion, AI-assisted control of reactor reactivity introduces tightly coupled, high-consequence decision-making contexts. In such cases, trust must be grounded not in full interpretability of decisions but in assurance of system behavior within bounded risk envelopes.

It is during design phases (D1–D4) that this trust can be slowly established. For those developing the system this trust can be grounded in verification, validation, and demonstrated predictive reliability of the models and simulations used to represent relevant physical processes and operational constraints. For stakeholders not working directly on the design process, trust can be grounded in transparency from designers, though "transparency" for different stakeholders requires different kinds of disclosures and explanations depending on their interests, levels of technical knowledge,

and other considerations [100, 101]. Trust is further reinforced through meaningful participation by stakeholders in relevant aspects of the design process when possible. The increase of trust is in addition to other reasons to want meaningful participation by stakeholders, such as improved outcomes from knowledge sharing, improved success of projects due to buy in by those who participated in its design, and participative justice concerns of those most affected by a decision having a voice [102–104].

6.3 Ethical Structure of Optimization and Decision-Making

The centrality of "optimization" across the applications described in Section 3 highlights a key insight from AI ethics: algorithmic decision-making may appear objective, but it encodes normative assumptions. Objective functions, constraints, and reward structures formalize trade-offs among competing values. In hypersonic trajectory design (Section 3.1) for example, optimization balances thermal load, structural limits, and blackout duration. In satellite operations (Section 3.2), fuel efficiency competes with responsiveness and collision avoidance margins. In nuclear propulsion (Section 3.3), efficiency must be subordinated to strict safety constraints governing reactivity and thermal stability. The AI ethics literature has emphasized that such trade-offs are not value-neutral [105, 106]. In reinforcement learning systems, reward functions act as formalizations of desired behavior but may fail to capture rare or extreme conditions that are critical in safety-critical environments [107]. Likewise, the values of resilience and redundancy are in tension with efficiency, as are the values of understandability of a system versus size and complexity. When values are in tension, optimizing for one comes at the expense of the other. Which of these values are prioritized over the other are not technical questions, but rather the domain of ethics: tradeoffs of values that must be understood explicitly and resolved collaboratively.

Engineering strategies discussed earlier in the paper can be interpreted as mechanisms for embedding ethical constraints directly into system design. For example, physics-informed neural networks and constrained solution spaces (Section 3.1) restrict optimization to physically admissible regions not only for the sake of efficiency but because those are the solutions people are interested in given their goals for the project. Hybrid modeling approaches in nuclear reactor design (Section 3.3) limit exploration within established safety boundaries, and what level of risk or safety people ought to accept is not an objective, technical issue. Sensor fusion and fault detection systems (T4) provide redundancy and mitigate the impact of erroneous decisions. These approaches align with emerging work in AI safety that emphasizes designing for robustness and constraint satisfaction rather than relying solely on post hoc evaluation of outcomes [108, 109]. In this sense, ethical considerations are operationalized through the structure of decision-making processes rather than treated as external criteria.

6.4 Grounding Ethical Considerations in Engineering Realities

A central conclusion of both this analysis and the broader AI ethics literature is that ethical principles must be translated into domain-specific engineering practices. High-level principles such as transparency, accountability, and safety require operationalization within concrete system architectures [95, 105]. This is illustrated in our use cases: communication blackouts in hypersonic reentry necessitates onboard autonomy, shaping feasible forms of human oversight. The combinatorial complexity of satellite constellation management requires distributed AI coordination, redefining responsibility as a system-level property. The tightly coupled dynamics of nuclear propulsion impose strict temporal and safety constraints, requiring bounded autonomy and conservative design.

These constraints define the ethical landscape. They determine what forms of autonomy are necessary, what levels of oversight are possible, and what risks must be accepted. As argued in systems safety engineering, safety and ethics in complex systems are emergent properties of socio-technical interactions rather than attributes of individual components [110]. Accordingly, the ethical implications of AI in safety-critical aerospace systems are best understood as properties of system design. Trust, accountability, and acceptable risk emerge from the interaction of algorithms, physical systems, human operators, and operational environments. The analysis presented in this paper demonstrates that ethical considerations are not ancillary to engineering practice but are integral to the design and deployment of AI-enabled aerospace systems.

7. Conclusion

The integration of AI into safety-critical aerospace systems introduces challenges that extend beyond technical performance to questions of governance, legal responsibility, accountability, and ethics. AI is taking a greater role in system-level decision-making, which will have an increasing impact on hypersonic re-entry, autonomous satellite

systems, and nuclear space propulsion Under the current space law framework, responsibility and liability continue to rest with the relevant states. On the other hand, increasing autonomy raises questions about whether these existing frameworks are sufficient for future AI-enabled space operations. This assessment therefore highlights the need to consider technical, legal, accountability, and ethical requirements together when designing and operating AI-enabled aerospace systems.

For such complex systems, accountability needs to be distinguished from legal responsibility and liability. Accountability should consider whether an AI-enabled system remained within a documented, tested, and approved operational envelope. Furthermore, accountability may also need to be distributed across subsystems while maintaining sufficient traceability to determine where a fault or decision originated. Hence the importance of clearly identifying where decision authority lies along the autonomy spectrum. These questions should be addressed during the design phase rather than during final certification or after an accident. Trust, accountability, safety, risk acceptance, and the trade-offs associated with autonomous decision-making are part of the system design itself. These elements also shape the ethical landscape by determining what forms of autonomy are necessary, what levels of oversight are possible, and what risks must be accepted.

Future work will focus on translating broad ethical principles into domain-specific engineering practices for safety-critical aerospace systems. This includes proposing new legal frameworks that are sufficient for AI-enabled space operations, developing structured approaches to distributed accountability and responsibility gaps in complex multi-agent systems, and defining decision authority more precisely across different levels of autonomy. Further work is also needed to determine how sufficient accountability records can be maintained alongside navigation, mission management, fault detection, and other safety-critical functions when onboard computing and storage are limited. Risk acceptance and system-specific ethical trade-offs will also need to be addressed explicitly and collaboratively for individual applications.

A central conclusion of this analysis is that ethical principles must ultimately be translated into domain-specific engineering practices. For AI to be used responsibly in safety-critical aerospace systems, accountability and ethics must be built into how these systems are designed, tested, and operated rather than treated as requirements added after deployment. The establishment of an ethical framework as a boundary to the rest of the elements considered in this report is thus a pending task as well.

References:

1. Austin (2026) Elon Musk's vision of the future. The Economist

2. Milligan D, Gestal D, Camino O, et al (2004) SMART1 Electric Propulsion Operations. In: 40th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit. American Institute of Aeronautics and Astronautics, Fort Lauderdale, Florida

3. Rutter D (2025) NASA Starling and SpaceX Starlink improve traffic coordination. In: NASA Explore. https://www.nasa.gov/centers-and-facilities/ames/nasa-starling-and-spacex-starlink-improve-space-traffic-coordination/

4. Forsythe W, Danisman H, Eissa MA (1985) A study of the closed-loop structure of multivariable feedback control systems. Trans Inst Meas Control 7:251–261. https://doi.org/10.1177/014233128500700505

5. Antsaklis P (2020) Autonomy and metrics of autonomy. Annu Rev Control 49:15–26. https://doi.org/10.1016/j.arcontrol.2020.05.001

6. Lin W-Q, Miao X-R, Chen J, et al (2025) Simulation-to-reality transferability framework for operating-parameter forecasting in nuclear reactors using domain adaptation: Simulation-to-reality. Nucl Sci Tech 36:. https://doi.org/10.1007/s41365-025-01643-1

7. Darrington J, Vainqueur BE, Ritter C (2024) The First Real-Time Digital Twin of a Nuclear Reactor. In: Grieves M, Hua EY (eds) Digital Twins, Simulation, and the Metaverse: Driving Efficiency and Effectiveness in the Physical World through Simulation in the Virtual Worlds. Springer Nature Switzerland, Cham, pp 225–247

8. Niculescu ML, Fadgyas MC, Cojocaru MG, et al (2016) Computational Hypersonic Aerodynamics with Emphasis on Earth Reentry Capsules. INCAS Bull 8:. https://doi.org/10.13111/2066-8201.2016.8.3.5

9. Gong Z, Yu J, Liu X, et al (2025) Deep Neural Networks Based Hypersonic Vehicle Reentry Impact Point Prediction With Fixed-Length Lightweight Training Data. IEEE Trans Aerosp Electron Syst 61:17029–17040. https://doi.org/10.1109/TAES.2025.3601096

10. Patel J, Subbarao K (2024) Nonlinear Reachable Set Computation and Model Predictive Control for Safe Hypersonic Re-Entry of Atmospheric Vehicles

11. Zhang D, Li S, Xu J, Hu X (2024) Internet of Things Based Digital Twin Model Construction and Online Fault-Tolerant Control of Hypersonic Flight Vehicle. Drones 8:460. https://doi.org/10.3390/drones8090460

12. Alilou S, Shahrassai L, Sobhanian S (2024) Optimizing Spacecraft Re-entry Communication: Plasma Electron Depletion Through Controlled Particulate Injection and Fresnel Coefficients Analysis. Braz J Phys 54:. https://doi.org/10.1007/s13538-024-01568-x

13. Kundrapu M, Loverich J, Beckwith K, et al (2015) Modeling Radio Communication Blackout and Blackout Mitigation in Hypersonic Vehicles. J Spacecr Rockets 52:853–862. https://doi.org/10.2514/1.A33122

14. Mao M, Peng K, Zhao Z, et al (2024) Modeling and Analysis on Dynamic Terahertz Channel Capacity in Hypersonic Plasma Sheaths. IEEE Trans Plasma Sci 52:240–247. https://doi.org/10.1109/TPS.2024.3369104

15. Zhao L, Wei H, Li J, et al (2021) Performance Analysis of Typical Telemetry Modulation Methods Under Downlink Integrated Channel of Hypersonic Vehicles. IEEE Trans Plasma Sci 49:3524–3531. https://doi.org/10.1109/TPS.2021.3119847

16. Wu N, Wang J, He F, et al (2018) Optimization transpiration cooling of nose cone with non-uniform permeability. Int J Heat Mass Transf 127:882–891. https://doi.org/10.1016/j.ijheatmasstransfer.2018.07.134

17. He Y, Zhu L, Guo Y, et al (2025) A physics-informed neural networks approach for coupled flow and heat transfer problems. Int Commun Heat Mass Transf 165:109085. https://doi.org/10.1016/j.icheatmasstransfer.2025.109085

18. Zhou D, Huang D, Tie M, et al (2025) A fault diagnosis method based on interpretable machine learning model and decision visualization for HVs. Appl Intell Dordr Neth 55:270. https://doi.org/10.1007/s10489-024-06219-x

19. Cao F, Zhu H, Du C, Lu P (2024) Deep learning based actuator fault identification for hypersonic vehicles: A zero-shot case. J Eng Sci 46:1613–1622. https://doi.org/10.13374/j.issn2095-9389.2023.09.27.001

20. Hu G, Zhou T, Liu Q (2021) Data-Driven Machine Learning for Fault Detection and Diagnosis in Nuclear Power Plants: A Review. Front Energy Res 9:. https://doi.org/10.3389/fenrg.2021.663296

21. Bao CY, Zhou X, Wang P, et al (2023) A Deep Reinforcement Learning-Based Approach to Onboard Trajectory Generation for Hypersonic Vehicles. Aeronaut J 127:1638–1658. https://doi.org/10.1017/aer.2023.4

22. Su L, Wang J, Chen H (2023) A Real-Time and Optimal Hypersonic Entry Guidance Method Using Inverse Reinforcement Learning. Aerospace 10:948. https://doi.org/10.3390/aerospace10110948

23. Xu H, Cai G, Zhang S, et al (2022) Hypersonic reentry trajectory optimization by using improved sparrow search algorithm and control parametrization method. Adv Space Res 69:2512–2524. https://doi.org/10.1016/j.asr.2021.12.030

24. Das PP, Pei W, Niu C (2023) Reentry trajectory design of a hypersonic vehicle based on reinforcement learning. IOP Publishing, p 012005

25. Zhang D, Zheng X, Xie Y, Hu X (2022) Angular-Accelerometer-Based Flexible-State Estimation and Tracking Controller Design for Hypersonic Flight Vehicle. Aerospace 9:206. https://doi.org/10.3390/aerospace9040206

26. Singh NK, Bhaumik S, Bhattacharya S (2012) Tracking of ballistic target on re-entry using ensemble Kalman filter. In: 2012 Annual IEEE India Conference (INDICON). pp 508–513

27. Arrighi S R (2024) NASA's SERT II: 'A Genuine Space Success Story'. https://www.nasa.gov/centers-and-facilities/glenn/nasas-sert-ii-a-genuine-space-success-story/. Accessed 28 Aug 2026

28. Dawson VP (1991) Engines and Innovation: Lewis Laboratory and American Propulsion Technology. NASA

29. (2017) Deep Space 1 - NASA Science. https://science.nasa.gov/mission/deep-space-1/. Accessed 28 Aug 2026

30. (2023) Planetary Defense - DART - NASA Science. https://science.nasa.gov/planetary-defense-dart/. Accessed 28 Aug 2026

31. European Space Agency End of the blue glow: BepiColombo turns off solar electric propulsion for Mercury arrival. https://www.esa.int/Enabling_Support/Operations/End_of_the_blue_glow_BepiColombo_turns_off_solar_electric_propulsion_for_Mercury_arrival. Accessed 28 Aug 2026

32. BepiColombo factsheet. https://www.esa.int/Science_Exploration/Space_Science/BepiColombo/BepiColombo_factsheet?utm. Accessed 28 Aug 2026

33. Possner MP, Almendra M, Garcia G (2012) A New Flight Dynamics Solution for Operations of the Boeing 702 Satellite. San Diego, CA

34. Harrigan M, Cundiff C, Rovey JL, Coverstone V (2025) Simulation-Based Evaluation of Dual-Mode Propulsion Strategies for Small Satellite Collision Avoidance

35. Talent DL (2007) An Assessment of the Impact of the January 2007 Chinese ASAT Test on the LEO Environment. In: Ryan S (ed). The Maui Economic Development Board, Wailea, Maui, Hawaii

36. National Aeronautics and Space Administration (2009) The Collision of Iridium 33 and Cosmos 2251: The Shape of Things to Come. Daejeon, Republic of Korea

37. Zhang W, Hu P (2025) An Edge AI Solution for Space Object Detection

38. Gonzalo JL, Colombo C (2021) On-board Collision Avoidance Applications Based on Machine Learning and Analytical Methods. In: Flohrer T, Lemmens S, Schmitz F (eds). ESA Space Debris Office, p n/a

39. Tafanidis NA, Banerjee A, Satpute SG (2025) Reinforcement Learning Based Intelligent Control for Low-thrust Tight Station Keeping in Low Earth Orbit. Int J Control Autom Syst 23:2105–2116. https://doi.org/10.1007/s12555-024-0648-z

40. Yu H, Dou L, Zhang X, et al (2025) Safe Reinforcement Learning: Optimal Formation Control With Collision Avoidance of Multiple Satellite Systems. IEEE Trans Cybern 55:. https://doi.org/10.1109/TCYB.2024.3491582

41. Harrison T (2021) Tactically Responsive Space: Emerging as a New Deterrence Tool

42. Wang X, Fang X (2023) A multi-agent reinforcement learning algorithm with the action preference selection strategy for massive target cooperative search mission planning. Expert Syst Appl 231:120643. https://doi.org/10.1016/j.eswa.2023.120643

43. (2018) Orbit Fab: Spacecraft Refueling Services and RAFTI Interface. In: OrbitFab Refuel Your Spacecr. https://www.orbitfab.com/

44. Abdul-Rahman S, Ozusaglam S (2025) Standards and Innovation: A Review of Empirical Evidence. Innovation and Research Caucus

45. Emrich Jr WJ (2023) Principles of Nuclear Rocket Propulsion. Elsevier

46. Wilson BA, Taylor C, Gray C, et al (2024) Space Nuclear Power Autonomous Control Algorithm and Control Element Test Bed. In: 2024 IEEE Aerospace Conference. IEEE, Big Sky, MT, USA, pp 1–7

47. Sharma A, Stopher M (2025) Nuclear Thermal Propulsion and Flight Safety: A Review. Am J Stud Res 800–811. https://doi.org/10.70251/HYJR2348.35800811

48. Holman R, Pierce B (1986) Development of NERVA reactor for space nuclear propulsion. p 1582

49. Sahadath MH, Cheng Q, Pan S, Ji W (2026) Surrogate Modeling for Neutron Transport: A Neural Operator Approach. Nucl Sci Eng 1–19. https://doi.org/10.1080/00295639.2026.2660591

50. Raissi M, Perdikaris P, Karniadakis GE (2019) Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys 378:686–707. https://doi.org/10.1016/j.jcp.2018.10.045

51. Guddanti SS, Fernandez-Tous M (2025) Control Systems for Nuclear Thermal Propulsion: Analysis of Past and Present Techniques for Future Space Missions. In: Space Vehicles for Exploration & Propulsion for Deep Space. International Astronautical Federation (IAF), New Delhi, India, pp 128–137

52. Schwarcz D, Schneider N, Oren G, Steinitz U (2024) Reactor Optimization Benchmark by Reinforced Learning. EDP Sciences, Les Ulis, p 17009

53. Zhu P, Zhao P, Liu X, et al (2025) Multi-objective conceptual design method for nuclear thermal propulsion reactor. Nucl Tech 48:110601. https://doi.org/10.3724/j.0253-3219.2025.hjs.48.240286

54. Maddock C, Parsonage B (2024) Multi-Objective Design Optimisation and Analysis of a Crewed Earth-Mars Transportation System using Nuclear Thermal Propulsion. Milan, Italy

55. Qi B, Liang J, Tong J (2023) Fault Diagnosis Techniques for Nuclear Power Plants: A Review from the Artificial Intelligence Perspective. Energ Basel 16:1850. https://doi.org/10.3390/en16041850

56. Ayodeji A, Amidu MA, Olatubosun SA, et al (2022) Deep learning for safety assessment of nuclear power reactors: Reliability, explainability, and research opportunities. Prog Nucl Energy New Ser 151:104339. https://doi.org/10.1016/j.pnucene.2022.104339

57. Zhang R, Zhu S, Wang K, et al (2025) Artificial Intelligence in Reactor Physics: Current Status and Future Prospects. https://doi.org/10.48550/arxiv.2503.02440

58. Zohuri B (2024) Applications of AI and ML in Nuclear Reactor Control Systems. 3:1–08

59. Fu J, Jin Z, Dai Z, et al (2024) Model predictive control for automatic operation of space nuclear reactors: Design, simulation, and performance evaluation. Ann Nucl Energy 199:110321. https://doi.org/10.1016/j.anucene.2023.110321

60. Gong A, Chen Y, Zhang J, Li X (2024) Possibilities of reinforcement learning for nuclear power plants: Evidence on current applications and beyond. Nucl Eng Technol 56:1959–1974. https://doi.org/10.1016/j.net.2024.01.003

61. Jendoubi C, Asad A (2024) A Survey of Artificial Intelligence Applications in Nuclear Power Plants. IoT 5:666–691. https://doi.org/10.3390/iot5040030

62. Wan EA, Van Der Merwe R (2000) The unscented Kalman filter for nonlinear estimation. In: Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373). IEEE, Lake Louise, Alta., Canada, pp 153–158

63. Bulusu R (2025) Systemization of Knowledge: Resilience and Fault Tolerance in Cyber-Physical Systems

64. Bak S, Chivukula DK, Adekunle O, et al (2009) The System-Level Simplex Architecture for Improved Real-Time Embedded System Safety. In: 2009 15th IEEE Real-Time and Embedded Technology and Applications Symposium. IEEE, San Francisco, CA, USA, pp 99–107

65. Obermaisser R, Peti P (2007) The Fault Assumptions in Distributed Integrated Architectures. Los Angeles, California, United States, pp 2007-01–3798

66. (1967) Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies

67. (1979) Agreement Governing the Activities of States on the Moon and Other Celestial Bodies

68. Aust A (2010) Handbook of International Law. Cambridge University Press, Cambridge

69. (1969) Vienna Convention on the Law of Treaties

70. United Nations Office for Outer Space Affairs (2026) Treaty Status. https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/status/index.html. Accessed 11 Aug 2026

71. (1928) Factory at Chorzów (Merits)

72. (1949) Corfu Channel (United Kingdom of Great Britain and Northern Ireland v. Albania)

73. United Nations General Assembly (2001) Articles on Responsibility of States for Internationally Wrongful Acts. United Nations

74. (1972) Convention on the International Liability for Damage Caused by Space Objects

75. United Nations General Assembly (1963) Declaration of Legal Principles Governing the Activities of States in the Exploration and Use of Outer Space

76. (1984) Commercial Space Launch Act, as Amended

77. License applications and requirements

78. Effective periods, and modifications, suspensions, and revocations, of licenses

79. COSPAR (2026) COSPAR Policy on Planetary Protection. Committee on Space Research (COSPAR)

80. Fleisher W, Cibralic B, Basl J, et al (2025) Responsibility and Accountability in an Algorithmic Society. Philos Technol 38:144. https://doi.org/10.1007/s13347-025-00970-w

81. Ryan P, Porter Z, Al-Qaddoumi J, et al (2024) What's my role? Modelling responsibility for AI-based safety-critical systems

82. Martin DA (2026) Defining the Responsibilities of Engineers: A Conceptual Framework for Engineering Ethics Education, Policy, and Practice. Eng Stud 18:105–130. https://doi.org/10.1080/19378629.2025.2609063

83. Porter Z, Zimmermann A, Morgan P, et al (2022) Distinguishing two features of accountability for AI technologies. Nat Mach Intell 4:734–736. https://doi.org/10.1038/s42256-022-00533-0

84. (2011) DO-178C: Software Considerations in Airborne Systems and Equipment Certification.

85. European Union Aviation Safety Agency (2024) EASA Artificial Intelligence Concept Paper Issue 2: Guidance for Level 1 & 2 Machine-Learning Applications. EASA

86. Raji ID, Smart A, White RN, et al (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, Barcelona Spain, pp 33–44

87. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x

88. Nissenbaum H (1996) Accountability in a computerized society. Sci Eng Ethics 2:25–42. https://doi.org/10.1007/BF02639315

89. Thompson DF (1980) Moral Responsibility of Public Officials: The Problem of Many Hands. Am Polit Sci Rev 74:905–916. https://doi.org/10.2307/1954312

90. Cooper AF, Moss E, Laufer B, Nissenbaum H (2022) Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In: 2022 ACM Conference on Fairness Accountability and Transparency. ACM, Seoul Republic of Korea, pp 864–876

91. Elish MC (2016) Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (pre-print)

92. Liu Y, Wang H, Wu T, et al (2022) Attitude control for hypersonic reentry vehicles: An efficient deep reinforcement learning method. Appl Soft Comput 123:108865. https://doi.org/10.1016/j.asoc.2022.108865

93. Parasuraman R, Manzey DH (2010) Complacency and Bias in Human Use of Automation: An Attentional Integration. Hum Factors J Hum Factors Ergon Soc 52:381–410. https://doi.org/10.1177/0018720810376055

94. Santoni de Sio F, van den Hoven J (2018) Meaningful Human Control over Autonomous Systems: A Philosophical Account. Front Robot AI 5:15. https://doi.org/10.3389/frobt.2018.00015

95. Floridi L, Cowls J, Beltrametti M, et al (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach Dordr 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

96. Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. Minds Mach Dordr 30:99–120. https://doi.org/10.1007/s11023-020-09517-8

97. Endsley MR (2017) From Here to Autonomy: Lessons Learned From Human–Automation Research. Hum Factors 59:5–27. https://doi.org/10.1177/0018720816681350

98. Lee J, See K (2004) Trust in automation: Designing for appropriate reliance. Hum Factors 46:50–80. https://doi.org/10.1518/hfes.46.1.50.30392

99. Jobin A, Ienca M, Vayena E (2019) Artificial Intelligence: the global landscape of ethics guidelines. arXiv.org. https://doi.org/10.48550/arxiv.1906.11668

100. Elliott KC (2022) A Taxonomy of Transparency in Science. Can J Philos 52:342–355. https://doi.org/10.1017/can.2020.21

101. Elliott KC, Werkheiser I (2023) A Framework for Transparency in Precision Livestock Farming. Anim Basel 13:3358. https://doi.org/10.3390/ani13213358

102. Shrader-Frechette K (2002) Environmental Justice: Creating Equality, Reclaiming Democracy. Oxford University Press

103. Werkheiser I (2016) Community Epistemic Capacity. Soc Epistemol 30:25–44. https://doi.org/10.1080/02691728.2014.971911

104. Miller GJ (2022) Stakeholder roles in artificial intelligence projects. Proj Leadersh Soc 3:1–15. https://doi.org/10.1016/j.plas.2022.100068

105. Mittelstadt BD, Allo P, Taddeo M, et al (2016) The ethics of algorithms: Mapping the debate. Big Data Soc 3:1–21. https://doi.org/10.1177/2053951716679679

106. Barocas, Solon; Hardt, Moritz; Narayanan, Arvind (2023) Fairness and Machine Learning: Limitations and Opportunities. The MIT Press

107. Amodei D, Olah C, Steinhardt J, et al (2016) Concrete Problems in AI Safety. https://doi.org/10.48550/arxiv.1606.06565

108. Seshia SA, Sadigh D, Sastry SS (2022) Toward verified artificial intelligence. Commun ACM 65:46–55. https://doi.org/10.1145/3503914

109. Datta A, Sen S, Zick Y (2016) Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In: SP. IEEE, NEW YORK, pp 598–617

110. Leveson N (2017) Engineering a safer world : systems thinking applied to safety. MIT Press, Cambridge, Mass. ; London