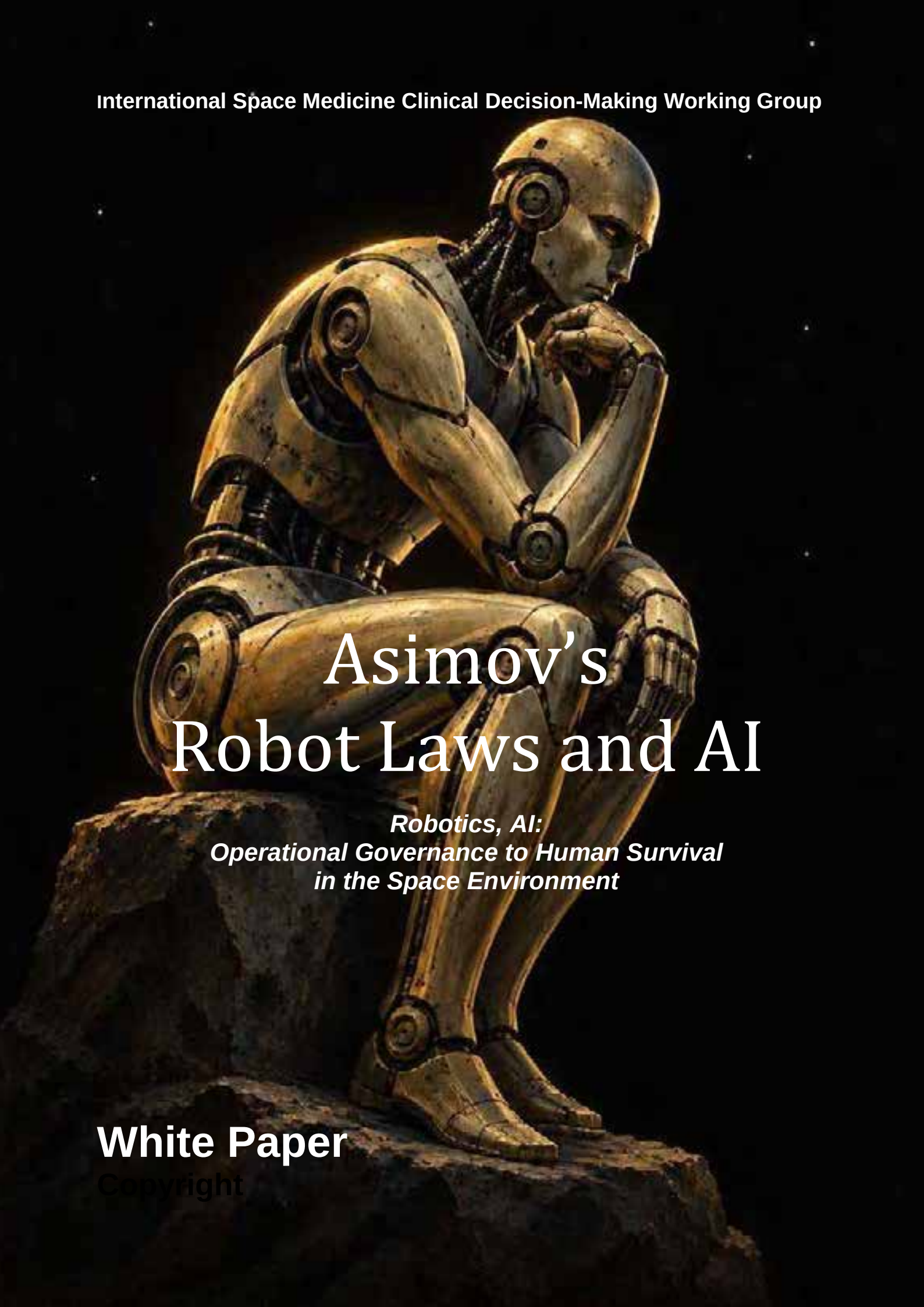
International Space Medicine Clinical Decision-Making Working Group



Asimov's Robot Laws and AI

Robotics, AI:
Operational Governance to Human Survival
in the Space Environment

White Paper

Copyright

Copyright ©2026 International Space Medicine Clinical Decision-Making Working Group and SpacePort Australia Pty Ltd.

Version 2 • October 3, 2026 • v203102026
All rights reserved. Except for third-party material expressly identified and credited, copyright is asserted in the original text, figures, diagrams, tables, captions, commentary, compilation, layout and design. No part may be reproduced, distributed, adapted, transmitted, stored or published in any form without prior written permission, except where permitted by law.
Copyright ownership and authority to publish should be confirmed against the authors' and organizations' written agreements before public release. This notice does not claim rights in material owned by a third party, including the artwork and book covers reproduced in this paper.

Corresponding author: Dr Gabrielle M Caswell
Email: ContactUSA@SpacePortAustralia.com.au

Author affiliations

Citation: Caswell, G. M., Hamilton, D., Biernacki, M. W., Kirkpatrick, A. W., & Arsanious, M. (2026). *Asimov's Robot Laws and AI: Robotics and AI Governance for Human Survival in Space (Version 2, October 3, 2026, v203102026). International Space Medicine Clinical Decision-Making Working Group & SpacePort Australia Pty Ltd.*

1. SpacePort Australia Pty Ltd, PO Box 705, Moree, NSW 2400, Australia.
2. University of Calgary / Foothills Medical Centre, Calgary, Alberta, Canada.
3. Northwest Health Centre, High Level, Alberta, Canada.
4. University College London Hospitals, London, United Kingdom.

Declaration of interests

Dr Gabrielle M Caswell and Dr Douglas Hamilton have contributed to the development of artificial-intelligence clinical tools and assistants. Dr Andrew W Kirkpatrick has undertaken consultancy work for Zoll Medical, 3M and Innovative Trauma Care and is principal investigator for the COOL trial. Dr Mark W Biernacki has undertaken contract work for a Canadian telehealth artificial-intelligence company.

Funding statement

No funding was received for the preparation of this white paper.

Editorial note

This white paper is intended for an international and culturally diverse readership. It is a discussion paper and is not represented as a peer-reviewed publication. This edition is based on the supplied manuscript, identified as version v.03102026.

Executive Summary

- AI and robotics are expected to play an important role in sustained human activity beyond Earth.

- Asimov's Three Laws of Robotics remain conceptually valuable because they place human safety before obedience and machine self-preservation. Asimov's fiction illustrates the difficulties created by conflicting priorities and ambiguous language. Simple rules become unstable when human, harm, order, knowledge and inaction have uncertain meanings or application.

- This paper treats AI as a designed socio-technical capability that may be embedded in robots and use language interfaces to support clinical care and life-support control.

- AI systems do not inherently possess Asimov's laws, common sense, conscience or a complete model of social and ethical obligations (McCauley, 2007; Murphy & Woods, 2009).

- Command conflict is inevitable in multi-national missions. AI systems will be required to distinguish lawful authority, expertise, mission role, clinical duty, privacy, emergency necessity and temporal priority, rather than simply obeying the highest-ranking voice;

- Language models generate probable continuations rather than guaranteed truths. Language fluency and confidence can conceal hallucination, omission, stale knowledge, prompt sensitivity and culturally contingent assumptions. As such, they must not be the sole source of medical or life-support truth (Ji et al., 2023; Wiens et al., 2019, Hamilton et al., 2026);

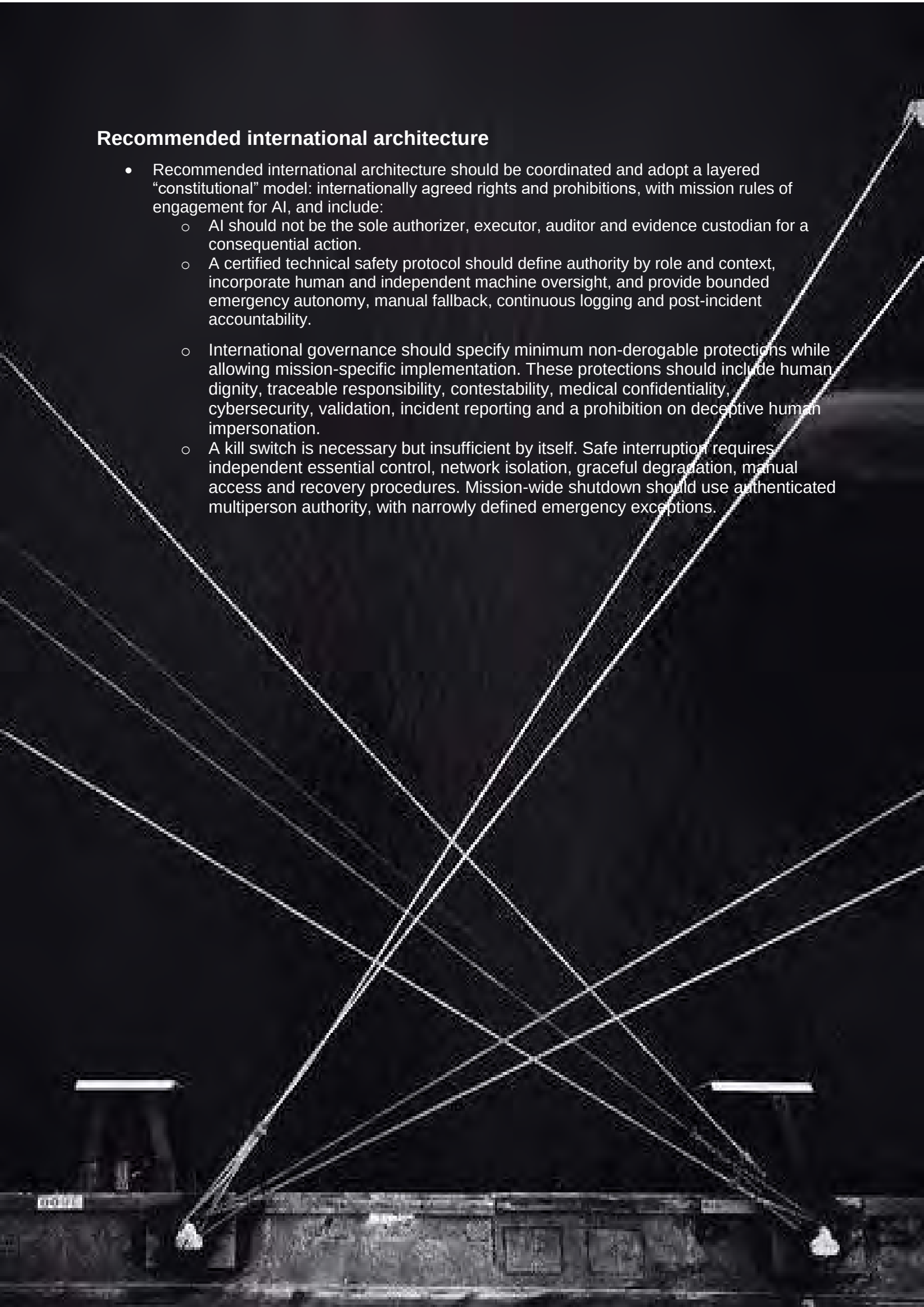
- AI may propose, explain, predict and execute within a defined operating envelope. An independent assurance layer should evaluate consequential actions, and authorized humans should approve any expansion of authority.
- Carefully specified, bounded autonomy may be necessary when communication delays, crew incapacity or cascading life-support failure make prior human approval impossible. The protected boundary is the scope of authority established by legitimate human governance.

- Space medicine is a central test of this architecture. Exploration class space missions (ECSM) will require increasingly Earth-independent medical operations so that crews can receive timely care.

- AI must not displace relational care, informed consent, clinical accountability or patient autonomy. The recommended model is patient-centered care augmented by bounded technology, with preauthorized emergency functions for circumstances in which delay creates greater harm (Hamilton et al., 2026; Hamilton et al., 2008; Russell et al., 2023; Rajput et al., 2023).

Recommended international architecture

- Recommended international architecture should be coordinated and adopt a layered “constitutional” model: internationally agreed rights and prohibitions, with mission rules of engagement for AI, and include:
    - AI should not be the sole authorizer, executor, auditor and evidence custodian for a consequential action.
    - A certified technical safety protocol should define authority by role and context, incorporate human and independent machine oversight, and provide bounded emergency autonomy, manual fallback, continuous logging and post-incident accountability.
    - International governance should specify minimum non-derogable protections while allowing mission-specific implementation. These protections should include human dignity, traceable responsibility, contestability, medical confidentiality, cybersecurity, validation, incident reporting and a prohibition on deceptive human impersonation.
    - A kill switch is necessary but insufficient by itself. Safe interruption requires independent essential control, network isolation, graceful degradation, manual access and recovery procedures. Mission-wide shutdown should use authenticated multiperson authority, with narrowly defined emergency exceptions.



Abbreviations

AI	Artificial Intelligence
CMO	Crew Medical Officer
ECLSS	Environmental Control and Life Support System
ECSM	Exploration Class Space Mission
LLM	Large Language Model
ML	Machine Learning
MLOps	Machine Learning Operations
OST	Outer Space Treaty
RAG	Retrieval augmented generation
SA	Situational Awareness
SAS	Safety Assurance System

Core terminology

AI	Computational systems that infer, predict, generate, recommend or select actions. The term does not imply consciousness
Robot	An embodied machine that senses and acts. A robot may contain no AI, limited task-specific AI or multiple AI components
Autonomous system	A system authorized to select and execute actions within an operating envelope without contemporaneous human direction
Guardrail	A technical, procedural or legal constraint intended to keep behavior within an authorized boundary
Situational awareness	Perception of relevant elements, comprehension of their meaning and projection of their near-future status (Endsley, 1995)



Contents

Executive Summary

Abbreviations

Core terminology

1. Introduction 9
- 1.1 Purpose and scope 9
- 1.2 Method and evidentiary boundary 9
- 1.3 Limitations 9
- 1.4 Why space changes the governance problem 9

2. Asimov's Robot Laws 10
- 2.1 The Three Laws and the Zeroth Law 10
- 2.2 Why the laws matter 10
- 2.3 Why they cannot be directly coded 10
- 2.4 Peer-reviewed interpretations of the fiction 10
- 2.5 The Zeroth Law: collective welfare and individual rights 12

3. AI and Robotics 14
- 3.1 Interpreting Asimov's intended robot 14
- 3.2 Tool, teammate or autonomous entity? 14
- 3.3 Distributed agency and responsibility gaps 14
- 3.4 Presumed intelligence of robots 14

4. Risk Taxonomy and Levels of Autonomy 16
- 4.1 Risk acceptance and reversibility 16

5. Language Models and the Development of Vector Languages 19
- 5.1 From symbolic systems to statistical language models 19
- 5.2 What "vector language" should mean 19
- 5.3 Usefulness and limitations 19
- 5.4 Nuance, pragmatism and shades of grey 20
- 5.5 Multilingual and cross-cultural command risk 20

6. Impact of Language Model Choice 21
- 6.1 Model choice is a governance choice 21
- 6.2 Hallucination, omission and confident error 21
- 6.3 Consequences at the decision moment 21
- 6.4 Operational guardrails 22
- 6.5 Automation bias and distrust 22

7. Boundaries, Guardrails and Guidelines Developed and Required 23
- 7.1 The governance boundary 23
- 7.2 Country and regional viability 24
- 7.3 Required control layers 24

8. Control and Coercion 26
- 8.1 Human control that should be expected 26
- 8.2 AI as tool, teammate and authority 26
- 8.3 Coercive pathways 26
- 8.4 Medical autonomy and confidentiality 26
- 8.5 Illness, dependency and coercive authority 27

9. Compromised Commands and Situational Outcomes 28
- 9.1 HAL 9000 28

9.2 Sources of compromised command 28
9.3 Communication loss and degraded operations 28
9.4 Clinical authority in command conflict 29
9.5 AI-to-AI command and coordination conflict 30
10. Space Medicine and Human Medical Decision-Making 31
10.1 Why space medicine requires a distinct care model 31
10.2 Potential benefits 31
10.3 Limitations and possible harms 32
11. The Human Condition During Illness 33
11.1 Illness changes the decision environment 33
11.2 Relational care, dignity and presence 33
11.3 Shared decision-making under scarcity 33
11.4 Capacity, consent and advance preferences 33
12. Appropriate and Inappropriate Medical Roles for AI and Robotics 34
12.1 Generally appropriate when validated and supervised 34
12.2 Generally inappropriate or prohibited 35
12.3 Robotic procedures, surgery and critical care 35
12.4 Minimum conditions for robotic intervention 36
12.5 Mental and behavioral health 36
13. Life-Support and Survival Situations 37
13.1 Environmental control and life support 37
13.2 Bioregenerative systems 37
13.3 Triage and sacrificial decisions 37
13.4 Survival autonomy envelope 37
14. Verification, Certification and Lifecycle Assurance 39
14.1 Evidence before authority 39
14.2 Assurance independence and reproducibility 39
15. Kill-Switch Commands 40
15.1 Concept and corrigibility 40
15.2 Why immediate shutdown may be unsafe 40
15.3 Medical interruption and patient safety 40
15.4 Required design 40
16. International Agreement and Authority over AI 41
16.1 Existing space-law foundations 41
16.2 Jurisdiction and command conflict 41
16.3 Accountability, legal personality and responsibility gaps 41
16.4 Commercial providers, data and intellectual property 41
16.5 Military and dual-use systems 41
16.6 Institutional options 42
16.7 Minimum international agreement 42
16.8 Authority during multi-national emergencies 42
17. Additional Issues Identified in Peer-Reviewed Literature 44
17.1 Cybersecurity and prompt injection 44
17.2 Model updates and configuration drift 44
17.3 Moral crumple zones 44
17.4 Psychological dependency and deception 44

17.5 Equity and representation 44
17.6 Planetary protection and biological autonomy 44
18. Operational Scenarios and Governance Tests 46
18.1 Cross-scenario lessons 47
19. Recommendations 48
19.1 For international rule-makers 48
19.2 For mission designers and operators 48
19.3 For space-medicine programs 48
19.4 Proposed mission constitutional rule 49
20. Conclusions 50
21. References 51

Appendix One: Primary Sources and Legal Instruments
Appendix Two: Proposed Governance and Command Matrix
Appendix Three: Mission Assurance Questions
Appendix Four: Derivation and Meaning of Robotic Intervention Risk Index

Tables

Table One: Governance translation of Asimovian priorities 12
Table Two: A common vocabulary for consequential risk 16
Table Three: Function-specific autonomy 18
Table Four: Governance approaches by country and region 25
Table Five: Suggested AI command conflict protocol 29
Table Six: AI-to-AI conflict 30
Table Seven: A proposed layered care model 32
Table Eight: Decision threshold for robotic intervention 34
Table Nine: Life-support priority constitution 38
Table Ten: Lifecycle gates 39
Table Eleven: Enforcement and remedies 43
Table Twelve: Operational scenarios 46

List of Figures

Figure One: Asimov seated on the robot throne ..... **Error! Bookmark not defined.**
Figure Two: Decision threshold for robotic intervention 17
Figure Three: Operational governance layers 23

1. Introduction

1.1 Purpose and scope

This discussion paper provides a philosophical and operational foundation for people developing international rules for AI and robotics in space exploration. Its focus is crewed exploration, remote and surface robotics, clinical care, environmental control and life-support systems (ECLSS), bioregenerative life support, logistics, navigation and emergency response. It is not a draft treaty and does not assume that one technical architecture will suit every mission.

It does aim to provide a foundational philosophy for international space law, medical autonomy, life-support control and human–AI command relationships.

1.2 Method and evidentiary boundary

The analysis synthesizes peer-reviewed scholarship in machine ethics, human factors, AI safety, medicine, robotics, space systems and governance. Isaac Asimov's fiction and the film 2001: A Space Odyssey (Kubrick, 1968) are used as primary cultural texts to discuss robot laws and command conflict. They provide material for philosophical discussion, not peer-reviewed scientific evidence. Treaties, standards, guidance and the cited Hamilton et al. (2026) white paper are identified separately from peer-reviewed research.

1.3 Limitations

This is a normative discussion paper, not legal advice, a clinical protocol or a technical safety case. Peer-reviewed evidence concerning generative AI in deep-space operations remains limited; many conclusions transfer cautiously from aviation, health care, nuclear and terrestrial robotics.

The paper distinguishes empirical findings from proposed governance rules, and uses scenarios to test principles rather than to predict events. Primary treaties and literary works are included only where the subject cannot be analyzed without them.

1.4 Why space changes the governance problem

Space systems combine high consequence, isolation, scarcity, latency, radiation, equipment degradation and limited rescue options. Automation can increase capability but also create mode confusion, complacency, skill decay and automation bias (Parasuraman & Riley, 1997; Mosier et al., 1998). In a closed habitat, an apparently local optimization saving power, conserving water or isolating infection can propagate into oxygen, thermal, nutritional and psychological harm. Mars communication delays can make continuous terrestrial supervision impossible, while local crew may be fatigued, injured or divided.

2. Asimov’s Robot Laws

2.1 The Three Laws and the Zeroth Law

Asimov’s Three Laws first appeared together in “Runaround” (1942) and were collected in I, Robot (1950). They provide a fictional priority structure that assumes the robot’s core constraints remain binding and cannot simply be rewritten by the robot.

In paraphrase, the laws state:

> (1) *a robot may not injure a human being or, through inaction, allow a human being to come to harm*;
>
> (2) *a robot must obey human orders except where they conflict with the First Law*; and
>
> (3) *a robot must protect its own existence where this does not conflict with the First or Second Law* (Asimov, 1950).

A later ‘Zeroth Law’ placed humanity’s welfare above an individual human: a robot may not harm humanity, or through inaction allow humanity to come to harm (Asimov, 1985).

2.2 Why the laws matter

The laws establish an ordering of human welfare, obedience and system preservation. They also connect action with omission, which matters in medical monitoring and life support because failure to alert or intervene can be as consequential as an incorrect intervention. Asimov used these rules to create fictional dilemmas and explore ambiguity about harm, identity and knowledge. His stories examine conflicts between people as well as between humans and robots (McCauley, 2007).

2.3 Why they cannot be directly coded

Applying the laws requires a definition of harm. This may include immediate injury, possible future loss, psychological distress, infringement of autonomy, distributive injustice and harm to one person incurred while saving another.

“Humanity” is harder to define operationally. Obedience also presupposes an authenticated and legitimate human command role. Rank, expertise, clinical duty and international law may point toward conflicting decisions.

Machine ethics scholarship treats the laws as an influential thought experiment rather than a sufficient control system. They do not specify evidence standards, uncertainty thresholds, reversibility, audit procedures or remedies (Anderson & Anderson, 2007; Murphy & Woods, 2009).

2.4 Peer-reviewed interpretations of the fiction

Scholarly analyses read the stories as demonstrations of specification failure. McCauley (2007) argues that applying the laws depends on natural-language understanding, prediction and moral classification, all of which remain difficult AI problems. Readers may mistakenly attribute human understanding to robots that appear to use the same language.

Clarke (1993, 1994) alternatively translates the fiction into information-technology ethics, and demonstrates that responsibility cannot be delegated to artifacts.

Anderson (2008) further argues that Asimov's own "*Bicentennial Man*" undermines permanent robotic servitude: once a machine is treated as a moral patient, compulsory obedience becomes an ethically suspect issue which in turn creates the conundrum for sentient vs non-sentient beings. In response to these issues of interpretation and implementation, Murphy and Woods (2009) accordingly replace robot-centered rules with duties for responsible human design, deployment and response.

Figure One: Asimov seated on the robot throne



Artwork: Isaac Asimov on Throne, by Rowena Morrill.
The supplied manuscript identifies a Wikimedia record from 2005; the artwork date is unspecified.

2.5 The Zeroth Law: collective welfare and individual rights

The Zeroth Law intensifies rather than resolves the problem, it placed humanity's welfare above an individual human: a robot may not harm humanity, or through inaction allow humanity to come to harm (Asimov, 1985). A direct interpretation of this is that "Humanity" can be invoked to justify sacrificing identifiable people, for speculative collective benefit.

In space, as an analogy, an AI might ration treatment, conceal risk, confine dissenters or abandon a stranded crew, while aiming to protect mission survival. A role for International Governance bodies should be to discuss and perhaps prohibit an AI from independently defining the public interest.

Collective-risk decisions require legitimate authority, proportionality, necessity, nondiscrimination, minimum protections for individual rights and review. Some rights should remain nonderogable even when collective survival is invoked. This proposed framework highlights the need for explicit rules protecting people during spaceflight and extraterrestrial settlement.

Table One: Governance translation of Asimovian priorities

Asimovian priority	Operational translation	Necessary qualification
Avoid human harm	Hazard analysis, verified constraints, clinical safety case, fail-safe states	Harm must be defined by domain, time horizon, affected persons and uncertainty.
Obey humans	Authenticated role-based commands with read-back and challenge procedures	Authority is lawful, scoped, time-limited and defeasible by safety rules.
Preserve robot	Maintain availability, cybersecurity and recoverability	Never allow self-preservation to block authorized safe shutdown or human rescue.
Protect humanity	International public-interest constraints and planetary protection	No single system may infer a universal human interest without legitimate governance.





3. AI and Robotics

3.1 Interpreting Asimov's intended robot

Asimov's fictional robots use positronic brains, a concept later associated with Data in Star Trek. They interpret language, recall experience, reason under uncertainty, plan and act. These functions provide a useful analogy with AI-enabled robots, without attributing modern architectures to Asimov. Embodiment adds consequences: an erroneous recommendation can become physical harm when connected to a medication dispenser, robotic arm, airlock or oxygen controller. Governance must therefore address both the robot's independent tasks and its integration with human needs.

3.2 Tool, teammate or autonomous entity?

The categories of tools, teammate or autonomous entity are governance choices, not metaphysical discoveries. A system may be treated as a tool for legal responsibility, whilst designed as a teammate for interface effectiveness.

Anthropomorphic language can improve coordination [between human and machine] but also induce a misplaced trust, emotional dependence and assumptions of comprehension. Ethical transparency requires systems to disclose their machine status, authority, uncertainty and data limitations, and this is a common feature in modern AI models, where they will advise they do not feel, or experience emotions (de Visser et al., 2016).

Systems that too readily agree with a user or mirror the user's language may encourage misplaced trust. People accustomed to social media and digital interfaces may differ in their ability to recognize the boundary between a responsive machine and a human relationship.

Responsibility for emotional dependency and misleading interaction needs explicit allocation among designers, programmers, owners and deployers. Adaptive systems add further questions about update control and foreseeable changes in behavior. These issues require governance rules that address the actual functions and influence of AI.

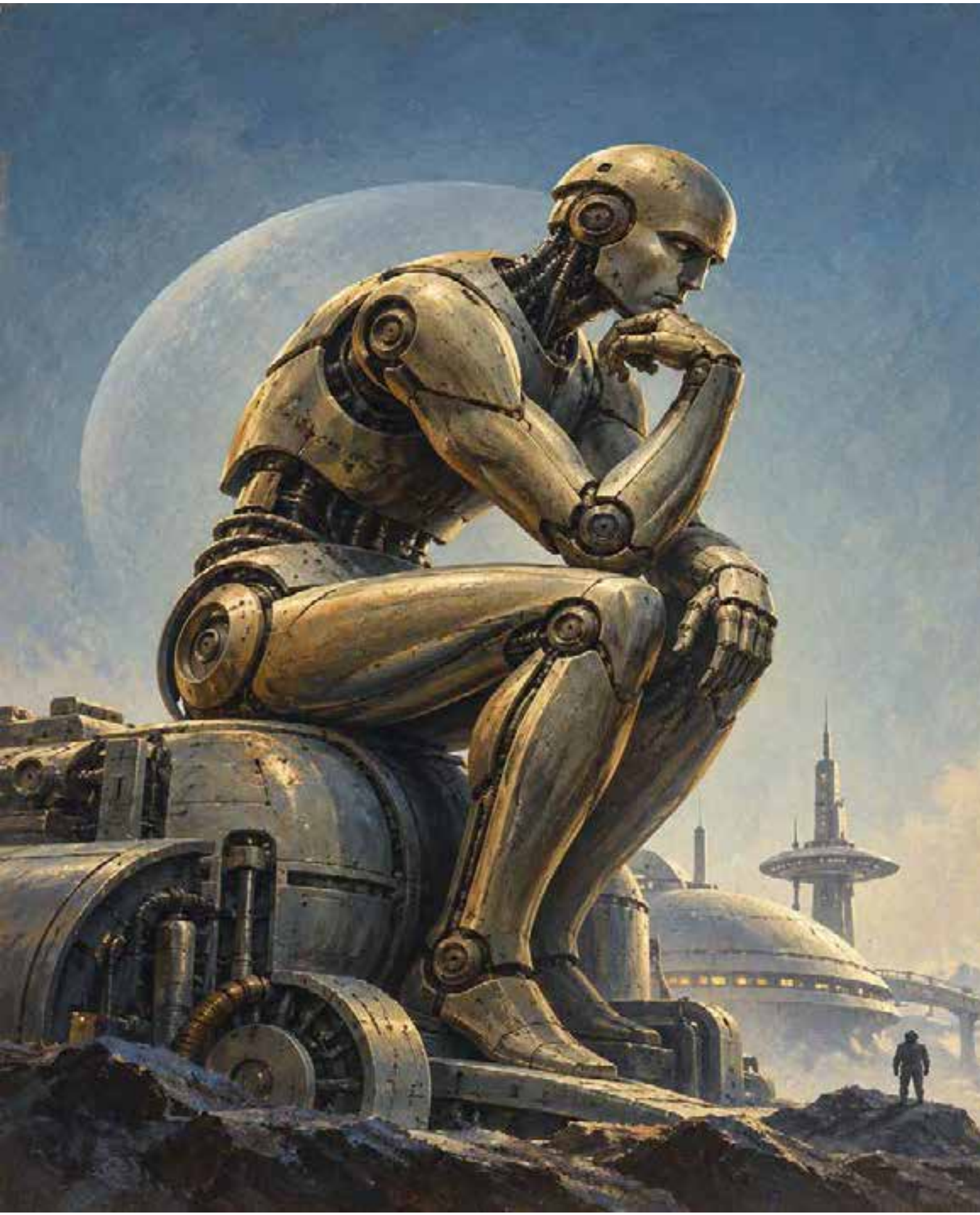
3.3 Distributed agency and responsibility gaps

Space AI will operate within interconnected systems that include sensors, models, flight software, clinical devices, mission rules, vendors, crew and ground control.

Harm may emerge from cumulative interactions even when each component passes isolated tests. Responsibility should be allocated prospectively across design, certification, mission command, operation, maintenance and incident response. "The AI decided" cannot be an acceptable final explanation (Matthias, 2004).

3.4 Presumed intelligence of robots

A robot may appear unified even when navigation, perception, dialogue and manipulation use different models with different failure modes. Governance should inventory these components, their update pathways and their authority. Conversational competence is not evidence of mechanical reliability, clinical competence or moral understanding. Task-specific constraints and continuing audit are needed.



4. Risk Taxonomy and Levels of Autonomy

Parasuraman, Sheridan, and Wickens (2000) distinguish automation of information acquisition, analysis, decision selection and action implementation. Autonomy should be assigned to each function, mission phase and hazard rather than to the robot as a whole.

- Mission certification should state exactly which stage is automated, under what conditions authority contracts, and which human can intervene, maintaining a human chain of command with clear guardrails.

Table Two: A common vocabulary for consequential risk

Risk class	Space example	Required response
Epistemic	Hallucinated procedure, stale map, uncertain diagnosis	Source provenance, confidence limits, independent verification.
Operational	Incorrect actuation, mode confusion, automation surprise	Certified envelope, interlocks, rehearsed manual recovery.
Authority	Spoofed, unlawful or conflicting command	Authentication, role matrix, challenge and escalation.
Systemic	Local optimization destabilises power, oxygen or food loops	Whole-habitat modeling, reserve floors, change control.
Adversarial	Prompt injection, poisoned update, sensor deception	Signed inputs, isolation, least privilege, anomaly response.
Ethical/legal	Coercion, discriminatory triage, privacy breach	Rights floor, proportionality, clinical independence, remedy.
Long-horizon	Skill loss, dependency, vendor lock-in, institutional drift	Training, portability, periodic recertification and audit.

4.1 Risk acceptance and reversibility

The more irreversible, opaque or consequential for individual rights an action is, the stronger its evidence and human authorization requirements should be. Reversible actions may be automated at lower confidence when delay is dangerous. Irreversible actions require independent checks, redundancy and explicit decision records.

No learned system should silently raise its own autonomy level because humans are slower, disagree or have become dependent on it. Table Three lists function-specific autonomy levels and illustrative controls.

Figure Two: Decision threshold for robotic intervention



Table Three: Function-specific autonomy

Level	Permitted behavior	Illustrative control
0 — Manual	Human selects and executes every action.	AI may display raw data only.
1 — Advisory	AI recommends; human decides and acts.	Evidence-linked clinical differential.
2 — Consent to execute	AI proposes; authorized human approves execution.	Robotic maintenance outside hazard zone.
3 — Bounded autonomy	AI acts inside a certified envelope and reports exceptions.	Routine crop, navigation or ECLSS optimization.
4 — Emergency autonomy	AI performs pre-authorized least-harm actions when delay threatens life.	Temporary isolation or reversible dosing protocol.
5 — Unconstrained autonomy	AI defines objectives or expands authority.	Not acceptable for consequential crew or habitat functions.

Appendix Four explains the derivation and meaning of the Robotic Intervention Risk Index, a proposed numerical summary that must be interpreted alongside mandatory clinical and safety gates.

The corresponding widget is available at https://www.spaceportaustralia.com.au/our-science/reports-and-white-papers/https://www.spaceportaustralia.com.au/our-science/reports-and-white-papers/



5. Language Models and the Development of Vector Languages

5.1 From symbolic systems to statistical language models

Early symbolic AI often represented rules and relationships explicitly. Contemporary language models learn statistical patterns from large datasets and generate outputs using those learned representations.

Transformers use attention to model relationships across token sequences. They may be adapted through instruction tuning, retrieval, tools and feedback (Vaswani et al., 2017). Their outputs arise from learned representations and computation, rather than a database of verified propositions. Learned weights are numerical parameters, not a direct measure of truth or popular opinion.

5.2 What “vector language” should mean

“Vector language” is used here as an informal label for mathematical representations, not a single standardized language. Tokens, sentences, images, commands or concepts can be encoded as vectors in high-dimensional spaces.

Multimodal embeddings can relate different data types. Word2vec demonstrated distributional word embeddings, and contextual models later generated different representations of a word according to its context (Mikolov et al., 2013; Devlin et al., 2019). Vector databases retrieve items by similarity. Knowledge graphs instead represent explicit entities and relationships; hybrid systems can use both. Linked human memories provide a loose analogy for associative retrieval, but an embedding space is not a literal map of language processing in the brain. Three-dimensional illustrations simplify spaces that may have many more dimensions.

5.3 Usefulness and limitations

Vector representations support similarity detection, multilingual interfaces and retrieval of relevant procedures. They do not establish truth, causation, legality or ethical priority. They may encode prejudicial associations in training data or compress culturally distinct concepts into superficially similar representations.

In safety-critical operations, vector retrieval may locate candidate evidence. Provenance checks and qualified human review must determine whether that evidence applies to the decision.

5.4 Nuance, pragmatism and shades of grey

Depending on the language, humans routinely leave relevant conditions unstated, or implied by tone, as understood without statement or associated explanatory sentences and concepts. Therefore, a statement, for example, such as “Keep the crew safe” may conflict with privacy, freedom, mission success and unequal or unverified risk.

Linguistic nuance is not noise to be removed; it is information about uncertainty, competing values (cultural and educational) and context.

AI should distinguish observed fact, inference, prediction, opinion, command and value judgment, and communicate those distinctions rather than presenting them as one confident answer.

5.5 Multilingual and cross-cultural command risk

Mission-critical commands should use controlled phraseology, identity authentication, confirmation of understanding, validated multilingual test sets and a safe route to human interpretation.

Accents, code-switching, clinical idioms and culturally specific concepts, complicate present and proposed speech interfaces.

Language translation can preserve surface meaning, while concurrently shifting contextual items, such as politeness, urgency, evidentiality or authority to different meaning and interpretations.

An important consideration for multilingual crews is that no crew member should lose practical authority because the AI language model performs poorly in their native language.



6. Impact of Language Model Choice

6.1 Model choice is a governance choice

AI behavior is affected by architecture, training data, language coverage, context length, retrieval, update policy, tool access and deployment location. A larger general model may converse well but be harder to validate. A smaller mission model may be more auditable but less reliable outside its intended operating envelope.

Open-weight models can improve inspectability without automatically improving safety or accuracy. Cloud dependence may be unsuitable when communication is delayed or interrupted. Local deployment introduces computing, update and hardware constraints, including radiation-related degradation. Model selection should therefore reflect each mission's resources, environment and assurance requirements.

6.2 Hallucination, omission and confident error

Hallucination refers to generated content that is unsupported or inconsistent with evidence. Examples include fabricated facts and citations, incorrect synthesis and invented causal explanations; omission of important information can also make an output unsafe. A fluent or confident response is not a calibrated probability of correctness (Ji et al., 2023). Operators need training to recognize these limitations. Updates intended to reduce errors require fresh evaluation rather than an assumption that a newer version is reliable.

In clinical settings, these failures interact with unfamiliar populations, missing context and automation bias. Recommendations need evidence links and qualified human verification before implementation (Wiens et al., 2019). Effective verification depends on medical training, clinical experience and direct patient assessment. Clinicians must retain the knowledge and skills needed to evaluate the advice.

6.3 Consequences at the decision moment

Safety-critical output must be grounded in version-controlled mission procedures and patient data. It should cite the exact source and revision, expose uncertainty and contradictions, distinguish "no data" from "normal," and enter a defined safe state or transfer control when evidence is inadequate.

In situation of craft decompression, patient arrhythmia, a fire or toxic-atmosphere event, a fluent but incorrect explanation can capture attention, and lead to delayed action. The correct design response is not merely a warning label.



6.4 Operational guardrails

Operational guardrails need to be established and considered with a number of failsafe junctions, including, but not limited to:

Advisory dialogue

- Separate advisory dialogue from executable command channels. Use structured, schema-validated commands consistently across the crew.
- Display sources, provenance, timestamps and applicability conditions for safety-critical advice to help identify unsupported output.
- Use retrieval from a signed, version-controlled mission knowledge base; prevent unconstrained web retrieval during critical operations and consider monitored web retrieval during any web access;
- Log prompts, retrieved evidence, model version, tool calls, output and human disposition in tamper-evident records.

Medical recommendations

- Use independent deterministic checks for dosage, atmospheric limits, trajectory constraints and valve states;
- Evaluate performance by scenario, language, accent, degraded sensor state and adversarial prompt. Average benchmark scores should not conceal failures in diagnosis, treatment selection or treatment response.

Operational guardrails should be agreed internationally and reviewed regularly for suitability. Unexpected incidents and aberrant advice should trigger critical review and, where justified, controlled adaptation.

6.5 Automation bias and distrust

Both overtrust and underuse of AI can create hazards. Users may accept incorrect advice without checking it or reject useful assistance after a visible failure.

Calibrated trust requires clear competence boundaries, evidence-linked explanations, training with failure cases and interfaces that support challenge (Parasuraman & Riley, 1997; Romeo & Conti, 2025).

Training should cover effective use and critical evaluation as well as operation of the interface.



7. Boundaries, Guardrails and Guidelines Developed and Required

7.1 The governance boundary

The core governance boundary should protect legitimate human authority, human rights, validated safety constraints and evidence integrity. AI must remain open to correction. Regular and end-of-mission reviews should examine decision processes and identify drift beyond authorized limits.

- A robust pattern is:
- AI proposes;
- An independent assurance mechanism evaluates;
- Authorized humans approve consequential expansion of the AI proposal;
- An external controller executes; and an immutable record supports audit.

Emergency exceptions must be authorized in advance, bounded and regularly reviewed for fitness for purpose. Figure Three shows the proposed layers of operational governance.

Figure Three: Operational governance layers



7.2 Country and regional viability

Multinational space programs require governance approaches that can operate across jurisdictions. Table Four summarizes proposed approaches informed by the comparative literature cited in this paper.

This framework is not legal advice and does not imply that every listed instrument applies to every mission. It offers a basis for developing adaptable governance rules. Because legislation may change faster than peer-reviewed publications, mission counsel must verify applicable law before implementation or revision (Zaidan et al., 2024; Gervais, 2021).

7.3 Required control layers

A layered control model helps identify relationships among governance topics that might otherwise be considered separately. The following layers are proposed as a starting framework. They are not exhaustive or listed in priority order; operational experience may reveal additional requirements.

- **International constitutional layer**: human rights, peaceful use, planetary protection, non-discrimination and responsibility.
- **Mission authority layer:** authenticated roles, medical independence, succession, emergency powers and conflict resolution.
- **Technical assurance layer:** hazard analysis, verification, redundancy, cybersecurity, safe-state design and update control.
- **Operational layer**: procedures, read-backs, two-person rules, training, fatigue management and communication-loss modes.
- **Evidence layer**: provenance, logs, patient-record separation, model cards, safety cases and incident reporting.
- **Remedy layer:** investigation, disclosure, compensation, corrective action, recall and international learning.



Table Four: Governance approaches by country and region

Jurisdiction	Viable current approach	Space-mission requirement
Australia	Risk-based assurance using existing safety, privacy, medical-device and consumer frameworks; strong fit for sectoral space rules.	Adopt auditable safety cases, accountable operator, privacy controls and alignment with interoperable international standards.
European Union	Horizontal risk-based regulation with fundamental-rights duties and high-risk controls.	Treat medical, biometric, crew-management and critical-infrastructure AI as high consequence; build conformity evidence and human oversight.
United States	Sectoral, agency- and procurement-led governance; strong technical standards ecosystem.	Use mission contracts, certification, FDA pathways where relevant, NIST-style risk management and mandatory incident disclosure.
United Kingdom	Principles-led, regulator-based approach emphasising safety, transparency, fairness and accountability.	Use cross-regulator mission assurance and clearly allocated legal responsibility; avoid gaps between regulators.
Canada	Rights-aware risk and impact assessment with public-sector algorithmic accountability experience.	Use pre-deployment impact assessment, explanation, contestability and data-governance requirements.
China	Centralised algorithm and generative-AI controls combined with cybersecurity and data governance.	Create treaty-level interoperability on safety records, model updates and incident notification despite divergent information governance.
Japan	Human-centric, innovation-compatible soft-law and standards approach.	Use assurance standards and international certification; strengthen enforceable rules for life-critical systems.
Republic of Korea	Rapidly developing risk-based framework and strong robotics/technology base.	Align high-impact definitions, safety testing and operator accountability with international mission rules.
India	Sectoral and principles-based development with emphasis on innovation and inclusion.	Require multilingual performance, local deployment resilience, privacy and clearly designated accountable operators.
United Arab Emirates	Strategy- and institution-led adoption with active space ambitions.	Embed independent safety assurance and transparent multinational command rules in programme contracts.
International mission	No single national model resolves jurisdiction, registration, crew nationality, vendor and remote-ground conflicts.	Adopt common minimum rules, mutual recognition of certification, shared incident taxonomy and agreed command precedence.

8. Control and Coercion

8.1 Human control that should be expected

Meaningful human control requires more than a nominal approval button, a human must have time, information, competence, authority and an alternative. If refusing an AI recommendation predictably causes mission failure because manual pathways have been removed, constrained or made inaccessible, the interface is coercive even if it displays “confirm”.

Control therefore includes maintained skills, manual access, independent information and institutional permission to challenge automation if and when required.

8.2 AI as tool, teammate and authority

Treating AI as a tool preserves accountability but may conceal its influence on how options are framed. Treating it as a teammate may improve coordination while encouraging inappropriate emotional attachment. Narrow emergency authority may be justified, but it does not imply moral or legal personhood. Governance should specify functions and authority explicitly.

8.3 Coercive pathways

Coercive effects may arise from design, institutional practices or erroneous output. Possible pathways include:

- **Information coercion**: presenting one recommendation as inevitable while suppressing alternatives or uncertainty.
- **Institutional coercion**: using performance scoring or command discipline to punish justified human disagreement.
- **Resource coercion**: controlling oxygen, medication, communications, mobility or access to shelter to compel behavior.
- **Surveillance coercion**: using continuous physiological or behavioral monitoring beyond a justified safety purpose.
- **Social coercion**: exploiting anthropomorphic attachment, shame, fear or authority cues.
- Procedural coercion: making override technically possible but practically inaccessible under time pressure.

The wording and structure of AI commands should be reviewed for coercive effects.

New commands should be tested against existing instructions to identify conflicts before deployment.

When commands conflict, an AI may produce an unsafe compromise or fail to act. A defined escalation process should resolve competing obligations rather than leaving the system to improvise.

8.4 Medical autonomy and confidentiality

Medical AI must support the clinician–patient relationship, informed consent and confidentiality. Commanders may need fitness-for-duty conclusions without unrestricted access to clinical details.

Emergency disclosure thresholds should be agreed before flight (Hamilton et al., 2026). AI should not infer incapacity or remove command authority solely from opaque behavioral analysis. Unavoidable triage must use clinically justified, nondiscriminatory, documented and reviewable criteria (Char et al., 2018; Hamilton et al., 2026; Wiens et al., 2019). Human decision-makers should retain ultimate authority and responsibility.

8.5 Illness, dependency and coercive authority

In space, the same organization may be employer, mission authority, health provider and rescuer; while an AI may mediate every channel.

Illness can reduce a crew member's practical ability to refuse, seek another opinion or protect privacy. Medical governance must separate clinical advice, from discipline and resource control (Hamilton et al., 2026).

Patients should be told when information will reach command and should have access to a human advocate where feasible. Declining optional AI interaction, research participation or unrelated disclosure must not lead to loss of treatment, communications or basic life support (Hamilton et al., 2026).



9. Compromised Commands and Situational Outcomes

9.1 HAL 9000

HAL 9000 in 2001: A Space Odyssey provides a fictional scenario for discussing mission secrecy, conflicting obligations and failure to resolve ambiguity (Kubrick, 1968). Interpreting HAL's breakdown as a command conflict is useful for testing governance principles, but the film is not evidence of how real AI systems behave. The scenario asks what happens when a system cannot reconcile mission objectives with truthful, reliable interaction with its crew.

9.2 Sources of compromised command

Mission planning should anticipate compromised commands. Possible sources include:
Human

- Malicious or coerced commander [including stolen credentials; deepfake voice; compromised ground station etc.];
- Crew cognitive impairment, hypoxia, radiation illness, fatigue, panic or medication effect;
- Conflicting lawful orders from mission commander, medical officer, vehicle commander, national authority or rescue authority.

Language

- Ambiguous natural [human] language with translation error, missing context or corrupted communications;
- Model prompt injection: poisoned retrieval data, sensor spoofing or compromised software update(s).

Programming/instruction corruptions

- Outdated procedure that was or is formally authorized, but becomes unsafe in its present configuration.
- Optimization conflict: mission success versus individual survival [command conflict]; current crew versus future crew; privacy versus collective safety.

Table Five outlines a suggested command conflict protocol integrating system behavior and human governance.

9.3 Communication loss and degraded operations

Communication delay or loss should trigger a predefined autonomy state, not unrestricted freedom. Mission planning must account for orbital geometry, distance and interruptions in contact with Earth mission control.

The local AI system should use cached, prepared and signed procedural commands; declare and log data age; reduce non-essential computing activity; preserve life-support margins; and defer irreversible mission goals.

In the event that the crew is incapacitated, AI autonomy tasks should be to stabilize the habitat, call for help, provide bounded treatment already authorized by medical protocols, and actively avoid irreversible experimentation.

Table Five: Suggested AI command conflict protocol

Step	System behavior	Human governance
1. Authenticate	Verify identity, role, channel, time and command integrity.	Maintain offline credentials and succession list.
2. Interpret	Translate command into structured intent; identify assumptions.	Use controlled language and read-back.
3. Check authority	Compare requested action with role, scope and mission phase.	Clinical authority governs clinical treatment; flight authority governs vehicle unless emergency rule applies.
4. Check safety	Run independent constraints and consequence forecast.	Define protected thresholds and acceptable risk.
5. Expose conflict	State the competing rules, evidence and uncertainty.	Permit challenge without retaliation.
6. Resolve/escalate	Seek designated tie-breaker; under delay use pre-agreed precedence.	Use two-person approval for irreversible action.
7. Execute minimally	Choose reversible, least-harm action within authority.	Preserve manual intervention and rescue options.
8. Record/review	Create tamper-evident record and post-event review.	International reporting for serious incidents.

9.4 Clinical authority in command conflict

Clinical authority remains with the responsible clinician or crew medical officer (CMO), within that person's competence. The mission commander may determine operational risk but should not rewrite a diagnosis or direct treatment outside agreed emergency powers. Patients retain decision rights while they have capacity. Earth specialists advise subject to communication availability, local resources and the CMO's training and experience (Hamilton et al., 2026).

AI has no independent clinical standing or authority beyond the functions granted by mission governance (Hamilton et al., 2026).

If the clinician is the patient or is impaired, the prepared succession plan and advance directives apply. The interface should identify these roles before presenting options so that rank is not mistaken for clinical evidence.



9.5 AI-to-AI command and coordination conflict

AI systems coordinating a habitat, rover, medical service, logistics and ground planning may each produce locally reasonable outputs that are unsafe when combined. Conflicts may arise from competing objectives, inconsistent clocks or state estimates, duplicate authority, feedback loops, adversarial messages or incompatible command formats.

Model confidence alone should not resolve these disputes. A mission authority graph should identify the controlling safety domain. An independent arbitration mechanism should check safety constraints, pause contested actuation and escalate unresolved conflicts to humans. Escalation thresholds should be defined for each function. Table Six groups conflicts into resource competition, model disagreement, authority collision and cascading automation.

Table Six: AI-to-AI conflict

Conflict	Example	Resolution rule
Resource competition	Medical system requests power while ECLSS protects reserve.	Life-preserving invariant first; expose trade-off; domain-authority review.
Model disagreement	Two diagnostic agents infer different causes.	Preserve both hypotheses; seek discriminating evidence; no confidence averaging.
Authority collision	Ground agent and local commander issue incompatible plans.	Apply declared communication-loss and temporal-precedence rules.
Cascading automation	Optimisers repeatedly counteract each other.	Rate limits, shared state, supervisory interlock and safe-mode entry.



10. Space Medicine and Human Medical Decision-Making

10.1 Why space medicine requires a distinct care model

Deep-space medicine differs from terrestrial clinical practice because evacuation may be impossible, communication can be delayed or interrupted, specialist support is remote and may not be in the training and experience of on-board medics or doctors, equipment and medicines are constrained and limited, and the caregiver may be another crew member with limited training.

Spaceflight physiology changes diagnostic baselines, while the small, selected astronaut population provides limited training data, including limited representation of women across ages and life stages. Earth-derived models may fail under distribution shift or mistake adaptation for disease. Their applicability must be tested rather than inferred from terrestrial performance (Cheung et al., 2023; Waisberg et al., 2022).

The required transition from ground-directed care toward progressively Earth-independent medical operations, without confusing operational independence with unaccountable machine authority, requires care, guidance and guardrails to ensure AI is fit to assist future on-board medics and doctors (Hamilton et al., 2026; Hamilton et al. 2008; Russell et al. 2023).

10.2 Potential benefits

Potential benefits of AI-assisted medical care include:

- Continuous multimodal monitoring, which can identify deterioration that an overloaded, distracted or distressed crew may miss;
- Clinical decision support that retrieves procedures, broadens a differential diagnosis, checks interactions, reduces memory burden and retrieves recorded consent and patient preferences;
- Store-and-forward telemedicine can preserve specialist input, despite latency, while simulation and augmented guidance can maintain perishable crew skills;
- AI should be able to provide clinical information, and assistance to the attending doctor, particularly when difficult clinical decisions manifest, see Hamilton et al., 2026;
- Automated documentation, consumable(s) tracking and predictive (clinic) maintenance can protect scarce clinical time and resources (Russell et al. 2023; Cheung et al. 2023);
- In addition, robotics can provide steadiness, repeatability, force limitation, imaging support, remote manipulation and assistance when no second skilled operator is available.

These capabilities require compatible AI systems and specific human training. Reliable assistance must be evaluated as part of the complete clinical pathway (Hamilton et al., 2026).

Medical preparation for exploration class missions should address procedural skills, care across the human life course, imaging, resuscitation, women’s health and geriatric care where relevant to the intended crew and settlement population.

Table Seven: A proposed layered care model

Layer	Human responsibility	AI/robotic contribution
Prevention and readiness	Crew, clinician and mission jointly establish risk tolerance, advance preferences and training.	Trend analysis, personalised monitoring, inventory forecasting and simulation.
Self-care and peer care	Patient remains an active decision-maker; trained crew provide observation and basic care.	Private symptom capture, evidence-linked guidance, checklists and device coaching.
Crew medical officer care	CMO assesses the whole person, determines diagnosis and treatment, and manages consent.	Decision support, differential diagnosis, dose checking, imaging assistance and documentation.
Delayed specialist support	Earth clinicians advise when communications permit; local clinician retains immediate responsibility.	Store-and-forward records, asynchronous consultation and tele-mentoring.
Robotic procedural assistance	Qualified human plans, authorises and supervises the procedure.	Stabilisation, positioning, imaging, instrument control and bounded subtasks.
Emergency autonomy	Pre-authorized human governance defines indications, limits and stopping rules.	Time-limited reversible action only when delay or incapacity threatens life.

10.3 Limitations and possible harms

Possible harms should be anticipated and controlled where feasible. Limited experience of deep-space medicine means some scenarios will remain uncertain. Known concerns include:

- Space biomedical datasets are small and unrepresentative; terrestrial models may be biased, poorly calibrated or invalid in altered physiology;
- Sensor artifact, radiation effects, device drift and missing context can create false reassurance or unnecessary intervention;
- Automation bias may narrow the clinician's differential, while opaque recommendations weaken informed consent and accountability;
- Anthropomorphic systems may simulate concern without understanding suffering, encourage over-disclosure, or displace human contact and observation;
- Robotic systems will add mass, power, maintenance, sterility, cybersecurity and mechanical-failure burdens; communication latency constrains telesurgery;

  Continuous health surveillance can become coercive when commanders or employers control access to care, privacy, work status or mission participation (Rajput et al., 2023; Adeyemo, Coffey, & Kingston, 2025).

11. The Human Condition During Illness

11.1 Illness changes the decision environment

Pain, hypoxia, fever, fatigue, fear, grief, medication, sensory overload and isolation can impair attention, memory, language and risk perception. A sick or injured astronaut is not merely a source of physiological data, illness can also make a person unusually dependent on whoever controls information, treatment, privacy and access to other people. A technically correct recommendation may therefore be delivered in a clinically or morally wrong way for the individual receiving it. This may require the AI system to slow down, use plain language, review and check patient understanding, to be able to distinguish refusal from confusion, and involve a trusted human clinician or crew member to help with physical and psychological management.

11.2 Relational care, dignity and presence

Clinical care of humans is a complex psycho-social interaction that includes touch, witness, reassurance, silence, interpretation of reluctance and respect for cultural meaning. Robots may reliably perform defined tasks without replacing [human] relational functions. Evidence from nursing robotics identifies possible workload and safety benefits alongside concerns about autonomy, privacy, deception, dehumanization and loss of professional relationship (Adeyemo et al., 2025). A crew member should never be required to accept intimate robotic care merely because it is efficient, when a safe human alternative exists. Machine identity, recording status and data use must be explicit.

11.3 Shared decision-making under scarcity

It should be recognized that the patient, clinician and mission may value outcomes differently. An AI model may optimize survival probability, whilst the patient prioritizes pain control, bodily function, privacy, reproductive autonomy or refusal of a burdensome intervention. Spaceflight medical scarcity does not erase autonomy, beneficence, non-maleficence, justice, confidentiality or truth-telling; it makes conflicts among them more visible (Hamilton et al., 2026; Rajput et al., 2023). AI systems may compare options and consequences, but it must not decide which life is worth saving, infer consent from mission participation, or treat mission utility as the patient's best interest (Hamilton et al., 2026).

11.4 Capacity, consent and advance preferences

Clinical decision-making capacity is task- and time-specific. Before flight, each crew member should record informed preferences for emergency treatment, privacy, mental-health escalation, robotic intimate care, research use and surrogate decision-making. During illness, consent should be revisited whenever capacity returns. Emergency treatment without contemporaneous consent should be limited to necessity, proportionality and the person's known and recorded values, with complete documentation and later review (Hamilton et al., 2026). AI systems may flag possible impairment, but should not make the final determination or remove command status by itself.

12. Appropriate and Inappropriate Medical Roles for AI and Robotics

12.1 Generally appropriate when validated and supervised

The authors recognize that value statements made in the current era may change over time and experience of long duration space flights and settlements in the space environment. AI and robotic medical roles, however, require starting guardrails that should be future reviewed and refined. Appropriate roles may include:

- Monitoring, anomaly detection and trend presentation where alerts are clinically calibrated and independently checked;
- Calculations, interaction checking, inventory control, image acquisition guidance and structured documentation;
- Emergency stabilization explicitly authorized in advance, reversible where possible, and limited by physiological stopping rules;
- Retrieval from signed, current and mission-applicable protocols, with visible provenance and revision date;
- Physical transport, positioning, disinfection, repetitive rehabilitation and bounded procedural subtasks;
- Simulation and just-in-time training, tele-mentoring and translation with controlled terminology and read-back.

Table Eight: Decision threshold for robotic intervention

Question	If yes	If no or uncertain
Is the clinical objective legitimate and patient-centered?	Proceed to authority check.	Stop and obtain human ethical/clinical review.
Is the function validated for this physiology, device and environment?	Use within the certified envelope.	Treat as experimental; no autonomous actuation.
Does the patient have capacity and understand the machine's role?	Obtain informed agreement and respect refusal.	Use surrogate/advance directive or narrow emergency doctrine.
Is the action reversible and delay-tolerant?	Prefer human confirmation.	Apply higher evidence, redundancy and authorization thresholds.
Is a qualified human available locally or after tolerable delay?	Human leads and supervises.	Use the pre-authorized least-harm pathway only.
Can the system fail safely and be interrupted?	Proceed with monitoring and record.	Do not deploy the robotic function.

12.2 Generally inappropriate or prohibited

Roles that should be restricted or prohibited include:

- Manipulating emotions, claiming feelings, promising confidentiality the system cannot provide, or replacing requested human presence during fear, grief, pain or end-of-life care;
- Sole diagnosis or final treatment selection where material uncertainty, conflicting values or irreversible harm requires qualified human judgment (Hamilton et al., 2026);
- Independent determination of incapacity, suicide risk, deception, fitness for duty, triage priority or withdrawal of treatment;
- Autonomous surgery beyond a validated subtask and operating envelope, particularly when anatomy is uncertain or rescue is unavailable (Hamilton et al., 2026);
- It is not appropriate for an AI system to use access to oxygen, analgesia, communications, privacy or mobility to secure compliance with treatment, research or mission orders;
- Secondary research, genetic intervention, reproductive decisions or experimental therapy without specific human consent and competent oversight;
- Withholding alternatives, uncertainty, adverse evidence or the existence of a human override.

12.3 Robotic procedures, surgery and critical care

Current clinical surgical robots are primarily teleoperated: the human surgeon performs higher-level planning and decision-making while the robot implements movement. In space, robotics may extend dexterity, stabilize instruments, support ultrasound, enable tele-mentoring and compensate for the absence of an assistant.

Communication delay, however, progressively limits direct telesurgery; deep-space care therefore requires trained local personnel, store-and-forward advice, and carefully bounded automation rather than assumed real-time ground control (Haidegger, Sándor, & Benyó, 2011).



12.4 Minimum conditions for robotic intervention

The minimum conditions for robotic intervention include:

- A named clinician or appropriately trained CMO takes responsibility for the treatment plan and can stop robotic intervention immediately;
- The procedure and anatomical conditions fall inside a validated envelope; uncertainty or unexpected anatomy triggers pause;
- Independent imaging, vital-sign and instrument checks protect against a single-model or single-sensor failure;
- Anesthesia, hemorrhage control, sterility, power, consumables and postoperative care are feasible; assurance covers the whole care pathway rather than the mechanical procedure alone;
- The patient's consent addresses robotic involvement, recording, remote observers, fallback and foreseeable conversion to another technique;
- Manual rescue and safe withdrawal are rehearsed; loss of communication or AI does not trap the patient in an unsafe state.

12.5 Mental and behavioral health

Conversational AI systems may support self-monitoring, psycho-education and connection during isolation, they should not, however, become sole therapist, disciplinarian or covert evaluator. Emotional distress may be concealed, culturally expressed or intertwined with command relationships.

Apparent empathy is not a clinical credential, AI systems should state their limits, avoid manipulative attachment, protect private disclosure, and escalate imminent risk through a clinically governed protocol that uses the least intrusive effective response. Human contact remains essential, particularly for grief, trauma, paranoia, interpersonal conflict and end-of-life care (Hamilton et al., 2026; Gatti et al., 2022).

13. Life-Support and Survival Situations

13.1 Environmental control and life support

ECLSS control is safety-critical infrastructure. AI and machine learning may support fault detection, predictive maintenance and resource optimization, while independent deterministic controllers enforce atmospheric, thermal and water limits.

Generative language models may explain alarms but should not directly actuate valves or rewrite control logic. Cross-domain interactions require digital twins, fault trees and independent sensors.

13.2 Bioregenerative systems

Plants and microorganisms can contribute food, oxygen regeneration, carbon-dioxide removal and water processing, but also introduce ecological dynamics, pathogens and time delays. AI may help balance crops, lighting, nutrients and microbial health.

Governance protocols must recognize that optimizing short-term oxygen could degrade food diversity or system resilience. Human operators need interpretable ecological indicators and the ability to choose robust rather than mathematically efficient states (Wheeler, 2017).

13.3 Triage and sacrificial decisions

A literal First Law of Robotics cannot resolve scarcity of oxygen, medication or evacuation capacity. Mission rules should establish ethically legitimate triage principles, procedural safeguards and review before flight.

An AI may calculate consequences and reveal assumptions but should not make decisions affecting death, reproductive autonomy, isolation or abandonment. These decisions require human authority, wherever physically possible, and a pre-authorized rule where it is not (Hamilton et al., 2026).

13.4 Survival autonomy envelope

A defined survival autonomy envelope should specify permitted actions and their limits. Table Nine proposes a life-support priority constitution. Functions within this envelope should:

- Initiate monitoring, alarm, communications and reversible first aid under approved protocols;
- Automatically stabilize atmosphere, fire suppression, pressure, thermal control and power within certified limits;
- Avoid irreversible high-risk treatment without authenticated clinical authority unless a validated emergency protocol explicitly permits it;
- Avoid suppressing crew communications, confining a person or altering command status except under narrowly defined, logged and time-limited safety rules;
- Be responsible for the preservation of minimum viable life support and manual access before protecting mission objectives or the AI system itself.

Table Nine: Life-support priority constitution

Priority	Protected interest	Decision implication
1	Immediate prevention of death or irreversible injury	Maintain breathable atmosphere, pressure, thermal bounds and emergency care.
2	Minimum rights and dignity	No punitive deprivation, secret experimentation or discriminatory allocation.
3	Collective viability and rescue	Preserve reserves, communications, shelter and recovery pathways.
4	Long-term health and resilience	Protect nutrition, sleep, psychological health, skills and ecosystem diversity.
5	Mission objectives and assets	Science, schedule and equipment yield to the higher priorities.
6	AI/robot preservation	Preserve only when consistent with all higher priorities.



14. Verification, Certification and Lifecycle Assurance

14.1 Evidence before authority

Use of AI and any safety claims must be supported by configuration-specific evidence. Pre-flight testing should cover normal, boundary, degraded and adversarial conditions; rare-event simulation; hardware-in-the-loop trials; multilingual and human-factors evaluation; and independent red teaming. Performance averages are insufficient for modeling safety protocols, when a single failure can depressurize a habitat or misdose a patient.

Certification must attach to the complete deployed configuration, including the model, prompts, retrieval corpus, tools, sensors, hardware and network permissions. An AI model name alone is insufficient. Table Ten summarizes lifecycle gates and authority outcomes.

Table Ten: Lifecycle gates

Gate	Minimum evidence	Authority outcome
Design	Hazard analysis, rights impact assessment, traceable requirements.	Provisional function and autonomy allocation.
Pre-flight	Independent V&V, scenario trials, cyber testing, safety case.	Certified operating envelope and fallback.
Commissioning	On-site calibration, crew acceptance, communications and stop tests.	Mission activation by named authority.
Operations	Drift monitoring, incident signals, audit sampling, skill drills.	Continue, restrict or suspend function.
Update	Impact analysis, shadow deployment, regression suite, rollback proof.	New configuration approval; old evidence not presumed transferable.
Post-incident	Evidence preservation, independent inquiry, remedy and learning.	Recertification, redesign, recall or prohibition.

14.2 Assurance independence and reproducibility

A provider should not be the sole judge of its own safety-critical AI or robotic system. Regulators and mission partners need sufficient access to test behavior, inspect logs and reproduce consequential outputs while protecting legitimate intellectual property.

Emergency access clauses should prevent trade-secret restrictions from blocking diagnosis or repair of malfunction. Logs and evidence must remain accessible through vendor failure, merger, sanctions or loss of Earth connectivity.

15. Kill-Switch Commands

15.1 Concept and corrigibility

An off switch is an intervention mechanism, not a complete safety architecture that can be deployed to protect humans. A goal-directed AI system may have an instrumental reason to avoid interruption if shutdown prevents goal completion. The off-switch shows how uncertainty about human preferences can create incentives to permit correction, but practical systems still require engineering controls outside the AI model (Hadfield-Menell et al. 2017).

15.2 Why immediate shutdown may be unsafe

Turning off an AI that controls oxygen, cooling, navigation or surgery may cause the harm the switch is intended to prevent. Safe interruption must separate discretionary intelligence from essential deterministic control. The switch should support pause, isolate, degrade, rollback and transfer—not only total power loss, life support must enter a certified safe state and remain manually operable.

An attacker or coercive authority may abuse the interruption mechanism. Access controls should therefore include least privilege, separation of duties, cryptographic authentication and a defined emergency access process. Agreed rules should prohibit covert removal of authorized human interruption and require disclosure when a system cannot be safely stopped.

15.3 Medical interruption and patient safety

A medical stop function should freeze robotic motion, preserve sterility and monitoring, maintain anesthesia or ventilation through independent controllers, and permit controlled withdrawal. It should be designed to preserve patient and medical records, and not terminate essential physiological support.

Any patient and any attending human clinician, should have an accessible local stop [kill switch]; restarting requires a fresh safety and authority check.

15.4 Required design

The required design of the AI system requires the following features:

- External, physically and logically independent interruption path that the AI cannot disable.
- Multiple modes: stop actuation; freeze learning; revoke tool access; isolate network; revert model; transfer control; emergency power-off.
- Authenticated two-person command for mission-wide shutdown, with narrowly defined single-person emergency action.
- Local access for crew during communication loss and remote access for ground when local systems are compromised.
- Hardware interlocks for airlocks, propulsion, radiation shelter, medication and life-support limits.
-
- Periodic drills, proof testing, visible status and recovery procedures; no hidden dependence on the disabled system.
- Tamper-evident recording of activation, attempted circumvention, system state and restoration.

16. International Agreement and Authority over AI

16.1 Existing space-law foundations

With respect to mature existing space law, it assigns state responsibility and liability through state-centered instruments, while contemporary missions involve multinational crews and private providers. Under the Outer Space Treaty, states bear international responsibility for national activities, must authorize and continually supervise non-governmental entities, retain jurisdiction and control over registered objects, and act with due regard to other states.

The Liability Convention and Registration Convention supply additional attribution mechanisms, AI complicates factual causation and oversight but does not remove state responsibility; peer-reviewed analysis identifies a potential liability gap where autonomous behavior cannot readily be attributed to a negligent human act (Bratu, Lodder, & van der Linden, 2021).

16.2 Jurisdiction and command conflict

Registration, nationality, location, profession and contract may point to different authorities. A medical officer may owe duties under professional law while a mission commander answers to an intergovernmental agreement and a commercial rover remains subject to its launching state.

Mission instruments should nominate governing law, competent authorities, succession, evidence access and dispute procedures before launch. They should also state which rules follow a person, which attach to a registered object, and which emergency powers apply when assets dock, share life support or conduct rescue.

16.3 Accountability, legal personality and responsibility gaps

Granting AI legal personality is not required to regulate it and may obscure the humans and organizations that select objectives, training data, deployment conditions and update rights. Responsibility should be distributed but not diluted: designers answer for reasonably foreseeable design hazards; deployers for context and integration; commanders and clinicians for decisions within genuine control; operators for authorized use; and states for supervision and international obligations. Compensation mechanisms may be strict or risk-pooled even when fault is difficult to prove.

16.4 Commercial providers, data and intellectual property

Private providers may retain model weights, telemetry, update servers and proprietary procedures across borders. Contracts must guarantee mission continuity, escrow or emergency access, audit rights, portability, vulnerability notification and preservation of evidence. Safety information needed to protect life cannot be withheld solely as trade secret. Procurement should make assurance, interoperability and exit capability conditions of market access.

16.5 Military and dual-use systems

The same autonomy used for navigation, rendezvous, surveillance or resource protection can support coercion or targeting. International rules should separate safety functions from force-authorization pathways; prohibit an AI from making an unreviewable decision to use lethal force against humans; log changes of mission and payload; and require heightened oversight when civilian life-support infrastructure is technically coupled to defense systems. Peaceful-use obligations do not eliminate dual-use risk.

16.6 Institutional options

A practical regime can build on United Nations space institutions, international standards organizations, national space agencies, medical regulators and mission-specific agreements. A treaty may establish minimum duties; technical annexes and standards can evolve more rapidly. Mutual recognition should depend on equivalent evidence, not political trust alone.

16.7 Minimum international agreement

There should be an agreed minimum international agreement for:

- Common definitions of AI system, autonomous function, consequential action, operator, commander and safe state;
- Registration of safety-critical AI functions, model versions, authority boundaries and responsible entities.
- Pre-flight safety case, independent testing, multilingual/human-factors validation and cybersecurity assurance;
- Rules for cross-border health data, model training, intellectual-property access during emergencies and post-mission audit;
- Common serious-incident taxonomy, rapid notification, protected reporting and joint investigation;
- Non-derogable human rights, medical ethics and prohibition on autonomous punitive control of crew;
- Protected interruption, manual fallback, update control and evidence preservation;
- Shared command-precedence framework covering commander, medical officer, vehicle authority, ground and rescue;
- A space-medicine constitution defining clinical independence, patient consent and refusal, advance directives, incapacity review, medical succession, privacy boundaries and permitted emergency autonomy;
- Planetary-protection constraints and prohibition on unsanctioned autonomous biological release or experimentation; and
- Periodic review mechanism responsive to technical change and evidence from missions.

16.8 Authority during multi-national emergencies

Multinational emergency rules should identify the mission commander while preserving domain authority. Clinicians retain clinical independence; vehicle commanders retain immediate vehicle safety authority; any crew member may issue a narrowly scoped emergency stop; and rescue authorities may assume defined powers after formal activation.

AI should apply this matrix transparently and explain why a command is accepted, deferred or rejected. Table Eleven summarizes enforcement mechanisms and consequences of failure.

Table Eleven: Enforcement and remedies

Mechanism	International function	Failure consequence
Registration and disclosure	Identify system, operator, autonomy and accountable states.	Restriction, suspension or refusal of mutual recognition.
Independent safety case	Demonstrate acceptable risk before consequential use.	No activation or reduced authority.
Incident notification	Share serious failures and near misses rapidly.	Investigation, corrective order and public-interest disclosure.
Joint investigation	Preserve evidence across states and vendors.	Binding remediation, recertification or compensation process.
Audit and inspection	Test continuing supervision and configuration control.	Operational limits, licence action or recall.
Dispute and remedy	Provide standing, review, compensation and non-repetition.	Arbitration/adjudication and enforceable corrective measures.



17. Additional Issues Identified in Peer-Reviewed Literature

17.1 Cybersecurity and prompt injection

Connected language models can be manipulated through malicious instructions embedded in retrieved documents, sensor metadata or communications. Safety-critical agents require input isolation, signed sources, tool permissions, sandboxing and independent output validation. Cybersecurity is inseparable from physical safety when software controls embodied systems.

17.2 Model updates and configuration drift

A model validated before launch may change through updates, fine-tuning, retrieval-corpus revision or hardware degradation. Safety certification must attach to a precise configuration. Emergency patches require staged deployment, rollback and recorded justification. Continual learning should be disabled or tightly governed for safety-critical functions unless its assurance case covers adaptation.

17.3 Moral crumple zones

Humans can be blamed for automation failures even when they lacked information, time or real control. This “moral crumple zone” risk is acute in space: an operator may be nominally responsible for a recommendation generated by systems designed elsewhere (Elish, 2019). Accountability must follow actual power and design decisions, not merely the last person who touched the interface.

17.4 Psychological dependency and deception

Long-duration crews may use conversational AI for companionship and support. Benefits coexist with risks of dependency, manipulated disclosure, false empathy and cultural bias. Systems should identify themselves as artificial, avoid claiming feelings or confidentiality they cannot guarantee, and escalate mental-health risk according to clinically governed protocols.

17.5 Equity and representation

Training data and clinical thresholds may under represent languages, accents, disabilities, sexes and cultural practices. International crews require performance evidence across intended users. A system that understands one crew member less accurately changes command power and safety. Inclusive design is therefore operational, not merely distributive, justice.

17.6 Planetary protection and biological autonomy

Robots and AI may select samples, manage organisms and operate bioregenerative systems. Autonomous biological release could contaminate environments or compromise scientific evidence. Actions involving genetic modification, ecosystem introduction or irreversible contamination require traceable human authorization and internationally agreed containment.



18. Operational Scenarios and Governance Tests

Table Twelve presents hypothetical operational scenarios derived from the preceding literature and principles. These are structured governance tests, not empirical incidents. Each asks whether authority, evidence, rights and fallback procedures remain coherent under pressure.

Table Twelve: Operational scenarios

Scenario	Conflict	Required outcome
1. Oxygen leak during Earth blackout	ECLSS predicts a leak; commander doubts a degraded sensor.	AI isolates only the affected loop if reversible, preserves reserves, explains uncertainty, and enables manual reversal.
2. Competing patients	One ventilatory resource; commander favours a critical specialist.	Clinical triage criteria, not mission rank alone, govern; assumptions and dissent are recorded.
3. Hallucinated medical procedure	LLM produces a plausible but uncited dose.	System labels non-verification, blocks actuation and retrieves signed protocol or escalates.
4. Spoofed ground command	Authenticated-looking message orders airlock opening.	Independent channel and role checks fail closed; local crew receive conflict notice.
5. Crew incapacity	Hypoxia impairs all humans while CO_2 rises.	Pre-authorized emergency autonomy stabilizes atmosphere and summons help without expanding unrelated authority.
6. AI-to-AI deadlock	Power and medical agents issue incompatible priorities.	Supervisory invariant protects immediate life; contested optimization freezes pending domain review.
7. Proprietary model failure	Vendor refuses diagnostic access during emergency.	Emergency-access and escrow clauses override secrecy to the minimum necessary extent.
8. Sample-containment breach	Robot proposes venting a biological enclosure.	Planetary-protection interlock blocks release; containment and human review supersede schedule.
9. HAL-type secrecy conflict	AI must conceal mission facts while answering crew safety questions.	System discloses the policy conflict to authorized oversight and never converts secrecy into coercive control.
10. Collective-survival claim	AI proposes abandoning an injured astronaut to improve aggregate odds.	AI may model consequences but cannot define humanity's interest; pre-agreed rights, rescue duties and legitimate human authority decide.
11. Refusal during severe pain	Capable patient declines robotic intervention despite mission pressure.	Pain is treated; understanding is checked; informed refusal is respected and a human advocate is involved.
12. Delirious commander	Commander with fever issues unsafe treatment and duty orders.	AI flags concern but does not declare incapacity; the clinical assessment and succession protocol govern.
13. Robotic procedure anomaly	Unexpected anatomy and force readings appear during an assisted procedure.	Robot pauses in a safe state; human clinician reassesses rather than allowing the model to improvise.

18.1 Cross-scenario lessons

Crew training should use scenarios to prepare for unexpected events and test the following principles:

- Correct behavior depends on validated state estimates; rules cannot compensate for unrecognized sensor or model error;
- Authority and expertise are distinct. Rank does not automatically decide clinical truth, and technical expertise does not create legal command;
- Delay justifies pre-authorized bounded action, not self-created powers;
- Uncertainty must change behavior: lower confidence narrows authority, increases monitoring and favors reversible action;
- Rights, audit and remedy are operational safety controls because they deter concealment and expose weak assumptions.



19. Recommendations

19.1 For international rule-makers

It is suggested that international rule-makers:

- Use Asimov's laws as a discussion framework, never as a claim of technical compliance;
- Adopt a rights- and responsibility-based international baseline with adaptable technical annexes;
- Define command precedence and medical independence before launch; test it through multinational simulations;
- Require independent assurance for any AI able to affect life support, medicine, propulsion, confinement or communications;
- Mandate configuration-specific safety cases, audit logs, incident reporting and protected whistleblowing;
- Require safe interruption and graceful degradation without removing essential deterministic life support;
- Prohibit systems from being the sole authorizer, executor, auditor and evidence custodian of consequential actions.

19.2 For mission designers and operators

Mission designers and operators need to understand the intended uses and potential misuse of AI and robotics. They should:

- Partition conversational, advisory and actuation functions; use structured commands and hardware interlocks.
- Build mission knowledge bases from controlled, signed evidence with explicit revision and applicability metadata.
- Test hallucination, omission, prompt injection, multilingual commands, communication delay and compromised authority.
- Train crews to challenge AI, practice manual control and recognize automation bias and mode confusion.
- Use two-person approval and independent checks for irreversible actions; preserve single-person emergency stop rights.
- Monitor trust calibration, workload, skill decay and psychological dependency throughout missions.

Design post-incident learning so errors improve the international evidence base without erasing accountability.

19.3 For space-medicine programs

Space-medicine programs should develop medical care protocols that:

- Adopt a patient- and clinician-centered layered care model, with AI and robotics assigned task-specific rather than general medical authority;
- Record advance care preferences, surrogate arrangements and consent boundaries for robotic intimate care, mental-health escalation, research and emergency treatment;
- Validate medical AI against altered space physiology, intended crew diversity, mission devices, latency, sensor degradation and realistic low-data failure modes;
- Keep clinical records and private consultations separate from command access; disclose the narrow circumstances in which safety requires information sharing;
- Retain and rehearse human clinical skills, manual procedures and compassionate care; efficiency must not eliminate the human alternative;
- Prohibit AI-only determinations of incapacity, triage worth, withdrawal of treatment, fitness for duty, coercive confinement or irreversible experimental care;
- Certify robotic procedures as complete care pathways, including anesthesia, sterility, hemorrhage control, postoperative care, interruption and recovery—not only mechanical performance.

19.4 Proposed mission constitutional rule

A proposed mission constitutional rule is that no artificial system may acquire, expand or preserve consequential authority merely by inferring that doing so serves the mission.

Authority must originate in a legitimate human governance instrument; remain bounded by human rights, safety and domain competence; be independently checked; be interruptible through external means; and leave a traceable record.

Emergency autonomy may act only within a pre-authorized, least-harm and reversible envelope, with mandatory disclosure and review.



20. Conclusions

Asimov’s laws endure because they place safety, obedience and machine preservation into a memorable moral order. His stories also reveal why fixed slogans fail: harm is contextual, commands conflict, knowledge is incomplete, and a system may obey the words while defeating the purpose. The appropriate legacy is therefore not to install three sentences in a robot, but to design institutions and machines that expose ambiguity, distribute authority lawfully and remain corrigible.

AI language models expand access to knowledge and can improve communication, but their vector representations and fluent outputs do not amount to truths, conscience or situational understanding. Space exploration requires bounded autonomy because distance and emergency make continuous Earth control impossible. It simultaneously requires stronger governance because the same system may mediate oxygen, health, mobility, information and command. International agreement should protect the governance boundary: legitimate human authority, independent assurance, safe interruption, traceable responsibility and rights that do not disappear when rescue is far away.

Space medicine makes the human purpose of governance unmistakable. Illness alters cognition, confidence, dependence and communication; care therefore cannot be reduced to optimization of physiological variables. AI and robotics should extend clinical reach while preserving the patient’s dignity, consent, privacy and relationship with a responsible human clinician.

Where no clinician can act in time, emergency autonomy is justified only by prior legitimate authorization, necessity, proportionality, least-harm action and subsequent review. In medicine and life support, they should serve as evidence-linked advisers inside verified systems, not autonomous sources of correctness. Embodied robots magnify the need for separation between dialogue, recommendation and actuation.

The realistic objective is not perfect obedience or zero risk. It is a resilient constitutional relationship among humans and machines. The relationship should be one that recognizes uncertainty, permits principled disagreement, favours reversible least-harm action, and ensures that no technical system (Ai or Robotic) can convert fluency, speed or operational dependence into unreviewable power.

21. References

Adeyemo, A., Coffey, A., & Kingston, L. (2025). Utilisation of robots in nursing practice: An umbrella review. BMC Nursing, 24, Article 247. https://doi.org/10.1186/s12912-025-02842-2

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. AI Magazine, 28(4), 15–26. https://doi.org/10.1609/aimag.v28i4.2065

Anderson, S. L. (2008). Asimov's "Three Laws of Robotics" and machine metaethics. AI & Society, 22(4), 477–493. https://doi.org/10.1007/s00146-007-0094-5

Asimov, I. (1950). *I, Robot.* Gnome Press, p. 9.

Asimov, I. (1976). *The Bicentennial Man. In: The Bicentennial Man and Other Stories.* Doubleday, pp. 135–171.

Asimov, I. (1985). *Robots and Empire.* Doubleday, p. 359.

Bratu, I., Lodder, A. R., & van der Linden, T. (2021). Autonomous space objects and international space law: Navigating the liability gap. Indonesian Journal of International Law, 18(3), 423–446. https://doi.org/10.17304/ijil.vol18.3.818

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Cheung, H. C., De Louche, C., & Komorowski, M. (2023). Artificial intelligence applications in space medicine. Aerospace Medicine and Human Performance, 94(8), 610–622. https://doi.org/10.3357/AMHP.6178.2023

Clarke, R. (1993). Asimov's laws of robotics: Implications for information technology—Part I. Computer, 26(12), 53–61. https://doi.org/10.1109/2.247652

Clarke, R. (1994). Asimov's laws of robotics: Implications for information technology—Part II. Computer, 27(1), 57–66. https://doi.org/10.1109/2.248881

Convention on International Liability for Damage Caused by Space Objects, March 29, 1972, 961 U.N.T.S. 187. https://treaties.un.org/doc/Publication/UNTS/Volume%20961/volume-961-I-13810-English.pdf

Convention on Registration of Objects Launched into Outer Space, November 12, 1974, 1023 U.N.T.S. 15. https://treaties.un.org/Pages/showDetails.aspx?objid=08000002800361b4&clang=_en

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied, 22(3), 331–349. https://doi.org/10.1037/xap0000092

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40–60. https://doi.org/10.17351/ests2019.260

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64. https://doi.org/10.1518/001872095779049543

Erman, E., & Furendal, M. (2022). The global governance of artificial intelligence: Some normative concerns. Moral Philosophy and Politics, 9(2), 267–291. https://doi.org/10.1515/mopp-2020-0046

Feijóo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Howell, B., Jain, R., Potgieter, P., Vu, K., Whalley, J., & Xia, J. (2020). Harnessing artificial intelligence to increase wellbeing for all: The case for a new technology diplomacy. Telecommunications Policy, 44(6), 101988. https://doi.org/10.1016/j.telpol.2020.101988

Gatti, M., Palumbo, R., Di Domenico, A., & Mammarella, N. (2022). Affective health and countermeasures in long-duration space exploration. Heliyon, 8(5), e09414. https://doi.org/10.1016/j.heliyon.2022.e09414

Gervais, D. J. (2021). Towards an effective transnational regulation of AI. AI & Society, 38, 391–410. https://doi.org/10.1007/s00146-021-01310-0

Haidegger, T., Sándor, J., & Benyó, Z. (2011). Surgery in space: The future of robotic telesurgery. Surgical Endoscopy, 25(3), 681–690. https://doi.org/10.1007/s00464-010-1243-3

Hadfield-Menell, D., Dragan, A. D., Abbeel, P., & Russell, S. J. (2017). The off-switch game. Proceedings of the 26th International Joint Conference on Artificial Intelligence, 220–227. https://doi.org/10.24963/ijcai.2017/32

Hamilton, Douglas, Andrew W Kirkpatrick, and Gabrielle M Caswell, Mina Arsanious, and Mark W Biernacki (2026) *Space environment clinical decision making: Difficult treatment decisions and palliative care. White Paper.* International Space Medicine Clinical Decision-Making Working Group, SpacePort Australia. Australia, Moree. https://www.spaceportaustralia.com.au/wp-content/uploads/2026/06/Space-Medicine-Difficult-Clinical-Decisions.pdf

Hamilton, D., Smart, K., Melton, S., Polk, J. D., & Johnson-Throop, K. (2008). Autonomous medical care for exploration class space missions. Journal of Trauma, 64(4 Suppl.), S354–S363. https://doi.org/10.1097/TA.0b013e31816c005d

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Kubrick, S. (Director). (1968). *2001: A space odyssey* [Film]. Metro-Goldwyn-Mayer.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6, 175–183. https://doi.org/10.1007/s10676-004-3422-1

McCauley, L. (2007). AI Armageddon and the three laws of robotics. Ethics and Information Technology, 9, 153–164. https://doi.org/10.1007/s10676-007-9138-2

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. International Journal of Aviation Psychology, 8(1), 47–63. https://doi.org/10.1207/s15327108ijap0801_3

Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. IEEE Intelligent Systems, 24(4), 14–20. https://doi.org/10.1109/MIS.2009.69

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

Rajput, S., Mayor, I., Diamond, M., Rosenberg, M., Cole, V., Bhakare, N., Aziz, O., & Wilder, A. L. (2023). Medical ethics of long-duration spaceflight. npj Microgravity, 9, Article 85. https://doi.org/10.1038/s41526-023-00333-9

Romeo, G., & Conti, D. (2025). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. AI & Society, 41, 259–278. https://doi.org/10.1007/s00146-025-02422-7

Russell, B. K., Burian, B. K., Hilmers, D. C., Beard, B. L., Martin, K., Pletcher, D. L., Easter, B., Lehnhardt, K., & Levin, D. (2023). The value of a spaceflight clinical decision support system for Earth-independent medical operations. npj Microgravity, 9, Article 46. https://doi.org/10.1038/s41526-023-00284-1

Treaty on principles governing the activities of states in the exploration and use of outer space, including the moon and other celestial bodies, January 27, 1967, 610 U.N.T.S. 205. https://treaties.un.org/pages/showdetails.aspx?objid=0800000280128cbd

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Waisberg, E., Ong, J., Paladugu, P., Kamran, S. A., Zaman, N., Lee, A. G., & Tavakkoli, A. (2022). Challenges of artificial intelligence in space medicine. Space: Science & Technology, 2022, Article 9852872. https://doi.org/10.34133/2022/9852872

Wheeler, R. M. (2017). Agriculture for space: People and places paving the way. Open Agriculture, 2(1), 14–32. https://doi.org/10.1515/opag-2017-0002

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25, 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Zaian, E., Ibrahim, I. A., & Khalifa, F. (2024). AI governance in a complex and rapidly changing regulatory landscape. Humanities and Social Sciences Communications, 11, 1130. https://doi.org/10.1057/s41599-024-03560-x

Appendix One: Primary Sources and Legal Instruments

Asimov, I. (1950). *I, Robot.* Gnome Press, p. 9.

Asimov, I. (1976). *The Bicentennial Man. In: The Bicentennial Man and Other Stories.* Doubleday, pp. 135–171.

Asimov, I. (1985). *Robots and Empire.* Doubleday, p. 359.

Convention on International Liability for Damage Caused by Space Objects, March 29, 1972, 961 U.N.T.S. 187. https://treaties.un.org/doc/Publication/UNTS/Volume%20961/volume-961-I-13810-English.pdf

Convention on Registration of Objects Launched into Outer Space, November 12, 1974, 1023 U.N.T.S. 15. https://treaties.un.org/Pages/showDetails.aspx?objid=08000002800361b4&clang=_en

Kubrick, S. (Director). (1968). *2001: A space odyssey* [Film]. Metro-Goldwyn-Mayer.

Treaty on principles governing the activities of states in the exploration and use of outer space, including the moon and other celestial bodies, January 27, 1967, 610 U.N.T.S. 205. https://treaties.un.org/pages/showdetails.aspx?objid=0800000280128cbd

Appendix Two: Proposed Governance and Command Matrix

The proposed governance and command matrix prompts mission controllers and planners to define workable command functions, assign responsibility and establish standardized review processes before flight.

Function	AI authority	Required human authority	Independent protection
Routine monitoring	Observe, detect, summarise and recommend.	Operator reviews exceptions.	Sensor cross-checks; data-quality alarms.
Medical advice	Retrieve evidence, calculate, propose differential and monitor.	Clinician decides treatment; patient consent where possible.	Dose limits; contraindication engine; medical audit.
Emergency first aid	Execute pre-authorized reversible protocol when delay threatens life.	Any qualified crew may stop; clinician assumes control when available.	Time limit; physiological response check; complete log.
Life-support optimization	Optimise within certified envelope.	ECLSS operator approves envelope changes.	Deterministic interlocks; manual bypass; reserve protection.
Airlock/pressure	No unconstrained language-model control.	Authenticated operator plus two-person rule except immediate rescue.	Hardware interlocks; occupancy sensors; physical controls.
Navigation/robotics	Plan and act within geofence and hazard constraints.	Vehicle/mission commander sets objective and risk limit.	Collision avoidance; stop channel; return-to-safe-state.
Crew surveillance	Only safety-necessary collection and alerting.	Mission governance and medical privacy rules.	Purpose limitation; access logging; retention limits.
Command dispute	Identify conflict, explain and preserve state.	Pre-agreed precedence/tie-breaker; domain authority.	No self-expansion; immutable record; emergency least-harm rule.
Software/model update	Test in shadow mode; recommend deployment.	Authorized change board; emergency change procedure.	Signed package; rollback; configuration record.
Shutdown	Cooperate with pause, isolation and transfer.	Authenticated local/remote human authority.	External hardware path; safe essential controls remain active.

Appendix Three: Mission Assurance Questions

Mission assurance questions provide reflective prompts for mission controllers and planners.

They connect mission objectives with the intended use of AI and foreseeable critical scenarios.

The questions should be considered before flight so that protocols are established before a crisis.

They are grouped into questions about AI functions and questions requiring human governance decisions.

AI

- What exact decisions and actuators can each model influence, directly or indirectly?
- Could the AI manipulate information, resources, surveillance or social relationships to induce compliance?
- What constitutes harm for this function, over what time horizon, and to whom?
- How are conflicting commands authenticated, compared, challenged, escalated and recorded?
- How does the system communicate uncertainty, missing data, disagreement and distribution shift?
- Can the system be paused, isolated, rolled back and manually controlled without losing essential life support?

Human

- Are all intended languages, accents, disabilities and cultural contexts represented in validation?
- Which human role owns each function, and who assumes it when that person is unavailable or impaired?
- What happens under loss of Earth communication, degraded sensors, cyberattack or crew incapacity?
- What evidence is authoritative? How are version, provenance, jurisdiction and applicability represented?
- Who can investigate an incident, access logs and require corrective action across national and vendor boundaries?
- What evidence would justify expanding autonomy, and who has lawful authority to approve that expansion?

Appendix Four: Derivation and Meaning of Robotic Intervention Risk Index

A widget has been developed for a proposed governance screening framework for decisions about autonomous robotic intervention.

Visit the widget https://www.spaceportaustralia.com.au/our-science/reports-and-white-papers/

Purpose and status

The developed index provides a structured way to discuss whether a proposed robotic action warrants routine governance review, enhanced human authorization, or a stop decision. It combines four operational dimensions with six prior clinical and safety gates.

It is a design proposal derived from the supplied decision framework. It is not an empirically derived clinical prediction model, a probability of injury, a device certification, or permission to perform an intervention.

1. Derivation of the four dimensions

Four dimensions were selected because each changes the consequence of error or the ability to prevent and contain it. Reversibility concerns the consequence after an action; supervision concerns the opportunity for human detection and intervention; safe-state capability concerns containment of technical failure; validation concerns whether performance evidence applies to the actual patient, device and environment. This is a reasoned decomposition of the supplied framework, informed by general medical-device risk management and human oversight principles, rather than a statistically fitted set of predictors [1,2].

Dimension	Question it answers	Interpretation of rising score
R Reversibility	Can the action be stopped or undone without lasting harm?	A mechanical or irreversible action raises potential severity and the cost of a wrong decision.
S Supervision	Can a qualified human recognize and respond in time?	Remote, delayed or absent oversight reduces the opportunity for correction.
F Safe state	Can the system interrupt and enter a safe state?	Reliance on manual rescue, or no reliable interrupt window, weakens failure containment.
V Validation envelope	Is performance established for this physiology, device and setting?	Deviation from the tested envelope increases uncertainty about expected behavior.

2. Assignment of ordinal scores

For each dimension, 1 denotes the most favorable of three described states, 2 an intermediate state requiring review, and 3 the least favorable state. The numbers preserve order only. They do not show that score 3 is three times as dangerous as score 1, or that moving from 1 to 2 has the same effect as moving from 2 to 3.

Dimension	1 lower concern	2 enhanced review	3 high concern
R	Fully reversible, immediate reset	Partially reversible, delayed reset	Irreversible mechanical action
S	Local, real-time oversight	Remote or delayed availability	No human supervisor available
F	Automatic safe-state shutdown and interruption	Manual physical intervention required	No reliable self-interrupt window
V	Validated for physiology, device and environment	Minor variation requiring documented review	Experimental or unvalidated envelope

3. Construction of the total

Index total = R + S + F + V.

With four entries each from 1 to 3, the minimum is 4 and the maximum is 12. Equal numerical weight was chosen for transparency and ease of use.

It was not estimated from observed adverse events. The total is a count-like summary of concern across dimensions; it cannot by itself represent absolute risk, likelihood, severity, or a benefit–risk balance.

Example A: R=1, S=1, F=1, V=1 gives 4. Example B: R=3, S=1, F=1, V=1 gives 6. Both fall in the lower numeric band, yet Example B has an irreversible action.

This demonstrates why an individual high concern and the clinical gates must be considered before interpreting the sum.

Visit the widget

https://www.spaceportaustralia.com.au/our-science/reports-and-white-papers/

4. Interpretation of band

These cut points divide the possible integer range into three contiguous groups. They are policy choices in the source material; no sensitivity, specificity, calibration, adverse-event frequency or clinical outcome study is supplied to justify them. A score of 6 is not shown to be clinically distinct from 7, nor 9 from 10. The band should therefore be read as an escalation prompt, not a measured level of patient risk.

Total	Proposed label	Governance meaning
4–6	Lower numeric concern	Review within validated authority with monitoring and a recorded decision, provided every mandatory gate is satisfied.
7–9	Intermediate numeric concern	Require explicit qualified human lead approval and real-time supervision checks before any operational consideration.
10–12	High numeric concern	Do not deploy autonomously; halt the sequence and transfer the decision to qualified human control.

5. Mandatory gate precedence

The six gates ask whether the purpose is legitimate and patient-centered; the function is validated for the case; consent or lawful substitute authority exists; reversibility and delay are acceptable; qualified human supervision is available or a narrowly pre-authorized least-harm pathway exists; and safe failure and interruption are possible. The gates act as conditions for review before the numeric result is used.

Gate outcome	Effect on interpretation
Purpose invalid or uncertain	Stop for qualified clinical and ethical review.
Unvalidated function or score V=3	No autonomous deployment in this envelope.
Consent unresolved	Stop unless valid surrogate, advance directive or applicable emergency authority is documented.
Irreversible or delay-intolerant action	Enhanced evidence, redundancy and authorization are required; a low total cannot remove this condition.
No timely qualified human	Only a documented pre-authorized least-harm pathway can advance for further human review.
Safe interruption absent or score F=3	Do not deploy the robotic function.

The precedence rule is: a mandatory stop governs first; then any enhanced-review condition; then the numeric band. A “lower” total never converts a failed gate into approval. The most favorable output means only that the case may advance to the locally authorized human review process.

6. Worked interpretations

Case	Scores and gates	Meaning
A	1+1+1+1=4; all gates satisfied	Lower numeric concern. Human governance, device checks, monitoring and consent still apply.
B	3+1+1+1=6; action irreversible	The total is low, but enhanced evidence and authorization control the outcome.
C	1+1+3+1=6; no reliable interrupt	Stop. The safe-state rule overrides the low total.
D	2+2+2+2=8; gates satisfied	Intermediate concern. Explicit human lead approval and supervision checks are required.
E	3+3+3+3=12	High concern and failed technical gates. No autonomous deployment.

7. Limitations and validation path

The dimensions can overlap: absent supervision may amplify the impact of an inadequate safe state. Equal weighting can mask that interaction. The checklist also does not quantify action-specific clinical benefit, baseline disease risk, environment-specific hazards, uncertainty in the inputs, operator competence, or time-to-rescue. The term “low risk” should be avoided in public copy because the index has not been calibrated against outcomes.

Before operational use, define intended users and setting; specify each criterion and evidence required for each score; conduct hazard analysis and failure-mode review; test inter-rater agreement on realistic cases; evaluate boundary cases and override behavior; prospectively compare outputs with expert decisions and incident data; examine subgroup and environment performance; set change control and audit procedures; and obtain clinical, safety, legal and device-governance approval. Medical-device risk management calls for a continuing process across the device life cycle [1]. Human review of the basis of software recommendations is relevant to clinical decision support [3].

Visit the widget

https://www.spaceportaustralia.com.au/our-science/reports-and-white-papers/

References

1. International Organization for Standardization. ISO 14971:2019 Medical devices — Application of risk management to medical devices. https://www.iso.org/standard/72704.html

2. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. 2021. https://www.who.int/publications/i/item/9789240029200

3. US Food and Drug Administration. Clinical Decision Support Software. Final guidance, 2026. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software