# Human-in/on-the-Loop AI: Enabling Human Controllability and Decision-Making in Spaceflight

Leonie Bensch[1*], Oliver Bensch[2], Tommy Nilsson[3], Joseph A. Paradiso[1], Scott Davidoff[4], Lancelot Blanchard[1], Pat Pataranutaporn[1], Melanie Weber[5], Nathan Kutz[6], Frank Flemmisch[7]

[1*]Responsive Environments Group, Massachusetts Institute of Technology, 75 Amherst Street, Cambridge, 02139, Massachusetts , USA.
[2]German Aerospace Center (DLR), Linder Höhe, Köln, 51147, North Rhine-Westphalia, Germany.
[3]European Space Agency (ESA), Linder Hoehe, Troisdorf, 51147, North Rhine-Westphalia, Germany.
[4]NASA Jet Propulsion Laboratory (JPL), 4800 Oak Grove Dr., Pasadena, 91109, California, USA.
[5]John A. Paulson School of Engineering and Applied Sciences, Harvard University, 150 Western Ave, Allston, 02134, Massachusetts , USA.
[6]Institut für Arbeitswissenschaft, RWTH Aachen, Templergraben 55, Aachen, 52062, North Rhine-Westphalia, Germany.

*Corresponding author(s). E-mail(s): lebensch@mit.edu;
Contributing authors: oliver.bensch@dlr.de;
tommy.nilsson@fit.fraunhofer.de; joep@media.mit.edu;
sd@scottdavidoff.com; lancelot@media.mit.edu; patpat@mit.edu;
mweber@seas.harvard.edu; kutz@uw.edu; f.flemisch@iaw.rwth-aachen.de;

**Abstract**

Establishing a long-term human presence on the Moon and beyond remains a central goal of space exploration, with programs like Artemis paving the way for Mars missions. These ambitious endeavors face challenges, including communication latency, bandwidth constraints, and autonomous operations in safety-critical environments. Future astronauts and mission control will rely heavily on Artificial Intelligence (AI) systems to process and integrate large, fragmented datasets

from spacecraft, rovers, and life support systems, making Human-Centered AI (HCAI) essential to ensuring these systems align with human goals.
This chapter explores the principles of HCAI, emphasizing Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL) frameworks. HITL systems prioritize continuous human engagement to adapt AI-driven processes in real-time, while HOTL systems focus on autonomous operations with periodic human oversight. Both paradigms maintain human accountability and agency, offering complementary approaches to managing dynamic, safety-critical tasks during long-duration spaceflight missions.
The chapter further delves into the backend and frontend components of HCAI HITL HOTL systems. The backend includes algorithms, data handling, and supervisory control mechanisms that underpin AI functionality. The frontend encompasses user interfaces, augmented reality (AR), virtual reality (VR), and explainable AI elements that enable intuitive and transparent human-AI collaboration. Case studies, including the CORE personal assistant for astronauts developed at ESA and interfaces for spacecraft science operations at NASA, illustrate practical applications of these principles. By addressing both theoretical and applied aspects, this chapter closes with challenges, future research topics, and a roadmap for designing HITL and HOTL systems that ensure robust human-AI collaboration in future (human) spaceflight, ensuring human oversight and adaptability.

**Keywords:** Human-Centered Artificial Intelligence (HCAI), Human-in-the-Loop (HITL), Human-on-the-Loop (HOTL), AI Transparency, AI Explainability, Spaceflight, Autonomous Systems, Mission Control, Augmented Reality (AR), Virtual Reality, Digital Twins, Astronaut Support Systems, Communication Latency

# 1 Introduction

Over fifty years after humanity's first steps on the Moon, the vision of a sustained human presence on other worlds continues to shape the direction of space exploration. Spearheaded by the international Artemis program, ambitious initiatives are currently underway aiming to establish permanent human bases on the Moon (Smith et al., 2020), which will eventually pave the way for crewed missions to Mars and further into our solar system.

Future astronauts embarking on these long-duration missions will rely on a wide array of advanced technical infrastructures, such as rovers, satellites, life support systems, and spacecraft sensors, to make critical decisions. The data from such systems is however, typically fragmented and collected in diverse formats, making it difficult for human users to interpret and process efficiently.

To make matters worse, with mission assets and participants potentially very large distances apart, communication delay will be an unavoidable aspect of these missions. For example, delays in communication between Earth and Mars can extend up to 24 minutes, depending on their relative alignment. Similarly, the Artemis program's focus on the lunar south pole will introduce recurring periods of radio blackout during future lunar expeditions. As a result, future astronauts will need to be able to make decisions

and operate autonomously, without continuous supervision or real-time support from mission control centers on Earth.

Ongoing advancements in Artificial Intelligence (AI) offer a promising avenue to tackle these challenges. By interpreting and synthesizing data from multiple sensors and datasets, AI systems could play a vital role in mediating key information to enhance astronauts' situational awareness and help guide their decisions. Indeed, people are already living in a hybrid presence, with reality brokered between an always-on and often-watched cellphone held at arm's length. This trend will only grow as wearable systems become commonplace, and eventually, implantable interfaces, now developed to alleviate physical and sensory disabilities, become more broadly useful. We already see AR sensory augmentation applied in fighter aircraft (Fetterolf, 2020), and it's only a matter of limited time before this technology finds its way into everyday life. Space operations, featuring integrated sensor networks, complex data processing, as well as what will probably be a distributed mission control architecture, will in this sense be at the forefront of advancing augmented perception and enhancing user decision-making (Von Hippel, 1986).

The attitudes behind automation and manned spaceflight have changed enormously over the past decades and will continue to evolve with the advent of modern AI and ever-advancing computation becoming flight-ready. Due to an emphasis on safety, NASA has historically avoided control or operational schemes for manned missions that are not completely predictable (Bergmann, Croopnick, Turkovich, & Work, 1979). These concerns have largely persisted into the era of AI-driven critical space systems such as autopilots. Although control research has long explored various techniques to put guardrails around such AI-actuated systems (Dong et al., 2024), they are by nature unpredictable. Science Fiction has lavishly explored this scenario for well over a half-century, perhaps best historically articulated in Stanley Kubrick's 2001 (Clarke, 2016), if not Isaac Asimov's Robot stories from decades before (Asimov, 2004). HAL from *2001: A Space Odyssey* has long been seen as a benchmark for AI (Stork, 1997), and the revolution in the performance of Large Language Models (LLMs) over the last few years has abruptly brought that vision well within sight. Now that this genie is out of the bottle, tremendous resources are devoted to AI technology, hence, we expect that AI systems will soon evolve to miniaturize and run on lower power (indeed, as this sentence is being written, such advances are rocking the stock market (Liu et al., 2024)), making space application practical. The benefits of AI-supervised spacecraft are myriad, from efficient operation and improved safety to being a stabilizing companion to the crew on long-duration missions. But although guardrails may keep the system largely in compliance, the complexity of AI is so vast that all corners of its 'intellect' can't be mapped, hence, risks ranging from control/system errors to social manipulation of the crew will be hard to rule out. As this is of immense terrestrial concern as well, work here on Earth to build dependable/trustworthy AI will find its way into space.

Ensuring that AI-driven actions align with human goals and ethical standards in the harsh conditions of space will demand extensive utilization of Human-Centered AI (HCAI) approaches, prioritizing people over algorithms in decision-making processes (Xu, Gao, & Dainoff, 2023). While advanced AI has attained a demonstrable ability

to enhance efficiency and manage routine tasks (see for instance the AI assistant CIMON on board the International Space Station (ISS), Figure 2), HCAI emphasizes the continued need for human oversight in complex, context-sensitive decisions that require empathy, creativity, and adaptability (Xu et al., 2023).

Such human oversight is typically implemented through *Human-on-the-Loop (HOTL)* approaches, where human operators serve as supervisors, ready to intervene when necessary to ensure the accuracy and safety of AI-driven processes. Effective HOTL mechanisms must be designed into systems from the ground up, with elements of human oversight integrated into their backend functionalities, including algorithms, data processing, and system architecture, to ensure the AI systems operate and evolve in alignment with human goals.

The unpredictable nature of human spaceflight requires maintaining a high degree of redundancy by avoiding overreliance on AI systems. This means that operators will need to be able to seamlessly transition from supervisory roles to active participation and, if necessary, assume full control of the AI system's operation. HOTL approaches will therefore need to be complemented with *Human-in-the-Loop (HITL)* solutions designed for continuous human engagement, enabling operators to actively shape AI-driven processes. For example, astronauts or mission controllers can provide ongoing input, refine the system's understanding, and guide the AI's actions to ensure that the technology aligns closely with mission objectives and the intent of the astronauts and/or mission controllers (see Figure 1). Previous work has demonstrated that such human-AI collaboration makes HITL particularly valuable in ambiguous and dynamically changing conditions (Gil, Albert, Fons, & Pelechano, 2020; Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos, Bobes-Bascarán, & Fernández-Leal, 2023; Wu et al., 2022).

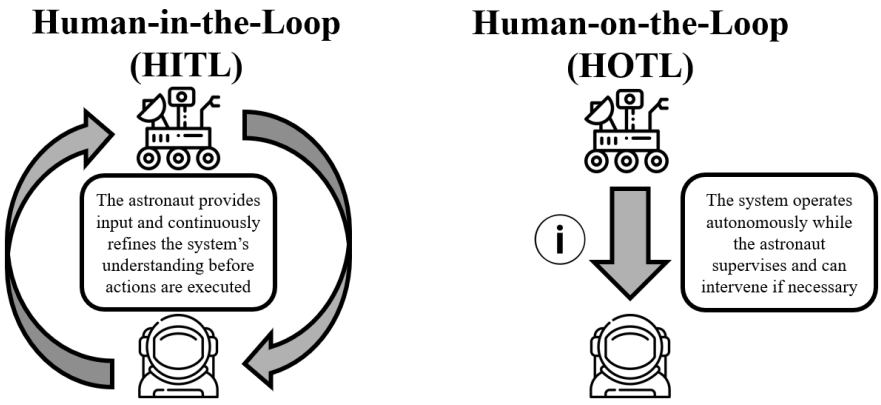


**Fig. 1** Comparison of Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL) interaction paradigms for astronaut–AI collaboration. In HITL, the astronaut actively provides input and continuously refines the system's understanding before actions are executed, ensuring direct human oversight at every step. In HOTL, the system operates autonomously while the astronaut supervises and intervenes only when necessary, allowing for increased autonomy with human oversight.

HITL systems will need to evolve to properly reflect these conditions, perhaps by showing predicted/probable outcomes that are informed and updated as information arrives (Bensch, Bensch, Nilsson, Saling, Bewer, et al., 2024). These ideas are also proposed to apply to social interaction. For instance, an astronaut at a large distance could interact in real-time with an LLM based around people socially close to him or her, which is updated as the remote participants also interact. This could be incarnated as a close synthetic "mutual friend" who knows the parties, or potentially a dynamic emulation of particular individuals. Technically, this could be achieved by generating steering vectors for a person based on their interactions with an LLM (Konen et al., 2024), which could be updated regularly.

Realizing such visions will require innovative front-end interfaces, such as virtual (VR) and augmented reality (AR) environments, infused with explainable AI mechanisms to improve human understanding and usability. Moving beyond traditional interaction paradigms, this could involve combining multiple sensory channels — not just visual and auditory feedback as in typical AR, but also, for example, distributed haptic stimulation (those of us who doubted the ubiquity and usefulness of haptics stand corrected by our smartwatches). Ongoing research (e.g., (Chin, 2024)) searches for the best combined sensory conduits and representations for information that wants to come into our awareness, and this will be important in future space missions (again, science fiction has already anticipated some of this, dating at least to 1987 in the Galactic Center novels from Gregory Benford (Benford, 1987)).

In exploring the issues above, this chapter offers the reader with a detailed overview of HCAI in the space exploration domain, focusing on HOTL and HITL principles, their design, and implementation from both frontend and backend perspectives. Through a set of case studies, we demonstrate the vital role of these concepts in supporting robust human-AI collaboration, ensuring seamless and meaningful interactions between the AI and mission control or astronauts. With that we aim to offer a basic roadmap for developing HITL and HOTL systems that promote human-centered values under the uniquely challenging operational conditions of human spaceflight. The chapter concludes with an outlook on future research.

# 2 Related Work

## 2.1 Trustworthy, Transparent, and Explainable AI

AI's ability to autonomously process and interpret large datasets makes it uniquely positioned to compensate for the frequent lack of direct supervision that future astronaut crews will face when managing key systems during long-duration missions. Rapid analysis and detection of anomalies in large sets of sensor data could prove vital for monitoring key infrastructure, such as life support systems. Through timely identification of potential malfunctions, AI could thus, for instance, alert crews and recommend contextually appropriate solutions, thereby significantly enhancing both the safety and operational efficiency of future missions.

In the safety-critical environment of spaceflight, endowing such processes with human-centered AI principles will be particularly important. Above all, for crews to make informed and effective decisions, future AI systems will need to exhibit a

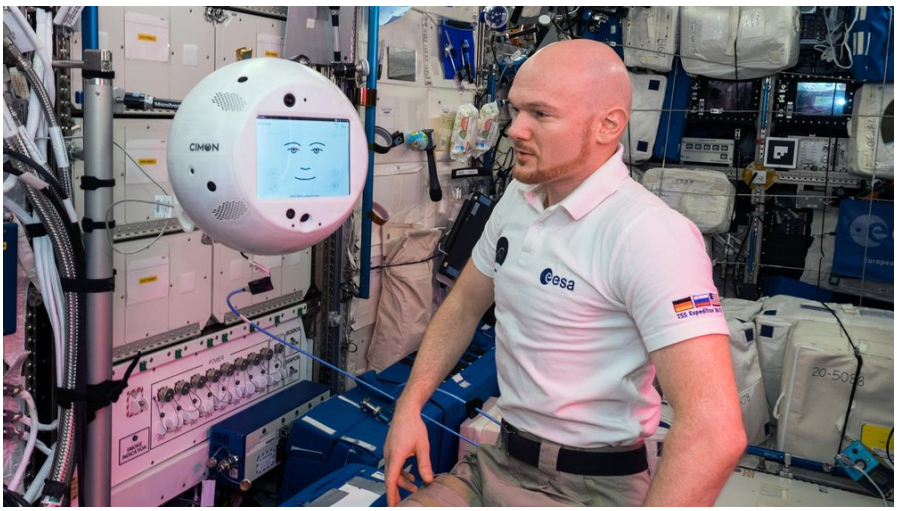


**Fig. 2** Astronaut Alexander Gerst interacts with CIMON (Crew Interactive Mobile Companion) aboard the ISS. CIMON is an AI-powered robotic assistant built to support astronauts with procedural tasks by providing hands-free information retrieval and voice-controlled interaction (Schmitz et al., 2020). Credit: DLR/T. Bourry/ESA

degree of transparency by articulating the reasoning behind their recommendations and actions. Achieving such explainability of AI behavior will be key to developing efficient partnerships between future astronauts and their AI systems.

The importance of AI transparency is already widely recognized in other domains. Schmidt et al. (2020), for instance, have described it as a necessary prerequisite for building trust in AI systems (Schmidt, Biessmann, & Teubner, 2020). Toy (2023) went a step further, arguing that AI ought to be more "comprehensible, defensible, and rationally grounded", which would ideally lead to system actions being rendered in a manner that is "reminiscent of human cognition". Accordingly, techniques such as feature importance analysis, decision trees, and rule-based systems have already been successfully employed to help users understand the logic behind AI decisions (Pillai, 2024). Similarly, visual tools such as graphs, charts, and heatmaps have been found to simplify complex datasets and provide intuitive insights into AI behavior (De, Saraf, Mishra, & Tripathy, 2024).

Such work might easily lead us to believe that maximizing the transparency of AI systems by fully revealing their opaque inner workings ought to be a central design goal. Nevertheless, studies have shown that this approach is not only technically challenging but can also backfire by overwhelming users with information. Indeed, as demonstrated for example by, Rodden, Fischer, Pantidi, Bachour, and Moran (2013) even comparatively simple systems, such as AI tools for monitoring and regulation of household energy consumption, involve intricate layers of both technical and social complexities, which, if fully exposed, would risk overwhelming users with excessive volumes of information, making efficient and informed decision-making impossible. In the context of spaceflight, where timely and accurate decision-making is critical, such information overload could prove to have devastating consequences.

In light of this, a key strand in research concerned with human-AI interaction has shifted towards the idea that it is *explainability*, rather than transparency, that ought to be the primary requirement for future AI systems. While the exact definitions may vary among researchers, explainability fundamentally refers to the need to render AI actions and decisions understandable. Where transparency aims to expose a system's internal processes and mechanisms, explainability focuses on articulating the outcomes of the system's behavior in ways that are actionable and meaningful to its end users (Phillips et al., 2021).

A significant milestone in the shift towards explainability came in 2016 when the American Defense Advanced Research Projects Agency (DARPA) launched the Explainable Artificial Intelligence (XAI) program, based on the premise that making the learned models and decisions of AI systems intelligible to the end users represents a key step towards trustworthy autonomous systems: *"Dramatic success in machine learning has led to a torrent of AI applications. Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own. However, the effectiveness of these systems is limited by the machine's current inability to explain its decisions to human users. [...] Explainable AI — especially explainable machine learning — will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners"* (Gunning & Aha, 2019).

Following DARPA's call, a growing number of experts from non-defense-related domains are currently in the process of advancing our understanding of how best to articulate and explain AI system behavior to human users. Perhaps the most immediate challenge faced by such efforts is the absence of any clear consensus regarding the exact set of criteria a system needs to meet in order to be considered explainable. Indeed, the contemporary research landscape is abuzz with diverse perspectives on *what* information an AI system should convey to its users, as well as *how* this information should be conveyed.

Much of this discourse is centered around everyday human-machine interactions, such as smart home AI solutions. In such conventional domains, explainability is primarily valued for enhancing usability by making AI behavior more predictable and interpretable, and for promoting ethical standards by ensuring AI systems and their providers are accountable.

In safety-critical domains, such as human spaceflight, the importance of explainability is perhaps even more far-reaching. A key principle in aerospace engineering is *redundancy*, the practice of minimizing potential failure points by avoiding overdependence on any single component. Making AI systems explainable will in this sense provide a layer of redundancy by making it easier for astronauts to manually intervene if a system malfunctions to maintain control and ensure mission continuity. In this way, explainability serves as a vital mechanism for keeping astronauts both on the loop and, eventually, in the loop, enhancing overall mission safety.

The benefits of strong user engagement extend beyond just redundancy, as elaborated by several studies. Whitworth (2005), for instance, stressed the need for at least partial human control over AI systems, pointing out that computers typically have a poor understanding of context, which may result in faulty actions if left unattended.

Other studies corroborate this, showing that human-robot teamwork has a strong tendency to yield better results than AI systems alone (Johnson, Bradshaw, Hoffman, Feltovich, & Woods, 2014). R. Yang and Newman (2012) argued that coordinating such human-AI partnerships often requires robust and innovative interfaces, whilst You et al. advocated the use of rich visualizations (You, Jin, Wang, Fang, & Luo, 2016). In the following section, we shall explore such prospective interfaces in greater detail.

## 2.2 Immersive Interfaces for AI Explainability

As researchers explore methods to enhance AI explainability, user interface design and implementation have become increasingly critical in space exploration, where any friction in human-AI interaction could compromise crew safety and mission success. Spaceflight presents unique challenges, including large amounts of mission-critical data, strict time constraints or potentially crippling health conditions, such as space motion sickness. Additionally, astronauts often face physical limitations, such as restricted hand or body mobility or a limited field of view due to wearing a spacesuit during extravehicular activities or the inability to interact with interfaces during high-G phases, such as launch and reentry. Microgravity further complicates interaction with conventional input devices, requiring alternative control methods. Cognitive load is another key concern, as astronauts must process multiple streams of information while operating in high-risk environments. For mission control, communication delays, especially in deep-space missions, demand interfaces that enable asynchronous decision-making and predictive modeling to anticipate astronaut needs before real-time responses are possible. These constraints call for interface solutions that optimize accessibility, situational awareness, and usability under extreme conditions.

In this context, traditional desktop interfaces remain essential due to their familiarity and flexibility. These interfaces typically include dashboards, control panels, and command consoles that display key information while allowing direct interaction with AI systems. Real-time data visualization through graphs, charts, and dynamic updates ensures that astronauts and mission control personnel have access to the most up-to-date information, which supports their decision-making in a familiar and reliable way. However, relying solely on desktop interfaces may restrict situational awareness in complex, rapidly evolving scenarios, such as when multitasking or navigating environments, where spatial understanding becomes essential.

In contrast, AR interfaces enable a spatial representation of critical information within the user's view of the actual environment (Milgram & Kishino, 1994) and could therefore effectively facilitate the spatial understanding of relevant mission data. By blending or overlapping real and artificial environmental data, AR ensures that the operator's gaze remains focused on the task. This enables hands-free interaction and reduces cognitive workload by keeping users focused on tasks while seamlessly supporting them (Milgram & Kishino, 1994). Such use of AR thus helps ensure that the human operator remains central to the decision-making process. Moreover, AR's capacity to transform complex datasets, such as telemetry or trajectories, into spatially intuitive visualizations makes them easier to comprehend and act upon by their

human users. Several studies have demonstrated AR's effectiveness in enhancing situational awareness by enabling users to track the past, present, and future states of operations (Bolstad, Endsley, & Hill, 2003). In the context of spaceflight, AR has been proposed to improve astronaut situational awareness during both extra- and intravehicular activities (EVAs/IVAs), such as navigation, task execution, and hazard awareness (Becker, Nilsson, Demedeiros, & Rometsch, 2023; Braly, Nuernberger, & Kim, 2019; Markov-Vetter & Staadt, 2013; Rometsch et al., 2022). By projecting procedural and safety-critical information directly into an astronaut's field of view, AR could reduce errors and improve task efficiency, especially when wearing a restrictive spacesuit. For mission control, AR offers additional opportunities to enhance human-AI collaboration by, for example, supporting operators with virtual spatial overviews or tables that intuitively interpret mission data, reducing the need to switch between multiple interface windows or screens. In this regard, a virtual mission control table can show the mission progress and tasks in a spatial and intuitive manner, similar to a "bird view", while the AI's predictions and suggestions are displayed in a visual manner on top of the 3D data. These systems have the potential to streamline decision-making processes, particularly in complex scenarios where quick, accurate interpretation of data is critical (Anandapadmanaban, 2020). By enabling humans to remain "in the loop," AR aligns with HCAI principles, ensuring that astronauts and mission operators retain oversight and agency in critical decision-making processes.

Another promising technology is the use of VR, which immerses users in a fully virtual environment, offering a controlled setting for training, simulation, and mission planning. In HCAI applications for space missions, VR could be utilized in various ways: VR environments could enable astronauts to practice difficult maneuvers, respond to anomalies, and familiarize themselves with mission protocols in a safe, controlled setting (Garcia, Schlueter, & Paddock, 2020). Such hands-on training enhances readiness and reduces the likelihood of errors during actual mission procedures. VR could prove particularly relevant in long-duration missions where astronauts need to be trained in new procedures to handle unforeseen situations, or old (hazardous) ones that require regular training and are expensive or dangerous to practice (Ennis et al., 2021).

On the other hand, by creating digital twins of mission environments, VR could allow mission control to experience conditions from the astronauts' perspective, improving understanding and oversight of mission activities. VR also enables real-time collaboration between astronauts and mission control, allowing both parties to interact within a shared virtual space, enhancing communication and coordination. In this regard, research has shown that VR environments can, for instance, significantly enhance teleoperation of remote technologies, such as rovers or robotic arms, by providing a virtual 3D space that mirrors the physical environment, which improves precision and responsiveness in remote space operations, especially when combined with haptic cues (Weber, Hulin, & Schiffer, 2022):

Haptic feedback systems complement visual and auditory information by providing tactile sensations, enriching the user's interaction with AI systems (Zhu et al., 2020). In HCAI applications, haptic interfaces offer several advantages. Subtle vibrations or forces can notify users of critical events or changes in system status, ensuring

important information is conveyed even in noisy or visually cluttered environments, where more information would not be understandable to the astronaut or mission controller (Wickens, 2002). Haptic feedback can also improve the precision of control inputs, such as the manipulation of virtual objects in VR or AR environments, by providing tactile confirmation of actions or haptic feedback indicating how virtual objects feel, enhancing the realism of the virtual environment experience.

The most effective HCAI interfaces likely integrate multiple interface technologies to leverage their respective strengths. For example, a mission control interface might combine classic desktop elements for data monitoring with AR overlays for spatial data visualization and haptic feedback for critical alerts. Similarly, astronauts on a mission might use VR for training and simulation while relying on AR for real-time task support and haptic feedback for tactile interactions.

It is important to note that AR and VR technologies can likewise be utilized in conjunction to support training and real-world space missions. Depending on the availability of tools or equipment and the specific environment, astronauts could, for example, choose between a complete EVA simulation in VR or an augmented real-world environment using AR technology. Such AR features could then seamlessly transition between real-world experiences and training scenarios by: (1) integrating the AR features used during the mission into the EVA training simulations conducted in a VR environment; (2) employing these features within the AR mission support tool as a training platform that offers extensive support instructions; and (3) gradually reducing the level of support by AI in training scenarios through skill estimation, in both AR and VR environments, to a minimum over time. This approach could ensure that astronauts are thoroughly familiar with relevant space environments, the tasks, and the use of AR features intended to assist them during the actual IVA or EVA. By providing only the necessary and useful AR cues, it could furthermore avoid information overload, ensuring the appropriate level of support during the mission. Seamlessly switching between AR and VR technologies in this manner thus enables the training of future EVA tasks regardless of location.

# 3 HOTL: Human-on-the-Loop AI

## 3.1 Backend Methodologies

When talking about human supervision in AI systems, we generally presuppose that these systems are applied to novel and never-before-seen types of input data, for which we expect them to make an “educated” decision based on their training. This decision, or the process thereof, is observed by a human agent, with the option to intervene if needed. As such, this scenario assumes that these AI systems are equipped with some kind of agency that another human agent can control and inhibit during operation. On the other hand, in the era of big data, the sheer quantity of information often overwhelms traditional analytical approaches, making AI-driven analysis essential for extracting meaningful predictions. For example, with close to exponential growth in research articles approaches like Literature-Based Discovery (Bensch & Hecking, 2021) aim to accelerate or completely automate research (Bensch, Hecking, & Kutz, 2024) with only keeping researches up-to-date with an option to adjust the research process

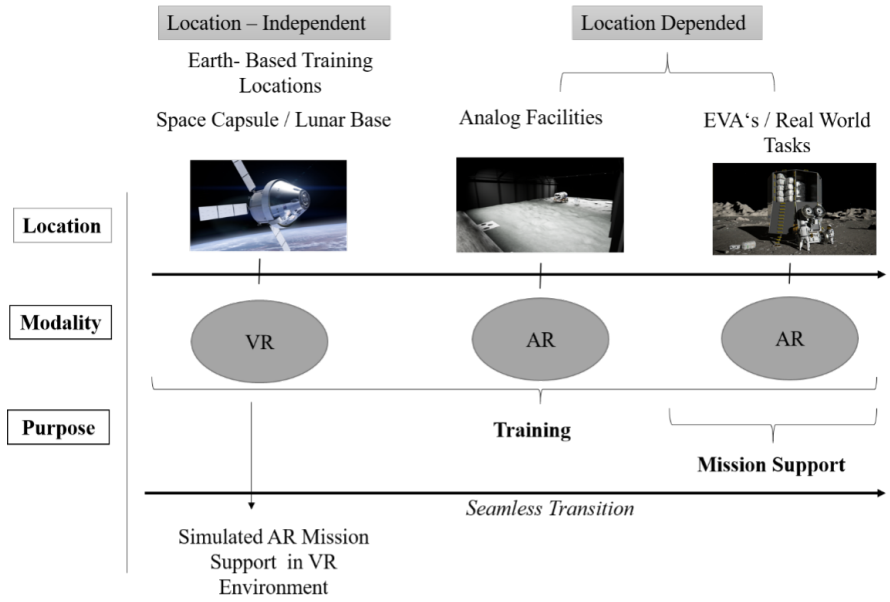


**Fig. 3** Seamless transition between EVA training in VR and AR for real-world mission support. Depending on the operational context, astronauts can engage in full-scale EVA simulations within VR environments or use AR overlays to augment real-world environments with AI-driven spatial information. This integration allows for seamless integration of training and operational support, enabling astronauts to familiarize themselves with mission procedures, optimize situational awareness, and gradually reduce AI assistance as their expertise increases. The approach ensures adaptive and context-aware human-AI collaboration, minimizing cognitive overload while maximizing effectiveness in both training and real-mission scenarios.

if needed. When considering AI systems for spaceflight, both scenarios must be taken into account. For instance, realistic rover data may not be available on Earth for training, and a small spacecraft crew may need to operate complex systems while simultaneously monitoring live telemetry data.

A common framework for HOTL supervision is Supervisory Human Control (SHC), which is generally applied to automated systems in aviation, self-driving cars, and robotics (Sheridan, 2021). In the context of HOTL machine learning, SHC can be applied to regulate the agency of AI models through interactive systems (Heer, 2019), by enabling human operators to oversee, intervene, and guide the decision-making process of AI models in real-time. This includes tools like visual data wrangling and natural language translations to create real-time feedback loops that allow users to understand and influence the AI's behavior. Such systems help ensure accountability, improving transparency, and enhance user trust by providing meaningful control over automated processes. In the realm of human spaceflight, these HOTL supervisory frameworks become even more critical. For example, autonomous spacecraft systems and planetary rovers must operate reliably in environments where rapid response to unexpected anomalies, such as system malfunctions, radiation-induced errors, or sudden terrain changes, is crucial. By integrating SHC, astronauts and ground control can

monitor and adjust AI-driven operations in real time, ensuring that decisions related to navigation, system diagnostics, and emergency interventions are executed safely. Advanced visual and natural language interfaces enable this immediate oversight, a capability that is crucial during long-duration missions when physical access to the remote system is limited.

Reinforcement Learning from Human Feedback (RLHF) (described in more detail in the HITL section 4.1), can also be used in the context of HOTL machine learning to integrate human oversight within models. In this scenario, a reward model is gradually trained by requiring human interaction for annotation only at specific times. In human spaceflight operations, this approach is particularly valuable because it minimizes the need for continuous intervention while ensuring that the AI's decision-making remains aligned with mission-critical objectives. Therefore, the HOTL AI system can become more resilient and adaptive, based on targeted annotations by astronauts and mission control when anomalies or unexpected behaviors occur.

## 3.2 Frontend Methodologies

While HITL solutions seek to enable robust human-AI interaction, the sheer complexity and scale of AI systems in future long-duration space missions will frequently make direct human participation in AI-driven processes difficult. Instead, a significant portion of data processing and actuation will have to be delegated to AI, operating independently of direct crew involvement. As a result, much of the computational activity involved in these missions runs the risk of becoming opaque to users, reducing their immediate awareness and understanding of critical system operations.

To prevent such a situation, HOTL interfaces are primarily concerned with making AI behaviour legible in a manner that would enable efficient human supervision without overwhelming users with excessive information. In theory, this means an AI would autonomously handle routine tasks while providing users with clear, concise insights that allow for rapid comprehension and timely user intervention if necessary.

Much of the existing research on AI explainability has focused on developing methods that enable users to retrospectively verify the nature and rationale behind specific system actions, often through post-hoc explanations. Weitzner et al., for instance, have previously used the term "information accountability" to describe the user's ability to check "whether the policies that govern data manipulations and inferences were in fact adhered to" (Weitzner et al., 2006).

Nevertheless, given the dynamic nature of space missions, it is likely that a viable HOTL interfaces would need to make system behaviour intelligible in real-time, rather than in hindsight. This notion is in line with the Technomethodology framework formulated by Dourish and Button's, emphasizing the importance of a system's ability to give accounts of its own work in order to reveal more of the mechanisms unfolding under its hood (Dourish & Button, 1998). System designers are in this sense encouraged to develop solutions capable of giving accounts of some of the complexity implicit in their actions while balancing transparency with necessary abstraction to expose critical aspects of decision-making. Dourish and Button conclude that such a translucent nature of a computer system is vital for achieving a sense of accountability, explaining that "just like social interactions in everyday life, we would like our interactions

with computer systems to be organized in terms of abstractions that are supported by their own unfolding, rather than opaque black boxes".

In practice, this means that AI systems ought to utilize a level of abstraction to produce representational models communicating relevant information to their users, while keeping much of the unnecessary complexity hidden. Any potentially important change of the system behaviour should then be accordingly reflected by a corresponding change in its representational model. It is this representational model, as Paul Dourish explains, that becomes a vehicle for conveying accounts to its users (Dourish, 1997). The role of an account, he argues, would be to allow the user to see not only what their system is doing, but also why it is being done and what is likely to be done next. This continuous offering of such accounts is then what lays the ground for an explainable AI system.

To achieve such AI explainability, immersive interfaces centered around AR and VR technologies could, for example, be used to present AI-generated insights in a spatially intuitive manner. AR can overlay real-time system updates into an astronaut's field of view, reducing the need to shift focus between different screens. For more complex tasks, AR-based mission overviews can allow operators to visually inspect AI-generated plans, making adjustments through direct interaction rather than relying on abstract data tables. These immersive interfaces could prove to play a key role in bridging the gap between automation and human oversight, allowing future astronauts to stay informed without being overloaded.

### 3.3 Case Study: Making Space AI Explainable (METIS)

This case study explores the implementation of the HOTL framework within the Mars Exploration Telemetry-Driven Information System (METIS), an AI-based Intelligent Personal Assistant (IPA) designed to enhance astronaut autonomy during long-duration space missions (Bensch, Bensch, Nilsson, Saling, Sadri, et al., 2024). The METIS assistant serves as an exemplary model for HOTL by autonomously managing complex tasks while ensuring human supervisors retain oversight to maintain safety and mission alignment.

METIS is an AI assistant developed to support astronauts by autonomously managing routine operations, monitoring spacecraft systems, detecting anomalies, and aiding in mission planning. Unlike current ground-based communication systems that astronauts on the International Space Station (ISS) heavily rely on for observing experiments, sensor data, and system statuses, METIS is designed to operate independently, thereby reducing the dependence on Earth-based mission control. This shift is crucial for long-duration missions to the Moon and Mars, where communication delays of up to 24 minutes make real-time support impractical. The architecture of METIS is based on a multi-agent system (MAS) comprising monitoring, reasoning, planning, and commanding agents. These agents work together to process and integrate vast amounts of fragmented data from spacecraft sensors, rovers, and satellites. METIS combines generative pre-trained transformers (GPTs) (Grattafiori et al., 2024), Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) and KGs (Pan et al., 2024).

By trying to replace the current ground communication infrastructure, METIS ensures that astronauts have immediate access to processed and integrated information, minimizing the need to wait for Earth-based responses. This autonomy is balanced with the HOTL framework, where METIS provides the astronaut with updates and confidence scores. The astronaut can either intervene manually or is asked to support the system if confidence levels fall and errors could occur.

METIS autonomously processes complex tasks using LLMs. However, relying on GPTs only can lead to hallucinations, where the AI generates plausible but incorrect information (Alkaissi & Mcfarlane, 2023), which is detrimental to trust and mission safety. Trust and reliability are the most important factors when developing AI systems for space, since incorrect information could lead to deadly situations. To address this, METIS integrates RAG (Lewis et al., 2021) on KGs (Pan et al., 2024). This combination allows METIS to access an easily updatable and structured database where every response can be traced back to grounded information within the KG, ensuring accuracy and traceability of the provided information. In contrast to single documents, KGs combine information in nodes and their relations, similar to a mind map.

Despite its autonomous capabilities, humans remain in supervisory roles with the option to intervene, ensuring that all actions align with safety protocols and mission objectives. METIS employs alert systems that notify astronauts of potential issues, allowing them to assess the situation using the information provided and decide whether to take autonomous action or seek further assistance. For instance, if METIS detects an anomaly that necessitates an adjustment to the mission timeline, it generates a proposed action plan and updates the astronaut, who can adjust the generated plan if needed.

Building and maintaining trust in METIS is essential, particularly in safety-critical environments like spaceflight. METIS employs several strategies to mitigate errors and enhance trust among human operators: 1) By integrating KGs and RAG, every AI-generated output can be traced back to its source data, providing explainability and allowing astronauts to verify the accuracy of the information. This traceability ensures that astronauts can understand the reasoning behind METIS's recommendations, improving confidence in its autonomous operations. 2) Additionally, METIS incorporates human correction mechanisms that allow astronauts to override or adjust its suggestions if discrepancies or errors are detected. For example, if an astronaut identifies a misinterpreted sensor allocation in the KG, they can provide feedback to METIS, which then updates the corresponding sections in the KG. Moreover, based on this KG and reasoning methodologies such as chain-of-thought (Wei et al., 2023) METIS features explainable AI elements that offer detailed explanations for its decisions and recommendations, further enhancing trust by making the AI's reasoning process transparent and explainable to the astronauts.

While METIS is designed for autonomous operation during deep space missions, it incorporates human accountability to ensure that all critical decisions align with mission objectives and safety protocols. METIS manages routine tasks and data processing autonomously, allowing astronauts to focus on more complex and mission-critical activities. However, for significant decisions, such as altering mission plans or addressing major system failures, METIS defers to human supervisors for final

approval. This decision-making hierarchy ensures that while METIS can efficiently handle everyday operations, the ultimate authority remains with the astronauts, who are responsible for making critical judgments. Additionally, the HOTL framework within METIS includes a detailed accountability mechanism that logs all AI decisions and human interventions. This comprehensive logging facilitates post-mission analysis and continuous improvement, ensuring accountability throughout the mission lifecycle. By balancing autonomous functionality with human oversight, METIS ensures that mission objectives are met while adhering to stringent safety standards.

### 3.4 Implications of HOTL for HCAI

HOTL systems embody the principles of HCAI by ensuring that automated functions are governed by flexible yet well-defined boundaries. Humans remain the final authority for major mission decisions or corrective actions, thus preserving moral and operational accountability. At the same time, the reduced frequency of interventions enables the AI to handle more tasks independently, which can be beneficial in resource-scarce environments such as space flight, facing challenges such as long communication delays or limited bandwidth. Careful interface design ensures that operators have sufficient explainability into the AI's intentions and performance without being overwhelmed by constant demands for attention. This overarching approach aligns with HCAI's vision of augmenting human capabilities through trustworthy automation, ensuring that efficiency does not come at the cost of human supervision or mission success.

## 4 HITL: Human-in-the-Loop AI

### 4.1 Backend Methodologies

In order to enable continuous human engagement within a HITL system, a methodology needs to be employed that can enable the integration of continuous human feedback within an AI system. This human engagement can either happen at training time, where astronauts and mission control collaboratively fine-tune the model using simulated scenarios and historical mission data, or inference time, or at both times sequentially, depending on the level of control that we require in the model. In the context of space flight, this could mean that mission-critical decisions—such as trajectory adjustments, anomaly detection, or system diagnostics—are continuously monitored and recalibrated by astronauts and ground control teams. For example, as a spacecraft navigates through unpredictable space environments, operators could provide immediate corrective feedback to the AI, adjusting navigation algorithms or safety thresholds in response to unexpected terrain or system irregularities.

It is important to note that human engagement is already present in most supervised machine learning systems since they rely on human-annotated data. As such, training a model using traditional Supervised Fine-Tuning (SFT) methods can already be seen as HITL compatible. Yet, issues arise from the long iteration cycles that result from using SFT. The training of large models can oftentimes be very long and costly, and, as such, it can be very tedious to perform several cycles of iteration on supervised

fine-tuning on human-annotated data based on when final results need to be obtained every time. When performing HITL AI, it is often desirable to develop shorter iteration cycles to accelerate the integration of human feedback inside the training of the model, and allow a more interactive refinement process that enables rapid adjustments based on human feedback. SFT also often depends on the availability of large amounts of annotated data, which might not be available in the space flight context. Another motivation for developing shorter iteration cycles is to avoid the high energy costs associated with the repeated training of large models, especially in the context of limited resources, which is a critical consideration given the resource constraints of space missions.

Several methods have been developed in recent years to enable a more continuous integration of human feedback within the training process of machine learning systems. For example, Low-Rank Adaptation (LoRA) has been proposed as an alternative to SFT to make fine-tuning faster and less dependent on large amounts of human-annotated data, making it an interesting candidate for HITL training scenarios (Hu et al., 2021). Instead of fine-tuning an entire model, LoRA only focuses on a small subset of the model's weights, also reducing training time and hardware requirements. LoRA's efficiency in fine-tuning only a subset of model parameters is particularly beneficial onboard spacecraft, where hardware limitations and the need for near-real-time processing require swift updates without extensive retraining.

Tackling the same objective, Reinforcement Learning from Human Feedback (RLHF) has been proposed to allow the distillation of important human-provided knowledge into large models, especially when SFT or LoRA is too tedious or intractable (Askell et al., 2021; OpenAI et al., 2024; Ouyang et al., 2022). For example, this methodology is particularly useful for integrating human feedback that cannot easily be provided in the form of annotated data, such as personal preferences or emotional reactions. In order to distill this knowledge into the model, RLHF first trains a reward model that aims to model the requested human-provided knowledge accurately (using PPO (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017) for example), before training the base model to optimize a loss provided by this reward model. Other similar methods have been proposed to distill this knowledge without the need to train an external reward model, but by simply using supervised learning to steer the model towards expected results. This method is used in algorithms such as DPO (Rafailov et al., 2023) and KTO (Ethayarajh, Xu, Muennighoff, Jurafsky, & Kiela, 2024). RLHF could, for instance, capture the unique expertise of astronauts by incorporating their intuitive assessments and situational judgments during anomalies or unexpected events—such as radiation-induced sensor irregularities or unanticipated terrain features on lunar or Martian surfaces.

Finally, Active Learning has been suggested as a semi-supervised learning approach in which a model can be trained on a large unannotated dataset, by requesting a human oracle to label only a select sample of it (Settles, 2009). This sample selection can be made through different techniques that use heuristics to a varying degree, ranging from *random exploration*, where samples are selected at random, allowing the model to explore a diverse range of data points without bias, to *near-decision boundary*, where samples are chosen based on their proximity to the model's decision boundary,

focusing on data points that the model is uncertain about or likely to misclassify. In the context of spaceflight, active learning becomes especially valuable because annotated data from space missions is often scarce and expensive to obtain. For instance, during a long-duration mission, a spacecraft's onboard systems may collect large amounts of unannotated sensor data under rapidly changing conditions.

HITL can also be performed at inference-time, where a decision is made collaboratively by the human and AI agents. By balancing the human and AI inputs during critical decision points during space missions, we can effectively reduce bias and improve both accuracy and fairness. Multiple methods have been proposed to adequately combine the decision-making process of both the human and AI agents. These methodologies mostly differ through their adaptivity, which refers to the ability of the system to integrate user feedback over time and improve its decision-making strategy. Rejection Learning (Chow, 1957; Cortes, DeSalvo, & Mohri, 2016), for instance, is a nonadaptive method that allows an AI system to reject the responsibility of making a prediction when its confidence is below a certain threshold. In this scenario, the human agent is asked to make a final decision in the place of the AI agent. Learning To Defer (L2D) (Keswani, Lease, & Kenthapadi, 2021; Madras, Pitassi, & Zemel, 2018; Mozannar & Sontag, 2020), on the other hand, is an adaptive framework that allows a model to directly learn the human user's expertise and delegate decisions dynamically, depending on the model's confidence and the specific characteristics of the task. L2D frameworks train the AI to assess when human intervention is necessary by modeling the comparative strengths of the human and AI in different decision contexts, ultimately improving collaboration and task outcomes. In the context of (human) space missions, such adaptive approaches ensure that critical decisions, ranging from system anomaly resolutions to emergency trajectory adjustments, are made by using both automated analysis and human judgment.

It is important to note that most of the aforementioned frameworks rely on mutual understanding between humans and AI agents to enable proper delegation of responsibilities. As a result, and as mentioned earlier, explainable AI (XAI) methods are required to create a common language between the model outputs and human judgment.

In this regard, for instance, the explainability of LLMs can be enhanced using methods like RAG. LLMs are trained to predict the most probable next word and may hallucinate when asked questions outside their training data. By employing RAG, generated answers can always be cross-verified with the original sources used during inference, ensuring transparency and reliability. Moreover, for more complex queries, RAG can be extended to operate on KGs, enabling the incorporation of intricate, interconnected knowledge from multiple documents or even multi-modal and live data sources. This extension supports the delivery of richer, context-aware responses while preserving traceability—a crucial factor for maintaining trust in HITL systems and HCAI in high-stakes environments, such as space flight.

## 4.2 Frontend Methodologies

HITL solutions must go beyond merely communicating AI behaviour in an explainable manner, as seen in HOTL approaches. They require continuous and adaptive user interaction, with humans actively participating in and shaping an AI system's behaviour throughout a given task (Gil et al., 2020; Mosqueira-Rey et al., 2023; Wu et al., 2022). Enabling this level of engagement demands intuitive interaction interfaces and a rich tapestry of relevant information to underpin efficient human-AI collaboration.

This problem was perhaps most famously explored by Eric Horvitz (Horvitz, 1999) in his landmark paper on the principles of *Mixed-Initiative* User Interfaces, which sought to address the tensions between developing new interface metaphors and tools on one hand and designing fully autonomous systems on the other. Instead of siding with either of these paradigms, Horvitz advocated finding *valuable synergies* between them. He went on to identify several factors which he argued are critical for an effective integration of autonomous systems with traditional user interfaces, including a system's ability to assess uncertainty - both in its own decisions and the user's goals - and the use of dialog to resolve such ambiguities.

Following this line of reasoning, HITL interfaces frequently incorporate dynamic confidence metrics, alerts, and decision rationales to enable continuous user validation or correction of AI actions. Given such continuous user input, HITL interfaces must offer streamlined means for annotation or other forms of user feedback, minimizing any friction in guiding or refining AI behavior in real-time. For instance, an interface could feature color-coded warning levels and intuitive toggles, allowing users to promptly override AI recommendations when sensor readings fall outside acceptable parameters.

Achieving HITL is however not merely a question of the AI system rendering its own actions explainable and amendable. In space operations, these systems are bound to frequently involve collaboration between dispersed user groups, including mission crews and control centers on Earth. HITL interfaces ought to reflect this by catalyzing collaboration between different users. This notion was first advanced by Patrick Brezillon, who argued that the explainability of interactive intelligent systems emerges, at least in part, from cooperative problem-solving between human users (Brézillon, 1994).

Previous research has shown that immersive interfaces can effectively visualize and facilitate interaction within such evolving socio-technical systems. AR is particularly well positioned to play a leading role in providing contextually relevant just-in-time assistance and clarifications. If an astronaut is conducting a multi-step procedure, AR overlays can, for example, highlight specific parts of the hardware in need of attention (Braly et al., 2019; Rometsch et al., 2022). Additionally, the system can prompt the operator for feedback whenever it detects a potentially risky condition in the astronaut's spatial surroundings.

By enabling such real-time interaction, AR overlays ensure that the operator's preferences and actions are continuously integrated into the AI's decision-making process and shared with other relevant users. This ongoing exchange helps refine subsequent interactions while improving user coordination (Bensch, Bensch, Nilsson, Saling, Bewer, et al., 2024).


**Fig. 4** Astronaut Megan McArthur uses a HoloLens HMD onboard ISS. Credits: NASA.

Similarly, fully immersive VR environments are uniquely suited for allowing users to experience spaces, distances, and scales within a digitally recreated real-world setting from a first-person perspective, enhancing their experiential understanding (Jetter et al., 2020). Tideman et al. also found that such VR environments are particularly suited for participatory visualizations of complex systems that involve many different stakeholders (Tideman, van der Voort, & van Houten, 2008).

Apart from real-time visualization of ongoing processes, VR is more commonly utilized for training purposes, to help acquaint users with a system's capabilities or to practice edge-case scenarios. In HITL training sessions, operators can, for instance, experiment with a system's boundaries, effectively instructing the AI on how to respond in ambiguous cases. By actively monitoring and correcting missteps in a realistic virtual environment, humans become integral teachers of the AI, ensuring the model evolves with each simulated mission (Garcia et al., 2020; Weber et al., 2022).

Additionally, for tasks that demand real-time awareness, haptic signals may prompt users to examine AI decisions or confirm critical actions. Rapid vibrations on a glove or console, for example, might indicate a low-confidence AI assessment, prompting the user to step in. By modulating the intensity or pattern of the haptic signal, the system can convey different urgency levels (Wickens, 2002; Zhu et al., 2020).

## 4.3 Case Study: CORE Personal Assistant

This case study explores the implementation of the HITL framework within the Checklist Organizer for Research and Exploration (CORE), an AI-based Intelligent Personal Assistant (IPA) designed to continuously assist astronauts during long-duration space missions (Bensch, Bensch, Nilsson, Saling, Bewer, et al., 2024).

In contrast to METIS, which operates autonomously by processing data, updating system statuses and mission parameters, and keeping astronauts informed, requiring interaction only in critical situations, CORE is designed as a truly interactive assistant that continuously engages with astronauts to provide guidance and support.

CORE is an assistant developed to support astronauts by continuously assisting them during experiments and routine tasks that are clearly defined in structured procedures. Each procedure is divided into several steps and is usually available for astronauts on a tablet, computer or as a printed version (see figure 5). That requires the astronaut to control the computer and shift focus from an experiment to the device or handbook and back. Especially during Extravehicular activities (EVAs), where astronauts' range of motion, field of view and mission time are limited, operation of an additional device is challenging.

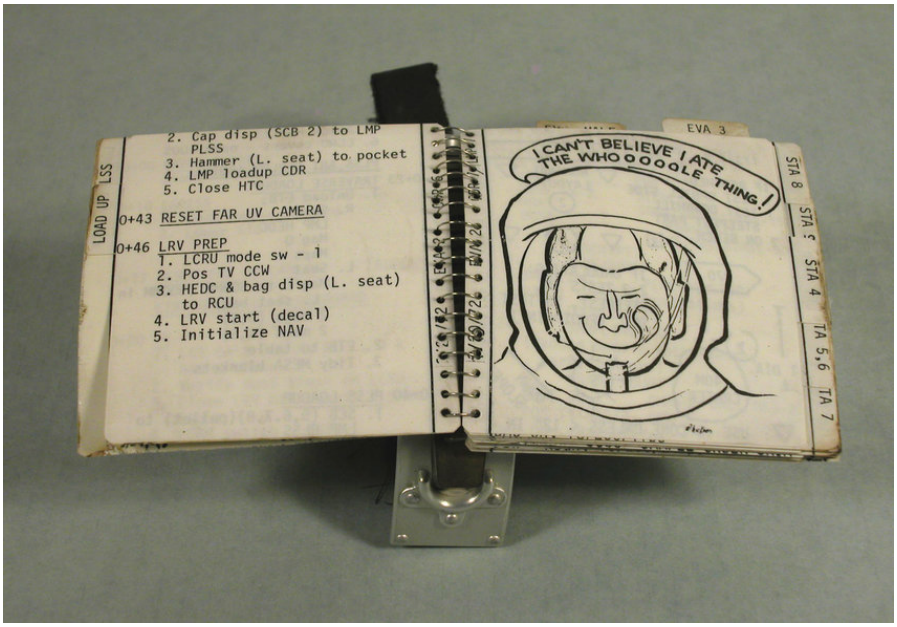


**Fig. 5** A small wrist-mounted notebook used by an astronaut during the Apollo 16 mission in April 1972. Designed for quick reference during extravehicular activities (EVAs), the notebook contained critical procedural checklists, navigation cues, and mission notes, ensuring astronauts could efficiently access key information while operating in the lunar environment. Credits: NASA.

Like METIS, CORE is based on a GPT (Grattafiori et al., 2024) for a seamless interaction in natural language and integrates RAG on KG. However, the CORE KG contains not only textual information, such as, procedure steps, components and manuals, but also multi-modal information such as images, video or live telemetry data (see figure 6). When integrated into an KG, live telemetry data of an airlock pressure sensor can be represented by a relationship connected to the manual information of that airlock and thus be directly provided to the astronaut when asking questions about that airlock (see another example and the RAG on KG workflow in figure 7).

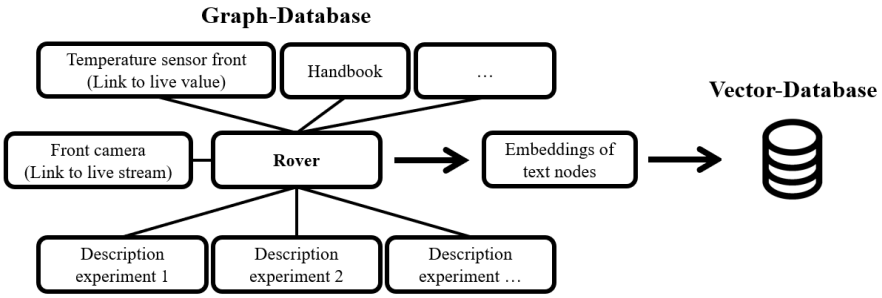


**Fig. 6** Example of a subgraph representing live telemetry data within a KG. This structure enables the integration of real-time mission data with procedural knowledge, allowing astronauts and mission control to access dynamic system states, correlate telemetry with relevant documentation, and receive AI-assisted insights for decision-making in space operations.

Since CORE is designed to provide a seamless interaction the system integrates speech-recognition (Radford et al., 2022) and synthesis to enable interaction in natural language. CORE's interaction style can be adjusted based on the situation and personal preference of the astronauts. CORE can interact in a friendly manner, or as mentioned, simulating the conversation style from, for example, a friend on Earth in regular situations using steering vectors (Konen et al., 2024). Using such vectors, the interaction style can be shifted either gradually or instantly to a strict tone and short sentences in critical situations. Simultaneous changes of CORE's voice and the user-interface can further help to shift the astronauts´ focus on crucial parts in dangerous situations.

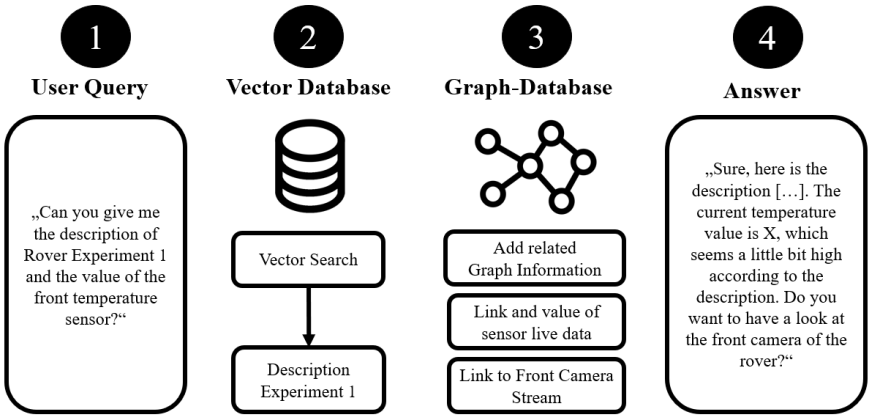


**Fig. 7** Example workflow of RAG applied to a KG for integrating live rover telemetry data. This approach enhances AI-driven decision support by dynamically retrieving relevant information, including live telemetry, from structured knowledge sources, enabling astronauts and mission operators to interact with mission data in a context-aware and explainable manner.

CORE utilizes AR (see figure 8) to display information in the form of a text, images, videos or graphs in front of the astronaut, making an additional device such as a tablet

or handbook obsolete. This hands-free access ensures that astronauts can continuously monitor vital information without interrupting their ongoing tasks. Furthermore, AR overlays can highlight specific areas of interest or concern, such as malfunctioning equipment or environmental hazards, enabling astronauts to respond promptly and effectively without requiring a translation of information from traditional displays to the real world, as the information is spatially and intuitively displayed within the real space. For instance, if CORE detects a potential leak in the life support system, it can highlight the affected module in the AR view, guiding the astronaut to inspect and address the issue in a fast and intuitive manner. By integrating AR, CORE not only enhances situational awareness but also allows for rapid intervention, improving overall mission safety and efficiency.

Upon request additional information can be visualized in the form of a text, such as procedure steps, or videos that explain the usage of a required tool. The HUD can be adjusted via voice, text, or drag-and-drop to present critical data at specific locations, ensuring that essential sections of the view remain unobstructed during experiments or in critical situations. Furthermore, AR elements do not have to remain static and can, for example, be overlaid on the corresponding experiment. For example, the temperature value of a specific sensor could always be represented at the sensor's location.

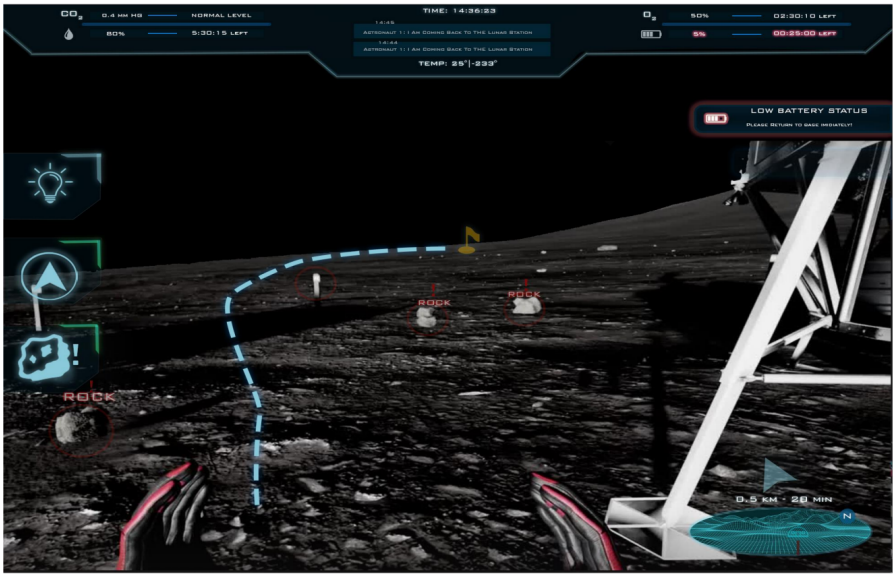


**Fig. 8** A CORE prototype illustrating the potential use of an AR interface displayed via HMD for astronauts during lunar surface operations. The AR system enables hands-free access to mission-critical information, such as procedural guidance, telemetry data, and interactive visual overlays, enhancing situational awareness and operational efficiency. By integrating AI-driven assistance with spatially anchored visualizations, CORE minimizes cognitive load and optimizes astronaut decision-making in extreme environments.

### 4.4 Case Study: Spacecraft Science Operations

Early work in spacecraft science operations autonomy focused on concepts of operations that would allow spacecraft to plan and execute activities opportunistically, based on previous observations and on the availability of on-board resources. Rather than relying on ground-in-the-loop operations, this approach allowed an onboard autonomous planning module to decide to perform observations of transient events of interest (e.g. a plume (Rossi et al., 2023; Thompson, Bunte, Castano, Chien, & Greeley, 2012), or possible biosignatures (Hand et al., 2022)) if sufficient thermal, power, and data volume resources are available. This work spans many of the broad technical areas of robotic operations, from Earth orbiters (Bernard et al., 1999; Troesch et al., 2020) to planetary missions, and includes planning and scheduling (Agrawal et al., 2021; Bernard et al., 1999; Gaines et al., 2022; Tran et al., 2005; Troesch et al., 2020), event detection (Francis et al., 2017), autonomous orbital (Bhaskaran, 2012) and surface (Carsten, Rankin, Ferguson, & Stentz, 2009; Rankin et al., 2021) navigation, and fault management (Mackey et al., 2021).

Previous generations of operational deep space and Earth observation spacecraft have been operated through a ground-in-the-loop and time-delayed process where operators on the ground uplink command sequences that the spacecraft performs at prescribed times (Sheridan, 1993). In the earliest missions, these sequences were entered manually (Standley, 2006), more recently, the process of generating command sequences on the ground has moved to more AI planning-based tools that develop and validate a command sequence based on high-level goals specified by the operators (Ai-Chang et al., 2004; Barreiro et al., 2012; Bernardi, Cesta, Finzi, & Orlandini, 2013; Verfaillie & Pralet, 2020).

As planning algorithms become more efficient and the available computational resources onboard increase, the planning process is gradually moving on-board, with the Mars 2020 onboard planner making its way into operations (Parjan & Gaines, 2024; Rabideau et al., 2020; Yelamanchili et al., 2021), and many more concepts finding incubation across research projects. In this case, ground-in-the-loop operations changes from uplinking command sequences to more complex concepts like engineering intentions (Castano et al., 2022; Rossi et al., 2023) or science goals (Gaines et al., 2020). Under this regime (Dvorak, Ingham, Morris, & Gersh, 2009), ground-in-the-loop autonomy-supported spacecraft activity planning requires a new, comprehensive approach to intent specification, that helps operators not only build the right set of goals (Castano et al., 2022), but also understand the possible outcomes (Alper Ramaswamy et al., 2019), how robots might perceive the world (Bae, Rossi, Hook, Davidoff, & Ma, 2020), and trust that the on-board planning and scheduling capability will achieve the desired intent (Rossi et al., 2023; Widder, Dabbish, Herbsleb, Holloway, & Davidoff, 2021).

A similar evolution is taking place, adding human-in-the-loop AI into the spacecraft downlink process. Downlink operators examine the downlinked channelized time-series measurements that represent key spacecraft state variables, to assess whether the spacecraft correctly executed the sequence that was uplinked. They also assess whether any errors were raised by the flight software, and whether the spacecraft state remains within nominal parameters. Historically these efforts have been supported by ground

tools to assess whether the data received matches prescribed patterns (Ko, Maldague, Lam, Bui, & McKinney, 2010; Quach, DeForrest, Klesh, & Schoolcraft, 2014; Trimble & Rinker, 2016), and by purpose-built programming environments (Kluyver et al., 2016; Pérez & Granger, 2007). Adding AI to this human-in-the-loop downlink analysis looks to enable operators to more easily identify and highlight deviations from nominal expectations (Bhamidipati, Rossi, & Castano, 2024), and to prioritize the highest-priority data during bandwidth-constrained passes (Dubay et al., 2022).

### 4.5 Implications of HITL for HCAI

HITL aligns closely with the goals of HCAI by retaining a high level of human agency and responsibility. As humans provide frequent input and validation, the risk of critical oversights from fully autonomous algorithms and thus out-of-loop risks can be reduced. Explainability becomes naturally integrated into the design since operators must have insight into how the system is reasoning in order to supply meaningful feedback and to understands the system's behavior at all times. Moreover, iterative improvements in performance, adaptability to different users and situations, and reliability can be achieved without losing human oversight, leading to a system that evolves while remaining accountable to its human operators. In environments such as spaceflight, where operational risks are considerable and tasks can be unpredictable, HITL demonstrates how interactive machine learning can augment rather than replace human expertise.

## 5 Discussion

In this chapter, we explored how HCAI frameworks, specifically HITL and HOTL, can be leveraged to enhance human autonomy, safety, and decision-making in the context of spaceflight. Building on case studies such as the *CORE Personal Assistant*, *METIS*, and *Spacecraft Science Operations*, we demonstrated how AI-driven tools can complement human expertise for tasks ranging from procedural support (HITL) to autonomous oversight with limited human intervention (HOTL).

The case studies and conceptual frameworks presented in this chapter illustrate that although HITL and HOTL systems offer substantial benefits for future spaceflight missions, they also introduce several challenges and trade-offs. This section highlights and identifies guidelines for developers, researchers, and designers on how to design HITL and HOTL systems in the human space flight context, while exposing potential risks and limitations, indicating how future research can address these concerns in an integrated HCAI context in future work.

### 5.1 Interface Design Guidelines for HITL and HOTL

HCAI frontend interfaces actively engage astronauts and mission controllers in the decision-making process. Instead of merely presenting information, these interfaces enable users to understand and operate with the AI's underlying assumptions, change settings or parameters of the AI system, and offer feedback that can influence future AI outputs. This focus on interaction supports the essential principles of explainability

and transparency, which are crucial for building trust and fostering effective collaboration between humans and AI systems. It is important to note that while trust and reliance are related, they are not synonymous: trust refers to a user's belief in the system's competence, benevolence, and integrity, whereas reliance refers to the practical decision to depend on the system, sometimes even in the absence of full trust when no alternatives are available (de Fine Licht & Brülde, 2021; Hoff & Bashir, 2015).

In this context, AI complacency, defined as overtrust in AI systems (Harbarth, Gößwein, Bodemer, & Schnaubert, 2024; Rodriguez et al., 2020), must be avoided at all costs. Overreliance on AI can lead to operators neglecting critical tasks or failing to recognize when the AI makes errors. Conversely, undertrust, characterized by skepticism or mistrust of AI capabilities, can result in operators underutilizing the AI's potential. In extreme cases, undertrust may lead to operator overload as humans compensate for tasks that could have been efficiently managed by the AI. Both overtrust and under-trust risk degrading situational awareness, which is particularly critical in high-stakes environments like space missions Both miscalibrated trust and inappropriate reliance risk degrading situational awareness, which is particularly critical in high-stakes environments like space missions(Endsley & Kiris, 1995).

Therefore, trust calibration between the operator and AI is essential for effective human-AI teaming. Shifts in control authority must be communicated effectively (Flemisch et al., 2012), as misaligned trust or inadequate communication can lead to out-of-the-loop situations where the operator cannot intervene effectively. In such cases, operators may struggle to understand the system's state and fail to retain situational awareness (Endsley & Kiris, 1995). Interfaces designed for HITL must emphasize reliability, safety, and trustworthiness to ensure effective human-AI partnerships (Shneiderman, 2020).

To allow for effective human-AI teaming, several key principles must be considered to enhance AI-human collaboration. Interface designs should incorporate strategies such as clear explanations of AI decisions, adaptive AI, and training to maintain mental model alignment between humans and AI.

Interfaces must not only display system status but also explain the reasoning behind AI-driven recommendations (explainability), providing detailed explanations of how the AI arrived at its recommendations helps users critically evaluate suggestions rather than blindly accepting them, reducing the risk of errors made by the AI-human team (Shneiderman, 2020). For example, if an AI system suggests altering a rover's path, the interface could highlight areas of potential risk, indicate where sensor confidence is low, or present historical data that influenced the decision. Additionally, saliency analysis could provide insights into how the model arrives at its predictions, potentially even during training, ensuring that the deployed model aligns with user requirements. Saliency analysis is a technique that evaluates the contribution of individual input features to a model's prediction, often by examining gradients or attention weights, to highlight which parts of the input data are most influential in driving the outcome. Recent research has also explored methods to improve model interpretability through saliency-guided training. For instance, a study introduced a saliency-guided training procedure that iteratively masks features with small and potentially noisy gradients while maximizing the similarity of model outputs for

both masked and unmasked inputs. This approach was shown to significantly enhance model interpretability across various domains while preserving predictive performance (Ismail, Bravo, & Feizi, 2021). While confidence scores can clarify the reliability of AI outputs (Bruzzese, Gao, Dietz, Ding, & Romanos, 2020; Zhang, Liao, & Bellamy, 2020), they must be carefully designed to avoid user misinterpretation or overreliance due to overtrust. For example, overly simplistic confidence scores without proper context might lead operators to misjudge the system's reliability.

The Prometheus Principles, as proposed by Shneiderman (2020), provide an essential framework for designing HCAI interfaces that advance explainability while ensuring automation aligns with user intent. These principles emphasize consistent interfaces that help users form, express, and revise their intent, continuous visual displays of relevant objects and actions; rapid, incremental, and reversible actions, error prevention, and informative feedback.

Regular training and simulations to gain experience with the AI system are also important to ensure operators remain vigilant and capable of intervening when necessary (Hoff & Bashir, 2015). These exercises should include scenarios where the AI provides suboptimal recommendations or fails, preparing users to retain their problem-solving skills under pressure (Bahner, Hüper, & Manzey, 2008).

Simulations should also account for communication delays inherent in space missions, which can significantly affect situational awareness and decision-making, which have been shown to disrupt ground-crew interactions and negatively impact performance (Parisi, Panontin, Wu, McTigue, & Vera, 2023).

Maintaining consistent operator engagement and using feedback mechanisms, HCAI interfaces can help prevent operators from losing situational awareness during critical tasks (Endsley & Kiris, 1995). Effective HCAI interfaces must also support smooth transitions of control between humans and AI. These transitions must be seamless and supported by clear communication mechanisms, such as visual alerts or auditory prompts, ensuring operators are fully aware of their responsibilities and system status during handovers (Flemisch et al., 2012).

Additionally, interfaces must adapt to evolving mission needs and the crew's growing expertise. Inspired by AI-enabled adaptive learning systems (Kabudi, Pappas, & Olsen, 2021), astronaut interfaces in spaceflight missions could initially provide step-by-step guidance to assist less experienced crew members in executing complex tasks without feeling overwhelmed. As the mission progresses and astronauts gain experience, adaptive assistance could dynamically adjust its level of support based on the operator's skills and task complexity. By balancing ability, authority, control, and responsibility, these interfaces could therefore ensure that both astronauts and AI systems contribute effectively to mission-critical decision-making.

Moreover, AI systems used in space missions must provide reliable guidance even when there is not enough data available to train highly accurate models. Such situations can occur when astronauts encounter rare mission events, unexpected technical failures, or completely new environmental conditions that the AI has never seen before. In deep-space missions, for example, AI models cannot always rely on large amounts of pre-collected training data because real-world conditions on the Moon or Mars may differ from simulations and past missions.

To ensure that AI remains useful in such unpredictable situations, design principles should incorporate additional sources of knowledge beyond standard training data. This means using side information, such as known physical laws, spacecraft system constraints, or expert-defined safety limits, to guide AI decision-making. For example, if an AI system predicts that an astronaut's oxygen supply will last much longer than physically possible, predefined safety rules should override that faulty prediction. Moreover, confidence values can be used to inform the astronaut about the likelihood of errors.

This approach prevents the AI from making dangerous or unrealistic decisions and helps astronauts maintain control in high-risk, uncertain environments. For instance, in a lunar rover operation, delegation interfaces might reduce AI control when terrain becomes unpredictable, allowing operators to manually navigate until the system recalibrates. These adaptive mechanisms ensure human operators remain actively engaged during critical tasks, even under challenging conditions.

## 5.2 Challenges for HITL and HOTL in Spaceflight

Looking beyond immediate missions, the scalability of HITL/HOTL frameworks for multi-crew or multi-mission operations needs to be ensured and tested in real-life contexts. While current designs typically address small-scale needs and currently only small (research-projects) have been proposed and only tested to a limited degree and in a controlled research environment (e.g., CORE, CIMON, METIS), future missions to the Moon, Mars, or beyond may involve larger teams and more complex tasks and likely more unexpected situations. Extending HCAI beyond pilot projects will require modular system architectures that support incremental upgrades and changes in mission planning. Moreover, collaboration between commercial and governmental stakeholders could accelerate the standardization of best practices in HCAI certification, software infrastructure, and training protocols for future space missions to ensure that human-AI collaboration is working well.

Spacecraft components undergo exhaustive verification and validation cycles. Incorporating adaptive AI systems poses complex questions for reviewers regarding how to certify systems capable of learning or evolving after launch. While classical "test-and-freeze" procedures ensure predictability, HITL/HOTL systems may rely on continuous human feedback and dynamic model updates, complicating safety certification. A potential pathway involves incremental validation, where AI models operate within well-defined boundaries (e.g., using RLHF or supervised checkpointing) and all updates are closely audited by mission control or onboard personnel. By establishing rigorous simulation campaigns and fail-safe mechanisms, it becomes feasible to align HCAI's flexibility with spaceflight's safety requirements.

Furthermore, incorporating side information and structured priors, such as physics-informed AI, can enhance the robustness of adaptive models by constraining learning within well-understood physical laws and mission parameters. AI systems can also maintain confidence thresholds for predictions, ensuring that operators are alerted when model uncertainty increases beyond acceptable limits. By monitoring confidence levels in real-time, the system can preemptively warn users of potential false

predictions, allowing for human intervention before erroneous outputs influence critical decision-making. This approach strengthens the trustworthiness of AI in space missions while maintaining operational flexibility.

At the same time, cybersecurity and automation failures pose serious risks, especially in long-duration missions where astronauts may heavily rely on autonomous or semi-autonomous systems. In isolated environments, protection against hacking or unauthorized access becomes challenging, since the capacity to patch vulnerabilities or receive timely support from Earth is limited. Similarly, redundancy in critical systems and adequate backup training is essential, so that astronauts have alternatives if the AI fails and are not fully reliant on it.

Deploying an AI assistant in space comes with serious technical hurdles. Hardware and power constraints are a major factor. While LLMs usually require powerful GPUs, they can run efficiently on CPUs using quantization (Xiao et al., 2024) and low-rank adaptation (LoRA) (Hu et al., 2021), reducing power consumption and memory needs. Still, space-rated processors are less capable than those on Earth and must be carefully optimized (Bensch, Bensch, Nilsson, Saling, Sadri, et al., 2024).

Noise interference is another challenge. Spacecraft are loud—life support systems and onboard equipment create constant background noise. On the ISS, noise levels reach up to 75 decibels, which can make voice-based AI unreliable without advanced noise cancellation (Limardo, Allen, & Danielson, 2015).

Finally, ensuring high degree of reliability while maintaining redundancy is essential. AI systems must handle hardware failures, radiation-induced memory errors, and data corruption. This means multiple backups, self-healing algorithms, and fallback models are essential to keep the system running on long missions (Bensch, Bensch, Nilsson, Saling, Sadri, et al., 2024).

## 5.3 Future Directions

Looking ahead, HCAI-based systems in space exploration may combine both HITL and HOTL paradigms in a hybrid fashion, adapting interaction levels according to the crew's workload and mission context. Although the AI might operate autonomously for routine monitoring or straightforward tasks, the interface could seamlessly transition to a more interactive mode when conditions become unpredictable or when the crew's specialized expertise is indispensable. To manage the variability of long-duration flights, upcoming research can also explore personalized dashboards that respond to each astronaut's physiological and cognitive state, ensuring that assistance modes and alert thresholds match real-time operator capacity. Here, for instance, technologies such as Electroencephalography (EEG), electromyography (EMG) and eye-tracking have proven to be useful tools to accurately estimate workload and situational awareness states in the context of human space flight (Becker, Nilsson, & Cowley, 2022).

Importantly, organizations such as NASA, ESA, and commercial providers (e.g., SpaceX, Blue Origin) may benefit from standardized guidelines on human-AI teaming, data governance, and moral/operational responsibility. Tying these guidelines into internationally accepted space law will be crucial as missions become more global and

commercialized therefore also common laws and guidelines for AI usage and responsibilities can be created that prioritize human needs and put the human at the core of the decision process. Yet, this is still an unresolved issue not only with regard to space flight but also on Earth, as a consensus on common guidelines or laws has yet to be reached (Corrêa et al., 2023).

While our primary focus in this chapter explored the human space flight context, much of the work on AI explainability, autonomy handovers, and user-interface feedback loops can benefit other safety-critical contexts: from nuclear plant monitoring to deep-sea exploration. The methods explored here, particularly high-level autonomy that can change back to human control when mission-critical changes occur, may improve complex operations in other domains where mistakes are costly or irreversible.

## 5.4 Cyborg Psychology

While HCAI proposes that human values and the human operator should stay at the core of the decision process a novel approach that we want to present in this chapter called *Cyborg Psychology* sees AI not only as a tool but as technology where the human works and lives in a form of symbiosis with intelligent systems. The relationship between AI and humans is therefore not only seen as one way but can also have consequences for human behavior. Cyborg Psychology, therefore, aims at identifying conditions under which human-AI systems can become more than the sum of their parts, while also seeking to understand and mitigate potential unintended consequences of such interactions (Pataranutaporn, 2024). While HCAI focuses on the concept that AI remains a tool that supports humans, Cyborg Psychology studies how AI integration fundamentally changes human cognition and interaction with the world.

This concept could become especially important when thinking about long-term space missions, where astronauts heavily rely on AI assistance and only limited human support is available. For instance, AI-based "companions" could mitigate loneliness on missions but might lead to new dependencies or require new forms of oversight.

The term "Cyborg Psychology" finds its roots in the discourse surrounding the investigation of the "Cybernetic Organism", commonly abbreviated as "Cyborg". In their seminal paper "Cyborgs and Space" published in 1960 (9 years before Apollo 11), Manfred E. Clynes and Nathan S. Kline introduced the scientific concept of the cyborg - envisioning a future where humans and machines would be seamlessly integrated to overcome the biological limitations of humans during space exploration (Clynes & Kline, 1960). This field emphasizes the complex, non-linear interactions between these two intricate systems, recognizing that both can influence and shape each other in dynamic and often unpredictable ways. Cyborg Psychology bridges the gap between various disciplines, situated at the convergence of human-computer interaction (HCI), AI, psychology, cognitive science, and human behavioral science. It also investigates potential negative outcomes, such as AI addiction, erosion of human agency, diminished empathy, manipulation of human behavior through AI-driven persuasion techniques, cognitive overload, and the exacerbation of existing cognitive biases. Cyborg Psychology utilizes insights from both human and AI behavior to design experiments and interventions that test various hypotheses. The primary goal

is to gain a comprehensive understanding of human-AI interactions, encompassing both beneficial and detrimental aspects. This includes studying how these interactions can potentially enhance human flourishing, while also investigating and addressing potential negative outcomes.

The field of HCI emerged around 1970 with the rise of personal computers (Myers, 1998), building upon earlier visionary concepts. Pioneers in the field had long been interested in computers as tools for supporting human cognitive processes(Ashby, 1956; Engelbart, 2021; Licklider, 1960). A significant focus in HCI is understanding the usability and design of human-intelligent system interactions. This area explores the design of intelligent tools, contrasting direct manipulation interfaces with interface agents for task delegation (Shneiderman & Maes, 1997). Human-computer integration (HInt) is a recent paradigm emphasizing the integration of computational and human systems, highlighting the need for a structured research agenda due to growing ethical and societal implications (Mueller et al., 2020).

The advancement of AI has led to new HCI paradigms, such as human-centered AI presented in this chapter, which balances human control and computer automation (Maes et al., 2021; Shneiderman, 2020). The concept of Homo Cyberneticus has been proposed, highlighting the era of human-AI integration and suggesting new research directions in human augmentation within HCI. Recent studies emphasize integrating Computer-Supported Cooperative Work (CSCW) perspectives into AI research. As AI becomes more prevalent in human-facing applications, the need for well-defined human-AI interaction principles has increased. Researchers have formalized and validated design guidelines for human-AI interaction, serving as a resource for practitioners and researchers. With the emergence of LLMs(S. Yang, Hu, Yu, Ali, & Wang, 2024), researchers are exploring more dynamic evaluation methods for the outcomes of human-AI interaction Ibrahim, Huang, Ahmad, and Anderljung (2024).

Within this evolving landscape of HCI and AI research, Cyborg Psychology emerges as a specialized area of study. It builds upon HCI foundations while introducing a distinct focus on cognitive phenomena and psychological concepts in human-AI interactions. Cyborg Psychology explores how to study and intervene in specific psychological concepts such as self-determination, motivation, reflection, and personal identity within these interactions. Crucially, it is oriented towards the broader goal of human flourishing, seeking to leverage AI technologies in ways that enhance overall well-being, personal growth, and the realization of human potential. By positioning itself at the intersection of HCI, AI research, and human behavior studies, Cyborg Psychology contributes to a more holistic understanding of the complex relationships between humans and AI systems. This interdisciplinary approach aims to guide the development of AI technologies that are not only functional and efficient but also psychologically beneficial and aligned with human values. Cyborg Psychology is therefore one key future research topic that could maximize the potential of AI-human teaming especially in the harsh conditions of spaceflight.

Future Cyborg Psychology research directions in the context of astronaut assistance could further investigate the interplay of astronauts with RAG on KGs. Astronauts must process vast amounts of information while managing a variety of experiments and tools, making efficient information retrieval crucial. One promising

avenue is the development of AI-driven note-taking systems that capture and structure mission-relevant knowledge in KGs. Personalized KGs, initiated during astronaut training, could serve as a long-term memory resource, enabling intuitive access to experiment protocols, tool usage histories and personalized settings, and procedural insights via natural language interactions. However, such AI integration also raises concerns regarding reliance and dependency on AI systems. It is essential to examine scenarios where AI assistance becomes unavailable due to system failures, network disruptions, or incorrect data processing. Understanding these risks and designing resilient human-AI collaboration frameworks will be critical in ensuring the safe and effective deployment of AI-powered support systems in space missions.

# 6 Conclusion

Successfully establishing a long-term human presence on the Moon - and eventually Mars - will depend on a robust synergy between human expertise and AI autonomy. As discussed throughout this chapter, HCAI frameworks, namely Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL), represent critical methodologies for achieving such synergy. These paradigms emphasize meaningful human oversight, clarity of AI recommendations, and responsible design of backend (algorithmic) and frontend (user interface) systems.

The HITL approach underscores continuous user engagement, incorporating ongoing feedback loops for tasks that demand immediate decision-making and frequent human intervention. This is particularly beneficial in dynamic or time-sensitive contexts, where astronauts or mission controllers can iteratively shape AI behaviors. Our case studies, such as the *CORE Personal Assistant* and *Spacecraft Science Operations*, demonstrate how HITL can provide astronauts with real-time interactive support to keep them actively involved in the AI-driven process. We have argued that such user engagement not only improves task efficiency but also mitigates the risk of operator complacency due to being out of the loop.

Conversely, the HOTL framework is tailored for scenarios where the AI operates with greater autonomy, with human operators positioned to oversee, review, and intervene only as needed. This approach enables AI to handle a wider array of low-level or data-intensive tasks while maintaining human authority for major decisions, serving as a safeguard in critical environments. Our *METIS* example illustrates how HOTL can be effectively employed in space systems to balance autonomous data analysis with human supervisory checks, thereby promoting a reliable and transparent mode of sensor operation during future spaceflight missions.

While HITL and HOTL each offer distinct advantages, future designs will likely combine both paradigms to address the evolving complexities of spaceflight. Hybrid systems may dynamically shift along a spectrum of human-AI collaboration based on situational demands, crew workload, and mission phase. Advances in immersive interfaces (e.g., AR/VR), KG integration, and reinforcement learning from human feedback will further refine these collaborative paradigms, promoting trust, safety, and efficiency. Meanwhile, Cyborg Psychology offers a new perspective on human-AI

teaming, considering AI not only as a support tool for tasks but as a technology that influences psychological processes when humans are deeply intertwined with it.

In the years and decades ahead, the silicon vs organic balance in space operations may well be dynamic. Humans voyage into space now for research and exploration, and over the coming years will increasingly work there, at least in low Earth orbit and on the Moon. Humans did not evolve to survive in space, however, radiation exposure away from the Earth's magnetic field can be lethal in long missions, along with complications from long-duration zero-gravity, extended social isolation, and other factors. Making hostile environments like Mars and the Moon into desirable places for people to live is also very complex, and with the world's population predicted to decrease (Spears, 2023), people will not need to go to these places to find a better life. Electronics, however, does not have these issues, and also historically follows exponential trends in capability over time. Humans will evolve in concert with their systems on the same timescales as envisioned space colonization. As we start to solve fundamental challenges in extended human spaceflight, humans will be increasingly augmented by their technology, including AI.

This chapter has explored human augmentation for spaceflight, but if we look further into the future, the distinction between human and augmentation will increasingly blur, leading to the question of whether it is our fragile biological selves that truly migrate into space or our AI progeny. Either way, it is a fascinating future, and we are all privileged to be alive at a time when seminal questions like this cannot only be posed but also witnessed in our lifetime.

**Acknowledgements.** Part of this work was conducted during a research visit at the NSF AI Institute for Dynamic Systems (CBET-2112085). We extend our gratitude to the team at the European Astronaut Center (ESA-EAC) and the MIT Space Exploration Initiative.

## References

Agrawal, J., Chi, W., Chien, S.A., Rabideau, G., Gaines, D., Kuhn, S. (2021). Analyzing the effectiveness of rescheduling and flexible execution methods to address uncertainty in execution duration for a planetary rover. *Robotics and Autonomous Systems*, *140 (2021) 103758*, , Retrieved from https://doi.org/10.1016/j.robot.2021.103758

Ai-Chang, M., Bresina, J., Charest, L., Chase, A., Hsu, J., Jonsson, A., ... Maldague, P. (2004, 01). Mapgen: Mixed-initiative planning and scheduling for the mars exploration rover mission. *IEEE Intelligent Systems*, *19*, 8-12, https://doi.org/10.1109/MIS.2004.1265878

Alkaissi, H., & Mcfarlane, S. (2023, 02). Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus*, *15*, , https://doi.org/10.7759/cureus.35179

Alper Ramaswamy, B.B., Agrawal, J., Chi, W., Kim Castet, S.Y., Davidoff, S., Chien, S. (2019). Supporting automation in spacecraft activity planning with simulation and visualization. *Aiaa scitech 2019 forum* (p. 2348).

Anandapadmanaban, E. (2020). *Vmcc: a virtual reality framework for augmenting mission control operations* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Ashby, W.R. (1956). An introduction to cybernetics.

Asimov, I. (2004). *I, robot* (Vol. 1). Spectra.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... Kaplan, J. (2021). *A general language assistant as a laboratory for alignment.* Retrieved from https://arxiv.org/abs/2112.00861

Bae, S., Rossi, F., Hook, J., Davidoff, S., Ma, K. (2020). A visual analytics approach to debugging cooperative, autonomous multi-robot systems' worldviews. *2020 ieee conference on visual analytics science and technology (vast)* (pp. 24–35). Los Alamitos, CA, USA: IEEE Computer Society.

Bahner, J.E., Hüper, A.-D., Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699,

Barreiro, J., Boyce, M.E., Do, M.B., Frank, J.D., Iatauro, M., Kichkaylo, T., ... Smith, D.E. (2012). Europa: A platform for ai planning, scheduling, constraint programming, and optimization. *International conference on automated planning and scheduling (icaps) – international competition on knowledge engineering for planning and scheduling.*

Becker, L., Nilsson, T., Cowley, A. (2022). Electroencephalography (eeg), electromyography (emg) and eye-tracking for astronaut training and space exploration. *arXiv preprint arXiv:2212.06139*, ,

Becker, L., Nilsson, T., Demedeiros, P., Rometsch, F. (2023). Augmented reality in service of human operations on the moon: Insights from a virtual testbed. *Extended abstracts of the 2023 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3544549.3585860

Benford, G. (1987). *Great sky river.* Bantam Spectra.

Bensch, O., Bensch, L., Nilsson, T., Saling, F., Bewer, B., Jentzsch, S., ... Kutz, J.N. (2024). *Ai assistants for spaceflight procedures: Combining generative pre-trained transformer and retrieval-augmented generation on knowledge graphs with augmented reality cues.* Retrieved from https://arxiv.org/abs/2409.14206

Bensch, O., Bensch, L., Nilsson, T., Saling, F., Sadri, W.M., Hartmann, C., ... Kutz, J.N. (2024). *Towards a reliable offline personal ai assistant for long duration spaceflight.* Retrieved from https://arxiv.org/abs/2410.16397

Bensch, O., & Hecking, T. (2021, Oktober). Towards open domain literature based discovery. *3rd international open search symposium.* Retrieved from https://elib.dlr.de/147093/

Bensch, O., Hecking, T., Kutz, J. (2024). Retrieval augmented generation and scientific knowledge graphs to support scientific hypotheses generation. *6th international open search symposium# ossym2024* (Vol. 30, p. 72).

Bergmann, E.V., Croopnick, S., Turkovich, J., Work, C. (1979). An advanced spacecraft autopilot concept. *Journal of Guidance and Control*, *2*(3), 161–168,

Bernard, D., Dorais, G., Gamble, E., Kanefsky, B., Kurien, J., Man, G., ... Tung, Y.-W. (1999). Spacecraft autonomy flight experience - the ds1 remote agent experiment. *Aiaa space technology conference and exposition.* Retrieved from https://arc.aiaa.org/doi/abs/10.2514/6.1999-4512

Bernardi, G., Cesta, A., Finzi, A., Orlandini, A. (2013). A knowledge engineering environment for p&s with timelines. *Workshop on knowledge engineering for planning and scheduling (keps), international conference on automated planning and scheduling (icaps).*

Bhamidipati, S., Rossi, F., Castano, R. (2024, March). Operations for autonomous spacecraft: Downlink analysis of onboard decisions and execution anomalies. *IEEE Aerospace Conference* (p. 1-9). Big Sky, MT. Retrieved from https://www.federico.io/pdf/Bhamidipati.Rossi.Castano.AERO24.pdf

Bhaskaran, S. (2012). Autonomous navigation for deep space missions. *Spaceops 2012* (p. 1267135).

Bolstad, C.A., Endsley, M.R., Hill, F. (2003). Tools for supporting team sa and collaboration in army operations. *Collaborative technology alliances conferences: Advanced decision architecture conference.*

Braly, A.M., Nuernberger, B., Kim, S.Y. (2019). Augmented reality improves procedural work on an international space station science instrument. *Human factors*, *61*(6), 866–878,

Brézillon, P. (1994). Context needs in cooperative building of explanations. *First european conference on cognitive science in industry* (pp. 443–450).

Bruzzese, T., Gao, I., Dietz, G., Ding, C., Romanos, A. (2020). Effect of confidence indicators on trust in ai-generated profiles. *Extended abstracts of the 2020 chi conference on human factors in computing systems* (pp. 1–8).

Carsten, J., Rankin, A., Ferguson, D., Stentz, A. (2009). Global planning on the Mars Exploration Rovers: Software integration and surface testing. *Journal of Field Robotics*, *26*(4), 337-357, https://doi.org/https://doi.org/10.1002/rob.20287 Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.20287 https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.20287

Castano, R., Vaquero, T., Rossi, F., Verma, V., Van Wyk, E., Allard, D., ... others (2022). Operations for autonomous spacecraft. *2022 ieee aerospace conference (aero)* (pp. 1–20).

Chin, S. (2024). *Towards perceptual augmentation* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Chow, C.K. (1957). An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, *EC-6*(4), 247-254, https://doi.org/10.1109/TEC.1957.5222035

Clarke, A.C. (2016). *2001: a space odyssey*. Penguin.

Clynes, M.E., & Kline, N.S. (1960). Cyborgs and space. *Astronautics*, *14*(9), 26–27,

Corrêa, N.K., Galvão, C., Santos, J.W., Del Pino, C., Pinto, E.P., Barbosa, C., ... others (2023). Worldwide ai ethics: A review of 200 guidelines and recommendations for ai governance. *Patterns*, *4*(10), ,

Cortes, C., DeSalvo, G., Mohri, M. (2016). Learning with rejection. R. Ortner, H.U. Simon, & S. Zilles (Eds.), *Algorithmic learning theory* (pp. 67–82). Cham: Springer International Publishing.

De, A., Saraf, S., Mishra, T.K., Tripathy, B. (2024). Interpretation and visualization techniques in ai systems and applications. *Explainable, interpretable, and transparent ai systems* (pp. 279–301). CRC Press.

de Fine Licht, K., & Brülde, B. (2021). On defining "reliance" and "trust": Purposes, conditions of adequacy, and new definitions. *Philosophia*, *49*, 1981–2001,

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., ... others (2024). Safeguarding large language models: A survey. *arXiv preprint arXiv:2406.02622*, ,

Dourish, P. (1997). Accounting for system behaviour: Representation, reflection and resourceful action. *Computers and design in context*, 145–170,

Dourish, P., & Button, G. (1998). On" technomethodology": Foundational relationships between ethnomethodology and system design. *Human-computer interaction*, *13*(4), 395–432,

Dubay, M.M., Johnston, N., Wronkiewicz, M., Lee, J., Lindensmith, C.A., Nadeau, J.L. (2022). Quantification of motility in bacillus subtilis at temperatures up to 84° c using a submersible volumetric microscope and automated tracking. *Frontiers in Microbiology*, *13*, 836808,

Dvorak, D.D., Ingham, M.D., Morris, J.R., Gersh, J. (2009). Goal-based operations: an overview. *Journal of Aerospace Computing, Information, and Communication*, *6*(3), 123–141,

Endsley, M.R., & Kiris, E.O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human factors*, *37*(2), 381–394,

Engelbart, D.C. (2021). Augmenting human intellect: a conceptual framework (1962).

Ennis, S., Rometsch, F., Saling, F., Fischer, B., Mittler, P., Ferra, L., ... others (2021). Astronaut training on-board the international space station using a standalone virtual reality headset.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., Kiela, D. (2024). *Kto: Model alignment as prospect theoretic optimization.* Retrieved from https://arxiv.org/abs/2402.01306

Fetterolf, C.S. (2020). *Using augmented reality to enhance situational awareness for aircraft towing* (Unpublished doctoral dissertation). Monterey, CA; Naval Postgraduate School.

Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., Beller, J. (2012). Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. *Cognition, Technology & Work*, *14*, 3–18,

Francis, R., Estlin, T., Doran, G., Johnstone, S., Gaines, D., Verma, V., ... others (2017). AEGIS autonomous targeting for ChemCam on Mars Science Laboratory: Deployment and results of initial science team use. *Science Robotics*, *2*(7), ,

Gaines, D., Chien, S., Rabideau, G., Kuhn, S., Wong, V., Yelamanchili, A., ... Lohr, C. (2022, June). Onboard planning for the Mars 2020 Perseverance rover. *16th symposium on advanced space technologies in robotics and automation.* Retrieved from https://ai.jpl.nasa.gov/public/documents/papers/M2020-OBP-ASTRA-2022-Final.pdf

Gaines, D., Doran, G., Paton, M., Rothrock, B., Russino, J., Mackey, R., ... Agha-mohammadi, A.-a. (2020, October). Self-reliant rovers for increased mission productivity. *Journal of Field Robotics*, *37*(7), 1171–1196, https://doi.org/10.1002/rob.21979 Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21979

Garcia, A.D., Schlueter, J., Paddock, E. (2020). Training astronauts using hardware-in-the-loop simulations and virtual reality. *Aiaa scitech 2020 forum* (p. 0167).

Gil, M., Albert, M., Fons, J., Pelechano, V. (2020). Engineering human-in-the-loop interactions in cyber-physical systems. *Information and software technology*, *126*, 106349,

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... Ma, Z. (2024). *The llama 3 herd of models.* Retrieved from https://arxiv.org/abs/2407.21783

Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, *40*(2), 44–58,

Hand, K.P., Phillips, C.B., Murray, A., Garvin, J., Maize, E., Gibbs, R., ... others (2022). Science goals and mission architecture of the europa lander mission concept. *The Planetary Science Journal*, *3*(1), 22,

Harbarth, L., Gößwein, E., Bodemer, D., Schnaubert, L. (2024). (over) trusting ai recommendations: How system and person variables affect dimensions of complacency. *International Journal of Human–Computer Interaction*, 1–20,

Heer, J. (2019). Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, *116*(6), 1844-1850, https://doi.org/10.1073/pnas.1807184115 Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1807184115 https://www.pnas.org/doi/pdf/10.1073/pnas.1807184115

Hoff, K.A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407–434,

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. *Proceedings of the sigchi conference on human factors in computing systems* (pp. 159–166).

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). *Lora: Low-rank adaptation of large language models.* Retrieved from https://arxiv.org/abs/2106.09685

Ibrahim, L., Huang, S., Ahmad, L., Anderljung, M. (2024). Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks. *arXiv preprint arXiv:2405.10632*, ,

Ismail, A.A., Bravo, H.C., Feizi, S. (2021). *Improving deep learning interpretability by saliency guided training.* Retrieved from https://arxiv.org/abs/2111.14338

Jetter, H.-C., Rädle, R., Feuchtner, T., Anthes, C., Friedl, J., Klokmose, C.N. (2020). " in vr, everything is possible!": Sketching and simulating spatially-aware interactive spaces in virtual reality. *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–16).

Johnson, M., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J., Woods, D.D. (2014). Seven cardinal virtues of human-machine teamwork: examples from the darpa robotic challenge. *IEEE Intelligent Systems*, *29*(6), 74–80,

Kabudi, T., Pappas, I., Olsen, D.H. (2021). Ai-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017,

Keswani, V., Lease, M., Kenthapadi, K. (2021). Towards unbiased and accurate deferral to multiple experts. *Proceedings of the 2021 aaai/acm conference on ai,*

*ethics, and society* (p. 154–165). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3461702.3462516

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., ... development team, J. (2016). Jupyter notebooks - a publishing format for reproducible computational workflows. F. Loizides & B. Scmidt (Eds.), *Positioning and power in academic publishing: Players, agents and agendas* (pp. 87–90). Netherlands: IOS Press. Retrieved from https://eprints.soton.ac.uk/403913/

Ko, A., Maldague, P., Lam, D., Bui, T., McKinney, J. (2010). The evolvable advanced multi-mission operations system (ammos): making systems interoperable. *Spaceops 2010 conference delivering on the dream hosted by nasa marshall space flight center and organized by aiaa* (p. 2303).

Konen, K., Jentzsch, S., Diallo, D., Schütt, P., Bensch, O., Baff, R.E., ... Hecking, T. (2024). *Style vectors for steering generative large language model.* Retrieved from https://arxiv.org/abs/2402.01618

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2021). *Retrieval-augmented generation for knowledge-intensive nlp tasks.* Retrieved from https://arxiv.org/abs/2005.11401

Licklider, J.C. (1960). Man-computer symbiosis. *IRE transactions on human factors in electronics*(1), 4–11,

Limardo, J., Allen, C., Danielson, R. (2015, 07). Status: Crewmember noise exposures on the international space station..

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... others (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, ,

Mackey, R., Nikora, A., Altenbuchner, C., Bocchino, R., Sievers, M., Fesq, L., ... Prather, M. (2021). On-board model based fault diagnosis for cubesat attitude control subsystem: Flight data results. *2021 ieee aerospace conference (50100)* (pp. 1–17).

Madras, D., Pitassi, T., Zemel, R. (2018). *Predict responsibly: Improving fairness and accuracy by learning to defer.* Retrieved from https://arxiv.org/abs/1711.06664

Maes, P., Shneiderman, B., Ren, X., Shi, Y., Wang, D., Wang, Q. (2021). Designing ai to work with or for people? *Extended abstracts of the 2021 chi conference on human factors in computing systems.*

Markov-Vetter, D., & Staadt, O. (2013). A pilot study for augmented reality supported procedure guidance to operate payload racks on-board the international space

station. *2013 ieee international symposium on mixed and augmented reality (ismar)* (pp. 1–6).

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, *77*(12), 1321–1329,

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054,

Mozannar, H., & Sontag, D. (2020, 13–18 Jul). Consistent estimators for learning to defer to an expert. H.D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 7076–7087). PMLR. Retrieved from https://proceedings.mlr.press/v119/mozannar20b.html

Mueller, F.F., Lopes, P., Strohmeier, P., Ju, W., Seim, C., Weigel, M., ... others (2020). Next steps for human-computer integration. *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–15).

Myers, B.A. (1998). A brief history of human-computer interaction technology. *interactions*, *5*(2), 44–54,

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., ... Zoph, B. (2024). *Gpt-4 technical report.* Retrieved from https://arxiv.org/abs/2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., ... Lowe, R. (2022). *Training language models to follow instructions with human feedback.* Retrieved from https://arxiv.org/abs/2203.02155

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X. (2024, July). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3580–3599, https://doi.org/10.1109/tkde.2024.3352100 Retrieved from http://dx.doi.org/10.1109/TKDE.2024.3352100

Parisi, M., Panontin, T., Wu, S.-C., McTigue, K., Vera, A. (2023). Effects of communication delay on human spaceflight missions. *14th international conference on applied human factors and ergonomics (ahfe).*

Parjan, S., & Gaines, D. (2024, March). "in obp we trust": Verification and validation of the m2020 on board planner flight software. *Ieee aerospace conference.* Big Sky, Montana. Retrieved from https://ai.jpl.nasa.gov/public/documents/papers/parjan-ieee2024-m2020.pdf

Pataranutaporn, P. (2024). *Cyborg psychology: The art & science of designing human-ai systems that support human flourishing* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Pérez, F., & Granger, B.E. (2007, May). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, *9*(3), 21–29, https://doi.org/10.1109/MCSE.2007.53 Retrieved from https://ipython.org

Phillips, P.J., Phillips, P.J., Hahn, C.A., Fontana, P.C., Yates, A.N., Greene, K., ... Przybocki, M.A. (2021). Four principles of explainable artificial intelligence.

Pillai, V. (2024). Enhancing transparency and understanding in ai decision-making processes. *Iconic Research and Engineering Journals*, *8*(1), 168–172,

Quach, W.L., DeForrest, L., Klesh, A.T., Schoolcraft, J. (2014). Adapting a large-scale multi-mission ground system for low-cost cubesats. *Spaceops 2014 conference* (p. 1634).

Rabideau, G., Wong, V., Gaines, D., Agrawal, J., Chien, S., Kuhn, S., ... Biehl, J. (2020). Onboard automated scheduling for the mars 2020 rover. *Proceedings of the international symposium on artificial intelligence, robotics and automation for space.* Noordwijk, NL: European Space Agency.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.* Retrieved from https://arxiv.org/abs/2212.04356

Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 53728–53741). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper$_f iles/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf$

Rankin, A., Maimone, M., Biesiadecki, J., Patel, N., Levine, D., Toupet, O. (2021). Mars curiosity rover mobility trends during the first 7 years. *Journal of Field Robotics*, *38*(5), 759–800,

Rodden, T.A., Fischer, J.E., Pantidi, N., Bachour, K., Moran, S. (2013). At home with agents: exploring attitudes towards future smart energy infrastructures. *Proceedings of the sigchi conference on human factors in computing systems* (pp.

1173–1182).

Rodriguez, S.S., Chen, J., Deep, H., Lee, J.J., Asher, D.E., Zaroukian, E. (2020). Measuring complacency in humans interacting with autonomous agents in a multi-agent system. *Artificial intelligence and machine learning for multi-domain operations applications ii* (Vol. 11413, pp. 258–271).

Rometsch, F.A., Casini, A.E., Drepper, A., Cowley, A., de Winter, J.C., Guo, J. (2022). Design and evaluation of an augmented reality tool for future human space exploration aided by an internet of things architecture. *Advances in Space Research*, *70*(8), 2145–2166,

Rossi, F., Allard, D.A., Amini, R., Vaquero, T.S., Dhamani, N.N., Choukroun, M., ... others (2023). Workflows. user interfaces, and algorithms for operations of autonomous spacecraft. *2023 ieee aerospace conference* (pp. 1–17).

Schmidt, P., Biessmann, F., Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278,

Schmitz, H.-C., Kurth, F., Wilkinghoff, K., Müllerschkowski, U., Karrasch, C., Schmid, V. (2020). Towards robust speech interfaces for the iss. *Companion proceedings of the 25th international conference on intelligent user interfaces* (pp. 110–111).

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017). *Proximal policy optimization algorithms.* Retrieved from https://arxiv.org/abs/1707.06347

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Sheridan, T.B. (1993). Space teleoperation through time delay: Review and prognosis. *IEEE Transactions on robotics and Automation*, *9*(5), 592–606,

Sheridan, T.B. (2021). Human supervisory control of automation. In *Handbook of human factors and ergonomics* (p. 736-760). John Wiley Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119636113.ch28

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504,

Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. interface agents. *interactions*, *4*(6), 42–61,

Smith, M., Craig, D., Herrmann, N., Mahoney, E., Krezel, J., McIntyre, N., Goodliff, K. (2020). The artemis program: an overview of nasa's activities to return humans to the moon. *2020 ieee aerospace conference* (pp. 1–10).

Spears, D. (2023). The world's population may peak in your lifetime. *The New York Times*, 10–10,

Standley, S. (2006). Cassini-huygens engineering operations at saturn. *Spaceops 2006 conference* (p. 5516).

Stork, D.G. (1997). *Hal's legacy: 2001's computer as dream and reality.* Mit Press.

Thompson, D.R., Bunte, M., Castano, R., Chien, S., Greeley, R. (2012). Image processing onboard spacecraft for autonomous plume detection. *Planetary and Space Science*, *62*(1), 153–159,

Tideman, M., van der Voort, M.C., van Houten, F.J. (2008). A new product design method based on virtual reality, gaming and scenarios. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *2*, 195–205,

Toy, T. (2023). Transparency in ai. *AI & SOCIETY*, 1–11,

Tran, D., Chien, S., Sherwood, R., Castano, R., Cichy, B., Davies, A., Rabideau, G. (2005). Demo: The autonomous sciencecraft experiment onboard the eo-1 spacecraft. *Proceedings of the fourth international joint conference on autonomous agents and multiagent systems* (p. 163–164). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1082473.1082832

Trimble, J., & Rinker, G. (2016). Open source next generation visualization software for interplanetary missions. *14th international conference on space operations* (p. 2348).

Troesch, M., Mirza, F., Hughes, K., Rothstein-Dowden, A., Bocchino, R., Donner, A., ... others (2020). Mexec: An onboard integrated planning and execution approach for spacecraft commanding. *Workshop on integrated execution (intex)/goal reasoning (gr), international conference on automated planning and scheduling (icaps intex/gr 2020).*

Verfaillie, G., & Pralet, C. (2020). A timeline, event, and constraint-based modeling framework for planning and scheduling problems. *Workshop on knowledge engineering for planning and scheduling (keps), international conference on automated planning and scheduling (icaps).*

Von Hippel, E. (1986). Lead users: a source of novel product concepts. *Management science*, *32*(7), 791–805,

Weber, B., Hulin, T., Schiffer, L. (2022). How to really get in touch with robots—haptic interaction technologies for vr and teleoperation. *Meaningful futures with robots* (pp. 102–111). Chapman and Hall/CRC.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models.* Retrieved from https://arxiv.org/abs/2201.11903

Weitzner, D.J., Abelson, H., Berners-Lee, T., Hanson, C., Hendler, J., Kagal, L., ... Waterman, K.K. (2006). Transparent accountable data mining: New strategies for privacy protection.

Whitworth, B. (2005). Polite computing. *Behaviour & Information Technology*, *24*(5), 353–363,

Wickens, C.D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, *3*(2), 159–177,

Widder, D.G., Dabbish, L., Herbsleb, J.D., Holloway, A., Davidoff, S. (2021). Trust in collaborative automation in high stakes software engineering work: A case study at nasa. In *Proceedings of the 2021 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381,

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S. (2024). *Smoothquant: Accurate and efficient post-training quantization for large language models.* Retrieved from https://arxiv.org/abs/2211.10438

Xu, W., Gao, Z., Dainoff, M. (2023). An hcai methodological framework: Putting it into action to enable human-centered ai. *arXiv preprint arXiv:2311.16027*, ,

Yang, R., & Newman, M.W. (2012). Living with an intelligent thermostat: advanced control for heating and cooling systems. *Proceedings of the 2012 acm conference on ubiquitous computing* (pp. 1102–1107).

Yang, S., Hu, L., Yu, L., Ali, M.A., Wang, D. (2024). Human-ai interactions in the communication era: Autophagy makes large models achieving local optima. *arXiv preprint arXiv:2402.11271*, ,

Yelamanchili, A., Agrawal, J., Chien, S., Biehl, J., Connell, A., Guduri, U., ... Towey, S. (2021, May). Ground-based automated scheduling for operations of the mars 2020 rover mission. *Proceedings space operations 2021.*

You, Q., Jin, H., Wang, Z., Fang, C., Luo, J. (2016). Image captioning with semantic attention. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4651–4659).

Zhang, Y., Liao, Q.V., Bellamy, R.K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 295–305).

Zhu, M., Sun, Z., Zhang, Z., Shi, Q., He, T., Liu, H., ... Lee, C. (2020). Haptic-feedback smart glove as a creative human-machine interface (hmi) for virtual/augmented reality applications. *Science Advances*, *6*(19), eaaz8693,