Titre: Title:	Symbiotic Human and Multi-Robot Planetary Exploration Systems
Auteur: Author:	Marcel Kaufmann
Date:	2024
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Kaufmann, M. (2024). Symbiotic Human and Multi-Robot Planetary Exploration Systems [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/59046/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/59046/
Directeurs de recherche: Advisors:	Giovanni Beltrame
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Symbiotic Human and Multi-Robot Planetary Exploration Systems

MARCEL KAUFMANN
Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Août 2024

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Symbiotic Human and Multi-Robot Planetary Exploration Systems

présentée par **Marcel KAUFMANN**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Michel DESMARAIS, président
Giovanni BELTRAME, membre et directeur de recherche
Philippe DOYON-POULIN, membre
Yue HU, membre externe

DEDICATION

To all spacekind, family, and friends who
always support(ed) my international endeavours,
and those that are watching from among the stars.

ACKNOWLEDGEMENTS

This work would not have been possible without the support of the Arbour Foundation, especially Marine Hadangue and Diane de Champlain, the support of the Canada Vanier Graduate Scholarship Program, the support of the Canadian Space Agency and my supervisor Giovanni Beltrame. Further, I would like to acknowledge Chantal Balthazard and Julie Doré for helping with administrative issues and trouble shooting wherever I was on this beautiful planet; sometimes the paperwork is overwhelming. I would also like to acknowledge the people of MIST lab that made working on all the projects and (field) experiments very enjoyable: Thank you Jacopo Panerati, David St Onge, Vivek Shankar, Ivan Svogor, and everyone that I have forgotten to mention. Thank you to Emily Coffey from Concordia University for your expertise and discussions around Psychology. Thank you Caltech, NASA, and the Jet Propulsion Laboratory for an unforgettable Visiting Student Program. Thank you for your mentorship and the continued collaborations throughout the course of the PhD studies, Ali Agha, Brett Lopez, Mike Milano, Ryan Stonebreaker, Tiago Vaquero, Maira Saboia da Silva, Kyon Otsu, and the rest of team CoSTAR. Living underground with you for the DARPA SubT Challenge has truly been enjoyable and rewarding. Thank you to Jen Blank for your support and for the continued collaboration during the NASA BRAILLE project and while writing proposals for future endeavors. Thank you for those that sparked my passion for space and created and/or supported what is known as United Space School today; Rob Alexander, Franceso Fusco, Tahir Merali, Laura Gibson ten Bloemendal & Daniel ten Bloemendal, Nicole & Chris Stott, Chris Hadfield, Justin Kugler, Karl Siebold, and those who are unfortunately no longer with us – Geoff & Annette Mules, Lynne Gibson, and Bernd Gliemann.

A special thanks goes to my Canadian friends, especially Linda Dao, Gabriel Dubé, and the Drudi family who made living in Canada very enjoyable creating a home from home; thank you Lisa, Eric, and Maria for all the turkey and for letting me be part of your family! Speaking of home, thank you to my family and friends in Europe that have supported me through both the ups and downs of being abroad – there truly have been good and challenging times but with the necessary support everything can be overcome; thank you Mom, Alex, Mirja, Thomas, Renè, Nicole, Michelle, Philipp, Kathrin, Niko, Elke, Klaus, Jan, Sebastian, Kevin, Florian, and Paul. Marlon, it has been great to see you grow from afar and I am curious to see who you will become, I am proud to be your godfather.

"Je me souviens" – Thank you & Merci.

RÉSUMÉ

L'exploration spatiale, la frontière finale, manque de capacités validées de systèmes multi-robots avec humain dans la boucle pour découvrir nos origines. À ce jour, juillet 2024, les missions d'exploration n'ont pas démontré de capacités autonomes de multi-robots en dehors de la Terre, malgré les récents progrès dans le développement de l'autonomie robotique. Les systèmes humains-multi-robots, en particulier, sont encore moins explorés dans des scénarios réalistes d'exploration planétaire et des lacunes subsistent dans la compréhension de la manière dont les humains et les robots peuvent collaborer efficacement et synergiquement. L'exploration planétaire et le secours en cas de catastrophe impliquent des environnements potentiellement dangereux qui limitent l'accès humain et nécessitent de nouveaux paradigmes d'interaction qui s'éloignent du paradigme de "une mission un robot".

Les objectifs globaux de la recherche de la dissertation sont (i) La définition de protocoles de collaboration entre humains et robots visant l'exploration d'environnements inconnus avec des équipes multi-robots. (ii) La mise en œuvre d'une architecture logicielle et des outils nécessaires pour réduire la charge de travail dans les systèmes humains et multi-robots. (iii) La création d'une interface intuitive pour le système robotique permettant un contrôle et une interaction efficaces avec une formation minimale et une faible surcharge cognitive pour un seul superviseur humain. (iv) La validation de nos approches à travers des exercices réalistes d'exploration et de secours en cas de catastrophe.

Ces objectifs sont abordés par la présentation de trois articles cohérents se concentrant sur le développement et le déploiement de technologies d'assistance pour les systèmes autonomes, la conception d'interfaces et leur intégration et validation avec les facteurs humains dans le contexte de l'exploration et de la recherche et sauvetage.

Les contraintes de temps et de risque peuvent introduire des défis supplémentaires pour les superviseurs humains de systèmes multi-agents. En considérant des capacités d'autonomie de plus en plus opaques et des systèmes robotiques complexes, nous introduisons un assistant d'autonomie appelé Copilot MIKE pour explorer des terrains extrêmes dans le premier article. Copilot MIKE vise à réduire la charge de travail de l'opérateur en (1) fournissant une conscience situationnelle par rapport aux équipes de robots et à l'environnement; (2) surveillant activement les aspects clés de l'avancement de la mission; (3) apportant un soutien aux processus de prise de décision concernant la planification et la programmation des tâches; et (4) aidant à créer une infrastructure de réseau de communication et à coordonner le déploiement des nœuds de communication pour construire une dorsale de communication. Les

résultats préliminaires de simulation indiquent que la charge de travail globale de l'humain dans la boucle a diminué, tandis que la concentration du superviseur humain a augmenté pour exécuter des tâches pendant les phases de préparation et d'exploration d'une mission.

Alors que de nombreux systèmes multi-agents sont traditionnellement déployés dans des environnements contrôlés et structurés permettant des tests structurés (par exemple, les entrepôts), le DARPA Subterranean Challenge (SubT) visait divers types d'environnements souterrains inconnus qui imposaient le risque de perte de robots en cas d'échec. Dans le deuxième article, nous introduisons une interface inspirée des jeux vidéo ainsi qu'un assistant de mission autonome qui répond aux lacunes du premier prototype. Nous testons et déployons ensuite le système en utilisant un système multi-agents hétérogène dans des environnements complexes.

Ce travail a abouti à une amélioration du contrôle supervisé par l'humain pour un système multi-agents, tout en réduisant la surcharge liée au basculement d'application, à la planification, à l'exécution et à la vérification des tâches. De plus, cette plateforme de collaboration humain-systèmes autonomes a augmenté le temps d'exploration disponible pour les agents déployés, ce qui a conduit à des zones explorées plus vastes. Les paradigmes d'interaction, les approches et les interfaces introduits dans ce travail ont été testés sur le terrain lors du SubT et d'une mission analogue sur Mars de la NASA BRAILLE. Lors des tests préliminaires, le système a été déployé avec jusqu'à 11 robots supervisés simultanément par un seul humain.

Des lacunes subsistent dans la compréhension de la manière dont l'autonomie et la conception des interfaces s'intègrent aux facteurs humains et aux performances lors de missions d'exploration à grande échelle. Le troisième article examine des équipes composées de deux robots (semi-)autonomes et d'un seul superviseur humain utilisant nos interfaces. Le système a été déployé dans des grottes au Monument National des Lava Beds (Californie du Nord) et lors d'une étude à facteurs croisés 2x2 entre sujets à Polytechnique Montréal, explorant à la fois des grottes réelles et simulées. Nous avons obtenu des résultats comprenant 38 participants qui ont évalué l'influence de l'autonomie (points de passage vs. autonomie complète avec interventions) et des interfaces (écran vs. réalité virtuelle) sur la charge de travail, la conscience situationnelle et la performance lors d'une mission d'exploration scientifique. Nous constatons que les mesures physiologiques continues correspondent aux métriques NASA TLX pour les différentes modalités d'interface. La condition des points de passage en réalité virtuelle donne le plus faible nombre de cibles scientifiques détectées, tandis que toutes les autres conditions obtiennent des niveaux de performance similaires. La conscience situationnelle, évaluée par les mesures de précision de la Méthode d'Évaluation de la Présence Situationnelle (SPAM), donne environ 90% de précision pour les deux interfaces.

Les résultats collectifs de ces trois articles fournissent de nouveaux paradigmes d'interaction, des approches pour concevoir des interfaces à faible charge cognitive, un modèle du monde cyber-physique en temps réel en réalité virtuelle, et des apprentisages sur l'interaction entre la conscience situationnelle, la charge de travail et la performance, validés lors de missions d'exploration planétaire.

ABSTRACT

Space exploration, the final frontier, is lacking validated human-in-the-loop multi-robot system capabilities to uncover our origins. As of yet, July 2024, exploration missions have not demonstrated autonomous multi-robot capabilities outside of Earth, despite recent leaps in the development of robotic autonomy. Human-multi-robot systems in particular, are even less explored in realistic planetary exploration scenarios and gaps remain in the understanding of how humans and robots can collaborate effectively and synergetically. Both, planetary exploration and disaster relief, involve potentially hazardous environments that limit human access, and thus require new interaction paradigms that break away from the one mission one robot paradigm.

The overall research objectives of the dissertation are (i) The definition of collaboration protocols between humans and robots targeting the exploration of unknown environments with multi-robot teams.(ii) The implementation of a software architecture and the tools needed to reduce workload in human and multi-robot systems. (iii) The creation of an intuitive interface for the robotic system that allows effective control and interaction with minimal training and little cognitive load overhead for a single human supervisor. (iv) The validation of our approaches through realistic exploration and disaster-relief exercises.

These objectives are addressed by the presentation of three coherent articles, focusing on the development and deployment of assistive technologies for autonomous systems, interface designs, and their integration and validation with human factors, in the context of exploration and search and rescue.

Time and risk constraints can introduce additional challenges for human supervisors of multi-agent systems. Considering increasingly opaque autonomy capabilities and complex robotic systems, we introduce an autonomy assistant called Copilot MIKE for exploring extreme terrains in the first article. Copilot MIKE aims to reduce operator workload by (1) providing situational awareness with respect to robot teams and the environment; (2) actively monitoring key aspects of the mission progress; (3) providing support for decision making processes regarding task planning and scheduling; and (4) helping to create communication network infrastructure and coordinate communication node deployment to build a communications backbone. Preliminary simulation results indicate that the human-in-the-loop's overall workload decreased, while the human supervisor's focus increased on executing tasks during the preparation and exploration phases of a mission.

While many multi-agent systems are traditionally deployed in controlled and structured environments that allow for structured testing (e.g., warehouses), the DARPA Subterranean Challenge (SubT) targeted various types of unknown underground environments that imposed the risk of robot loss in the case of failure. In the second article, we introduce a video game-inspired interface, an autonomous mission assistant which addresses shortcomings of the first prototype and test and deploy the system using a heterogeneous multi-agent system in challenging environments.

This work resulted in improved human-supervisory control for a multi-agent system, while reducing overhead from application switching, task planning, execution, and verification. Further, this human-autonomy teaming platform increased the available exploration time for the deployed agents, which in turn lead to larger explored areas. The interaction paradigms, approaches, and interfaces introduced in this work have been field hardened and field tested during SubT and a NASA BRAILLE Mars analog mission. During preliminary tests, the system was deployed with up to 11 robots simultaneously supervised by a single human.

Gaps remain in the understanding of how autonomy and interface design integrate with human-factors and performance during large-scale exploration missions. The third article investigates teams comprising two (semi-)autonomous robots and a single human supervisor using our interfaces. The system has been deployed in caves at the Lava Beds National Monument (Northern California) and during a two-by-two factor within-subject study at Polytechnique Montréal, exploring both real and simulated caves. We obtained results including n=38 participants evaluating the influence of autonomy (waypoint vs. full autonomy with interventions) and interfaces (screen vs. virtual reality) on workload, situational awareness, and performance during a scientific exploration mission. We find that continuous physiological measurements align with NASA TLX metrics in the interface modalities. The virtual reality waypoint condition yields the lowest number of science targets detected, while all other conditions achieve similarly good performance levels. Situational awareness, assessed by the Situation Presence Assessment Method's accuracy measures, yields approximately 90% correctness for both interfaces.

The collective findings from these three articles provide novel interactions paradigms, approaches to design low cognitive workload interfaces, a real-time cyber-physical world model in virtual reality, and significant insights on the interplay of situational awareness, workload, and performance, validated during planetary exploration missions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

ACM	Association for Computing Machinery
ANOVA	Analysis of Variance
BRAILLE	Biologic and Resource Analog Investigations in Low Light Environments
CADRE	Cooperative Autonomous Distributed Robotic Exploration
DARPA	Defense Advanced Research Projects Agency
DLR	Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Center)
ECG	Electrocardiography
EEG	Electroencephalography
ERB	Ethics Review Board
ESA	European Space Agency
HCI	Human Computer Interaction or Interface
HFE	Human Factors and Ergonomics
HMI	Human Machine Interaction or Interface
HMRS	Human-Multi-Robot System
HRI	Human Robot Interaction
HRV	Heart Rate Variability
HSI	Human Swarm Interaction
JPL	Jet Propulsion Laboratory
NASA	National Aeronautics and Space Administration
MR	Mixed Reality
MRS	Multi Robot Systems
NeBula	Networked Belief-aware Perceptual Autonomy
rmANOVA	aclrmANOVA
ROS	Robot Operating System
THRI	Transactions on Human Robot Interaction
TLX	Task Load Index
VAM	Virtual Augmented Mixed Reality
VR	Virtual Reality

CHAPTER 1 INTRODUCTION

"Look again at that dot. That's here.
That's home. That's us."

— Carl Sagan, *Pale Blue Dot*

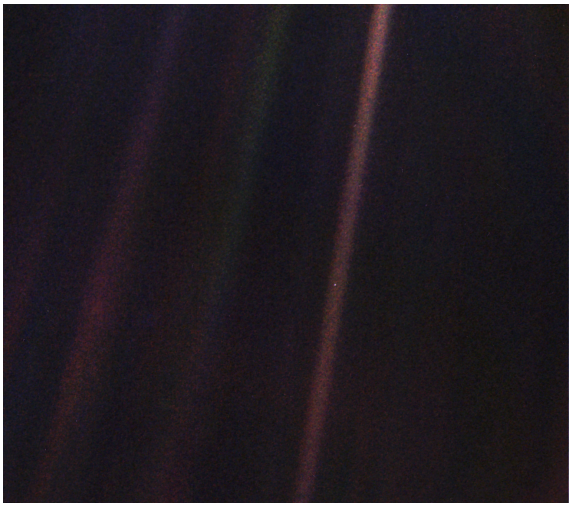

Figure 1.1 The Pale Blue Dot [Credit: NASA/JPL-Caltech]

This thesis by articles has been written in partial fulfillment of the requirements for the degree of Philosophiæ Doctoctor in computer engineering. The research was conducted under the supervision of Prof. Dr. Giovanni Beltrame, Polytechnique Montréal. We present three articles, either published or submitted, in Chapter 4, Chapter 5, and Chapter 6, respectively.

1.1 Context and Motivation

Looking at that dot, a speckle of dust in a ray of sunshine, this is Earth. Figure 1.1 shows the famous "Pale Blue Dot" photograph taken by the space probe Voyager 1 and it puts us in the perspective how little our vast planet is. The scientific and philosophical questions "Where did we come from?", "Are we alone in the universe?", and "What is our future beyond the Earth?" [1] are fundamental questions in the fields of astrobiology and space exploration. If we change the point of view looking outward from Earth into the night sky, only the Moon has been set foot on by humans to this date. Then there is Mars, "the

only planet we know of inhabited entirely by robots. [2]" In February 2021, the Mars 2020 flagship mission landed the Perseverance rover and its technology demonstration payload, the Ingenuity helicopter, on the red planet. Perseverance was used as communications relay and base station for the helicopter that flew an incredible 72 flights on Mars. All roving robotic space missions that have flown to date abided by the one mission, one rover paradigm. The first official multi-robot system scheduled to operate on another planet, the Moon, is called CADRE – Cooperative Autonomous Distributed Robotic Exploration. This mission (at the time of writing) is scheduled to launch in 2025 and will demonstrate distributed off-world measurements with ground-penetrating radar systems that a single robot could not achieve alone [3]. Similarly to collaborative robot agents, the integration of human capabilities with multiple robot agents promises interesting synergy effects. Human-robot interaction is a large area of research and this thesis sets out to help create symbiotic human and multi-robot systems for planetary exploration and terrestrial applications.

1.1.1 Symbiosis

In nature, symbiosis is a close and often long-term interaction between two different species [4]. There are three main forms of symbiosis called mutualism, commensalism and parasitism. Mutualism benefits both species from the interaction and can be seen in bees and flowers [5] where nectar benefits the bees as food in exchange for pollination. An example of commensalism, where the organism benefits itself without significantly harming the other, can be observed between barnacles and whales [6]. The barnacles profit from the whales' mobility and can enhance their chances of survival by feeding in various locations. The whales do neither benefit, nor suffer. Parasitism on the other hand benefits one and harms the other. This form of symbiosis can be observed looking at ticks [7]. They attach to their hosts, feed off them, and potentially transmit diseases. Figure 1.2 depicts these three nature examples.

(a) (b) (c)

Figure 1.2 Three Forms of Symbiosis in Nature. (a) Mutualism: Bees and Flowers, (b) Commensalism: Whales and Barnacles, and (c) Parasitism: Ticks and Their Hosts [Credit: iStock.com/kmatija]

Licklider [8] introduced the concept of man-computer symbiosis, which can be extended to humans and robots. In achieving symbiosis, the humans – in our scenarios, a single human – and robots become more intertwined and augment each other's capabilities, leading to outcomes that neither could achieve alone. To break the one robot one mission paradigm, it is crucial to understand how human-multi-robot systems (HMRS) can benefit from symbiotic concepts [9], or how these relationships emerge, and how they can be leveraged to achieve one of nature's forms of symbiosis – beneficial or harmful.

1.1.2 Human-Multi-Robot Systems

"Divide et impera" is latin for divide and rule, also known as divide and conquer, often quoted from Julius Caesar. This principle can be applied to human-multi-robot systems to divide complex tasks, distribute (sub-)tasks, and complement capabilities among the system's agents. Thus, they present a way of tackling large problems together. Together as teams, however, there are several characteristics to consider: team size, team composition, the interaction flow, proximity of agents, and control methods [10]. Figure 1.3 shows the connectivity graph of a single-human multi-robot team. While single-robot multi-human and multi-robot multi-human are also possible interaction models, we focus on the interaction flow for single-human multi-robot systems. Each node in the graph presents an agent or member of the team. Arrows between them show the interaction flow. Here, interactions are bi-directional to depict a system in which the human can send commands to the robot and the robots sends back information to the human. Robot agents can exchange information among each other, and the system could be extended to more than three nodes as indicated by the light blue additions in the figure. While some works suggest that more humans should be used per robot to divide control and mission task execution, thereby avoiding overburdening the human-in-the-loop [11], we present works that extend this interaction graph. Our approach incorporates the base station computer (the human "access point" to the system) as an additional agent for autonomous task planning, execution, and verification. This strategy helps reduce the burden on a single human operator (see Chapters 4 and 5).

Considering complex missions, for example exploration or search and rescue, there are two major task categories that exist among human and robot agents: *humans-better-at-it* and *machines-better-at-it* [12]. Robots and machines are great at processing large quantities of data, while humans are extremely versatile when it comes to solving unforeseen problems. Task performance on the other hand can be degraded when an operator does not give their undivided attention to a single robot [13]. This shows how important it is to understand the human node in HMRS.

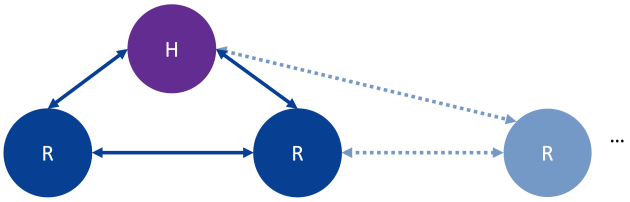

Figure 1.3 Human-Multi-Robot Interaction Graph. Each node represents a human (H) or robot (R). In human-multi-robot systems at least one Human and multiple robot nodes are present, so that there are at least three nodes in the graph. Arrows and their directionality depict the interaction flow. In this example, robots can be sent commands by the human and send information back, while robot agents might exchange information among each other. More robots would add more complexity to the graph as indicated by the light blue addition and the three dots.

1.1.3 Human Factors and Ergonomics

The fields of human factors and ergonomics (HFE) combine multiple disciplines in understanding interactions between humans and elements of a system [14]. How people interact with technology, how these interactions can be optimized, and how performance can be increased are often studied in this field, including interactions between humans and automation [15]. Metrics in the field can often be obtained by subjective and objective measurements of mental workload, operator performance, and situational awareness. Subjective measurements include questionnaires or other assessment techniques that either quantitatively or qualitatively assess a system. In cognitive ergonomics, some bio-physiological signals have shown to be a reliable measurement for workload, which includes EEG, ECG and derived measurements such as heart rate variability (HRV), pupil diameter, and functional near-infrared spectroscopy (fNIRS) [16]. In the field however, it is not practical to deploy clinical grade measurement equipment which can often be bulky and sometimes prohibitively costly. Thanks to miniaturization and technological advancements, wearable technologies have become readily available with, e.g., smart watches and chest-strap-based heart rate monitors. Subjective workload measurements cannot be obtained continuously without interrupting the task at at hand, which would lead to unrealistic scenarios. However, the relationship between objective and subjective measurements has to be validated, especially when using wearable sensors during real field deployments and real operations scenarios. Besides automation, human factors and performance can also be influenced by how information is presented to the human, which needs to be considered when designing human-robot interfaces.

1.2 Problem Statement

A multi-robot system, in contrast to a single all-purpose robot, or a single human, is potentially more efficient and resilient for exploration and mapping tasks in vast, potentially hazardous, and unknown environments. Higher levels of redundancy and the ability to explore more risky areas can be achieved by teams of robots. Adding humans to a multi-robot system during future missions, however, bears challenges when interfacing with the whole team of robots. Developing a "symbiotic human and multi-robot planetary exploration system" addresses research areas at the same time: (i) interaction with a team of multiple (semi-)autonomous robots, (ii) human-robot interaction and interfacing in caves, cavities and lava-tubes, as well as (iii) human factors measurements and validation.

- *Human-autonomy teaming:* a topic of particular interest for human space exploration, where effective teaming requires to consider the human as part of the system. The system should be designed with human limitations in mind and leverage (semi-)autonomous capabilities to increase team performance. We investigate the effectiveness of a single human supervisor within the team and implement higher autonomy in an effort to increase performance and reduce the human's workload in realistic exploration and search and rescue scenarios.

- *Human-robot interfaces:* the current state of the art is that there is no consensus on how to develop a user interface for a multi-robot system. Some attempts were made in the recent past (e.g. [17–20]) but they usually exploit traditional screen interfaces and Virtual Augmented Mixed Reality (VAM) only are used in simulation or small-scale and laboratory settings. A system for controlling a team of robots that deploys real-time cyper-physical interaction capabilities without relying on heavily pre-processed data or environment models does not yet exist (to the best of the authors' knowledge). Real-time approaches proposed in the literature have been limited to small models and some achieved live streaming of RGB-D camera data with limited fields of view [21]. However, none used multi-robot mapping capabilities and resulting point clouds. In this work, we will develop a prototype interface for the exploration of large-scale environments based on Virtual Reality (VR) technology that creates such cyber-physical space in real-time for safe virtual co-location.

- *Human-centered design:* Integrating human factors engineering into the design of autonomous systems is often an after thought. However, to optimize the interaction between humans and technology, ensuring systems are both functional and user-friendly, human factors need to be considered in early design stages already. Iterative testing

and direct user feedback ensure that systems are intuitive, reduce cognitive workload, and enhance team performance. Understanding how both autonomy and interfaces are influenced by human factors in realistic missions is often limited to laboratory experiments. Moreover, there is a gap in the understanding of real-time objective workload assessment with wearable sensors and the measurable interplay between autonomy and interfaces in exploration missions.

The author believes that addressing above challenges and gaps has great potential to extend the knowledge of the field and provide valuable results for the implementation of human and multi-robot systems for planetary exploration and terrestrial applications in mutually symbiotic ways.

1.3 Research Objectives

The problems presented in the previous section are addressed in the research articles contained in Chapters 4 to 6, focusing on the following overall research objectives:

I. The definition of collaboration protocols between humans and robots targeting the exploration of unknown environments with multi-robot teams.

II. The implementation of a software architecture and the tools needed to reduce workload in human and multi-robot systems.

III. The creation of an intuitive interface for the robotic system that allows effective control and interaction with minimal training and little cognitive load overhead for a single human supervisor.

IV. The validation of our approaches through realistic exploration and disaster-relief exercises.

1.4 Research Contributions

To the best of the authors' knowledge, this dissertation presents significant novel contributions to multi-robot systems, human-robot interaction, human-autonomy teaming, and human factors aligned with our research objectives. The contributions, categorized and presented per respective article, are as follows:

1. Kaufmann, M., Vaquero, T. S., Correa, G. J., Otstr, K., Ginting, M. F., Beltrame, G., & Agha-Mohammadi, A.-A. (2021). Copilot MIKE: An Autonomous Assistant for Multi-Robot Operations in Cave Exploration. In 2021 IEEE Aerospace Conference (50100) (pp. 1-9). IEEE.

   This work contributes an approach to a Multi-robot Interaction assistant for unknown cave environments. It establishes a human-robot interaction paradigm based on an adaptation of Sheridan's automation levels [22–24] and implements a novel system architecture that utilizes task planning and scheduling for collaborative human and multi-robot systems with the base station as dedicated agent in the Human-Multi-Robot System (HMRS). The system has been verified in simulation to qualitatively reduce operator workload and quantitatively decrease task execution times compared to scheduled task durations.

2. Kaufmann, M., Trybula, R., Stonebraker, R., Milano, M., Correa, G. J., Vaquero, T. S., Otsu, K., Agha-Mohammadi, A.-A., & Beltrame, G. (2022). Copiloting Autonomous Multi-Robot Missions: A Game-inspired Supervisory Control Interface. In ICAPS 2022, the 32nd International Conference on Automated Planning and Scheduling, 2022 Workshop on Scheduling and Planning Applications (SPARK), Singapore, Republic of Singapore, 7-12 June 2022.

   A game-inspired user interface for multi-agent robot missions, which expands the work of the previous article, identifies shortcomings, and addresses these by (i) integrating an automated planner for task planning and scheduling with resource constraints, (ii) creating a framework for verifiable task execution for increased reliability, and (iii) presenting verification results on how the overall system performed over the course of several real-world deployments, including the DARPA SubT Challenge final.

3. Kaufmann, M., Beltrame, G. (2024). Influence of Autonomy and Interfaces on Human and Multi-Robot Teams: a Study on Planetary Exploration". Submitted to ACM Transactions on Human Robot Interaction (THRI).

   This study contributes (i) a method for creating a fielded VR interface with real-time rendering capabilities that enables the exploration of ad-hoc and dynamically created cyber-physical spaces with multiple robots in large-scale environments, (ii) an evaluation of the influence of autonomy and interface design on multi-robot operations, comparing performance in a controlled study fielded system, (iii) an assessment of objective workload with a low-cost wearable HRV sensor and a comparison to the

subjective NASA TLX questionnaire, and (iv) a questionnaire for Situation Awareness, Immersion, and Trust (SAIT).

1.5 Impact

It is the author's believe that this work has and will continue to impact the field of human-multi-robot systems. The two already published articles presented in this thesis have achieved 14 citations at the time of writing. However, the principles and systems introduced in these articles have become part of the larger NeBula Autonomy framework, which partially reproduced and presented the human-robot interaction and interfacing capabilities as part of a larger NeBula systems paper [25] as well; it achieved 168 citations thus far. Throughout the PhD studies, the author's work has accumulated a total of 339 citations, including contributions beyond those presented in the thesis.

The contributions of this work lead to effective human-robot interaction paradigms, approaches, and interfaces that have been field hardened and field tested during the DARPA Subterranean Challenge (SubT) and the NASA BRAILLE Mars analog mission. In preparatory runs of SubT, up to 11 robots were simultaneously supervised by a single human using the interaction paradigms, autonomy assistant, and interfaces presented in this work.

The results presented in Chapter 5 have also, in part, inspired design elements for the first collaborative autonomous multi-robot systems' ground data and mission operations systems. The mission is scheduled to launch with the Intuitive Machines IM-3 lander (planned for 2025) and explore a small region on the Moon; Cooperative Autonomous Distributed Robotic Exploration (CADRE).

In addition, preparations (a pilot in the wild) for work on Chapter 6 resulted in a best poster award at the IROS 2022 HMRS Workshop – the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems Human-Multi-Robot Systems Workshop. The poster can be seen in Appendix A. Further, the author believes that the work presented in Chapter 6 will have an impact on the HRI and ergonomics community. We identified an implementation issue of how other works [26] decoded the Polar H10 heart rate monitor and provide our code that allows for easy access to hardware-derived heart rate variability of this device; this could result in better and more available data that is much needed in the field to even better understand human-machine systems.

Overall, especially combined with future works, the presented symbiotic human-multi robot systems could lead to the advancement of space exploration, help with disaster relief missions, and spark new academic ideas to come.

CHAPTER 2 LITERATURE REVIEW

This chapter reviews the existing literature on four key domains towards realizing symbiotic human and multi-robot systems, focusing on exploration tasks or broader relevant applications. Section 2.1 gives an overview of existing multi-robot systems in general. Section 2.2 presents the state-of-the-art in human and multi-robot interaction. Specific focus on the state of augmented, virtual, and mixed reality robotic interfaces is given in Section 2.3. Finally, Section 2.4 is providing an overview of human factors in the context of situational awareness and workload assessment. Further related works are presented within the context of each article in Chapter 4, Chapter 5, and Chapter 6.

2.1 Multi-Robot Systems

Multi-robot and swarm systems are promising and versatile tools for many tasks. Applications include planetary exploration, surveying, extravehicular missions, and terrestrial applications such as search and rescue or industrial tasks [27].

In contrast to single agent approaches, multi-robot systems (MRS) offer higher levels of redundancy and increased parallelism [28, 29]. More redundancy allows for higher risk missions, because the loss of a single agent would not jeopardize an entire mission. Furthermore, information sharing among agents can increase the reliability of MRS [30]. Heterogeneous multi-robot systems benefit from complementary capabilities [29], especially when performing collaborative tasks. In a lunar analog mission at Mount Etna, the German Aerospace Center (DLR) recently demonstrated the capabilities of their heterogeneous robot team. They conducted an abseiling experiment with two autonomous agents [31] proving that wheeled robots can access steep terrain that is beyond the individual robots' abilities.

While tele-operated systems that use augmented virtual reality interfaces [32] have potential for on-orbit servicing capabilities and other direct robot interactions, we focus on multi-robot systems that interact with human agents while exploring unknown environments. Human factors such as workload, situational awareness, human-machine trust and human reliability will play an important role [33] in creating reliable symbiotic planetary exploration systems.

However, supervising multi-robot systems is a difficult task for a single human operator [34] and only a limited number of works conduct testing on real-world systems.

2.2 Human and multi-robot Interaction

Multi-robot and swarm interfaces have only recently emerged as lively research topics. Methods from human-machine interaction (HMI), human-computer interaction (HCI) and computer vision can be used in human-swarm interaction (HSI). Interfaces span from visual systems that can use machine learning to understand human commands [35] to on screen planning interfaces [36] and virtual and augmented reality environments [37]. Gesture recognition, as well as natural language processing [38] have become common methods to interact with robotic systems. However, screen based interfaces, including tablets with touch input, are predominantly used in the literature which leaves a gap to research virtual, augmented, and mixed reality (VAM) devices. The amount of information a human has to hold in their head increases with the number of agents [36]. Thus, it is critical that interfaces provide high-level information that is clearly labeled, easy to understand, and notify a user about critical issues when attention is needed. Multi-robot systems require carefully designed, advanced interfaces to achieve effective human-robot teaming [10]. Three major categories of implementation challenges for human and multi-robot systems are identified in [10]. All three of which might be constraint by particular tasks or system specifications. Safety, scalability, and transparency are key issues to be considered in designing effective and performant MRS with HRI and they need to be validated beyond theoretical setups.

Many robotic systems implementations take advantage of the Robot Operating System (ROS) and its large set of tools, including the visualization tool RViz [39,40]. In a qualitative assessment of interfaces, Ikeda et al. [40] compare RViz, augmented reality, and a 2D graphical user interfaces and identify that AR can benefit even for debugging related works, particularly in the perception domain. In fact, RViz is often used to debug robotic systems and visualize internal states and understandings of a robot. The German Aerospace Center (DLR) developed an RViz-based interface to create shared situational awareness between roboticists and robot operators during field deployments, which allows for customization by individual developers who prefer to load their own domain-specific visualization configurations [19]. They conducted a pilot user study in a simulated environment involving nine participants, focused on a Mars surface sample collection mission. Results show that the interfaces were subjectively intuitive, but non-expert users experienced higher workloads overall, particularly on the frustration sub-scale of the NASA Task Load Index (TLX). This raises questions about whether customization of the interfaces (or the lack thereof) contributed to these higher workloads for non experts.

Robotics competitions, such as the DARPA Subterranean Challenge, have advanced the capabilities of autonomy and human-multi-robot systems (HMRS). The challenge's limitation

of HMRS to a single human supervisor in the loop drove significant investments in interface design improvements [20], including the work presented in this thesis.

In their survey on HSI, Kolling *et al.* [41] review existing literature, focusing on different swarm properties and discuss cognitive loads for operators with respect to different control models.

Sheridan and Verplank identified ten levels of HCI automation (see also [42, 43]). For a HSI system, it is important to define which tasks can be performed autonomously and which tasks should be performed solely by a human operator, as this determines the human role for certain tasks. The defined level of autonomy might indeed change with time, if for instance human cognitive load and situational awareness are taken into account. There are multiple ways to interact with a robotic team: As supervisor, a human could take control on a team level, or commands could be issued on an individual robot base [44]. An early model of human-computer interaction has been proposed by *Norman et al.* [45]. It consists of seven stages:

1. Formulation of the goal
2. Formulation of the intention
3. Specification of the action
4. Execution of the action
5. Perception of the system state
6. Interpretation of the system state
7. Evaluation of the outcome

This approach is iterative and is repeated until the goal is reached or a human-in-the-loop decides that the intention or goal have to be changed. While this HCI model is also applicable for agents within a multi-robot team, understanding the interplay between added autonomy capabilities and human-robot team performance remains an open question.

Task engagement and situational awareness are important to safely control a robotic team. Depending on the task engagement of a human, the autonomy level or goal might have to be adjusted. This can be severely affected by the user interface that is used to connect humans and a robot team [46]. The authors of [47] use a limited range of electroencephalography (EEG) signals to determine user engagement in a learning task, while recognizing cognitive capabilities as a critical factor for using robots. Podevijn *et al.* [48, 49] show that the "reality

gap" impacts the human psychophysiology when transitioning from simulated swarms to real swarms. Their experiments show that augmented reality can be a way to smoothen the transition. They also find that participants of their study experience higher arousal with a higher number of robots in the swarm. There are still many open research questions within the field of HRI: the authors of [47], for instance, see future work in the real-time evaluation of human task engagement with a robotic team. Wearable sensors could be one way of reaching such integration.

With respect to physiological measurements, pupillometry has recently gained new interest in the field of Psychology. Pupils have been found to dilate as a response to increased cognitive activity and other influences, like brightness and fixation [50]. In a different study [51] dilatation of the pupil was recently shown to be a good measurement of a pilot's cognitive load when dealing with auditory-visual interferences on a visual flight task. The main limitations of pupil dilatation measurements are that the diameter varies across subjects and that it is highly sensitive to luminance variations [52], which encourages future (interdisciplinary) work on this topic. Section 2.4 gives an additional overview of workload classification techniques used in this context.

Current space related human and multi-robot missions involve free floating spherical and cubical robots, aboard the International Space Station (ISS). The corresponding projects are called SPHERES (Synchronized Position Hold Engage and Reorient Experimental Satellite) [53] and Astrobee [54], respectively. SPHERES consists of three robots that were flying onboard the ISS until recently. Each robot has a size of approximately 21 cm in diameter and is equipped with different sensors like cameras. The platform can be remotely controlled through a computer based user interface [53] and has been used for guest experiments, educational purposes and contests. In [55], the authors take this platform and design a medical first-responder as a team mate for the astronauts aboard a space craft which is capable of assisting astronauts in cardiac arrest events. The Astrobees project is the SPHERES successor and has been launched to the ISS early 2019. The Astrobee platform is used to develop and test technologies required for autonomous operations (e.g., navigation aboard the ISS [56]), remote operation, and human-robot interaction with crew members. Latter includes user interfaces that can be used in proximity and remotely [54].

2.3 Augmented, Virtual, and Mixed Reality Robotic Interfaces

Augmented Reality (AR) is a technology that projects computer generated graphics onto real environments in real-time. In contrast, Virtual Reality (VR) places the user in a completely virtual environment and Mixed Reality fuses both AR and VR and integrates digital

objects into the real world. AR, VR, and MR raised a wave of interest in the early 1990s, but only recently—more than 25 years later—new applications and consumer products have emerged [57]. In the context of planetary or space exploration, there is little recent literature on collaborative Human-Robot interfaces which exploit AR, VR, or MR. The authors of [58] use a VR-based simulation and investigated how an Astrobee robot could guide non-expert users during an evacuation scenario aboard the ISS.

A purely virtual, browser-based interface for multiple rovers and astronauts has been presented in [59]. Here, the authors have implemented different view points, kinematic and environment models to simulate a Mars exploration Mission. They combine this work with their previous implementation of a task planning software—their system is centralized. Neither quantitative, nor qualitative evaluations of the interface have been conducted. AR, VR, and MR have become more advanced and popular in recent times. Especially with commercially available head mounted displays like the HTC Vive, Microsoft HoloLens, Meta Quest, or Apple Vision Pro which allow the creation of virtual experiences at high frame rates (approximately 60 fps to 144 fps).

Aukstakalnis [57] published a practical guide to AR and VR applications guiding the reader through the topics of human senses (e.g., vision, tactile, and audio), VR applications in Gaming, Architecture and Construction, Science and Engineering, Health and Medicine, Aerospace and Defense, Telerobotics and Telepresence, and Education. In the space engineering context, they state that Lockheed Martin has been using VR software for the development of the Orion space craft. A mixed reality approach has also been followed in [60], where virtual information is used to create a collaborative environment for an industrial robot and multiple humans in an aircraft manufacturing process. The authors of [60] still lack a high quality interface which enables an intuitive integration of smart devices and manufacturing equipment.

Kohn et al. [61] present a system for VR teleoperation of a robotic arm. They focus on an efficient method to render point clouds which allow near-realtime operations of the robotic arm (even for long distance tele-operations). The main trick here is to render known objects from CAD drawings and distinguish unknown elements using their model-based background segmentation. This of course does not work in unstructured and unknown environments, which are targeted by our symbiotic human and multi-robot planetary exploration system.

An AR-based user study with 60 participants and the goal to improve the understanding of robot motion intentions has been conducted by [62]. They make use of a head-mounted display (Microsoft HoloLens) to test the effectiveness of four different ways to represent such motion intent (navigation points, arrows, gaze, and utilities). The utilities mode, which

incorporated mini-maps and radars like they can be found in video games or pilot interfaces, was perceived significantly less helpful than the other modes. Their work also comes with practical and theoretical imitations: a motion tacking system is used for collocation of robots and user. In addition, participants of their study found the limited field of view of the HoloLens uncomfortable.

Astronaut Scott Kelly got to test the HoloLens as part of his International Space Station training during "Project Sidekick". This project was a technology demonstration to help astronauts communicate with ground operators, as well as receive mission relevant information on this head mounted display. Kelly himself said: "[...] now we look at the computer or an iPad to look at procedures. And if you could have a procedure right in your field-of-view, something that was command-able with your voice, you know where you could scroll through the different steps, that would be helpful." This shows that there is a wish for an audio visual interface to handle procedures more easily.

Augmented Reality has profited from advancements in various fields like computer vision, human-computer interaction, computer graphics, and others. One AR related problem is the pose (position and orientation) estimation of an AR user or AR device within a given coordinate frame. Existing solutions can be camera based and make use of camera calibration techniques which allow the reconstruction of a pose [63]. When such AR device is moved, continuous pose estimation can be seen as a tracking problem which can be solved using probabilistic approaches [64]. This addresses the question how to integrate human operators into a multi-robot system, which in the AR case is linked to the Multi-Robot navigation and simultaneous localization and mapping problems.

A limitation that can be introduced by AR/MR systems is the latency to process the environment and locate the user. The authors of [65] show how latency can directly decrease task performance in teleoperated scenarios. Ji et al. [66] compared different decision-making levels to direct and navigate a robot in a 20 m^2 analog environment mimicking an astronaut-rover setting (with non-astronauts as users). They measured workload using the NASA Task Load Index questionnaire finding that higher decision transparency decreases workload. Sending direct control commands (teleoperation) resulted in very high workloads compared to a point-by-point input. This work has several limitations regarding scalability. Firstly, only single-human single-robot interaction is evaluated and secondly, the environment is limited to a very small exploration area. Further, no assessment of workload was conducted during the fully autonomous mode, as they did not consider any human interactions for this part of the experiment. How an astronaut is affected by fully autonomous decisions and how they would interact with a multi-robot system remains an open question. While AR and

MR are great for in-situ experiments, there might also be use cases in which a robot team is sent into unknown and hazardous areas that would limit human access. Multiple interface modalities should be compared for exploration missions, and there is a need to address the lack of realistic or fielded deployments [10, 34] in complex, large-scale terrains. Furthermore, as low-latency is key for high task performance, real-time capable interfaces will be crucial to achieve mission success.

Reconstructing 3D scenes from sensor data is not an easy task [67], even with current improvements in computation and networking, rendering large-scale environments is often replaced by classical video streams, simulation models, or pre-processed virtual environments. Much of the VAM research involves industrial robots and oftentimes only a single robot agent [68–71]. This emphasizes that there is a need for real-time rendering capable interfaces that utilize ad-hoc sensor data. Interacting with multi-robot systems in such cyber-physical spaces allows for the co-location of humans and robots in real-time, while it prevents putting humans in hazardous situations. NASA's Artemis missions could greatly benefit from such interfaces.

2.4 Situational Awareness and Workload Assessment

In a planetary exploration setting where humans and (multiple) robots collaborate, high workload levels can be induced. Unknown environments require planning towards mission and task success. Mars exploration, for example, is currently performed remotely and humans teleoperate expensive single rover systems sending daily sets of commands to the remote rovers with limited autonomy capabilities. Real-time interaction is not possible due to the communication delays between Earth and Mars, which can last up to 26 minutes one way when the planets are at opposition [72]. Sending erroneous commands comes at high stakes in this setting (costs of approx. 400 Million, if the mission is compromised) and thus demands high task-performance [73]. A multi-robot system could reduce the cost risk with respect to single robot units, but controlling multiple robots or swarms of robots has higher workload demands [73, 74]. To make use of workload and situational awareness, these measures have to be evaluated —in real time. Various methods are used to analyze physiological objective signals. Some techniques involve vision and sensor based measurement methods and quantify workload using logic, pattern recognition, machine learning, and probabilistic methods [75].

Mental workload specifically refers to the demands placed on the brain's restricted information processing capability. As an analogy, physical workload can be characterized by the energy requirements for muscles [76]. There is two important classes of workload: overload, which occurs when there is more cognitive demand than resources available and underload,

which occurs in the opposite situation where there is little cognitive demand and high cognitive capacities are available [77]. Underload is often not considered when assessing task performance, but can negatively influence task performance and situational awareness as well [78].

A formal definition of *situational awareness* has been given by Endsley which states: "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [79, 80]. Further, Endsley and Jones describe three levels of situational awareness where level 1 refers to perceiving elements in the environment, level 2 SA implies comprehension of the current situation, and level 3 SA is the ability to project a future status from the current state of SA.

In a scenario where humans and robots work together, subjective workload measurements do not provide continuous output and are thus not feasible. To mention a common subjective measurement which is often used to compare objective and subjective workload measurements, we want to mention the NASA Task Load Index (NASA TLX). The NASA TLX assesses workload as a weighted mean of subjective ratings like effort, own performance, temporal demand, physical demand, mental demand, and frustration [81, 82]. These ratings are collected after a task has been completed and thus might not reflect specific workloads at specific times during a task. In terms of objective measurements, some have shown to be more reliable and capable to discriminate between different workload types (e.g., visual, speech, cognitive) than others: EEG, heart-rate variability, pupil diameter and functional near-infrared spectroscopy (fNIRS, which measures blood oxygen levels in the brain) are good workload indicators, as opposed to respiration rate, blink duration, or skin temperature (to mention a few) [16].

Common techniques to classify workload are machine learning based: linear regression, linear discriminant analysis, support vector machines, neural networks, model or ensemble based techniques are used. Linear regression is used to identify a line, plane or hyperplane to divide input features (e.g., heart-rate and pupil diameter) into different workload classes, while linear discriminant techniques project the input features into a smaller feature space to minimize the in-class distance and maximize the between-class distance [83–85]. Support vector machines (SVMs) project features into a feature space and separate the classes using linear or non-linear functions (comparable to linear regression). Some SVM techniques can incorporate a margin to separate classes, use a soft margin which leads to non ideal separation or use a kernel-based trick to reduce computational demand in higher dimensions [84]. Neural networks often consist of an input, processing and output layer. For each feature, the input

layer contains a node. So does the output layer for each class (e.g., underload, medium workload, high workload). Adaptive weights are learned during training and used to weigh each input feature in order to determine the output (classification) [83]. Ensemble techniques combine machine-learning algorithms and reduce each individual classifiers variance and bias by doing so. An ensemble can use the same classifier (e.g, bag of decision trees) or combine different techniques. Generalizability, however, remains an issue and large user studies are required to understand complex human-robot interactions better.

The authors of [86] used a neural network to assess workload in real time. As input they used electroencephalography (EEG), ECG (from which they derived heart-rate, heart-rate variability, and respiratory rate), and Electrooculography (e.g., eye blink frequency). In total, seven participants took part in the study and 43 features were used as input for the neural network. The participants were asked to conduct a baseline, low cognitive and high cognitive task. The collected data was split into 75% training data and a classification has been performed on the remaining 25%. The training set itself was split into 10 second epochs and a participant specific classifier was trained. The accuracy for individual participants ranges from 69% to 97%. The mean accuracy for all participants is reported as 84.3%. *Christensen et al.* [87] use these results and conduct similiar measurements but repeat the experiment over several days. They collected the data for eight participants and segmented their data into 40s epochs for training purposes of a neural network. Again, one classifier for each participant has been trained. The neural network has an accuracy of 99% for same-day classification of low and high workloads, while 83% are reached for inter-day classification. A support vector machine and linear discriminant classifier have been trained for the inter-day classification as well and reach 68% and 65% respectively. The decrease in inter-day accuracy suggest overfitting of the data and requires further validation.

In [88] real-time workload is assessed in an air traffic monitoring context. Four air traffic controllers with an average of twenty years of experience took part in this study for which EEG and the subjective NASA TLX measurements were used. Four test scenarios with different levels of adaptation (no adaptation, adaptation based on task complexity, adaptation based on cognitive complexity, and adaptation based on both task and cognitive complexity) are tested in sessions of about 75 minutes each. The subjective measurements indicate higher mental, physical and temporal demand with activated adaptation, but they rate their performance as best when EEG based adaptation was activated. Details on classification accuracy are not reported.

A direct comparison of human-multi-robot systems is difficult due to task specific and non standardized environments and constraints. We observe that many algorithms train partic-

ipant specific classifiers and thus are not generalizable across multiple subjects, operators or participants; which might be a necessity. Further, many studies deploy clinical-grade equipment to obtain objective workload measurements. However, low-cost wearable sensors have increasingly become available and should be investigated as viable options for fielded experiments.

Large numbers of user studies are needed to obtain and evaluate interactions schemes, such as point-to-point based input methods, autonomous modes, and teleoperation with direct inputs [89]. Performance of robotic systems can degrade quickly, when an operator neglects them, especially so in a teleoperated scenario. Crandall et al. [89] point out the important role of interfaces which are needed to gain (or maintain) sufficient situational awareness, make a plan, and execute by giving inputs to the robot.

Situational awareness can be assessed with different methods. However, situational awareness is not the same as simply providing a map [20]. Two common approaches to assess SA are the Situation Awareness Global Assessment Technique (SAGAT) and the Situation Present Assessment Method (SPAM). The SPAM method measures user's response times as the primary dependent variable while neither requiring a memory component nor to stop the current experimental task [90]. Fast response times indicate good SA [91]. As experimental interruptions introduce additional SA needs to get back to the previous task, assessing SA without the need for a memory component seems beneficial. However, the interplay of SA, workload, and performance in real cave exploration scenarios remained unanswered, which is why we conducted a user study investigating the influence of autonomy and interfaces on these human factors and human-multi robot system metrics.

CHAPTER 3 RESEARCH APPROACH AND THESIS ORGANIZATION

In this Chapter, we describe the three main phases to our research approach and present an overview of this thesis' document structure.

3.1 Approach

Phase 1-3 cover analysis, prototyping and validation, respectively. Each phase consists of work packages that we implemented.

3.1.1 Phase 1: Analysis

WP1: Identification of collaboration protocols between humans and robots. This work package focused on assessing existing multi-robot operations paradigms and protocols needed for HMRS. We identified and adapted techniques that presented themselves valuable to achieve a mutually beneficial symbiotic integration of a human supervisor in the robotic system. By declaring the base station to be a separate agent within the robotic team, we adapted the interaction graph of our system unlocking new interaction capabilities that allowed for advancements in system automation.

WP2: Identification of interface requirements. Two mission scenarios will be defined along with the performance goals and metrics at the user interface level. Existing standard metrics can be used for this purpose (e.g. [50, 92, 93]). The user interfaces of a number of existing hardware and software products will be assessed for their usability (based on heuristic assessments) and how they address mission scenarios relevant to cave exploration. This will inform our own design choices and design procedure. Our goal is to provide an intuitive interface that minimizes cognitive load for the operator, while taking real-time capabilities into account. The Unity [94] game engine had been chosen as potential technology to be integrated, but this was subject to a thorough analysis and prototyping phase with different hardware which is not presented. Several reliable physiological measurements for cognitive load have been identified, necessitating the selection of appropriate hardware for prototyping and integration. The choice of mission scenarios guided the user interface development, which in our case have become search and rescue in form of the DARPA SubT challenge, and scientific multi-robot space exploration, for now in analog mission scenarios like BRAILLE.

3.1.2 Phase 2: Prototyping

WP3: Development of a prototype software infrastructure. We co-developed parts of NeBula [25] and contributed to an evolving system. Hence this autonomy platform has been chosen to implement and run experiments with. Initially, we performed simulations using ROS and Gazebo, followed by the implementation on real robots (e.g., Boston Dynamics Spots and Clearpath Huskys). Higher levels of autonomy and interaction primitives were implemented and exported to be used by the HCI, implemented in WP4. The robots were tested in underground spaces to collect data and verify the reliability of the overall infrastructure. At this stage, physiological measures were already integreated in order to measure real-time workload under different autonomy levels.

WP4: Implementation of a prototype and integration with the human-multi-robot system. Designing user interfaces for a dynamic, challenging system as the one envisioned demands to harmonize user requirements with new technologies, implementation constraints and new possibilities.

An initial user interface was designed based on the latest specifications of the human-multi-robot system and along the lines of the two mission scenarios (Chapter 4). The overall capabilities were improved in a second iteration of this design following initial requirements and insights of WP1 and WP2. We took into account the progress of the system's autonomy capabilities as those constantly advanced as part of the DARPA SubT challenge. This required early and often evaluation with users (and/or beta testers) to understand automation needs and tasks to be performed by either the robot, the human, or a mix thereof.

3.1.3 Phase 3: Validation

WP5: Evaluation via field study. In this WP, we planned to refine our design and perform a large scale system demonstration at different available test-sites. We used NeBula-powered robots for testing since the focus of this research revolves around the software architecture, ergonomics and interfaces. A large scale test of a human and multi-robot exploration system, combined with the assessment of a user interface based on physiological inputs and varying autonomy, had not been performed yet. User studies like [16, 86–88] were used as a role model in terms of measuring different objective and subjective metrics. We formulated a set of research question with respect to workload measurements, performance, situational awareness and trust, and assess these factors.

This approach was not without flaws as all imagined risk and mitigation strategies have

not foreseen a global pandemic and its impact on various planned field tests nationally and internationally; however, we continued the research under these constraints and present the works in this thesis.

3.2 Document Structure

This dissertation titled “Symbiotic Human and Multi-Robot Planetary Exploration Systems”, is submitted to Polytechnique Montréal in partial fulfillment of the requirements for the degree of Philosophiæ Doctor in computer engineering. It follows the approved “thesis by articles” layout for dissertations and includes published works as part of the thesis body. Chapter 1 introduces the background and motivation for this thesis’ work briefly. Chatpter 2 provides a literature review of relevant works, and Chatpter 3 outlines the research approach breaking it down into workpackages. Chapter 4 introduces the autonomy assistant Copilot MIKE as part of the published works body and addresses elements of WP1, WP2 with respect to screen systems, and WP4 as a first prototype. Chater 5 introduces a game-inspired interface which was deployed at the DARPA SubT challenge, addressing WP4 and, in part, WP5. Chapter 6 introduces a method for creating a fielded VR interface with real- time rendering capabilities and validates the approach with a user study during realistic large-scale planetary exploration scenarios in accordance with WP2 to WP5. Finally, Chapter 7 discusses the overall work, outlines potential future impacts, presents opportunities for future works, and provides concluding remarks.

CHAPTER 4 ARTICLE 1 - COPILOT MIKE: AN AUTONOMOUS ASSISTANT FOR MULTI-ROBOT OPERATIONS IN CAVE EXPLORATION

Preface: This work contributes an approach to a Multi-robot Interaction assistant for unknown cave environments. It establishes a human-robot interaction paradigm based on an adaptation of Sheridan's automation levels [22–24] and implements a novel system architecture that utilizes task planning and scheduling for collaborative human and multi-robot systems with the base station as dedicated agent in the HMRS. The system has been verified in simulation to qualitatively reduce operator workload and quantitatively decrease task execution times compared to scheduled task durations.

Declaration of Contributions: As the main author of this article, my contributions included: Conceptualization and implementation of a front- and back-end architecture for the proposed autonomy assistant for multi-robot operations in challenging environments, reviewing existing works in the literature on Human-Multi-Robot Interaction and Interfaces, Automation, Autonomy, and Planning and Scheduling, writing code, conducting simulation experiments, analyzing the data, preparing figures, and writing most of the article.

Full Citation: Kaufmann, M., Vaquero, T. S., Correa, G. J., Otsu, K., Ginting, M. F., Beltrame, G., & Agha-Mohammadi, A.-A. (June 7, 2021). Copilot MIKE: An Autonomous Assistant for Multi-Robot Operations in Cave Exploration. In 2021 IEEE Aerospace Conference (50100) (pp. 1-9). IEEE.

Available online: `https://doi.org/10.1109/AERO50100.2021.9438530`

Abstract – Operating a team of robots under time and risk constraints can be challenging for a human operator. Environmental conditions, extrinsic risks, and accessibility might restrict humans from directly partaking in exploration tasks altogether. Hence, robotic systems with autonomous exploration and disaster response capabilities have evolved over the past years and help keep human explorers and emergency response teams from harm. In this work, we introduce Copilot MIKE, an autonomous assistant for human-in-the-loop multi-robot operations. Copilot MIKE assists a single operator in monitoring robot teams, strategic planning, and communicating high level commands to the robots. During complex and potentially stressful exploration missions, Copilot MIKE helps to maintain a bearable workload and

high situational awareness. In this work, we mainly focus on cave exploration tasks in the context of the DARPA Subterranean Challenge (SubT), but we designed a generic assistant that can be used in other domains, such as search and rescue, science, and (space) exploration missions as well. Experimental mission runs were conducted in preparation for the SubT cave challenge and Copilot MIKE has been tested in realistic cave exploration simulations. We show that Copilot MIKE has the potential to reduce workload, while our operators place trust in the system. They report that they focused on important parts of a mission, rather than planning, adopting and memorizing a complete mission strategy themselves.

4.1 Introduction

Enabling initial human exploration of Mars and unmanned exploration of other planetary bodies has motivated space agencies and private companies to develop new robotic autonomy and exploration capabilities. The *2020 NASA Technology Taxonomy* [95], for example, outlines key technical challenges that must be addressed if we are to sustain a long-term human presence in space. In preparation for future human exploration of the lunar surface, *NASA's Plan for Sustained Lunar Exploration and Development* [96] proposes the use of robotic precursor missions to perform scientific investigations on the lunar surface, use the obtained scientific data to inform the design of in-situ resource utilization technologies, and prove grounds for future Mars missions. In addition, analogue missions and robotic competitions on Earth seek to advance new technologies and demonstrate capabilities for exploring harsh and challenging environments.

NASA JPL's Team CoSTAR [97], is developing new technologies that are critical for enabling autonomous multi-robot exploration of large and unknown underground voids. One example of the application of these technologies is the DARPA Subterranean Challenge (SubT) [98] where terrestrial cave exploration can be seen as an analogue exploration mission for planetary subsurfaces (e.g. Lunar and Martian caves), and as an application domain to prove grounds for future space technologies.

During SubT, the goal is to map and explore the structure of an a-priori unknown subsurface void in a time-critical scenario while identifying and localizing specific objects/artifacts that are placed in the environment (e.g. mannequin survivors, backpacks, cell phones, helmets, etc.). Figure 4.1 depicts a typical SubT scenario in which a mannequin survivor has to be detected by a robot. In the SubT competition, only a single human operator is allowed to interact with the robotic team. CoSTAR's robotic team consists of more than four robots with wheeled, legged, and flying mobility. Although controlling and overseeing the actions of each robot may seem feasible, operating multiple robots with different capabilities

Figure 4.1 One of CoSTAR's Spot quadruped robots deployed at Lava Beds National Monument with the NEBULA perception and computing payload on board (Credit: NASA BRAILLE)

in kilometer-long underground environments can go beyond the cognitive capacity of a single human supervisor – even with advanced autonomy in place. SubT operations may involve cognitively demanding tasks such as monitoring 3D mapping of the environment and localization accuracy, establishing communication links between robots, assessing location and health of all robots, and submitting detected artifacts within the allotted competition time.

Our overarching goal is to develop a complementary autonomous system that guides and assists a human operator to explore extreme terrains by means of autonomous multi-robot exploration. An autonomous system can assist to reduce operator workload and perform tasks that would otherwise be ignored in situations where the operator is occupied with assigning high-level exploration goals, and manually solving issues that require human intervention. Given the risks associated with exploring unknown and hazardous environments, we are motivated to develop new human-robot architectures that integrate autonomy and robotic assistants into operations (e.g., construction, habitation, exploration) that are constrained by time, human and robot personnel, and available in-situ resources. Implementing a distributed multi-robot system that astronauts can use and collaborate with during these operations may

reduce an astronauts' overall workload for future missions [95, 99].

In this work, we introduce Copilot MIKE (Multi-robot Interaction assistant for unKnown cave Environments) as an autonomous assistant for exploring extreme terrains. Our contribution is a novel system architecture that enables collaborative human and multi-robot systems. Copilot MIKE aims to reduce operator workload by (1) providing situation awareness with respect to robot teams and the environment; (2) actively monitoring key aspects of the mission progress; (3) providing support for decision making processes regarding task planning and scheduling; and (4) helping to create communication network infrastructure and coordinate communicate node deployment. These features are integral as exploration missions increase in complexity and difficulty.

4.1.1 Related Work

Human-Multi-Robot Interaction and Interfaces

When it comes to developing machine interfaces and interaction modalities for human-robot systems, it is important to consider domain specific and capability constraints. Search and rescue, disaster response, as well as space exploration missions can pose similar requirements on such systems. Similarly, the availability of power, portability, and size constraints have to be considered [12]. The SubT competition rules impose requirements on all teams regarding the machine interface design. Two major requirements for SubT and our work are (1) to have a single non-interchangeable robot operator, and (2) to maintain all mission operations within a designated area for the entire mission duration. Additional competition requirements limit the support the operator may receive from the pit-crew to only readying robots. The operator is therefore the only member allowed to see all robot parameters and information collected about an a-priori unknown environment. The number of simultaneously deployed robots impacts operator's performance [100], hence user interface (UI) development aims at improving the operator-robot ratio [101]. In search and rescue scenarios studied by [13], operator performance is at its peak when a single operator gives their full attention to a single robot. Additional works suggest that more personnel per robot is needed to split tasks between robot control and mission task execution so as to not overwhelm the primary human-in-the-loop [11].

Automation and Autonomy

Achieving full or high levels of autonomy in deployed multi-robot systems is a key motivation in today's robotic research, but is difficult to achieve [101]. Autonomy capabilities in

such systems, including space systems, are usually classified in a range between complete tele-operation and full autonomy [99]. Such classification scheme has long been introduced in Sheridan et al.'s ten levels of automation ranging from no automated assistance to full autonomy that ignores human input [22, 23].

Planning and Scheduling

Schedulers have been used for several rover and satellite missions. Examples include OASIS, Mars Science Laboratory, Deep Space One, and the recent Mars 2020 Perseverance Rover that is currently in transit to Mars [102]. Different scheduling techniques, such as fixed cadence, event based scheduling, and hybrid approaches are discussed in [102]. While computational constraints can dictate planning and scheduling frequency for robotic missions, a common scheduling objective is to regain dead mission time which in return maximizes mission performance.

The remaining paper is structured as follows. First we provide an overview of the problem that we address in section 4.2. Next, we describe the proposed approach and provide details on Copilot MIKE's architecture and its underlying components in section 4.3. In section 4.4 we then describe preliminary experimental results from a representative simulated cave exploration scenario. section 4.5 finalizes the paper with a conclusion and promising future work.

4.2 Objectives

In robotic subterranean exploration, a typical goal is to autonomously and accurately map an unknown environment. Finding and locating phenomena (e.g., bio-chemical traces of life) or objects of interest can be a parallel goal during such an exploration mission. As robots map and identify features of interest in the environment, the resulting data products have to be retrieved and transmitted to a base station for further analysis or dissemination (for instance, in a cave environment, a base station could be located at the entrance). Ideally, human intervention during exploration is kept to a minimum - in the case of the SubT Challenge for example, only a single supervising operator is allowed to interact with the team of robots.

Current technology still requires some human intervention during exploration. Examples include deploying and coordinating heterogeneous robots with multi-modal motion capabilities (driving, walking, flying), resolving system failures and critical events, monitoring system health, and managing strategic mission planning and scheduling. All of these intervention

tasks can be overwhelming during operations, especially in time-critical missions. As an example of time-criticality, a typical robot exploration mission in the SubT challenge has a duration of 60 minutes with an additional 30 minutes of setup and preparation time. It is crucial for the operator to be fully focused and alert during the whole mission duration; this includes the preparation period where the operator must not forget important tasks. Results from previous SubT circuits, namely the tunnel and urban circuits, have shown that a single operator can be preoccupied with certain tasks while omitting others completely [103]. Not all tasks are equally important and brief multi-tasking interactions might be required due to risk, safety, resource, and time constraints. This imposes challenges on an operator's work and cognitive load as mission strategies are kept in mind and adapted for multiple robots in highly dynamic scenarios. A swift understanding of the current robot system and mission statuses is key for strategic decision making and mission success. When designing autonomous systems, it is important to consider that autonomy can be unintuitive to an operator and thus lead to confusion [99].

The long term objectives for developing Copilot MIKE are: (1) to reduce the operators' cognitive workload and working memory load, (2) to increase mission performance with respect to the explored area, maintaining communication links and submitting artifacts, while keeping interventions to a minimum, and (3) to increase situational awareness.

We aim to develop Copilot MIKE to present crucial information (tasks) to the operator while taking over some decision-making processes regarding planning and scheduling, spatial coverage and positioning, deployment of communication nodes, and monitoring of the robots' and system's health. When to prompt the operator and with what urgency is determined by Copilot MIKE, while the operator can always access system-wide information through a dedicated user interface. Copilot MIKE can convey information and suggestions and will intervene when critical decisions have to be made promptly.

4.3 Technical Approach

In this section, we describe key features of the proposed autonomous assistant for robotic subsurface void exploration, the Copilot MIKE. We first present its system architecture and information flow.

4.3.1 System Architecture

Our autonomy assistant's system architecture is illustrated in Figure 4.2. Copilot MIKE operates using five main components: (1) a task manager (2) assistive capabilities (3) a

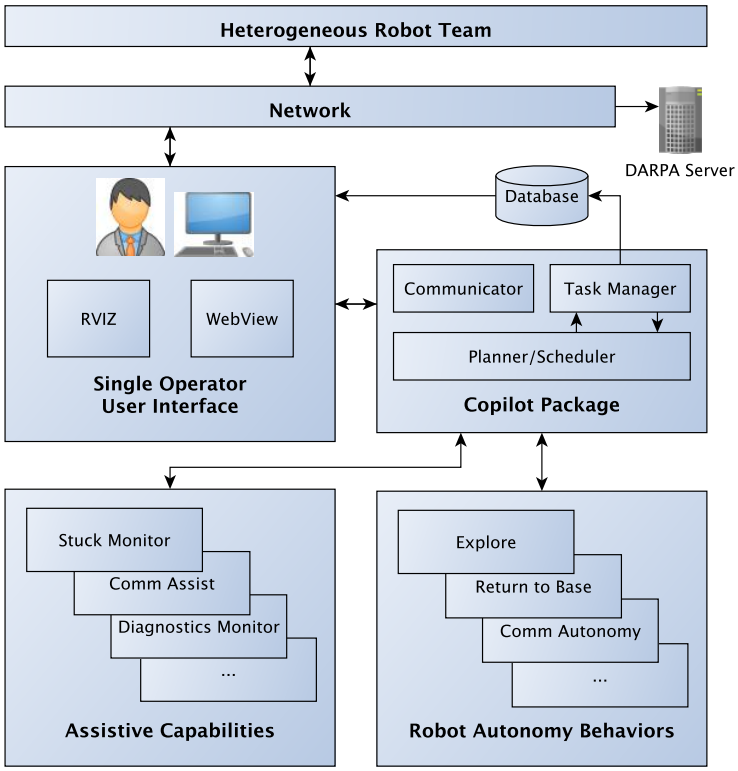

Figure 4.2 Copilot Architecture and Task Information Flow

scheduler (4) a communicator and (5) a user interface. Each component is implemented using ROS (Robot Operating System) and web-based technologies similar to [103]. The following sections describe each core component in detail.

4.3.2 Task Definition and Task Manager

A “task" has been defined as an atomic piece of work that is managed and distributed by the Copilot system. Each task is defined using the following properties:

(1) an ID. (2) a name. (3) a description. (4) an optional prompt message for notifications. (5) a status indicating if a task is pending, in-progress, succeeded, postponed, or failed. (6) the requesting agent. (7) the requesting assistive capability is also encoded to identify the origin of a task. (8) the agent for which this task assignment is addressed (e.g., operator, pit-crew, or robot). (9) a priority ranging from one (low) to ten (high). (10) a start time-bound (i.e. earliest and latest start time). (11) a deadline by which the task should be completed. (12) an estimated duration of the activity suggesting that the operator or the assistant will have to perform the task within this time. For example, when the assistant

asks/suggests the operator to do something that may take some time (e.g., check the status of *robotA* given its high-risk tilt). (13) whether a response is needed from the operator for the system to carry/conduct a particular activity, including (14) a set of pre-defined user responses (options). (15) a set of resources that are consumed or occupied during task execution. (16) an autonomy level, defining to which extend the task can be automatically performed by the assistant (e.g. if the operator does not respond by the deadline).

Tasks can either be generated statically for the different mission phases or dynamically by assistive capabilities at certain system states. Predefined tasks can be initialized at the beginning of an exploration mission (e.g., check sensors, start exploration mission, enable system logging) and executed later on. Copilot executes a task either by prompting and interacting with the operator or by autonomous execution if certain system conditions are met. All task statuses (e.g., pending, in-progress, succeeded, failed, delayed) are managed by the *task manager* module and dynamically updated on the database. The task manager interfaces with the task planner and scheduler to retrieve time-critical information that determines when the operator or Copilot will execute the given task.

4.3.3 Assistive Capabilities

An assistive capability is a distributed component used across the entire system. It is responsible for (1) detecting anomalous events, (2) requesting a resolution task to the Copilot core module, (3) monitoring the resolution of an anomaly, and (4) resolving the anomaly if requested by the Copilot or operator. Each assistive capability runs independently without knowing other system status, allowing a modular extension of autonomy functionality. Assistive capabilities are implemented to augment the operator's capabilities and address human shortcomings through Copilot. Examples of assistive capabilities include monitors for communication links between the robots and the base (e.g. communication loss), robots proximity (i.e. risky distances between robots), robot's unsafe tilts, robot's sensors status, and robot's mobility status (e.g. stuck or undesirable oscillations).

The Copilot system is designed to support varying levels of system automation with respect to the assistive capabilities. Table 4.1 is a modified definition of Sheridan's automation levels [22–24] described using terminology coherent with our work. Each autonomy capability can be assigned to one automation level to support operations, and the Copilot scheduler determines the final automation levels depending on the operator's load. For example, if the operator load is low, Copilot reduces the task automation levels to achieve an accurate and flexible resolution of issues by the human operator. On the other hand, if the operator is busy with critical tasks, Copilot increases the automation level of trivial tasks so that the

Table 4.1 Ten Levels of Copilot Autonomy

High	**10**	Copilot decides everything, full autonomy, disregards the operator
	9	Copilot executes, informs the operator only, if it decides to
	8	Copilot executes, informs the operator only if requested
	7	Copilot executes autonomously, then informs the operator
	6	Copilot allows operator input until a deadline, then executes default
	5	Copilot executes a suggestion, if operator approves
	4	Copilot suggests an alternative to the operator
	3	Copilot narrows down the options presented to the operator
	2	Copilot offers a complete set of decisions/actions
Low	**1**	Copilot offers no autonomy: all decisions and actions taken by operator

operator can focus on more important tasks.

Allowing flexibility in task automation has multiple benefits. First, it helps the scheduler to find a feasible plan that fits within the designated task deadline. Different automation levels typically require different amounts of time and resources for the same task. That gives the scheduler the freedom to allocate less time on simple low-priority tasks to open up time for complex important tasks without sacrificing the overall mission performance.

Next, it makes it easier to support various task types needed for complex robotic missions. Two major task categories in human-in-the-loop operations are *humans-better-at-it* and *machines-better-at-it* [12]. Most monitoring tasks fall into the *machines-better-at-it* category. For instance, the *tilt monitoring* assistive capability is a safety-critical capability that can consume a lot of attention and cognitive load if it is performed by the operator. The tilt monitor capability can be implemented with higher automation levels, such as level 7 (autonomously pausing a mission execution), or level 5 (prompting the operator to initiate a safety stop) to keep the vehicle safe with minimal operator's attention. On the contrary, there are tasks where a human's input is highly valuable. For example, finding the best places to drop communication nodes is a highly strategic decision-making process that requires careful consideration of mission progression, resource management, wireless signal coverage, traversability, and so on. Relying on human's high-level reasoning capability can drastically reduce the complexity of the system.

Increasing the automation levels is straightforward as the autonomy technology matures over time. In [104], we reported the autonomous communication drop capability that attempts

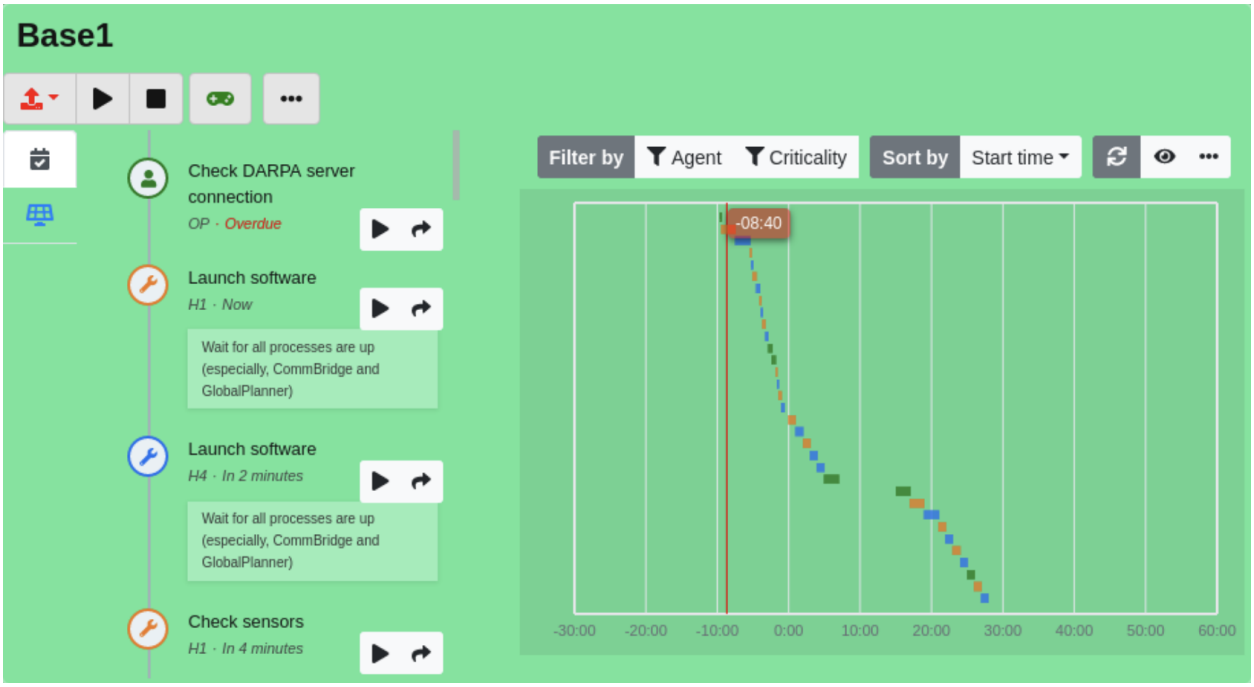

Figure 4.3 The user interface component for Copilot MIKE

to increase the automation level to 8 or higher. Copilot is flexible to accept individual task specification changes due to technology advancement.

4.3.4 Scheduler

One of the key autonomy features of Copilot MIKE is the ability to dynamically schedule tasks, based on their temporal constraints, dependencies, resource constraints, priorities, autonomy level as well as the current execution status (e.g. tasks that are in progress).

In this work, we designed and developed the scheduler component to be modular, meaning that one can plug and use different scheduling techniques and algorithms and apply any desire utility function. For example, one of the scheduling approaches we provide refers to a simple scheduler based on a Linear Program (LP) (implemented using the PuLP [105] library) that consider tasks temporal constraints and dependencies only while minimizing the makespan. Other approaches would consider for example prioritization and resources while minimizing both makespan and operators workload (e.g. deciding to autonomously execute tasks on behalf of the operator).

To handle uncertainty on task execution (e.g. delays and failures) and new task arrivals, the scheduler can either reschedule the tasks on a fixed cadence or when certain events occur (e.g. new task arrives or task is delayed beyond a certain threshold). In the default setup,

we run the scheduler in fixed-cadence mode, e.g. every 2 seconds.

The development of different scheduling approaches and the study of how they change Copilot's overall behavior and performance is an active research topic in this effort and is out of the scope of this paper.

4.3.5 Communicator

The communicator portion of Copilot handles communication modes that go beyond the UI capabilities. For instance, it handles direct user notifications via system messages for critical events. Communicator also relays information back to the assistive capabilities and other system components that require information. For instance, it handles operator task responses and disseminates those to trigger the execution of new activities, tasks and notifications. An important feature of communicator is that it prioritizes and limits the number of prompt messages that are provided to operators to avoid overloading.

4.3.6 User Interface

Improving the operator-robot ratio and cognitive load can be partially achieved through user interfaces (UIs). It is important to effectively communicate critical mission information to an operator, as well as the operator's intent and commands back to the multi-robot system. We build upon an existing infrastructure that leverages ROS and web-based technology as described in [103]. The previous interface seemed to fight for the operators attention, especially when monitoring multiple robots.

Figure 4.3 illustrates Copilot MIKE's new and improved UI interface component. This component is placed on top of the existing web interface and reduces the need to interact with each individual robot component in the web view. Located on the left of the Copilot UI component, is an interactive and chronological task list that includes execution instructions and a button interface to execute or stop the task. On the right is a complete task schedule displayed in the form of a dynamically generated Gantt chart. Symbols and color coding assist to reduce workload and enable effective machine-human communication. Representing the complete mission in a single view also allows an operator to maintain their working memory capacity at a lower level. Although the interface is adaptive and updates frequently, the new design aims to prevent the operator from being overwhelmed as new information appears. Figure 4.4 shows the single operator single screen setup that is used during field experiments and simulations.

4.4 Experimental Results

Copilot MIKE was demonstrated in a series of robotics operations to understand the influence of adding an autonomous assistant to the overall system. In this paper, we mainly focus on Team CoSTAR's multi-robot operations in realistic cave simulations which serve as preparation for the next DARPA Subterranean Challenge occurring in natural caves.

4.4.1 Cave Simulation Environments and Setup

The simulation experiments are conducted in synthetic cave models from the DARPA Subterranean Challenge repository and in a reconstructed real-world cave model from the NASA BRAILLE cave data. The simulation course covers multi-km^2 areas and requires rapid robot team exploration to explore the whole course under limited time constraints and to search for artifacts. Figure 4.5 shows one of the cave simulation environments.

In this experiment, a human operator supervises a robot team in the SubT operation. We experimented with different robot configurations that consist of ground and aerial vehicles. We ran 10 simulated cave exploration mission scenarios where the human operator is tasked to execute a set of predefined and dynamically generated tasks guided by the Copilot. The mission operation duration in this experiment is set to 30 minutes with 10 minutes of robot preparation time.

4.4.2 Results

Our system architecture can update and record time-critical information of ongoing tasks that are scheduled and managed during a mission. This information is used to understand how well the operator works *together* with the Copilot during simulation and influences the baseline capability to be used during real-world exploration missions. Statistical results obtained from our simulated experimental mission runs also assist to influence the user-experience of new and existing features. Herein we focus on one of the representative runs.

Figure 4.6 compares the scheduled and actual task start times in subplot (a) and the task duration of predefined and dynamically generated tasks in subplot (b). In this analysis, we can easily identify misalignments between scheduled and planned task start times, as well as the planned and actual task execution durations. A small time difference between the scheduled and actual start times in subplot (a) indicates that Copilot is successful in scheduling feasible start times since the operator can execute the tasks according to the schedule and on time. Another example includes pre-scheduling the tasks to start the logging mechanisms on the robots ("Start robot logging") appears to be timed within an acceptable

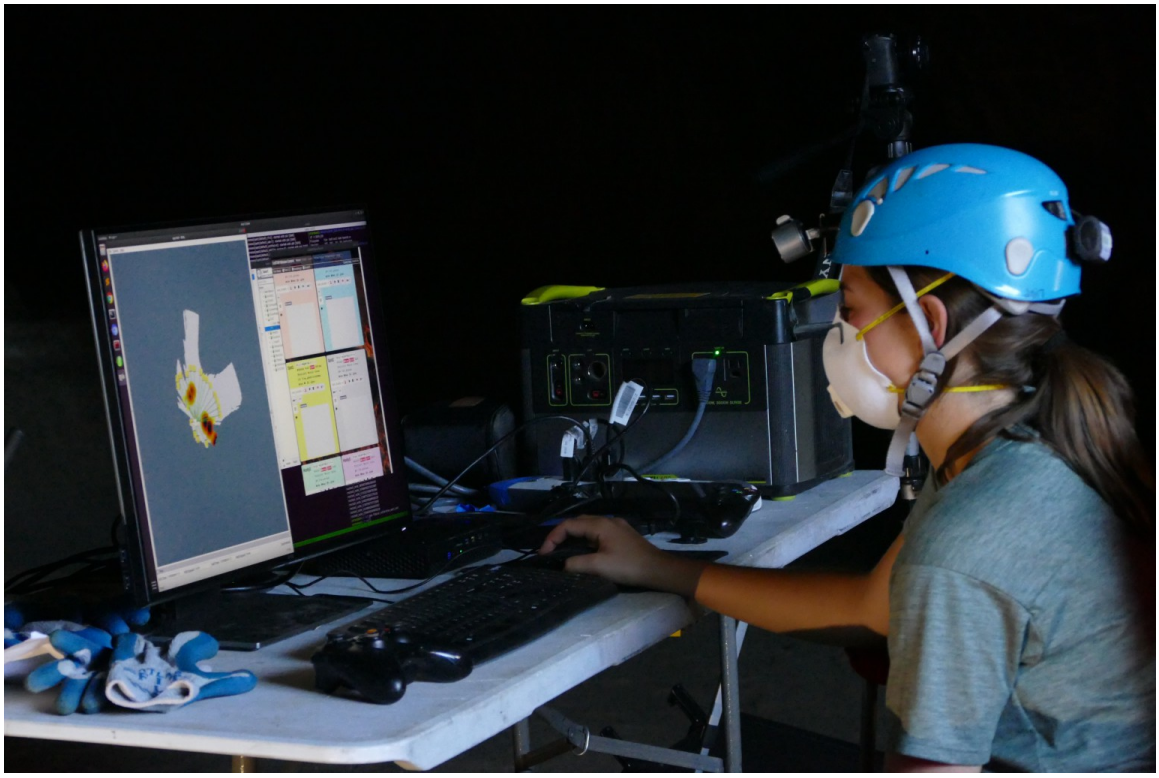

Figure 4.4 Single Operator Base Station Setup with 3D map and web interface. (Credit: NASA BRAILLE)

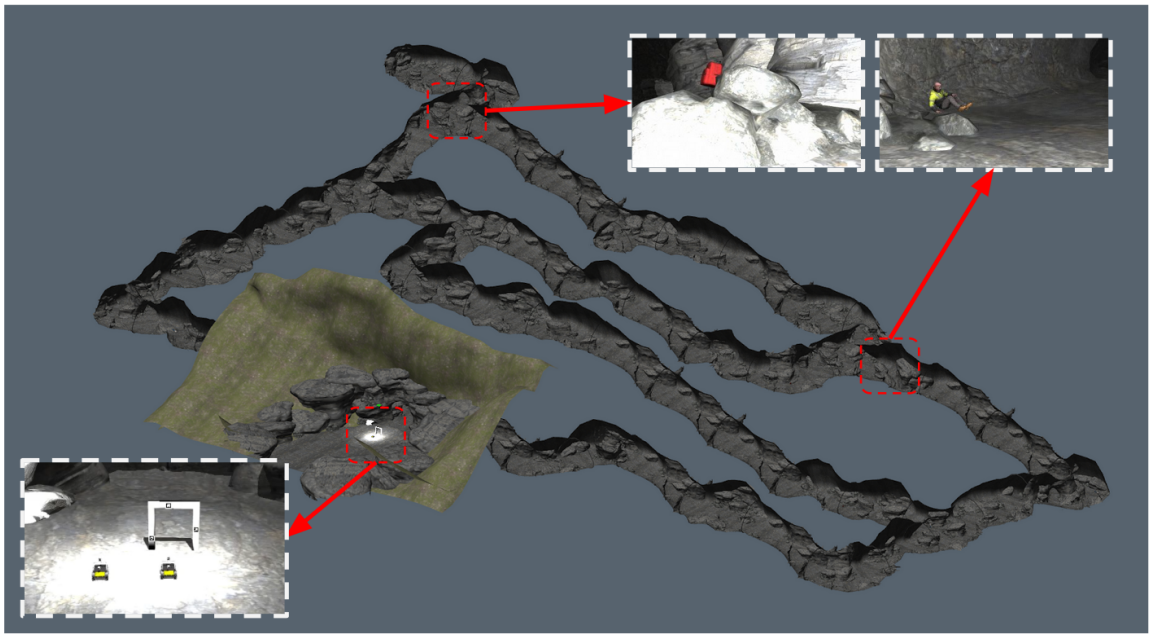

Figure 4.5 Multi-robot operation in a simulated cave environment

buffer; actual execution duration and planned duration match well. In comparison, the task of selecting a leader robot ("Set leader robot") requires much less time than planned. In other cases, such as when the Husky robot is tilted, an assistive capability can decide to resolve the task autonomously or change the task status so that the task can be disregarded (see auto resolved "Tilted! husky#"). The analysis also reveals, that in this particular run tasks to stop some logging mechanisms were skipped and hence have no actual start time.

These preliminary empirical results are crucial to further optimize our mission strategy and redefine timing specifications as needed. Learning methods and on-line optimization techniques could eventually be implemented to tune time-critical task parameters throughout multiple mission runs.

4.4.3 User-experience Feedback

After performing the mission runs, the operators provided feedback on their experience using Copilot MIKE to assist with scheduling and executing tasks. The operators' feedback was positive overall and gave early insights into how we can design autonomous assistants that help reduce the operators' experienced workload. The following are some of the operators' comments about their first experiences working with Copilot MIKE:

- "Copilot helps to focus on the task at hand, and not much on (remembering) what needs to be done next."
- "It decreases the chance to avoid missing information, especially while commanding multiple robots."
- "Copilot makes the operator rethink mission preparation strategies."
- "I trust the system and just want to follow the Copilot!"

These comments bring into light key aspects of the human-multi-robot experience such as ease of operation, actual utility in comparison to systems without copilots, and trust in the autonomous assistant's capabilities; such components demand further empirical investigation as the operator becomes dependent on offloading tasks to the Copilot.

The following observations suggest areas of improvement on the user-experience and multimodal interface which must also be investigated further:

- *It is important to adjust the schedule and dynamically reason about what tasks can be done autonomously.* Because task execution time may vary during exploration, it

is essential to have the Copilot provide flexible rescheduling and execution capability. During overwhelming scenarios where an operator cannot catch up and falls behind in executing a large number of simultaneous tasks (or a few tasks with large cognitive demand), it is also required for the system to decide when to autonomously execute the task on behalf of the operator, or if time is limited, to re-allocate it.

- *When issues occur (e.g. robot system failures), the operator might need to intervene to resolve them and refrain from executing upcoming tasks in the list.* There is an interesting need here to help the operator keep focus on the task at hand, but at the same time help maintain awareness of the mission progress and other eventual critical tasks. Other communication modalities could support this process.

- *Using the Copilot system with an initial predefined list of tasks (checklist) makes the operator think about the timing of tasks more explicitly.* In most cases, operators' have an underlying strategy of how to approach a new exploration scenario; which in our system is in the form of a predefined set of tasks. Feeding those tasks to Copilot allows the system to support the user's unique execution style and keep track of task execution and timing information. Such information helps to adjust task duration estimations and avoids over- or under-estimation of temporal constraints.

4.5 Conclusions

In this work, we introduced Copilot MIKE, an autonomous assistant designed to support human operators in multi-robot exploration of large subsurface and extreme terrain environments. We described the overall architecture of our system, its main components, and the process by which the proposed system manages and autonomously schedules tasks. We presented preliminary experimental results obtained by using Copilot MIKE in realistic simulated cave exploration scenarios with multiple robots. The experiments and operators' feedback highlight promising initial results of how the system can reduce overall workload and improve the operator's focus for executing tasks during mission preparation and exploration activities.

The current state of this work prepares Team CoSTAR with new autonomy capabilities and features. Copilot MIKE has the potential to be deployed in several human and multi-robot domains, including search and rescue operations and future space missions. In the near future, we plan to explore a variety of scheduling algorithms and flexible execution approaches to handle environmental changes including event-driven, and hybrid approaches between fixed

cadence and event-driven rescheduling. Moreover, we will work on the modeling of human physiological and robotic resources that can be incorporated into Copilot's reasoning.

Acknowledgments The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This work was conducted in collaboration with the Making Innovative Space Technologies Laboratory (MIST Lab) at Polytechnique Montreal. The first author would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for their generous support in form of a Vanier Canada Graduate Scholarship. Further, the authors would like to thank our former team member Dr. Michael (MIKE) Wolf for his contributions to the SubT project and all members of team CoSTAR.

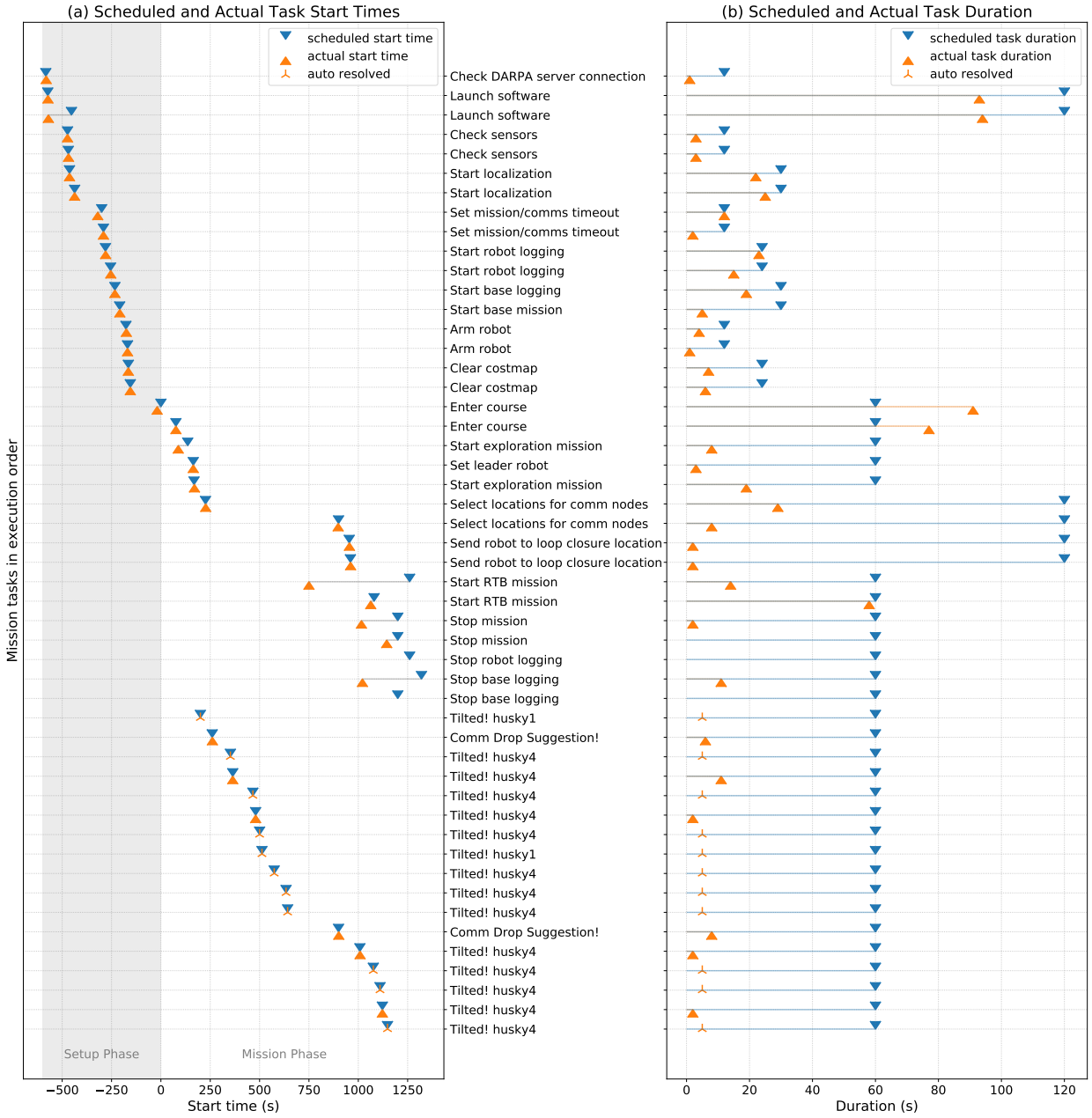

Figure 4.6 (a) Comparison of scheduled and actual task start times for a single experiment run and (b) the corresponding planned and actual task durations.

CHAPTER 5 ARTICLE 2 - COPILOTING AUTONOMOUS MULTI-ROBOT MISSIONS: A GAME-INSPIRED SUPERVISORY CONTROL INTERFACE

Preface: This article presents a game-inspired user interface for multi-agent robot missions, which expands the work of the previous article. It identifies shortcomings, and addresses these by (i) integrating an automated planner for task planning and scheduling with resource constraints, (ii) creating a framework for verifiable task execution for increased reliability, and (iii) presenting verification results on how the overall system performed over the course of several real-world deployments, including the DARPA SubT Challenge final.

Declaration of Contributions: As the main author of the article, my contributions included: Reviewing prior and existing works, proposing and implementing a verifiable and generic task framework, informing the design and implementation of the frontend, writing the code for the back-end and some frontend modules, gathering and analyzing the data, creating figures, and writing a first draft of the document and reviews.

Full Citation: Kaufmann, M., Trybula, R., Stonebraker, R., Milano, M., Correa, G. J., Vaquero, T. S., Otsu, K., Agha-Mohammadi, A.-A., & Beltrame, G. (May 19, 2022). Copiloting Autonomous Multi-Robot Missions: A Game-inspired Supervisory Control Interface. In ICAPS 2022, the 32nd International Conference on Automated Planning and Scheduling, 2022 Workshop on Scheduling and Planning Applications (SPARK), Singapore, Republic of Singapore, 7-12 June 2022.

Available online: `https://icaps22.icaps-conference.org/workshops/SPARK/papers/spark2022_paper_8.pdf` (last accessed 08/2024), `https://ai.jpl.nasa.gov/public/documents/papers/kaufmann-et-al-SPARK2022.pdf` (last accessed 08/2024), `https://arxiv.org/abs/2204.06647` (last accessed 08/2024)

Abstract – Real-world deployment of new technology and capabilities can be daunting. The recent DARPA Subterranean (SubT) Challenge, for instance, aimed at the advancement of robotic platforms and autonomy capabilities in three one-year development pushes. While multi-agent systems are traditionally deployed in controlled and structured environments that allow for controlled testing (e.g., warehouses), the SubT challenge targeted various types of

unknown underground environments that imposed the risk of robot loss in the case of failure. In this work, we introduce a video game-inspired interface, an autonomous mission assistant and test and deploy these using a heterogeneous multi-agent system in challenging environments. This work leads to improved human-supervisory control for a multi-agent system reducing overhead from application switching, task planning, execution, and verification while increasing available exploration time with this human-autonomy teaming platform.

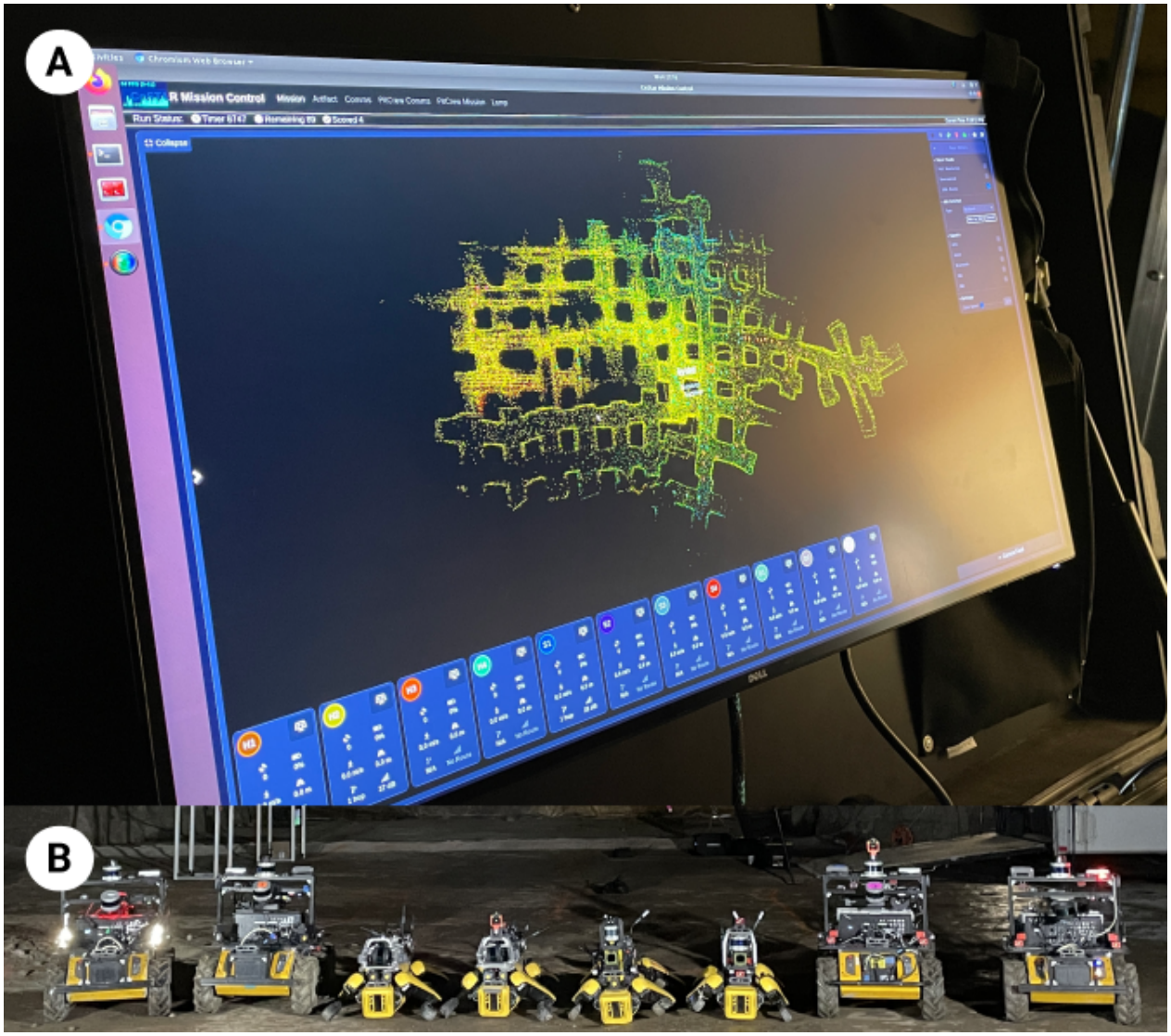

Figure 5.1 Team CoSTAR's Mission Control user interface (A). (B) a subset of CoSTAR's ground robots showing four customized Boston Dynamic's Spot and Clearpath Husky powered by JPL's autonomy platform NeBULA. Typically a deployment of 4 to 6 ground vehicles was targeted during SubT, but the number of agents is extendable (e.g., see A with 11 robots).

5.1 Introduction

Autonomous Exploration and SubT: Robotic exploration and the advancement of autonomy offer new ways to explore potentially dangerous and hard-to-access underground environments. Multi-agent systems have matured in controlled and structured environments like warehouses, factories, and laboratories, while current robotic challenges seek to advance these technologies for search and rescue scenarios, planetary prospecting, and subsurface exploration [106–108]. Motivated by the search for life on other planets, NASA JPL's team CoSTAR [109] took part in the Defense Advanced Research Projects Agency's (DARPA) Subterranean Challenge (SubT) seeking to advance robotic multi-agent systems and their technology readiness for potential future missions. If brought to other planets (e.g. Mars), subsurface missions could bring new insights into their geologic past as well as on their potential for supporting life in the environmentally protected undergrounds [110]. In contrast to traditional exploration missions where a team of operators and scientists controls one rover, SubT introduced the challenging requirement that only *a single human supervisor* can directly interface with the deployed multi-agent team in real-time and when a communication link is established. SubT is divided into three, one year development pushes with major field testing demonstrations. This work focuses on the advancements in our supervisor autonomy and game-inspired user interface that were developed under the restrictions of a worldwide pandemic and deployed during the SubT final competition comprising two preliminary missions (P1 and P2) and the final prize run (F).

Human-Robot Collaboration: Achieving man-computer symbiosis [8] has been a long-time goal of the community to promote a close coupling of human and machine capabilities and ultimately inspire the evolving field of human-robot interaction [111]. This work improves collaborative human multi-robot exploration and search performance fusing our extended autonomy assistant Copilot [112] that uses automated planning techniques with a game-inspired interface design for effective robot deployment, operations, and single operator supervision to create a more symbiotic interaction.

We present key design choices that are breaking away from common robot interfacing strategies that were deployed in similar challenge contexts [113,114] and used interfaces based on the Robot Operating System's (ROS) visualization tool RViz. Further, we leverage human-robot inter-dependencies to inform the design and development of supervised autonomy and interaction paradigms to achieve our set interaction objectives. The latest results from the SubT competition "Finals" are compared to a baseline from previous competition runs, namely the "Urban Circuit", which deployed earlier interface and system implementations and interaction paradigms that we improve with our combined game-inspired interface and enhanced

supervisory autonomy.

5.2 Related Work

Human-Robot Interaction and Interface Design: More than sixty years after the introduction of man-computer symbiosis by *Licklider* [8], *Chen and Barnes* [111] conclude that the boundaries of long-term human-robot symbiosis are still to be pushed by interdisciplinary collaborations. *Szafir and Szafir* [115] have identified best practices in the field of data visualization as a key driver to advance both HRI and data visualization. Complex visualizations and renderings have become achievable with off-the-shelf hardware, which allows the integration of visualization principles such as sensemaking [115] that helps a human digest information. In human-space systems *Rahmani et al.* [116] identified that interface technologies are currently in development, but their technology readiness levels are not very mature. Multiple design methods have been introduced in the literature, for instance, Coactive Design [117] which is a structured approach to analyze human and robot requirements and was used in the context of the 2015 DARPA Virtual Robotics Challenge that aimed at advancing disaster response capabilities. *Roundtree et al.* [118] found that abstract interface designs that visualize collective status over single agent information could increase performance; however such designs depend on the task at hand, team size and mission goals [111]. A common testing strategy in computer game development is Playtesting [119], which is comparable to simulation and field testing in the multi-robot domain. The game-inspired development technique RITE, which was introduced in the context of interface development for the computer game Age of Empires [120], was used and adapted for fast development sprints. Additionally, we drew inspiration from real-time strategy games like Age of Empires, which guided the design of the 3D portion of the interface.

Robot Challenge Interfaces: During 2013's DARPA Robotics Challenge, team ViGIR leveraged ROS to control a humanoid robot. The team decided to implement their interfaces using RViz and built an Operation Control Center consisting of at least six screens. Robot challenges are found to typically influence human-robot interaction design and interfaces [115] and for DARPA's SubT teams, the common design practice was based on RViz and ROS plugins ([114, 121–124]). Even our team started off using RViz as a quick way to prototype interfaces [125] and used it as the main way to interact with the robot agents due to its tight integration with ROS and ability to access robot data for debugging purposes. We shifted away from this approach for the final competition, and the resulting HRI modalities and supervisory interface are presented in this work.

5.3 Background and Objectives

Challenge Requirements: The overall SubT goals are two common problems faced by real-world multi-agent systems: first, the autonomous exploration of unknown environments, and second, the search for objects of interest hidden within. While exploration and search provide a need for specific capabilities, DARPA further introduced a set of guidelines and rules to motivate higher levels of autonomy for the deployed systems: (i) only a single human operator is allowed to interact, supervise, and interface with the robots; (ii) each mission is bound by a fixed *setup time* limit of 30 minutes and an *exploration time* limit between 30 and 60 minutes; (iii) a pit crew of four (Finals) or nine (Urban Circuit) can support the supervisor by setting up hardware in a designated area without access to wireless data streams, robot control, or interface; (iv) there is a limited number of attempts to submit discovered objects of interest; (v) the final challenge environment comprises tunnel, urban, and cave terrains to be explored.

Objectives: Deploying and operating large teams of robots like Team CoSTAR's robot fleet, shown in Figure 5.1B, are complex real-world problems. Addressing this set of problems creates the need for a resource-efficient and robust human and multi-agent system to i) not overwhelm the single human supervisor, ii) meet the timing requirements, and iii) increase the performance of both exploration and search tasks.

To tackle this challenge and develop a system that can deploy reliably even beyond the SubT challenge, we embed the following interaction objectives into our system design: (1) Reducing overhead and human workload (e.g., from application switching and manual task execution) (2) Creating and maintaining situational awareness (3) Managing large teams of robots (from setup, deployment to exploration) while allowing for a flexible configuration (4) Accessing critical information in a single unified interface (5) Maintaining an enjoyable performance that can visualize the complete robot team (6) Collaborating with autonomy and trusting automation.

5.4 Supervised Autonomy

5.4.1 Copilot

Motivation: After SubT's "Urban Circuit", the allowed personnel in the competition staging area was reduced from ten to five team members which includes the main supervisor. This required a shift in how robots were strategically and physically handled (minimum 2 people are needed to lift and stage a single robot). Task coordination was done by a pit crew

member directing the operator and influencing their actions while following static paper *checklist procedures*. Developing and deploying a computerized assistant that could take over this role was soon desired.

Original Implementation: A first version of Copilot, "an autonomous assistant for human-in-the-loop multi-robot operations" was introduced in [112]. This early Copilot was only tested in realistic cave simulations or during preparatory missions with one deployed robot. Copilot supports a single human supervisor in monitoring robot teams, aids with strategic task planning, scheduling, and execution, and communicates high-level commands between agents and a human supervisor if a communication link exists. The autonomy assistant aims at keeping workload acceptable while maintaining high situational awareness that allows rapid responses in case system failures are observed.

Task Interaction: Copilot takes over the decision-making processes regarding planning and scheduling, which reduces the need to memorize tasks and task sequences or the need to delegate a team member to take over such checklist-like tasks. Some tasks were implemented with higher autonomy levels and automatically executed limited actions, but most required the human to start the task, manually execute parts of it, and confirm that the task had been completed successfully or unsuccessfully while monitoring the system. On one hand, it reduced the need to remember tasks; on the other hand, more interactions with the newly introduced system were needed.

Scalability Limitations: Due to computational limitations, a full mission simulation could not be achieved with more than three robots at reduced real-time and not more than two in real-time. However, upon tightly integrating Copilot with multiple real robot platforms, we noticed that the current concept of operations didn't scale well when adding more robots to a mission. We learned that task execution on the real hardware requires different timing and introduces many sources for machine and human errors (e.g., if cables are loose, sensors don't power up, or unknown unknowns occur).

Visualization Limitations: In robotics interfaces, scheduling, and timeline views are often presented in a robot- or task-centric way, focusing on who or which agent is scheduled for a certain task and when, respectively [126]. The main task-centric approach that was used in early Copilot tests showed a vertical list view with a scrollable timeline. This timeline showed the four tasks closest in time on top. As the number of tasks scaled linearly with the number of deployed robots this list view became inefficient — especially when tasks had to be deferred and worked on in a non-sequential order.

5.4.2 Improved Copilot

The identified shortcomings motivated a redesigning and rethinking of Copilot's back-end and front-end to reduce and not just shift workload; thus, we implemented higher levels of automation.

Architecture Changes: Figure 5.2 provides a simplified overview of Copilot's updated task management architecture. A multi-robot task auto-generator and verifiable task executor have been added to the system, and the underlying planner has been replaced. All modules access a centralized task database which stores pending, active, successful, or failed mission tasks for setup, deployment, and during exploration.

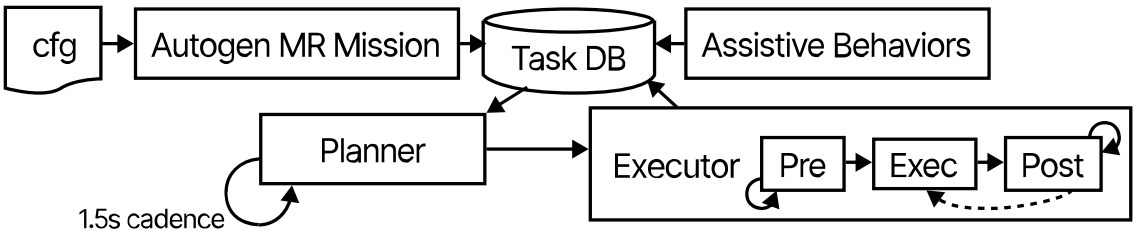

Figure 5.2 Copilot's task management architecture. Auto-generator, Planner, and Executor have been added or updated and access a centralized task database which stores pending, active, successful, or failed tasks.

Task Dependency Graph: A robot mission can be fairly complex, even when looking at the deployment of a single robot. In Figure 5.3 such a single robot mission is shown as a directed graph indicating the temporal constraints and execution dependencies with arcs between the nodes that represent a pre-defined set of mission tasks. Each task is defined by its duration, earliest start time, latest end time, and its dependency relations with other tasks.

To deploy multiple robots without the need to hard-coding all possible agent combinations and graphs, we use a scalable auto generator. The preceding superscript O in the graph (see Figure 5.3) indicates that human inputs or actions are required for the task. In the case of the `Launch base software` task, this means that the operator has to initiate the software launch as a pre-condition and is prompted to select the robots that they would like to deploy for the upcoming mission. Similarly, superscripts at the end of a task indicate that human action is needed before the next task can begin. Tasks without either have been fully automated for nominal cases in this newer Copilot version.

Task Planning and Scheduling: The aforementioned task dependency graph for the selected robots forms the input for Copilot's task planner and is stored in the MongoDB

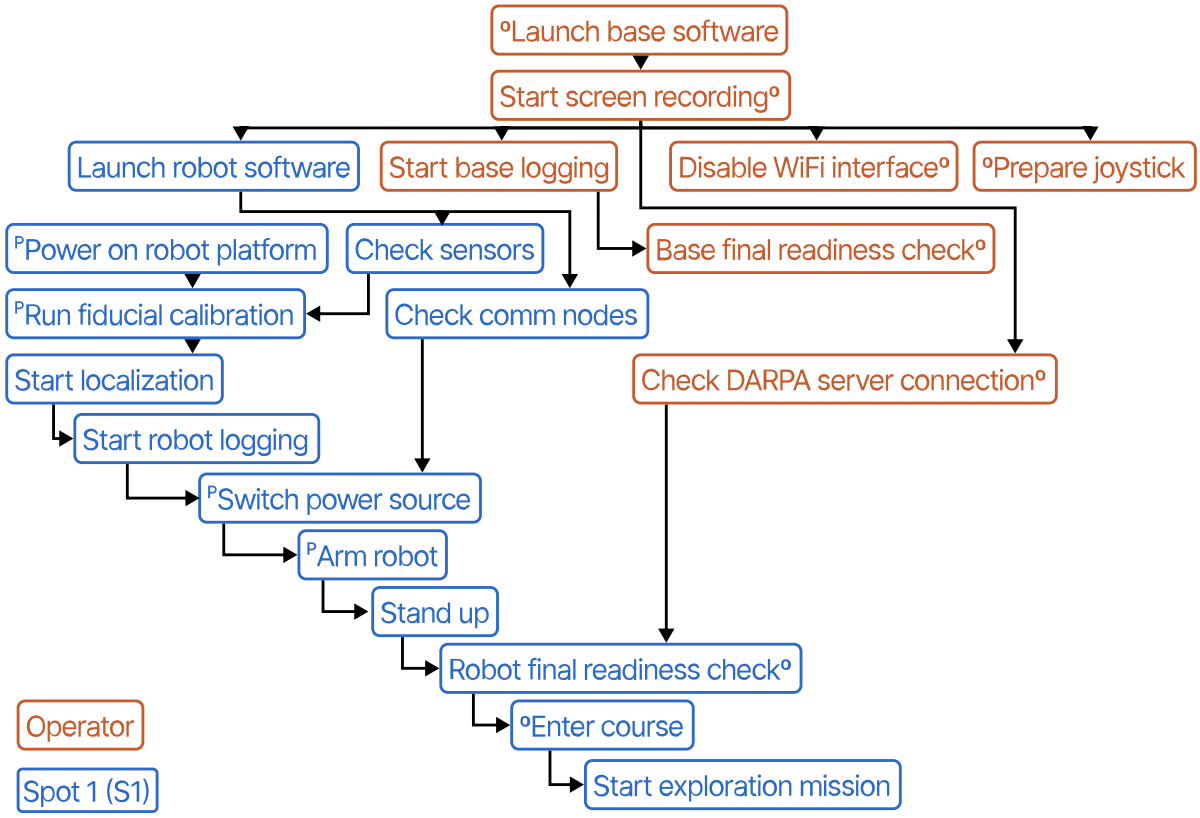

Figure 5.3 Pre-defined Copilot tasks for a single robot mission indicating task dependencies. The number of tasks scales linearly with the number of deployed robots. Spot1 related tasks are depicted in blue and operator tasks in orange. A superscript O or P at the beginning of a task indicate that the operator or pit crew has to manually fulfill some pre-condition. A superscript at the end indicates that a human sign-off is implemented before proceeding with the next task. For instance "Power on robot platform" requires a physical push of the robot platform startup button.

task database. The generation of a task plan for setting up, deploying, and assisting the operator during exploration is framed as an automated temporal planning problem. In the first version of Copilot, we formulated such problem as a Simple Temporal Network (STN), encoded as a linear program. In the improved version of Copilot, deployed in the final events of SubT, we moved to a PDDL temporal planning formulation to allow 1) flexibility on task representation with respect to state constraints, resources, and planning, and 2) use the body of planners available in the literature. Herein we integrated the OPTIC planner [127], a PDDL temporal planner that handles time window specification (timed initial literals), and discrete and continuous resources.

To perform planning, OPTIC uses both a PDDL domain file and a problem file. The domain file has been designed to represent tasks (modelled as operators) and its dependencies (pre-

conditions). The problem file is generated prior to calling the planner, and it is built based on the current state of mission and tasks execution. For example, if a task is ongoing, the PPDL file would represent the task as ongoing and add constraints to ensure it continues the execution to meet the necessary constraints. As a notional example of the scale of the planning problem, a mission with four robots would have approximately 60 tasks to be scheduled during setup and deployment. Planning is performed at a predefined cadence (e.g., every 1.5 seconds), but it also follows an event-based approach when task execution is late, or the human-in-the-loop changes their strategy — this helps mitigate execution uncertainty. The generated plan is parsed and stored in a Task Database (for logging and visualization across the system); each task is then dispatched for execution.

If a plan is not found by OPTIC due to temporal constraint violations (e.g., delays in task execution), Copilot will attempt to increasingly relax some of the key temporal constraints, such as the latest end time of certain activities (e.g., allowing setup tasks to end a few minutes after the setup time, overlapping with the beginning of the exploration time window). In critical scenarios, Copilot would notify the operator of a schedule relaxation to allow for further strategy changes.

Task Verification and Execution: A verifiable and generic task framework is introduced to Copilot, allowing for quick implementations and standardized task automation. Each task follows a strict precondition, execution, and post-condition template. Condition checks and execution can be triggered across agents, including the base station at which the human can oversee all automated processes at a high level in the new Copilot interface, which is described in Section: Game-Inspired Interface. The task template execution covers both fully automated tasks and semi-automated tasks where an operator's confirmation is required (e.g. deploying a robot into a cave requires a Go/No-go decision from the supervisor — deploying itself is an automated process). If a task fails during execution or post-condition checking, Copilot will try to resolve the issue by retrying tasks several times and allowing for more execution time. Failed tasks will be reported to the supervisor, who can choose to debug the issue at hand or trigger another automated retry. Retries and resets are possible at all levels, and completed tasks can be reset during an active mission in case a robot platform has to be rebooted.

5.5 Game-Inspired Interface

Game Inspiration: Inspiration for multi-agent interaction and interface design is partially drawn from real-time strategy games such as Age of Empires, StarCraft, and Command & Conquer. When played competitively, these games require a high sense of micro and macro-

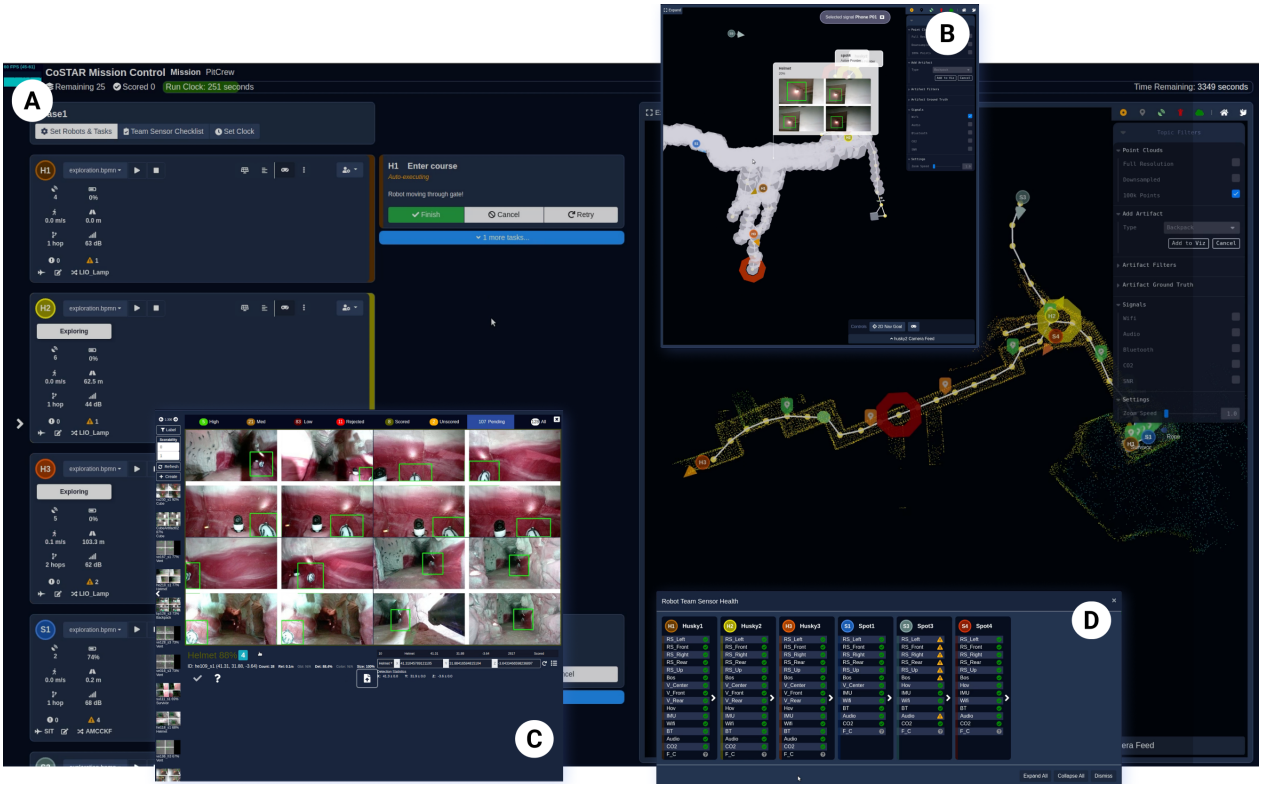

Figure 5.4 An overview of the major UI components. (A) The Robot and associated Copilot task cards. (B) The split-screen 3D visualization view with view controls, WiFi signal strength overlay, and an artifact card showing on the map. (C) The artifact drawer. (D) The robot health systems component.

management of units and their environment and the ability to efficiently switch between these two ways of managing a team. Micromanagement involves short-term strategy and decision-making, where individual units may require critical attention to win a battle, overcome an obstacle, or navigate to the next point of interest, while macromanagement refers to longer-term strategizing that involves resource gathering, unit production over time, and overall exploration and control of the map [128]. Parallels can be applied to the management of a robot team in the SubT competition. Even autonomous robots can benefit from or require human intervention and commanding, especially if critical attention towards failing subsystems is needed. Supervised multi-agent control draws from the human's situational awareness regarding the environment and robot states to effectively coordinate multi-agent behaviors, successfully locate artifacts, and score points.

Mission Phases: The user interface is designed to be adaptable to the overall mission and two major phases of an individual robot's competition run in particular:1) setup and deployment, and 2) mission execution with its exploration and search components. Across

these phases, the visibility and abstraction of information need to be flexible to facilitate focus on the anticipated operator interactions. In deployment, the user interface uses the Copilot-generated tasks and status information to guide the sole operator through the multitude of individual tasks while allowing them to maintain their situational awareness, manage the entire robot team, and coordinate with the pit crew.

Three Column Layout: The Mission Control interface is organized into different view components. Figure 5.4 shows the main split-screen with three columns aiming at creating reliable locations for the operator to look at when needing to accomplish functionally distinct tasks (A). The aim here is to reduce the amount of visual scanning, application switching, and to parse robot needs on an individual or team level swiftly. Individual robot information pertinent to monitoring health systems is available on the left, planned and actively re-scheduled Copilot tasks for individual robots are placed alongside each agent in the middle, and a 3D interactive visualization of the robots in their environment is anchored to the right. During mission execution, the primary goal of the user interface is to keep the operator situationally aware of a multitude of individual robot health systems and data sensed from the surrounding environment while presenting the most important information and thus reducing their cognitive workload. In Figure 5.1 the 3D visualization is expanded, and robot sensor and status information is minimized to select mission-critical information.

Health Systems and Robot Status: In order to effectively survey the status of any individual robot in the team, visibility into over 30 unique sensors and statuses needed to be surfaced to the operator per robot. This required identifying which indicators were critical to display at all times, which could be hidden within a sub-view, which were good candidates to be combined and abstracted, and which would be prioritized across either the deployment (split) or mission execution (split and expanded visualization) modes of the user interface. In addition to sensors visible at an individual level, an additional view was created to organize sensors compactly across the team, providing easy visual scanning for the operator during macro-management and deployment, as shown in Figure 5.4D. An abstraction of robot behaviors (e.g., exploring, dropping a communications node) and mobility states presents an overall status of each robot to the operator by color and a high-level description. This status is prioritized based on criticality to ensure the operator's attention will be requested for the most important issue at any given time.

Previously, Copilot tasks resided in an entirely separate module of the interface with limited screen estate, requiring the operator to move other related and necessary sensor and status information out of physical view. A reorganization where Copilot tasks are paired alongside their respective robots is utilized to reduce context loss and pair necessary information to

complete the tasks together, as shown in Figure 5.4A. Over time during the development roll-out, this pairing of health, sensor, and status indicators alongside Copilot tasks facilitated a level of trust from the operator where focus on a particular robot was not necessary unless a critical task requiring operator intervention appeared.

3D Visualization View: A 3D interactive visualization leveraging React Three Fiber (a React-based renderer for three.js) was created within the UI with the aim of achieving a significant reduction in operator task and application switching. Prior to this version of the interface, the operator was required to switch between a web browser to view robot health systems and status information and RViz (a visualizer for ROS) to view the robots within the 3D environment and command them. In the split view of the UI, the operator can have the full context of robot sensors and status information along with any outstanding Copilot tasks. When in the expanded visualization view, the layout shares similarities with layouts of traditional Real-Time Strategy (RTS) games, where content is functionally organized from the corners of the view and leave the center-most screen real estate where the operator will primarily interact with robots and information unobstructedly. From this view, the operator can take on any of the following tasks: surveying the mapped environment and robot positions for locations to scout, locating, and submitting object or signal artifacts, directing or course-correcting robot autonomy with manual navigation commands, viewing signal strength of the communications backbone within the environment, and assigning robots to drop communication nodes manually. The visualization allows the operator to navigate the 3D environment through panning, zooming, and filtering points of interest categories. To effectively manage the switching between micro and macro-level interactions, a single-click shortcut was implemented on each robot status card for the operator to quickly focus on any robot that requires attention. An additional shortcut is provided to zoom back out to an overview of the map.

Improvements over traditional RTS commanding controls were also made to minimize the amount of mouse control and coordination necessary. Instead of requiring to select or drag a bounding box prior to commanding a robot, the operator could simply interact with the visualized information roadmap (IRM) — a breadcrumb trail used for safely navigating the environment constructed by the team of robots [129] — and assign any robot with a high priority navigation point or communications node drop location through a context menu, regardless of whether the particular robots are currently in view or not.

To help with artifact management, the locations of detected artifacts are visualized and interactivity is added to allow the operator to quickly hover into a thumbnail and click to navigate to the dedicated Artifact Drawer Figure 5.4C for deeper analysis and submission.

Additional interactions are, for example, manually adding and manipulating detected artifact locations within the 3D space, by dragging its location across a plane for fine-tuning if a submission location was deemed incorrect and needed adjustment.

While in the expanded visualization view, compressed versions of the robot status modules are shown horizontally in the bottom left of the view with the mission status indicator made more prominent and placed above each module. These overall status indicators were given visual priority to ensure grabbing the operator's attention. For instance, the indicator would flash red when a robot had fallen over, was low on battery, or required assistance. The operator could immediately click the respective robot module and be oriented over it for micromanagement.

Artifact Drawer: Artifact submission was a critical part of SubT that also has many real-world parallels, for instance, in search and rescue. Especially under time constraints, it is necessary to quickly identify artifacts of interest in the environment, whether these be human survivors or other objects of interest. Detecting and localizing artifacts automatically is done using a state-of-the-art image processing pipeline [130], but no AI system is infallible, especially in unknown environments, so having a system for an operator to manually review artifacts efficiently was critical considering mission time and submission attempts.

In the old system [130], a manual artifact review system did exist, but it was built with a focus on only basic functionality and a high reliance on initially accurate artifact detections. Each artifact report took roughly 90 seconds to review. In redesigning this component, we wanted to focus on improving the review process from an ease of use perspective and decrease the time spent to confidently review an artifact report down to 15 seconds. Beyond simply making the system more intuitive for the operator, this actually had a major functional benefit from a trustability standpoint in that it allowed us to decrease the confidence threshold for flagging artifact detections and have the operator go through and verify nearly 6 times more potential artifact reports while not increasing total time spent.

To better design the new system for speed, it was important to understand which areas of the old one were slowing the process down the most. Testing the old system in simulation and conducting operator interviews revealed that the artifact review process needed too many clicks. Then, time had to be spent zooming in on and reviewing images and checking with RViz separately to verify that artifact coordinates were correct. No visual aid was given if corrections were necessary, and coordinates had to be updated by manually entering them for each axis in $\mathbb{R}^3$. Borrowing from game interface design, integrating the 3D visualization view directly into the web UI removed the need for application switching, and drag controls were added to adjust locations providing correctly scaled coordinate updates from the 3D

environment. A minified list that provides an overview of all artifact reports by confidence levels, plus maximizing the screen real estate of a single selected artifact helped increase efficiency. Finally, adding keyboard shortcuts as commonly used in gaming made meeting our target goal of 15 seconds possible.

5.6 Results

Over the course of the last challenge year, we conducted a limited series of field tests in three testing locations, including the abandoned tunnels at the Los Angeles Subway Terminal building, the Lava Bed National Monument in Northern California, and the Kentucky Underground lime-stone cave for which we applied our rapid development and testing strategy. We experimented with different robot configurations and in different stages of readiness as our system's capabilities matured. We deployed up to 11 vehicles simultaneously during these tests stressing the overall system (including Copilot and all the UI elements) and learning about its technical limitations like bandwidth and computing resources which will be presented in upcoming work.

We deployed the presented game-inspired user interface and supervised autonomy system during the SubT challenge using four to six ground robots nominally. While we could have exceeded the number of six robots using the newly designed interface and autonomy, six became the preferred number of agents to explore large-scale environments while allowing reliable communication links that would not exceed bandwidth limitations when robots disseminated information from autonomously explored out-of-comms areas. This allowed meeting the set interaction objectives, especially maintaining an enjoyable performance that can visualize the complete robot team while contributing to a lower workload due to fewer deployed agents.

In what follows, we analyzed screen recordings and log files collected during the SubT final competition, for which we extracted time-to-task information, robot deployment times, mouse locations, and application usage from runs P1, P2, and F that consist of a setup-time and mission phase of 30+30 and 30+60 minutes, respectively. Robots were only allowed to leave the setup area and enter the course when the mission time began. Readying the team of robots and not bleeding into the mission time was a crucial effort to maximize available mission and exploration time. The results are compared to an earlier state of the system that did not implement Copilot and used different interfaces, namely the SubT "Urban Circuit" similar to [125]. During the "Urban Circuit" task, coordination was done by humans only.

Robot Deployment: Figure 5.5 shows the robot deployment times that were achieved by

deploying Copilot and compares them to the baseline. We can see that during run P1, we achieved sending one robot in less than 60 seconds each, deploying a total of 6 ground vehicles in 5 minutes and 31 seconds. In runs P2 and F, we achieved staying below the one minute mark for the first three robots. Deploying the robots without Copilot and the new interface in the 'Urban Circuit" runs A1, B1, and B2 took more than 5.5 min per robot on average, thus significantly reducing the time available for exploration and consequently reducing ground coverage and information gain regarding the search task.

Application Usage: The new interface resulted in a shift in application usage and reduced switching between different applications and computers with a second set of peripherals, as RViz was running on a second device during the "Urban Circuit". Figure 5.6 presents the relative usage of applications for six SubT runs. Designing a unified interface resulted in a shift in application usage that reduced the use of RViz significantly. While more than 50% of time was spent on RViz during the "Urban Circuit" runs, we were able to unify user interactions and situation awareness in a single Mission Control interface. Only run F uses RViz for some time as a debugging tool that gave access to the robot's cost maps depicting the perceived risks around them. This information was not visualized by the new interface, but presents valuable key information in case of unexpected and off-nominal operations.

UI Feature Usage: With the main Mission Control interface being the main interaction point for human supervisory control, we then look at the feature usage within the interface itself. Figure 5.7 shows the relative interaction times with the split-screen view, the 3D full-screen console view, the sensor health overview, artifact submission drawer, and the BPMN modal that gives a detailed overview of a robot's inner state machine (which was relied upon during the "Urban Circuit"). We see that, especially during runs P2 and F, large amounts of time were spent on the artifact drawer and thus performing the search task analyzing the artifact reports that were generated by the multi-agent system. To gain situational awareness and potentially interact with the robot team, the human supervisor primarily relied on the split-screen view of the Mission Control app that is shown in the background of Figure 5.8 overlaid by a heat map that indicates the most active areas derived from mouse cursor positions sampled at 1.5 Hz. In this analysis, an area is deemed inactive if the mouse has been stationary for more than ten seconds. *Huang et al.* [131] found that the median difference between human gaze and mouse position during an active task is 77 pixels with a standard deviation of 33.9 pixels at 96 dpi screen resolution. A Gaussian kernel with $\mu = 98$ and $\sigma = 43$ adjusting for 122 dpi is used to derive our heat maps. Figure 5.8 indicates that the robot cards, Copilot tasks and the 3D view were all crucial tools while overseeing the robotic system and performing the exploration and search tasks.

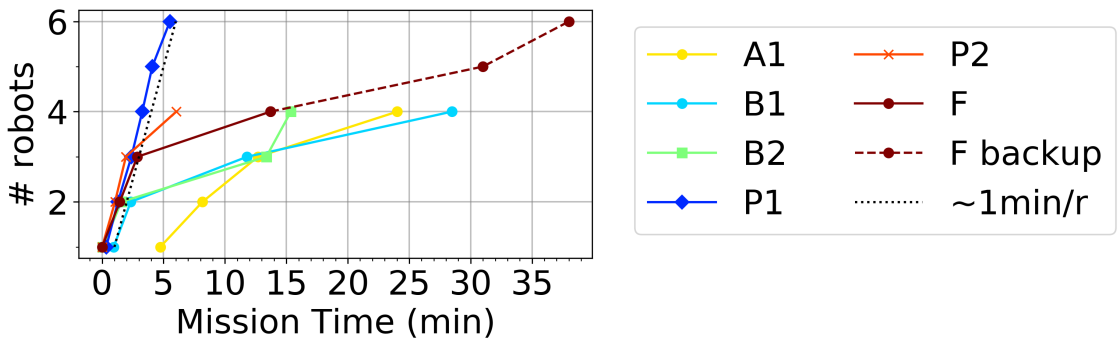

Figure 5.5 Robot deployment times per game run measured upon entering the course. The black dotted line (∼1min/r) indicates the team's internal goal for robot deployment and represents a deployment of one robot per minute. F backup marks insertion points of 2 robots that were not part of the initial deployment strategy but were added ad-hoc to compensate for robot failures during run F.

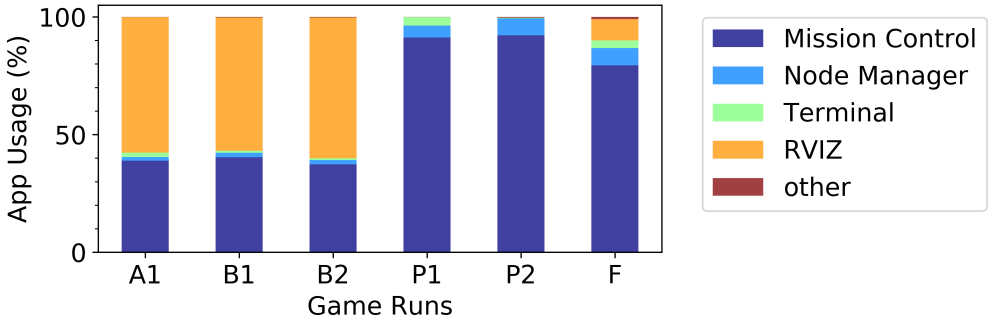

Figure 5.6 Application usage (foreground application) for six SubT mission runs in percent. A1, B1, and B2 represent the usage before the redesign that integrated 3D visualization and interactions for P1, P2, and F in a single Mission Control application using only one computer and screen. Note that node manager and terminal usage are underrepresented in runs A1, B1, and B2 because the initial setup phase of up to 10 minutes was not recorded for these runs due to different logging procedures.

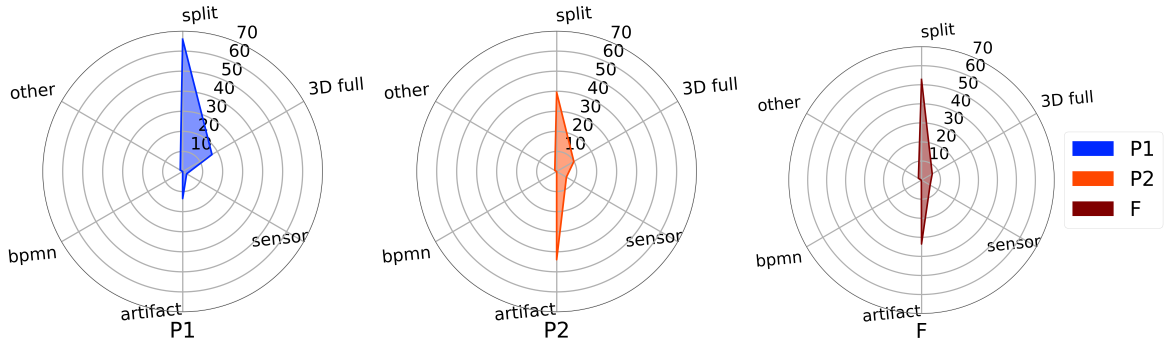

Figure 5.7 Analysis of the redesigned user interface interaction by view component in percent for runs P1, P2, and F.

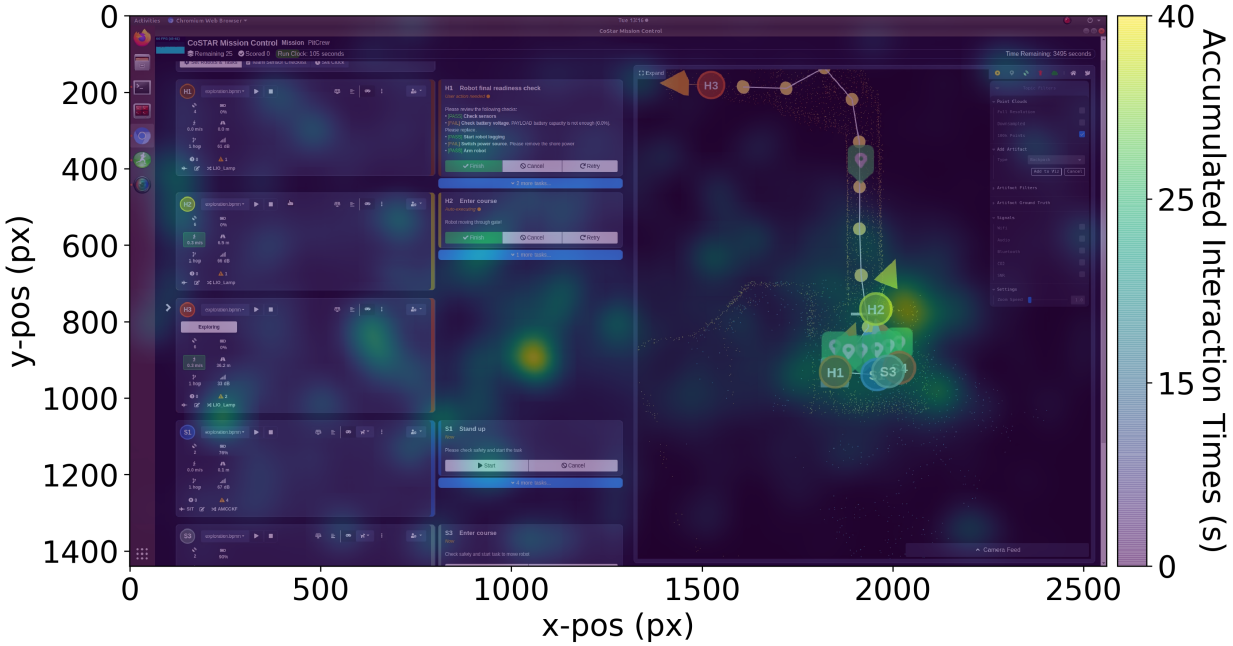

Figure 5.8 Activity heat map showing the x and y positions of cursor interactions (and indirectly gaze) overlaid on the Mission Control Split-Screen view exemplary for game run P2. The view consists of robot cards, a column for Copilot tasks, and the split-screen 3D view. A brighter heat map indicates higher interaction times in this area. Stationary cursors for more than 10 seconds are classified as inactive.

5.7 Conclusions and Future Work

In this work we (i) create a game-inspired user interface for multi-agent robot missions (ii) integrate an automated planner for task planning and scheduling, (iii) add a verifiable task framework for increased reliability, and (iv) present results on how the overall system

performed over the course of several real-world deployments, including the DARPA SubT Challenge final. In future work, we plan to deploy our interface and Copilot during scientific exploration missions to autonomously map and identify geological features and assess exploration strategies in lava tubes. This will lead to further validation of the subsystems and a structured assessment of a supervisor's workload outside the realm of the SubT challenge with experts and potentially non-expert users. Ultimately, we would like to assess operator workload from wearable sensors in real time and consider such constraints in Copilot's task planning.

Acknowledgment The work is partially supported by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004), and Defense Advanced Research Projects Agency (DARPA). This work was conducted in collaboration with the Making Innovative Space Technologies Laboratory (MIST Lab) at Polytechnique Montreal. The first author would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for their generous support in the form of a Vanier Canada Graduate Scholarship. Thank you to all members of Team CoSTAR for their valuable discussions and support.

CHAPTER 6 ARTICLE 3 - INFLUENCE OF AUTONOMY AND INTERFACES ON HUMAN AND MULTI-ROBOT TEAMS: A STUDY ON PLANETARY EXPLORATION

Preface: This study contributes (i) a method for creating a fielded VR interface with real-time rendering capabilities that enables the exploration of ad-hoc and dynamically created cyber-physical spaces with multiple robots in large-scale environments (ii) an evaluation of the influence of autonomy and interface design on multi-robot operations, comparing performance in a controlled study fielded system. (iii) an assessment of objective workload with a low-cost wearable HRV sensor and a comparison to the subjective NASA TLX questionnaire. (iv) a questionnaire for Situation Awareness, Immersion, and Trust (SAIT).

Declaration of Contributions: As primary author of this study, I conceptualized the research questions and objectives, designed the study, determined the methodology, and obtained ethics approval by the institution's Ethics Review Board (ERB) with approval number CER-2122-50-D. I designed and implemented the proposed architecture, including the real-time virtual reality interface, conducted the data collection for all participants, performed the data analysis, and interpreted the results. I created figures, wrote the entire article, including literature review and revisions based on received feedback.

Full Citation: Kaufmann, M., Beltrame, G. (2024). Influence of Autonomy and Interfaces on Human and Multi-Robot Teams: a Study on Planetary Exploration. Submitted to ACM Transactions on Human-Robot Interaction (THRI) on July 9, 2024. The version below has been revised in response to feedback from the doctoral committee.

Abstract – Recent advancements in robotic autonomy and interfaces have revolutionized human-robot interaction in many fields, including disaster response, industry, and scientific exploration. Missions exploring complex, unstructured environments on Earth, the Moon, Mars, and other celestial bodies remain challenging and often limit human access or presence. Gaps remain in the understanding of how autonomy and interface design integrate with human-factors and performance during large-scale exploration missions. We investigate teams comprising two (semi-)autonomous robots and a single human supervisor using our interfaces. The system has been deployed in caves at the Lava Beds National Monument (Northern

California) and during a two-by-two factor within-subject study at Polytechnique Montréal, exploring both real and simulated caves. We obtained results including n=38 participants evaluating the influence of autonomy (waypoint vs. full autonomy with interventions) and interfaces (screen vs. virtual reality) on workload, situational awareness, and performance during a scientific exploration mission. We find that continuous physiological measurements align with NASA TLX metrics to measure mental workload. The virtual reality waypoint condition yields the lowest number of science targets detected, while all other conditions achieve similarly good performance levels. Situational awareness, assessed by the Situation Presence Assessment Method's accuracy measures, yields approximately 90% correctness for both interfaces.

Figure 6.1 User study setup with virtual reality participant and screen interface in the background (far left), followed by an example of VR operations during a field test at the Lava Beds National Monument's Valentine Cave in Northern California. The right images show a real-time VR rendering of the robots and cave geometry, and a part of the Valentine cave with a legged robot on a mission, respectively.

6.1 Introduction

The search for life and potential resource utilization of planetary bodies such as the Moon, Mars, and other celestial objects has motivated space agencies and private entities to investigate and develop new space exploration capabilities. Increasingly, robotic exploration systems are used to search for life, for in-situ resource utilization, and to discover potential colonization grounds.

The *2020 NASA Technology Taxonomy* [132], for instance, presents technical challenges that need to be addressed to enable and sustain a long-term human presence in space and on other planets, including human-robot collaborative systems and autonomy technologies to facilitate and augment science and exploration missions [133]. The decadal strategy for planetary science and astrobiology [134] also identifies the value of human-scientific and human-robotic partnerships for future missions. In addition, the first off-world demonstration of cooperative

multi-agent autonomy CADRE (Cooperative Autonomous Distributed Robotic Exploration) is scheduled to fly to the Moon as a payload on the IM-3 mission [3], which could enable the age of extraterrestrial multi-robot exploration.

However, Human-robot interaction, especially when operating multiple (autonomous) robots at once, leaves much to be desired. In fact, large-scale multi-robot systems have been identified as one of the decade's biggest challenges [135]. While teleoperation is a well-known application in the human-robot-interaction domain, often multiple humans are involved in operating a robotic system [136] from a distance. Teleoperation is not always possible, and in many cases, robots remain difficult to control and require expert knowledge. The human-in-the-loop, while being invaluable at solving unforeseen problems, can also be performance-limiting as workload increases [20,137]. Simpler interaction designs and user-centered interfaces could benefit both expert and non-expert human users [136, 138] in attempting complex missions in a variety of domains ranging from structured warehouses to unstructured and hazardous environments that are unsafe for humans during planetary exploration missions.

Supervisory control and task coordination are properties that limit the performance of large robotic teams with single human supervisors. As the robot team size increases or is distributed in large environments, the amount of information instantly available to a human supervisor decreases; thus increasing the level of autonomy and situational awareness while decreasing cognitive workload is desirable. Both the way information is presented, and how robot intentions are perceived, influence situational awareness and workload.

In this work, we present a real-time capable virtual reality (VR) interface for LiDAR-based robot teams and conduct a user study (n=38 included participants) investigating the influence of autonomy (waypoint and full autonomy with interventions) and interfaces (screen and VR) on a team of two autonomous robots supervised by a single human operator. The human-robot team has been tested in the real world during the NASA BRAILLE [139, 140] analog mission and in the user study presented in this work. We collect heart rate variability and NASA Task Load Index (TLX) measurements [141, 142], and introduce a questionnaire to measure immersiveness and trust. For the robotic system, we leverage the NeBula autonomy framework [25] and extend it with our interfaces, filters, and a data collection pipeline, achieving real-time rendering capabilities in VR. Figure 6.1 shows an overview of the study setup, the cave exploration base station, the real-time VR rendering, and an example image inside the Valentine cave showing a deployed NeBula-powered Spot robot in Northern California.

The following research questions are addressed: **(RQ1)** Do continuous physiological measurements, i.e, heart rate variability from portable devices, differ from self-assessed (NASA

TLX) metrics? **(RQ2)** Is the 3D environment in virtual reality influencing performance positively or negatively, if at all? **(RQ3)** Is higher autonomy reducing or increasing situational awareness, if at all? **(RQ4)** Do higher levels of trust correlate with higher autonomy and/or physiological measurements?

6.2 Related Work

Human-Robot Interaction

HRI has become a part of many domains, including self-driving, "manufacturing, space, aviation, undersea, surgery, rehabilitation, agriculture, education, package fetch and delivery, policing, and military operations" [27], however, robots are expected to be controlled or supervised by humans for the intermediate future. There is a great need to integrate human factors in research and design. Knowing the human mental state and using it to plan tasks and avoid conflicts has been determined to be a remaining challenge in the field [27]. The authors of [143] concur that we are far from removing human supervisors and interventions to aid with unpredictable events despite the recent advances of autonomous systems. They advocate for systems that use shared control and shared autonomy paradigms. Deciding which level of autonomy a user wants is declared to be a complex problem. Some users might prefer high levels of autonomy while others do not want autonomy to be invasive.

Interfaces, Virtual, Augmented, and Mixed Reality

Rea and Seo [136] argue that teleoperation still needs to be solved as it is often a domain-specific task. They state that there is a need for user-centered interface design and more tests in the real world.

The authors of [144] use post-processed digital terrain models and visualize those in a virtual reality application and present an independent path planner for Mars rovers. However, their VR solution is limited as it does not incorporate the rover model or visualizations for the planned paths, declaring integration efforts as future work. Further, this solution is not real-time capable.

Walker et al. [67] review the terminology used in Virtual, Augmented, and Mixed Reality (VAM) in the context of HRI. They found an under-utilization of novel 3D input techniques and established that images and video streams are dominant visualization techniques, possibly due to ease of use.

An example of recent work using first-person video stream capabilities can be found in [145],

where Team MARBLE utilizes it during the DARPA SubT challenge.

Creating 3D reconstructions from sensor data is labeled as difficult, and despite improvements in computation and networking, rendering high-volume visual data of large environments is often replaced by simulated or virtual environments to create a cyber-physical interaction space [67]. A vast body of VAM research involves industrial robots and, more often than not, only a single agent [68–71]. This indicates a gap for real-time rendering capable cyber-physical interfaces that allow the co-location of humans and robots in a cyber-physical space. Keeping humans outside of potentially hazardous environments is needed, especially when performing missions off-planet. NASA's Artemis missions could greatly benefit from such interfaces.

The experimental setup in [68] consists of a robotic arm and a static RGB-D camera, which looks at a cluttered industrial environment. They deploy the Robot Operating Systems' visualization tool RViz, testing human-robot collaboration and collision avoidance with different algorithms and scene representations. Occupied parts of a scene are depicted either as large primitive shapes (similar to bounding boxes) or an octomap [146] occupancy gird. The scene is visualized on screen, and the maximum viewing range is actively limited to 0.9 m or 3 m.

The authors of [147] process point cloud for change detection experiments in a warehouse setting. Changes are indicated by renderings of large spheres utilizing a mixed reality headset. Mixed reality has the advantage that it can be used in the same physical environment without a rendering overhead to create the scene. This requires the person to be physically present in the surveyed area.

Search and rescue missions are similar to space exploration missions as environments can impose high risks. The work in [148] presents a VR-based interface in the context of a rescue mission allowing the control of a single robot. The rendering in VR uses sparse, incremental LiDAR points accumulated over five consecutive frames. The user must actively set the viewing angle in a fly-through mode and teleoperate the robot by giving direct velocity commands. A drawback of the presented method is that fly-through modes are prone to introducing motion sickness, and there is no physics-based interaction with the virtual environment in this application.

In contrast, Patterson et al. [149] create a realistic VR interface from post-processed LiDAR scans of a lava tube that allows limited physical interactions in the form of teleportation pads. The primary purpose of the developed interface is to create a VR-based log book for scientific data. They display high-resolution 2D images of bio markers and their positions within the 3D cave model. Multiple manual steps are involved in creating a realistically looking cave model with artificial textures. Registration and cleaning steps are performed

using open-source tools manually. Fixed teleportation pads within the environment allows users to jump within the environment, and change their viewpoints. This ensures users cannot accidentally move their heads outside of the model. At the same time, this method restricts visiting narrow corridors and does not allow walking around freely.

A user study on on-site collaboration for lunar exploration using shared mixed reality (MR) has been published in [66]. They compare four astronaut-rover configurations with different decision-making transparency levels. Deployed metrics are task performance and workload assessed using NASA TLX. Key findings include that higher decision transparency reduces workload and that sending direct control commands results in very high workloads compared to a waypoint input; workload has not been measured for a full autonomy mode as they anticipated no human supervision. To obtain overlays on the MR headset, a 2.7 s long parsing process is initialized to process the environment. This process is done once to align with the environment and again at the user's request. The test environment is limited to a 20 m^2 analog environment. Thus, scaling this solution to larger environments would require additional parsing steps, introducing high latencies that have been detrimental for task performance [65].

Workload, Situational Awareness, and Performance

Semi-autonomous systems are identified to reduce operator workload and allow time for other supervisory tasks [136]. However, task switching could reduce situational awareness, delay task completion, and add to the workload. St-Onge et al. [44] deployed a team of uncrewed aerial vehicles in a realistic planetary analog where operators were tasked to explore an area with five agents. They measured perceived workload, assessing mental demand and objective workload using pupillometry. Subjective and objective measures were found to disagree; however, the study was limited to only five participants in the field.

Heart rate variability as an indication of stress and workload has been studied for a long time. Early definitions and standardization efforts are documented in [150, 151]. The authors of [152] investigate whether two conditions of workload (load and rest) can be distinguished. Measuring HRV and pupil diameter, users performed an n-back task under load and relaxed in a chair under the rest condition. NASA TLX values were (as expected) reported to be higher for the n-back condition and significant differences were found for both HRV and pupillometry results in one of their test conditions in a bright room.

Kosch et al. [153] identify that measuring cognitive workload in the human-computer interaction (HCI) domain should be reassessed. The NASA TLX was originally developed to assess workload in pilots and astronauts and Kosch et al. [153] state that we currently do

not know its correctness for assessing workload in the HCI domain. The same holds true for assessing workload measurements in the context of multi-robot operations with varying levels of autonomy, resulting in our first research question RQ1.

Mixed-initiative systems seem to improve performance in navigation tasks and improve cognitive performance for secondary tasks and overall workload [154]. While the NeBula autonomy [25] allows us to deploy a fully autonomous robot system as one of our independent measures, we allow for human interventions and supervision at all times, giving the system mixed initiative capabilities.

Contributions

Addressing some of the shortcomings identified in this related work section, our main contributions are:

- A fielded VR interface with real-time rendering capabilities enabling large-scale exploration of cyber-physical spaces with multiple robots
- An evaluation of autonomy and interface design influence on multi-robot operations, comparing performance in a controlled simulation environment resembling the fielded system.
- An assessment of objective workload with a low-cost wearable HRV sensor and a comparison to the subjective NASA TLX questionnaire.
- A questionnaire for Situation Awareness, Immersion, and Trust (SAIT).

6.3 Methodology

In this section, we describe the deployed human multi-robot system, the interfaces, as well as the user study design, and the collected data. The goal is to provide insights into the effects of using different interfaces and autonomy levels when exploring unknown environments. In the context of a planetary exploration mission, finding hidden science targets that are autonomously identified by the robots is used as a means to evaluate performance.

6.3.1 The Human and Multi-Robot System Architecture

In preparation for this user study, we extended the modular autonomy framework NeBula [25] and developed a real-time capable VR interface that allows us to explore new environments while virtually being co-located with the robots.

In contrast to many VR applications, we do not rely on prior knowledge and do not deploy post-processed and pre-loaded world models. We introduce an architecture that is based on LiDAR input, leverage NeBula, and connect the Robot Operating System (ROS) with the Unity game engine to create the VR experience in real time. This architecture renders a true-to-scale representation of the environment and allows for physical interaction with the virtual world. Figure 6.2 shows the system's architecture for deployment with simulated and real systems. In either case, sensor data in the form of LiDAR points is produced by each robot agent while keyed scans are sent to a base station computer, merging the point clouds of multiple agents. Both robots and the base station run NeBula and are networked. For this study, we assume network connectivity at all times (while the system is capable of going in and out of communication range).

The merged point cloud is filtered to remove spurious artifacts and is sent to either the screen or VR for rendering via the ROS bridge. The screen interface leverages roslibjs [155] to connect to a web-based screen interface, and ROS Sharp (ROS#) [156] is used to connect to the gaming engine Unity in which we create our VR experience. Unity and ROS Sharp are running on a Windows PC that is tethered to an Android-based Meta Quest VR headset.

Figure 6.2 also shows the operator/supervisor in the loop and how their interactions are relayed back to the individual robot agents via different input methods. Throughout the study, objective and subjective workload measurements are collected, and participants fill out questionnaires.

This is a simplified representation of the data flow. The system exchanges much more information among agents via a mesh network [157,158], and processes data on- and off-board the robots. We do not depict the detection pipeline for science targets or science proxies [139] as this is not the scope of this work.

6.3.2 Interfaces, Interaction, and Rendering

Screen interface

The deployed interfaces are shown in Figure 6.3. The screen interface depicted in (A) uses a 2.5D false-color point cloud representation to visualize the scanned world to the user. Different heights are indicated by varying colors. The robot poses are shown by labeled circular markers in blue and purple for the here deployed robots. If the robots detect a science target in the environment, a spherical white marker is shown. Users can interact with either robot by clicking on the robot marker or selecting the robot card on the lower left of the screen. To send a manual goal, the 2D Nav Goal button has to be enabled (preventing

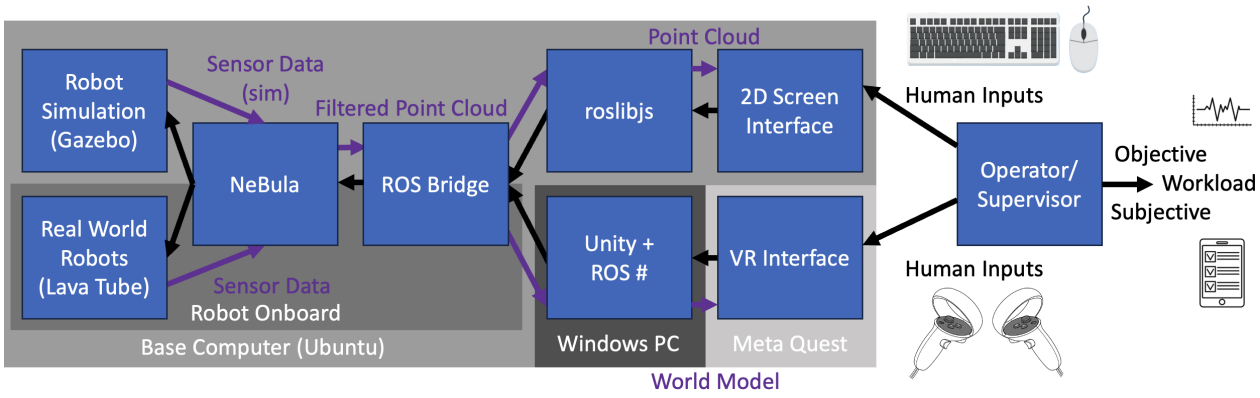

Figure 6.2 System architecture and simplified input/output data flow (from left to right): Robots are either operated in the simulation environment or during real-world deployment in a lava tube using the ROS-based NeBula Autonomy Framework. The sensor data (LiDAR) is then filtered and fed into either the screen or virtual reality (VR) visualization pipeline. For the screen setup, roslibjs is used to interface with the on screen web interface and human inputs are given via mouse and keyboard. On the VR side, we use ROS# to interface with Unity, where our real-time rendering pipeline is deployed to generate a cyber-physical world model that can interacted with using the Meta Quest headset and controllers. We measure workload by recording heart rate variability (objective) and have incorporated questionnaires after each experiment (subjective).

accidentally sending goals), and an octagonal marker in the color of the corresponding robot will appear as pointer to select the goal. Users can zoom in and out of the point cloud with the mouse wheel, or click and hold to move the map around in either 2D mode (left click) or 2.5D (right click) to change the viewed area and viewing angle. The 2D screen interface that we compare to was initially developed for the DARPA Subterranean Challenge to supervise large teams of autonomous robots efficiently [20, 159]. For this study, which includes non-expert users, we limit the interface to the full-screen mode, which is presented in (A), and limit interactions as described above in an effort to keep VR and screen interface similar and comparable while simplifying the training process for the users.

VR Interface

Screen captures of the VR interface renderings are shown in Figure 6.3 (B-I). The interface was designed to utilize the screen color coding to represent the robots (B-D), while their actual poses are shown by a digital twin. It is important to note that the screen captures do not accurately convey the immersive experience compared to wearing the VR headset. Users typically perceive the VR renderings in three dimensions with depth. They are able to change their viewing angle by moving their head like in the real world and can move around

in the environment. The resulting motion effects and changes in view point can only be fully appreciated when viewed through an immersive headset. Distance is represented by different shades of brown in the terrain. Looking down, the user can use a mini-map for orientation in the environment and see where the robots are relative to them (E). The user is always located in the center of the mini-map while the North faces towards the current view direction. Detected science targets are represented as white spheres that appear in the environment and mini-map, similar to the screen interface. In VR there is different means of transportation to get around in the environment. (F) and (G) show visualizations of a valid and invalid teleportation goal indicated by a green and red laser pointer, respectively. Pointing there and letting go at a valid goal pose moves the user's character in the environment. The robots can be given new goals in a similar fashion as shown by the blue and purple laser in (H) and (I). The robots then plan a path in an effort to reach the user specified goal.

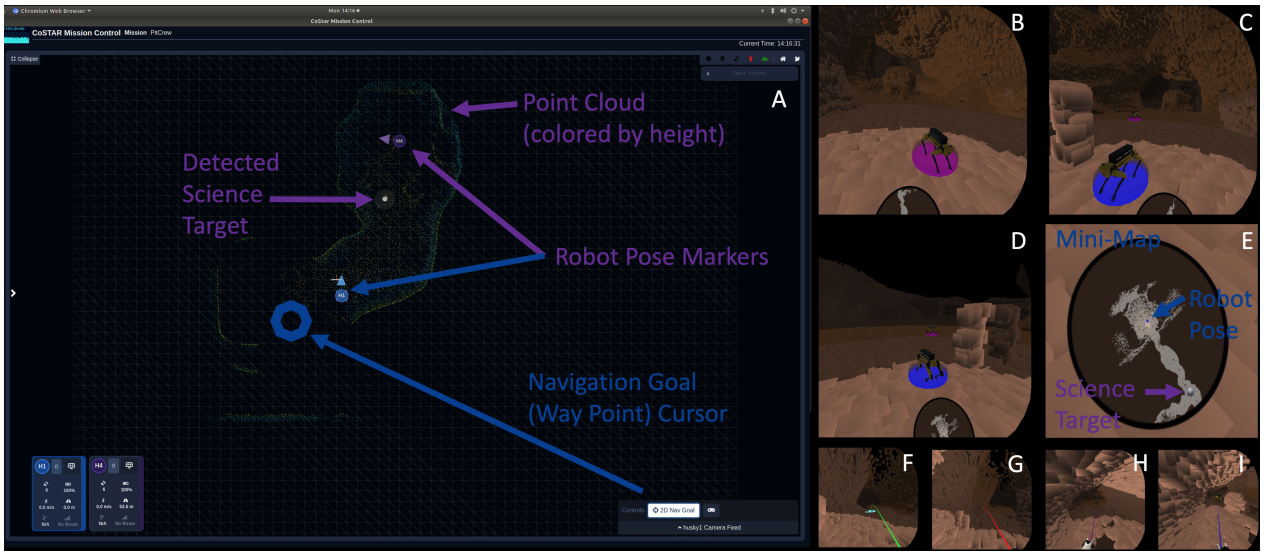

Figure 6.3 Screen (A) and VR (B-I) real-time interfaces. Both interfaces use similar color coding to indicate robot and target positions. Science target detections are visualized as white spherical markers, while robot locations are indicated by color-coded labels. The point cloud on screen (A) is colored by height, whereas the VR interface uses different shades of brown to indicate distance/depth. (B-D) shows views of actual robot positions in the real-time rendered environment with a mini-map at the bottom. (E) is an enlarged view of the mini-map that a user can look at for spatial/situational awareness in VR at all times. In VR, a user can walk or teleport their character to different positions in the environment. (F) shows a valid teleportation goal, while (G) cannot be reached due to collisions or being out of range (red). The blue and purple rays in (H) and (I) are used to input manual goals for the robots.

The interaction with the environment in VR differs from traditional screen, keyboard, and mouse setups, as the user is wearing a headset and uses a pair of joystick controllers. The

Meta Quest controllers and our key bindings are shown in Figure 6.4. Users were familiarized with all controls during a dedicated training phase to practice sending goals to robots and getting around in the environment via teleportation, walking, or jumping directly to a robot position. The key bindings were printed and placed on the users' desk so that they could review it during off times.

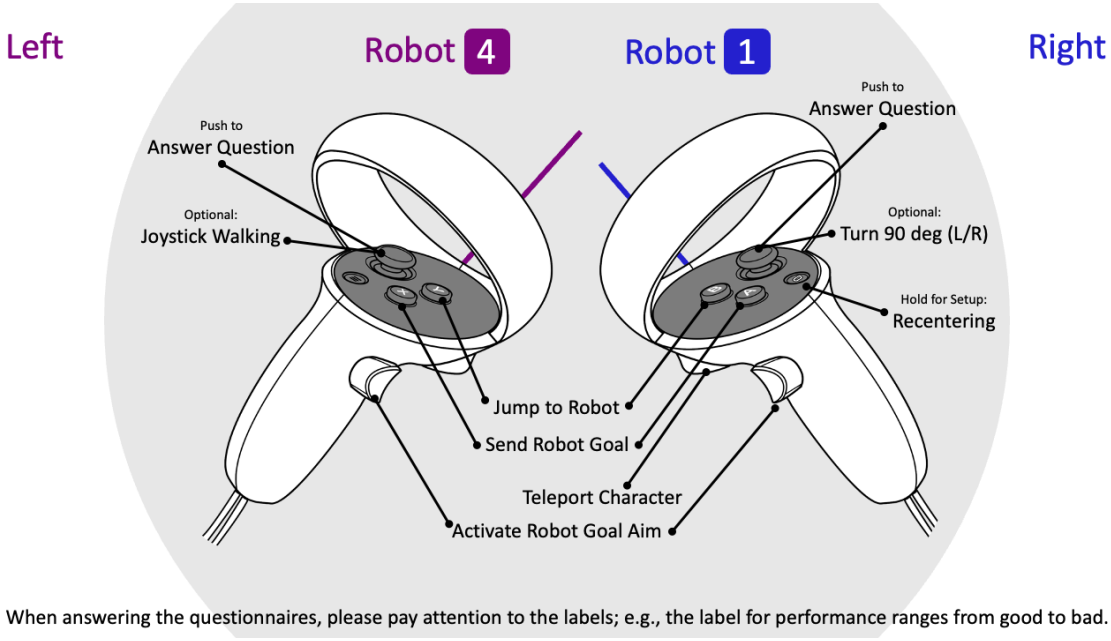

Figure 6.4 Meta Quest 2 Controllers and Input Modalities for Left and Right Hands to Interact With Robots Number 1 and 4. This chart has been used for participant training and was provided throughout the study.

Mathematical Formulation of Cube Rendering

To achieve a real-time rendering capable system, we process the merged point cloud of the robot system using the following pipeline for VR:

Let $\mathcal{P}$ represent the original point cloud. Define the filtered set $\mathcal{Q}$ as:

$$\mathcal{Q} = \{\mathbf{q} \in \mathcal{P} \mid |\mathcal{N}(\mathbf{q}, r)| \geq k\}$$

where $r = 0.5$ meters is the radius within which neighbors are counted, and $k = 15$ is the minimum number of neighbors required for a point $\mathbf{q}$ to be included in $\mathcal{Q}$. This filtering step is designed to remove noisy measurements from the LiDAR data while preserving the structural integrity of the environment.

For each point $\mathbf{q} \in \mathcal{Q}$, construct a cube centered at $\mathbf{q}$ with a uniform scaling factor S applied to each cube's dimensions. The vertices $\mathbf{v}_i(\mathbf{q})$ of each cube are determined by:

$$\mathbf{v}_i(\mathbf{q}) = \mathbf{q} + S \cdot \mathbf{v}_i^{\text{rel}}$$

where $\mathbf{v}_i^{\text{rel}}$ denotes the predefined relative positions of the vertices of a standard unit cube. Figure 6.5 illustrates the used vertex naming convention to define a cube.

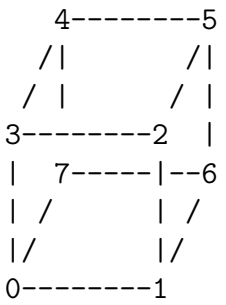

Figure 6.5 Vertex notation for cube surface generation

Define the set of triangles $\mathcal{T}$ representing the cube surfaces for the entire point cloud $\mathcal{Q}$ as follows:

$$\mathcal{T} = \bigcup_{\mathbf{q} \in \mathcal{Q}} \{T(i, j, k, \mathbf{q}) \mid (i, j, k) \text{ are indices defining a triangle as part of a cube face}\}$$

Each triangle $T(i, j, k, \mathbf{q})$ is composed of the vertices $\mathbf{v}_i(\mathbf{q}), \mathbf{v}_j(\mathbf{q}), \mathbf{v}_k(\mathbf{q})$. For example, the front face of the cube at point $\mathbf{q}$ is represented by the triangles:

$$T(0, 2, 1, \mathbf{q}) = \text{Triangle formed by vertices } \mathbf{v}_0(\mathbf{q}), \mathbf{v}_2(\mathbf{q}), \mathbf{v}_1(\mathbf{q})$$

$$T(0, 3, 2, \mathbf{q}) = \text{Triangle formed by vertices } \mathbf{v}_0(\mathbf{q}), \mathbf{v}_3(\mathbf{q}), \mathbf{v}_2(\mathbf{q})$$

The order in which the triangles are defined is important, as Unity uses a clockwise winding order, which determines that a triangle is visible when viewed from the front.

To reduce the computational load of our Unity VR application, we limit the activation of mesh colliders for physical interaction based on distance to the user's position within the world:

$$\forall T \in \mathcal{T}, \quad \text{activate collider on } T \text{ if } \min_{\mathbf{v} \in T} \|\mathbf{v} - \mathbf{x}\| < r_c$$

where $\mathbf{x}$ represents the position of the user as an interaction point, and $|| \cdot ||$ denotes the

Euclidean distance. A collider is added to a triangle if the minimum distance from any of its vertices to the interaction point is less than the collider range r_c.

Vertex blending is used to define the color of each triangle, interpolating the color of each surface using their defining vertices. In this experiment, we assign shades of brown to visually represent bins of distances, as traditional infrared LiDAR does not capture color information. Each vertex is assigned a false color $\mathbf{c}_i$ in a brown color spectrum to arbitrarily represent the cave scene. The color is modulated based on the vertex's distance d from the headset, enhancing spatial perception for the user over using uniform coloring, which would make it harder to discern distances. The following bins have been implemented for the coloring:

$$\mathbf{c}_i(d) = \begin{cases} \text{far\_color} \times \mathbf{c}_i & \text{if } d \geq 20 \\ \text{near\_color} \times \mathbf{c}_i & \text{if } 10 \leq d < 20 \\ \text{closer\_color} \times \mathbf{c}_i & \text{if } 5 \leq d < 10 \\ \text{closest\_color} \times \mathbf{c}_i & \text{if } d < 5 \end{cases}$$

Examples renderings of the resulting 3D world model are shown in Figure 6.3 (B-I) as 2D representation for reference, but as mentioned earlier they are not accurately depicting the immersed user experience as viewed from within an immersive headset.

6.3.3 Autonomy

In order to explore unknown and GPS denied caves and lava tubes within this study, we deploy two autonomy modes which allow users to interact with the robots via high-level commands: waypoint autonomy and full autonomy with interventions. While this simplifies the study design and allows for comparisons over the runtime of a full mission, an expert user might prefer to be able to actively switch between multiple autonomy modes and exploration strategies. The NeBula-powered robots utilize a world representation called Information Road Map (IRM) that encodes the world belief in a compact and scalable way [160]. The representation is graph-based, where nodes are discrete areas in space and actions are encoded in the edges; there are two tiers of IRM, a local and a global IRM. Nodes in the local IRM are represented as a dense grid centered around the current robot position encoding risk and the path length to the node. The global IRM is a sparse but connected graph representation of the world that may stretch multiple kilometers. Visited and safe-to-traverse spaces are saved as sparse breadcrumbs representing visited spaces in the pose graph, while uncovered and traversable areas are represented as frontiers. NeBula then deploys the Probablistic Local

and Global Reasoning on Information roadMaps (PLGRIM) hierarchy planners described in [160] to explore large-scale unknown environments. The full autonomy mode deploys this exploration strategy while taking into account both global and local IRMs exploring unknown frontiers. In the waypoint control mode, however, the robots utilize only the local IRM and its encoded risk to traverse to user defined navigation goals, meaning the user defines the exploration frontiers with their input. A navigation goal can be given to the robot within the range of the local IRM representation, while we do not actively expose the robots' understanding of risk to the users in this study – only the collected map representations (point cloud) in their screen and rendered VR representations are used. This high-level commanding removes the burden of local path planning from the user while interacting with the multi-robot system, which is in contrast to direct teleoperation. Teleoperated robots are often controlled in real time via direct velocity commands input via joysticks or dedicated software. Despite higher level autonomy, users might try to send the robot to goals within the local IRM that the fully autonomous system would have de-prioritized due to high risk, or the chosen goal might not be reachable at all. This means manual interventions come at the cost of being responsible for the actions that follow. When users decide to manually intervene during the full autonomy mode, this effectively reduces the autonomy and puts the robot in waypoint mode until the chosen user goal was reached with a certain margin of error. It is up to the user to recognize if a robot gets stuck in either autonomy mode, or if the robot cannot fully reach the chosen goal when operating in waypoint mode. To achieve this switching, we integrate a decision tree that prioritizes user inputs over global and local planner goals.

6.3.4 Study Design

Participants

A total of 40 participants, primarily but not exclusively students from Polytechnique Montréal, were recruited for this study. Each participant received an informed consent form in their preferred language (French or English) and signed it before participating (approved by IRB CER-2122-50-D). An important inclusion criteria of the study was that participants are proficient in English to fill out the questionnaires which were only provided in English. We were able to recruit 40 participants and include the data of n=38 participants. One participant withdrew not completing the experiments and another participant was excluded due to insufficient recorded data. In total, 32 male and 6 female participants were included in the analysis. Their ages are distributed around a mean of 25.34 with a standard deviation of 4.47, a minimum of 19, and a maximum of 36 years.

Sixteen (42.1%) participants have used VR headsets before, while the remaining 22 used VR for the first time. Most participants reported that they do not use VR on a weekly basis (86.8 %), while others stated using VR between 15 min to 7 hrs per week. More information on the participants' media consumption and gaming preferences can be found in Figure 6.6. Fifteen participants (39.5 %) indicated they have expertise in robotics. In this study, 17 participants (44.7%) declared not to have corrected vision, while 17 (44.7%) used glasses, one person used lenses, and one person shared that they have glasses but do not usually wear them, they conducted the experiments without. One single participant disclosed color blindness, one did not answer, and the majority (36) declared to not have a diagnosed or known color deficiency.

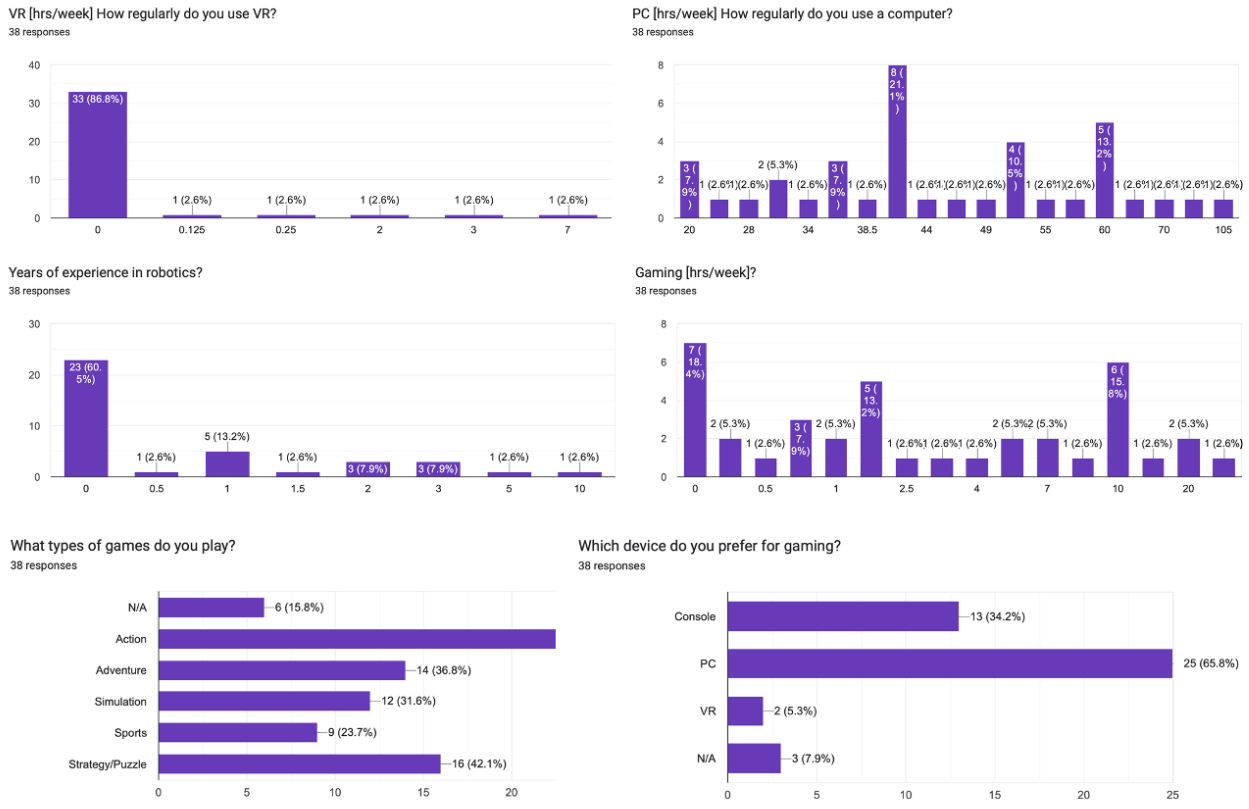

Figure 6.6 Demographic Media Consumption. Note that multiple choice was possible for *What types of games do you play?* and *Which device do you prefer for gaming?*

Exploration Task

The participants assumed the role of a robot supervisor and scientist who is interested in exploring a set of caves using a multi-robot system that is capable of locating science targets of

interest for potential follow-up missions. Participants were given the background knowledge that such caves, especially lava tubes, are of particular interest to future planetary exploration missions and to finding traces of life. Each participant was tasked to explore four caves with the goal to ensure thorough exploration.

Procedure and Independent Variables

We deploy a two-by-two within subject factorial design, with level of autonomy and interface used as the two independent variables. Specifically, we vary between waypoints (**WP**) vs. Full Autonomy (**FA**) with intervention capabilities and a 2D computer screen (**S**) versus a virtual reality (**VR**) interface, resulting in four conditions:

- **Condition 1 - SWP**: The autonomy level is low, and the 2D interface is being used
- **Condition 2 - SFA**: The autonomy level is high, and the 2D interface is being used
- **Condition 3 - VRWP**: The autonomy level is low, and the VR interface is being used
- **Condition 4 - VRFA**: The autonomy level is high, and the VR interface is being used

All subjects participate in all four conditions. Each operations run lasts approximately 20 minutes, followed by 5 minutes of questionnaires and 5 minutes of break. This method provides maximum control over extraneous participant variables (i.e., IQ, socioeconomic status, age, etc.). Figure 6.7 shows the flow of the experimental procedure, which includes a training phase in which the participants are exposed to each of the interface and autonomy modes. Participants were encouraged to ask all questions regarding the experiment during this phase. To ensure sufficient familiarity with all modalities, each participant had to find the hidden science targets in a training environment (see Figure 6.8) and demonstrate familiarity with the controls before they were allowed to proceed. After the training and a subsequent break, we collected baseline measurements of heart rate variability with participants in a resting state, focusing on a marker on the regular computer screen. During each condition run, participants were asked SPAM questions at three to five random times. Each exploration run (condition) was followed by administering the NASA TLX and our Situation Awareness, Immersion, and Trust (SAIT) questionnaires. After cycling through all conditions, including questionnaires and breaks, we conducted a debrief session where participants could ask further questions and discuss their experiences during the experiment.

The subjects' starting condition was randomized to reduce or avoid order effects. In our example, it might have been that people get used to the interface and robot system becoming

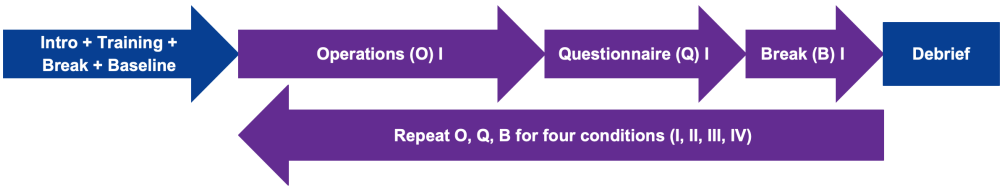

Figure 6.7 Study Procedure after Informed Consent. Note that SPAM questions were inserted at random times during the operations phase of each condition.

more comfortable using it the second time around regardless of the condition – an effect completely unrelated to the conditions that we investigate (familiarity effect). A factorial design was chosen over a simpler latin square design, as a latin square would limit the number of interaction effects that can be captured and balanced. All 24 permutations, that is the order in which the four conditions can be surveyed, were assigned to at least one participant, assigning random experimental orders of this set to the remaining participants. To further reduce effects caused by boredom and familiarity resulting from exploring the same cave multiple times, we stage each run in a different cave. Different caves naturally introduce topological differences as can be seen in Figure 6.8. The study uses two synthetic cave models (A,C) and two real-world models (B,D) with hidden science targets depicted as red circles. In (1-3) of Figure 6.8, different placements of science proxies can be seen in the real world, while (4) shows a real scientific target of interest illuminated in ultra-violet (UV) light. The exposure to UV light produces bio-luminescent markers, signs of life, that might be harder to detect using other wavelengths. However, to minimize topological effects, the caves were assigned to each run and participant analogous to how the conditions were assigned for each run.

Dependent Measurements

Objective Workload

Heart rate variability (HRV) measurements are often used as a noninvasive and objective measurement to determine stress levels and workload [153, 161, 162]. HRV can be obtained by measuring successive time differences between two regular R peaks in the QRS complex (electrocardiogram) of healthy individuals. Both inter-beat interval (IBI) and RR interval (short RR) are used as synonyms throughout the literature. Lower heart rate variability usually indicates higher stress, higher workload levels, or higher cognitive demand [162–164], when compared during similar levels of activity (not including exercise-induced changes).

The Polar H10 chest strap has proven to be a reliable and affordable sensor [161] which

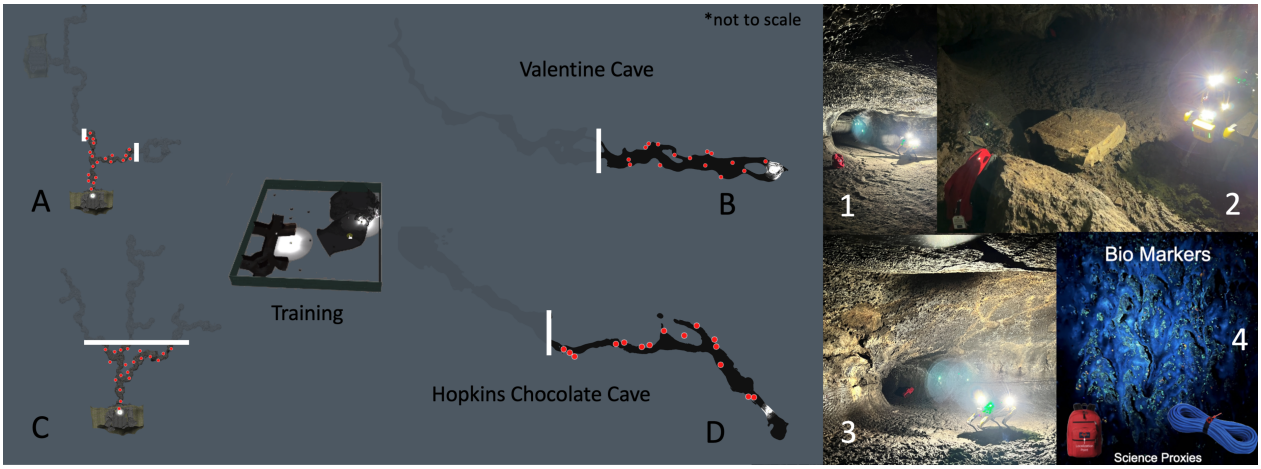

Figure 6.8 Cave topologies including training world depicting areas that are reachable within 20 minutes of exploration time (on a direct path). Greyed-out parts blocked off by white barriers cannot be reached. Red circles indicate science targets/proxies of interest. On the right (1-3) there are several example placements of science proxies (red backpacks) shown as deployed in the Valentine cave. The fluorescent bio markers (4) are actual targets of interest, here illuminated by UV light to enhance their visibility. Science proxies have been used instead of bio markers as this trip was used to collect training data for the development of an automated classification system.

provides IBI measurements as direct output. Hence, we decided to deploy the H10 for our study. To ensure comparability of results, we had our participants sit on a desk chair throughout all conditions and experiments that were conducted at Polytechnique Montréal (see Figure 6.1 left). The Polar H10 can output its data at a rate of 1 Hz, however it has to be extracted correctly. Jo et al. [26] present an ecosystem for wearable sensors and provide code in their git repository to extract data from the Polar H10. Their script seems to extract a battery state that is not encoded in the data, and they extract only a single RR interval per measured heart rate characteristic. To correctly unpack data from the Polar H10 heart rate sensor, the heart rate characteristic can be requested via the low energy Bluetooth standard GATT [165] from the sensor, specifically using handle 0x0010. The first byte of the data received indicates the presence of various data types, with bit 4 confirming whether RR intervals are included. If RR intervals are present, indicated by bit 4 being set (0x10 = 0b00010000), the heart rate in beats per minute (bpm) is contained in the second byte. Subsequent bytes contain RR interval data, where each RR interval comprises two bytes formatted in little-endian order. Each RR value is originally in units of 1/1024 seconds and must be converted to milliseconds or another appropriate unit for analysis. If multiple RR intervals are captured due to shorter inter-beat intervals, additional byte pairs are included, potentially up to nine RR intervals in a single heart rate characteristic, far exceeding typical

human physiological needs. We provide our ROS node and scripts to decode the Polar H10 online at `https://git.mistlab.ca/mistlab/ros_polar_h10`.

Subjective Workload

To assess the subjective workload of each condition, we ask users to fill the the NASA Task Load Index (NASA TLX) [141, 142] questionnaire after each experiment session. Individuals assess several psychological and task-related variables and rank these according to their importance. We have participants conduct this ranking only after the first experiment, maintaining the same weights for consecutive runs. For ease of use, we deployed the official iOS App [142] that is publicly available online. The NASA TLX questionnaire assesses the six factors: mental demand, physical demand, temporal demand, performance, effort, and frustration, to quantify the users' perceived workload. Both a weighted and raw (unweighted) average is reported, while lower values indicate less overall perceived workload.

Situational Awareness

The Situation Present Assessment Method (SPAM) is a method of measuring situation awareness (SA) online during an experiment. SPAM is based on the assumption that SA may sometimes involve simply knowing where in the environment to find some information rather than remembering what that information is exactly. In contrast to SAGAT (Situation Awareness Global Assessment Technique), the SPAM method uses response latency as the primary dependent variable while not requiring a memory component nor halting the current task [90]. Participants were asked to answer task-related questions that were interjected as a dual task at three to five random times during each experiment. The answers were given using keyboard or controller inputs. While quick response times are an indication of good SA [91], we instructed participants to prioritize correctness over response time when answering the questions. All questions were pre-recorded and played back to the participants only once. The keyword *operator* preceded each question to warn the participants of an incoming question. Questions were chosen randomly, included instructions on how to answer, and asked about information that can be seen in the interfaces (e.g., if a robot is near or far to a certain position, if they are stuck, moving, or how many science targets have been found). The full set comprises a list of ten questions shown in the appendix Table B.1.

We recorded the time to answer and the experimenter noted if the question was answered correctly with a true/false evaluation. We expect correctness of the questions to be high as the information can be deducted from the interfaces, however response times are expected to vary based on SA and potentially, because of the information that is immediately available

to the users in either interface. Moreover, VR might require a physical change of location to answer a question if the operator is present in a different part of the environment.

Subjective SA, Immersiveness, and Trust (SAIT)

In an effort to capture subjective SA and a measure of immersiveness and trust in autonomous systems, we ask participants to rate five additional statements. The questions are asked alongside the NASA TLX evaluation using a similar scale (ranking answers from 0 to 100 in 5% increments from do not agree to strongly agree). The statements of our subjective SAIT questionnaire are:

- I felt like I was exploring the cave together with the robots.
- The visualization helped me understand the terrain.
- I knew what the robots were doing approx. x% of the time.
- The robots' level of autonomy made it easy to explore with multiple robots when compared to operating multiple RC cars.
- I can rely on the system.

For the statement asking about the percentage of time the operator knew what the robots were doing, participants were required to provide their own estimate as a percentage, using the provided scale.

Körber [166] discusses if a single item is enough to measure trust in automation and concludes that it is sufficient for a global assessment. Hence, we chose to limit trust-related questions to the single statement of whether the participant can rely on the system without having to define or explain the complex and often differently perceived concept of trust with the participants.

Performance

To evaluate and compare the performance of the human-robot team under each condition, we utilize the total number of detected and inspected science targets. Finding more targets indicates that a larger part of the cave has been explored, thus providing a measure of coverage and potential scientific knowledge gain. The locations of the science targets are indicated in Figure 6.8. Results are reported as the percentage of the total targets seen by the robots, where 100% means that both robots have covered the complete cave system and detected all science targets.

Human Inputs and Interventions

Users are introduced to both interfaces and autonomy levels during a dedicated training phase. As part of the training users get to familiarize themselves with the controls and practice how to send goals to the robots. We record the number of navigation goals each user sends in the respective conditions and track how many human interventions occurred. Interventions in the way point autonomy mode are needed to direct the robots to their next goal, while interventions in the full autonomy mode turn off full autonomy until the robot has reached the manually set goal, if at all possible.

6.3.5 Statistical Analysis

To determine statistical significance, we are using two-way repeated measures ANOVA (rmANOVA) and non-parametric Wilcoxon signed rank posthoc tests reporting significance at alpha levels of $\alpha < 0.05$. Using rmANOVA, we first analyze the main effects and interactions between the two factors autonomy level and interface on the dependent variables. The method of rmANOVA is used because of its robustness to violations of the normality assumption [167] under the condition that sphericity is met. For a 2x2 factorial study design, the assumption of sphericity is automatically satisfied because there is only one pairwise within-subject variance per level that can be compared against. Despite this robustness, we assess normality using Q-Q plots. In case of significant rmANOVA results, we conduct Wilcoxon Signed-Rank tests where applicable. Significant results are reported with and without adjusted alpha levels for multiple comparisons using Bonferroni correction. Bonferroni reduces Type I errors (false positives) which could suggest that significant differences exist where none exist.

6.4 Results

6.4.1 Subjective SAIT Questionnaire

The results of our introduced questionnaire for subjective situational awareness, immersiveness and trust in autonomous systems are presented after repeated measures ANOVA and posthoc testing in the following. Users felt significantly more like they were *exploring the caves together with the robots* (rmANOVA: $p < 0.001, F = 19.25, \eta^2 = 0.142$; posthoc: $p < 0.001, W = 513$) when immersed in the VR headset ($\mu = 81.97 \pm 20.02$) compared to using the screen interface ($\mu = 60.99 \pm 31.25$). Further, it was reported that the VR *visualization helped* ($\mu = 81.25 \pm 19.82$) significantly more (rmANOVA: $p = 0.038, F = 4.64, \eta^2 = 0.042$; posthoc: $p = 0.002, W = 599.5$) in *understanding the terrain* when com-

pared to the on screen point cloud visualization ($\mu = 73.03 \pm 20.25$). The statement *I knew what the robots were doing approximately x% of the time* reports no significant differences for which interface was used, but the levels of autonomy resulted in a significant (rmANOVA: $p =< 0.001, F = 18.25, \eta^2 = 0.084$; posthoc: $p < 0.001, W = 576.5$) difference where FA users reported ($\mu = 65.53 \pm 23.06$) lower percentages than WP ($\mu = 77.89 \pm 17.95$). Depending on which interface modality is deployed, a significant difference (rmANOVA: $p = 0.042, F = 4.44, \eta^2 = 0.016$; posthoc: $p = 0.002, W = 694.5$) was observed when responding to whether the level of autonomy made it easy to accomplish the exploration task when comparing it to controlling (teleoperating) multiple RC cars, where WP ($\mu = 72.43 \pm 25.53$) responses were lower than FA ($\mu = 78.22 \pm 21.13$).

No significant differences were found for the statement *I can rely on the system*, and overall, no interaction effects were observed. Figure 6.9 visualizes the SAIT questionnaire main effects after posthoc testing as bar plots.

6.4.2 NASA Task Load Index

The results of the NASA TLX assessment are summarized in Table 6.1 after posthoc testing. Four out of the six factors that comprise the combined TLX scores show significant differences. Both raw and weighted TLX scores show significant differences for the deployed interface method while the VR workload is consistently rated approximately 11 points higher than the screen workload. Figure 6.10 depicts these results as box plots for visual comparison.

6.4.3 Heart Rate Variability

Multiple HRV measures can be derived from the collected data. Figure 6.11, for example, shows the average baseline-corrected RR intervals across all conditions per participant. Outliers in the HRV data captured by the Polar H10 have been removed for analysis using a 2-sigma threshold, and the resulting gaps in the data have been filled through linear interpolation.

Other derivatives and commonly used measures are SDNN, RMSSD, and the measured RR intervals itself which have been calculated during the analysis using functions that we implemented. Performing a repeated measures ANOVA on all of these leads to the values in Table 6.3. All measurements, derived or raw, indicate a significant difference for the interface condition. The baseline corrected RR intervals for example are significantly different with (posthoc: $p < 0.001, W = 8432.0$) and indicate higher average baseline deviation while using the screen ($\mu = 14.587 \pm 55.813$ ms) when compared to VR ($\mu = 10.081 \pm 63.758$ ms).

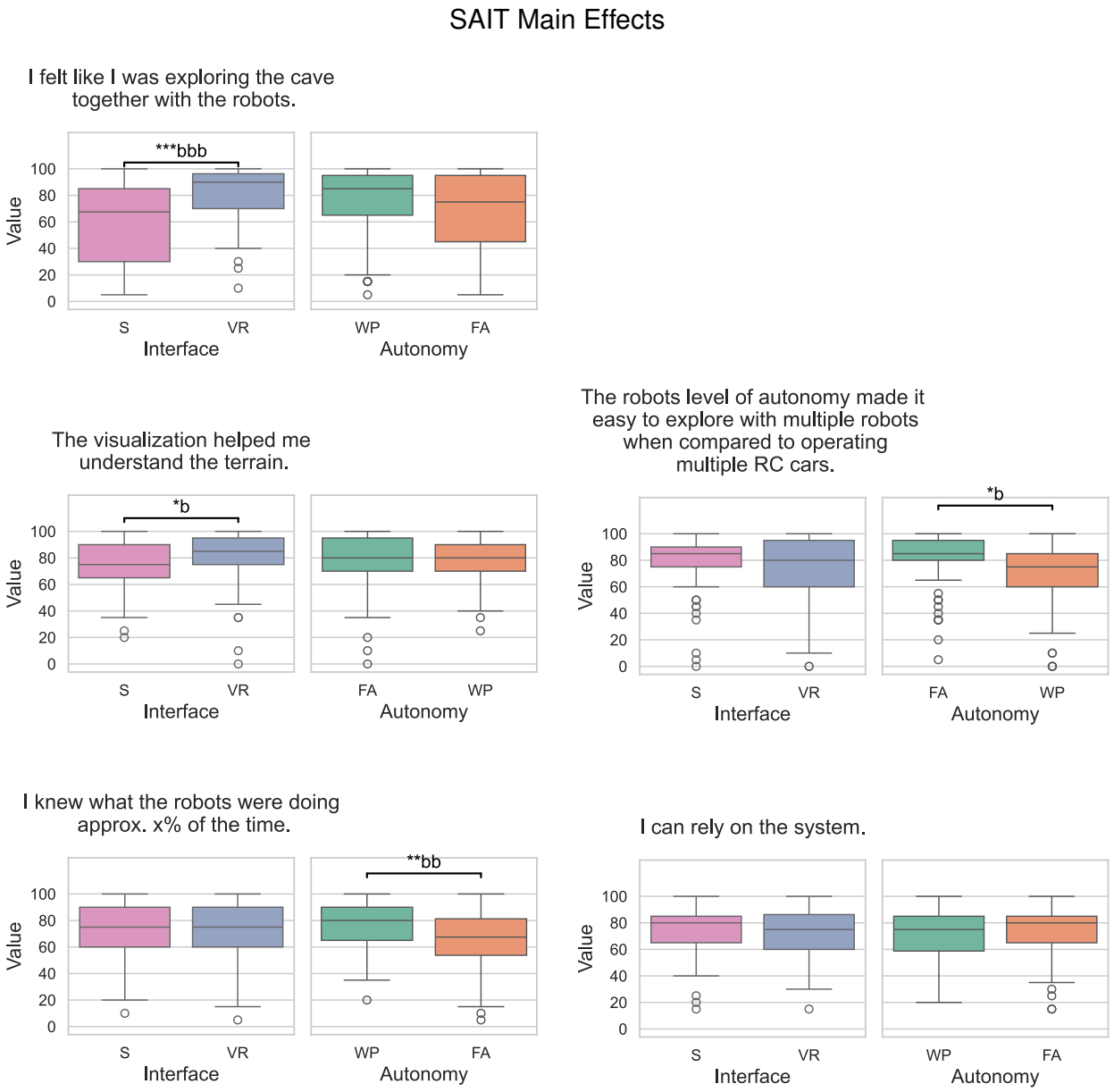

Figure 6.9 Pairwise Wilcoxon posthoc testing by independent variable with significance levels indicated by * and b after Bonferroni correction. Applied alpha levels are 0.05, 0.01, and 0.001 for one, two and three significance indicators, respectively. For interaction effects see rmANOVA tables in the appendix. N=38 included participants.

Table 6.1 Wilcoxon Test Results for NASA TLX measures. The corrected p-value is adjusted using the Bonferroni method. Significant results ($p < 0.05$) are in bold. N=38 participants.

Measure (Ind. Var.)	**W**	**p-value**	**p-bonf**	**Mean ± Std (I)**	**Mean ± Std (II)**
Mental Demand					
Interface	831.0	**<0.001**	**0.002**	47.570 ± 23.560 (S)	61.780 ± 23.050 (VR)
Autonomy	857.0	**0.007**	**0.013**	60.070 ± 21.440 (WP)	49.280 ± 25.880 (FA)
Physical Demand					
Interface	719.0	**<0.001**	**0.002**	20.860 ± 20.630 (S)	35.720 ± 27.560 (VR)
Autonomy	872.0	**0.020**	**0.040**	32.760 ± 26.420 (WP)	23.820 ± 23.630 (FA)
Temporal Demand					
Interface	1053.0	0.071	0.143	43.030 ± 23.710 (S)	50.200 ± 23.680 (VR)
Autonomy	966.0	**0.034**	0.069	50.720 ± 22.360 (WP)	42.500 ± 24.800 (FA)
Performance					
Interface	736.5	**0.002**	**0.004**	26.640 ± 17.900 (S)	37.760 ± 21.530 (VR)
Autonomy	1170.0	0.985	1.000	31.910 ± 20.720 (WP)	32.500 ± 20.420 (FA)
Effort					
Interface	708.5	**<0.001**	**<0.001**	44.140 ± 22.410 (S)	57.630 ± 21.810 (VR)
Autonomy	552.0	**<0.001**	**<0.001**	57.700 ± 19.970 (WP)	44.080 ± 24.020 (FA)
Frustration					
Interface	1014.5	0.182	0.363	43.090 ± 26.140 (S)	48.620 ± 25.690 (VR)
Autonomy	1255.0	0.599	1.000	44.470 ± 24.840 (WP)	47.240 ± 27.160 (FA)
Raw TLX Score					
Interface	689.5	**<0.001**	**0.001**	37.550 ± 15.020 (S)	48.620 ± 16.610 (VR)
Autonomy	976.5	**0.027**	0.054	46.270 ± 15.670 (WP)	39.900 ± 17.250 (FA)
Weighted TLX Score					
Interface	815.0	**<0.001**	**0.002**	41.830 ± 17.560 (S)	52.670 ± 17.340 (VR)
Autonomy	1101.5	0.088	0.175	49.940 ± 16.980 (WP)	44.570 ± 19.120 (FA)

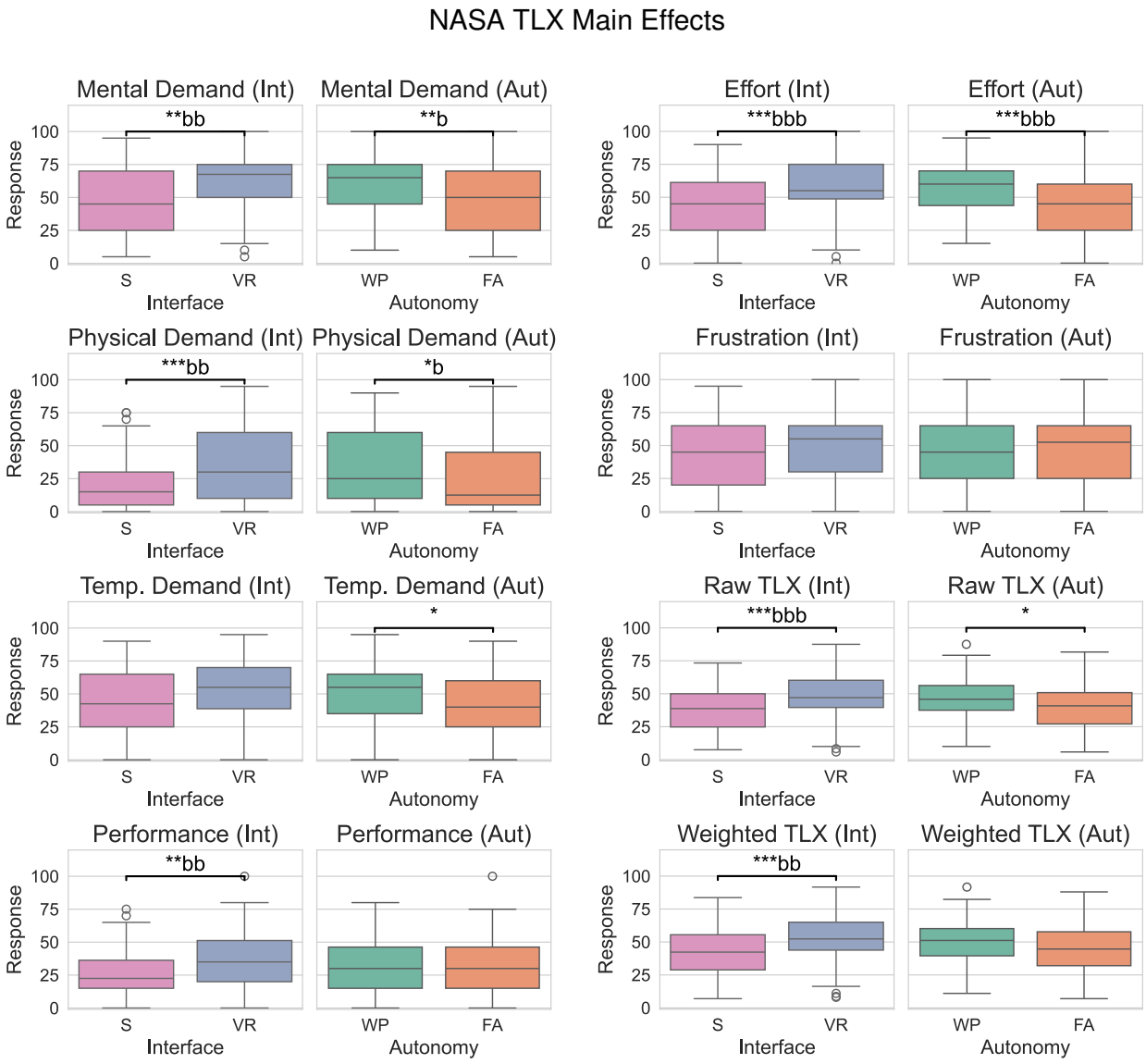

Figure 6.10 Main effects after rmANOVA and pairwise Wilcoxon posthoc testing for NASA TLX results by independent variable. Significance levels are indicated by * (uncorrected) and b (Bonferroni corrected). Applied alpha levels are 0.05, 0.01, and 0.001 for one, two and three significance indicators, respectively. Note that lower scores are better; full rmANOVA results can be found in the appendix. N=38 included participants.

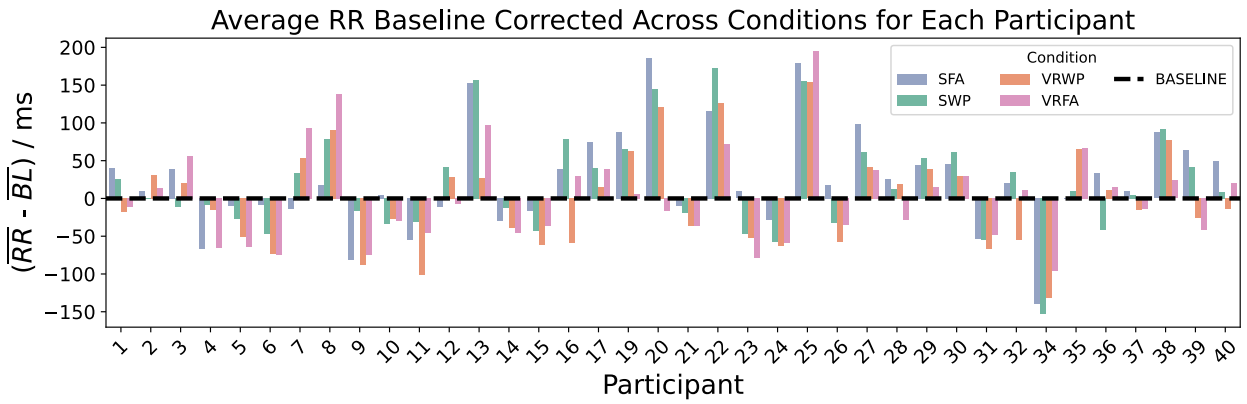

Figure 6.11 Baseline Corrected HRV per Participant and Condition

Table 6.2 summarizes the descriptive statistics for a variety of HRV measurements.

Table 6.2 HRV Interface Statistics. N=38 included participants.

Interface	Average RR	RMSSD	SDNN	RR BL Corrected
S	868.54 ± 134.30	43.25 ± 26.38	62.07 ± 28.19	14.59 ± 55.94
VR	854.04 ± 127.99	39.83 ± 23.84	55.58 ± 22.68	0.08 ± 64.04

The posthoc box plot for the baseline corrected averaged RR is included in Figure 6.18. This figure later compares the baseline corrected RR values obtained during the controlled study to pilot runs that were conducted during the BRAILLE analog exploration mission in real lava tubes.

6.4.4 Situation Presence Assessment Method SPAM

SPAM Answer Times

For SPAM, we measured response times and accuracy for the three to five randomly asked questions in each experimental run. Results include response times for all answered questions regardless of whether they were answered correctly or incorrectly. Questions that were not answered before the next question was asked have been excluded. Performing a repeated measures ANOVA (see Annex Table C.3) yields significantly different results for the interface independent variable. Wilcoxon signed rank posthoc testing confirms the significance ($p < 0.001, W = 1537.0$) and answer times are lower when using the screen ($\mu = 6.156 \pm 1.245$ s) compared to ($\mu = 8.335 \pm 4.937$ s) in VR. Thus, VR is found to be approximately 2 seconds slower on average.

Table 6.3 Two-way Repeated Measures ANOVA for HRV Metrics. N=38 participants.

Source	SS	F	p-unc	p-GG-corr	eps
Average RR					
Interface	9391.41	7.64	**0.009**	**0.009**	1.0
Autonomy	242.20	0.24	0.627	0.627	1.0
Interface * Autonomy	861.53	1.04	0.315	0.315	1.0
RMSSD					
Interface	476.83	7.42	**0.010**	**0.010**	1.0
Autonomy	9.09	0.21	0.646	0.646	1.0
Interface * Autonomy	1.78	0.046	0.831	0.831	1.0
SDNN					
Interface	1538.71	10.82	**0.002**	**0.002**	1.0
Autonomy	45.79	0.46	0.500	0.500	1.0
Interface * Autonomy	4.27	0.035	0.853	0.853	1.0
HRV Baseline Corrected					
Interface	9391.41	7.64	**0.009**	**0.009**	1.0
Autonomy	242.20	0.24	0.627	0.627	1.0
Interface * Autonomy	861.53	1.04	0.315	0.315	1.0

SPAM Accuracy/Correctness

Assessing the correctness of answered SPAM questions, we obtain significant differences using the repeated measures ANOVA ($p = 0.0170, F = 6.25\eta^2 = 0.0418$) for the interface method. Figure 6.12 shows the percentages of correctly answered questions and it can be seen that the data is skewed towards high percentages of correctness. Thus, we use a non-parametric Friedman test to confirm the robustness of the rmANOVA results. Friedman yields a significant difference for interface usage as well ($p = 0.022, \chi^2 = 9.679$). On average screen S ($\mu = 93.969 \pm 14.771$ %) yields a slightly higher percentage of correct answers than VR ($\mu = 86.952 \pm 18.810$ %). SPAM performance for both answer time and correctness are summarized as box plots in Figure 6.13.

6.4.5 Detection Performance

To gauge exploration coverage and overall system performance, we evaluated how many of the hidden scientific targets were found in each run. A significant difference has been identified by rmANOVA for the interface modality ($p < 0.001, F = 25.84, \eta^2 = 0.163$) and interactions between interface and autonomy ($p = 0.012, F = 7.026, \eta^2 = 0.040$). Figure 6.14 summarizes the detection performance after posthoc testing. Significant differences are confirmed for the

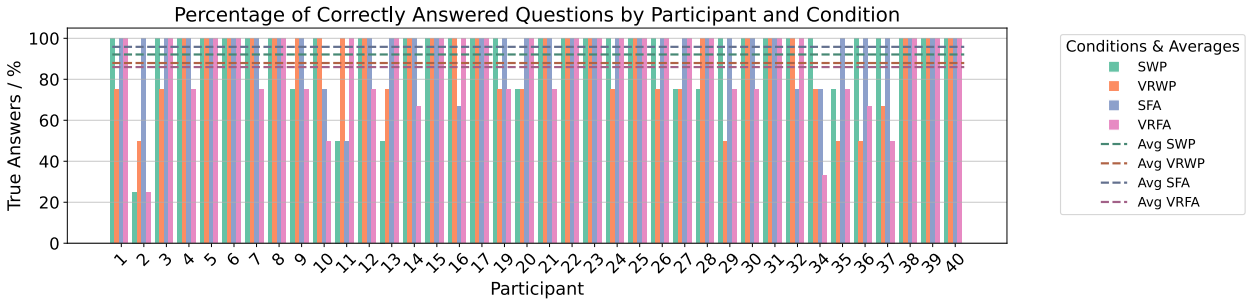

Figure 6.12 Percentage of Correctly Answered Questions by Participant and Condition.

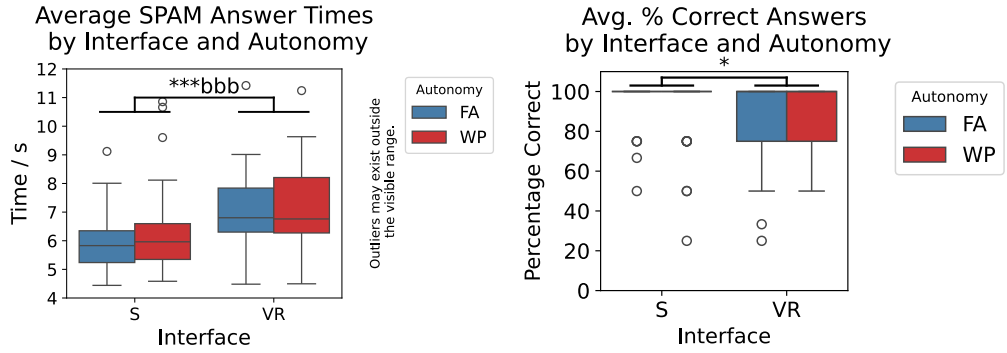

Figure 6.13 SPAM Performance box plots showing Answer Times and True Answer Percentage by Interface and Autonomy with significant differences between Screen and VR interfaces. The variance of correct answers for the screen interface is very low and almost always 100%. Thus, the boxes have collapsed only showing a median line and outliers. The median for both interface levels is identical. Significance levels are indicated by * (uncorrected) and b (Bonferroni corrected). Applied alpha levels are 0.05, 0.01, and 0.001 for one, two, and three significance indicators, respectively. N=38 participants.

detection performance with the screen interface ($\mu = 0.405 \pm 0.189$, max=0.90), leading to higher results than the VR interface ($\mu = 0.261 \pm 0.147$, max=0.675). Further, significant interaction effects can be observed between conditions SWP and VRWP, VRWP and SFA, SWP and VRFA, as well as VRWP and VRFA. The lowest performance is obtained for VRWP ($\mu = 0.205 \pm 0.105$), then VRFA ($\mu = 0.317 \pm 0.163$) followed by SFA ($\mu = 0.394 \pm 0.182$) with no significance between them and on top SWP ($\mu = 0.415 \pm 0.196$), without significant difference between SWP and SFA. After Bonferroni correction VRFA, SWP, and SFA performed at similar levels without significant differences between them.

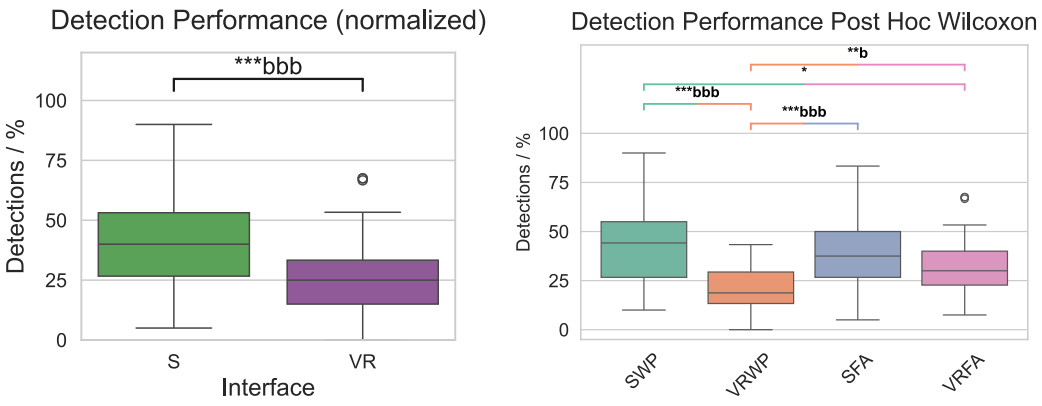

Figure 6.14 Significant Post Hoc Results with Wilcoxon for Detection Performance. N=38 participants.

6.4.6 Human Inputs and Interventions

Figure 6.15 shows the significant differences obtained for the number of interventions that were recorded. No significant difference was found between the usage of different interfaces; however, the level of autonomy significantly influenced how many human inputs were made. The full autonomy mode averaged at ($\mu = 31 \pm 18$, min $= 3$, max $= 82$) inputs, while waypoint mode required more goals to be sent, averaging at ($\mu = 92 \pm 42$, min=35, max=305) user goals. The achieved significance is high with ($p < 0.001, W = 12$). In addition, the interaction effects between interface and autonomy yielded highly significant differences between SWP and SFA, SWP and VRFA, VRWP and SFA, as well as VRWP and VRFA. The total number of interactions per participants and condition are depicted in Figure 6.16.

6.4.7 Pilot Study in the Wild

Prior to conducting the here presented user study, we deployed the system at the Lava Beds National Monument in Northern California with real robots in the loop. Three pilot participants with expert robotics knowledge tested the interfaces under different autonomy modes in the same four conditions. The overall raw NASA TLX results for the four investigated conditions in the real-world deployment are: SFA=35, SWP=62, VRFA=42, and VRWP=50 [140]. Figure 6.17 shows all NASA TLX sub-scales and individual measurements for completeness.

During the pilot study we also collected HRV data from the participants. In Figure 6.18 we compare the baseline corrected RR between the study (left) and the field deployment (right). The limited number of pilot participants allows only for a qualitative comparison, and it has to be noted that due to a sensor disconnect, no HRV data was collected for one

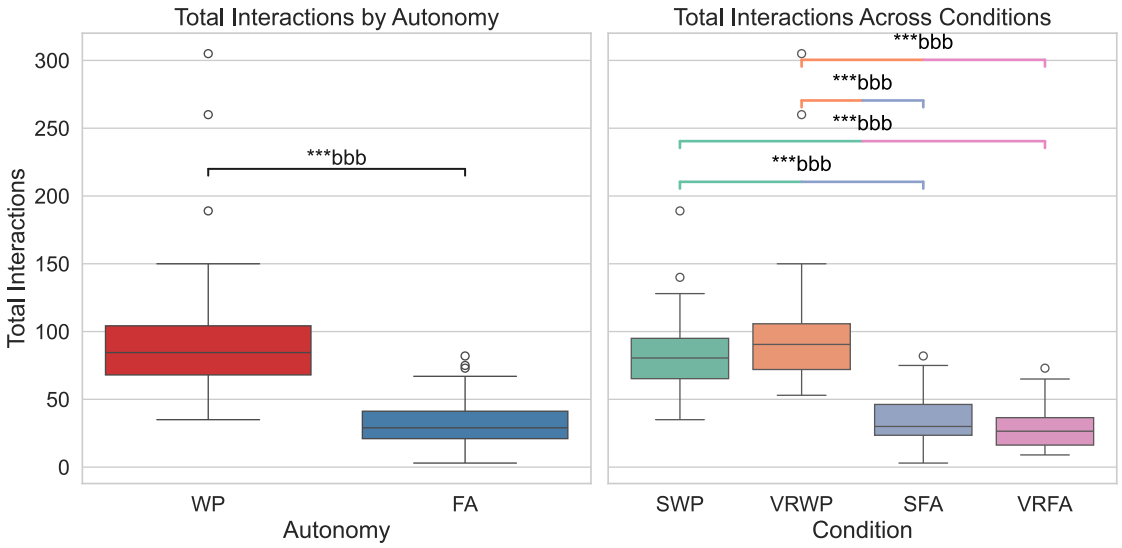

Figure 6.15 Significant differences for Human Inputs and Interventions by autonomy level (left) and per condition (right). Bonferroni corrected significance levels are indicated by 'b'. N=38 participants.

of the BRAILLE participants during their VR experiments. Additionally, operators in the real world conducted all experiments while standing at either the base station screen setup or using VR, which allowed for more movement than during the controlled study. Visually, there is very little difference in the baseline corrected RR measurements when compared to the fielded experiment. The RR measurements for VR are significantly lower than for the screen within the study. In contrast, VR yields higher baseline corrected RR in the field pilot visually, albeit within the error bars of the controlled user study.

Finally, we can look at SPAM answer times between the lab and field study. Table 6.4

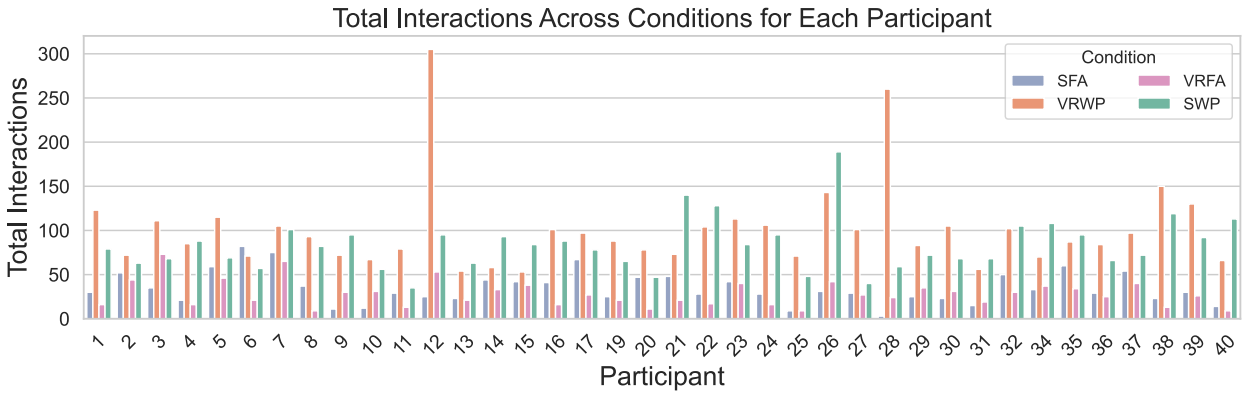

Figure 6.16 Total Interactions per Participant and Condition.

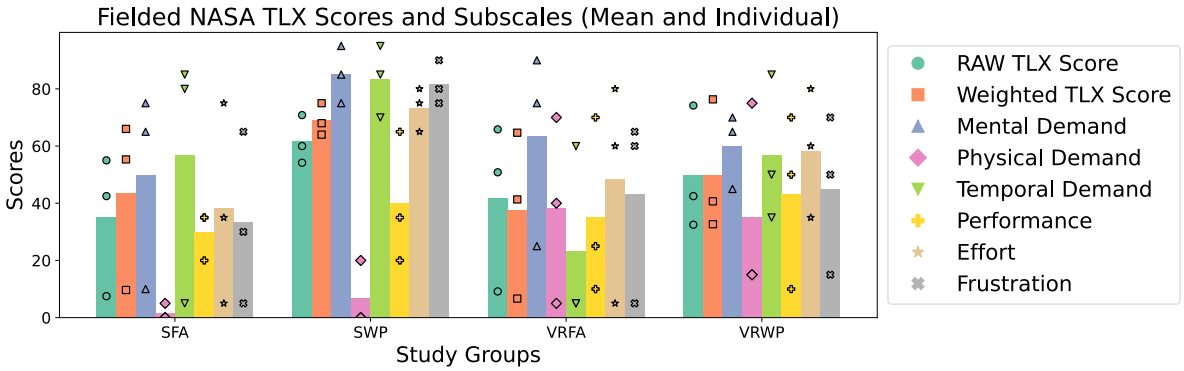

Figure 6.17 NASA TLX Scores Obtained During a Pilot in the Field with n=3 Participants.

contains descriptive statistics for the measured SPAM answer times for correctly answered questions. The lab study yields slightly higher answer times. Response times ordered from fastest to slowest conditions for the lab conditions are $SFA < SWP < VRWP < VRFA$. The fielded experiments result in a different order $SWP < VRWP < SFA < VRFA$, however VRFA has the slowest average response time in both cases. Note that during the field trial, response times were measured by the experimenter, marking the starting time of a correct verbal answer.

Table 6.4 Descriptive statistics of correct answer times, comparing Lab and Field data

Statistic	**SWP**		**VRWP**		**SFA**		**VRFA**	
	Lab	Field	Lab	Field	Lab	Field	Lab	Field
Participants	38	3	38	3	38	3	38	3
Mean	6.315	2.493	7.868	3.889	5.996	4.056	8.802	6.667
Std Dev	1.446	1.362	2.875	1.018	1.009	3.853	6.404	2.333
Min	4.584	1.000	4.496	3.000	4.441	1.667	4.483	5.000
Max	10.856	3.667	16.783	5.000	9.120	8.500	37.395	9.333

6.5 Discussion

Our study aims at finding insights whether the use of different interface modalities and different autonomy levels influence workload, situational awareness and performance.

6.5.1 Agreement of Objective and Subjective Workload

The first research question we asked was **(RQ1)**: *Do continuous physiological measurements, i.e., heart rate variability from portable devices, differ from self-assessed (NASA TLX) met-*

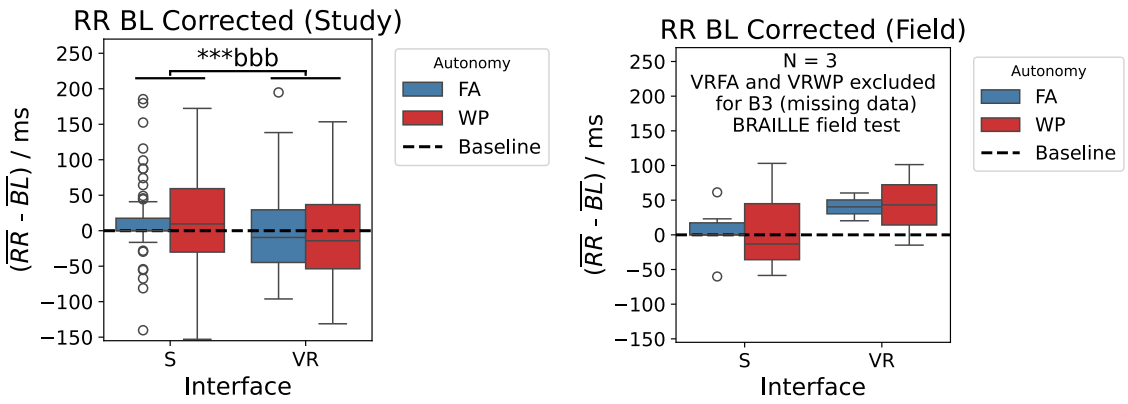

Figure 6.18 Baseline Corrected HRV for the Interface Factor for both the laboratory-controlled user study with n = 38 included participants (left) and the field deployment in the lava beds national monument (right). The significance indicator (left) results from pairwise Wilcoxon tests with Bonferroni correction. The field deployment plot serves as qualitative comparison with only n = 3 participants (2 for VR due to a sensor failure in the wild). VR participants in the field performed the experiment while standing, whereas the bigger study constrained participants to a desk chair.

rics? To answer this question, we deployed the portable Polar H10 heart rate monitor to objectively measure HRV as a proxy for workload and stress. The NASA Task Load Index subjectively assesses the workload from user provided ratings. Both objective and subjective measurements identified significant differences for the levels of interface that were used. Lower NASA TLX scores indicate less workload, and the screen conditions resulted in about 11% lower scores overall. Higher heart rate variability indicates lower levels of workload. Our subjective measurements (Table 6.2) indicate greater variability in the screen conditions, which aligns with the objective HRV data showing lower workload (greater variability) for screen interfaces. In understanding why less autonomy caused a higher workload, the number of human interventions (user goals) could be a contributing factor. Participants used approximately three times more interventions for the waypoint exploration mode ($\mu_{WP} = 92 \pm 42$, min=35, max=305 vs. $\mu_{FA} = 31 \pm 18$, min = 3, max = 82) and reported significantly higher ratings for effort and physical demand on the NASA TLX sub-scales. In contrast, the objective HRV measurements yield no significant differences in the autonomy levels. In [168] users were found to be more efficient teleoperating a Baxter robot in VR using waypoints (positional control) over using a method resembling traditional keyboard inputs (trajectory control). However, they did not find statistical differences in NASA TLX scores between conditions. Grier et al. [169] conducted a comprehensive meta-analysis on global NASA TLX Scores covering 237 studies across a variety of domains. In the realm of robot operations,

they included 167 studies and reported a range of weighted and raw NASA TLX scores: a minimum of 9.59, 25th percentile of 41.00, median of 56.00, 75th percentile of 63, and a maximum of 80. Our screen interface received a mean TLX score of 37.55 ± 15.02, positioning it at the low end of the ratings within the 25th percentile of scores. Similarly, our VR interface with a raw TLX of 48.62 ± 16.16, placing it comfortably within the low-moderate workload range as defined by Grier et al.'s findings.

6.5.2 Visualization Influences on Performance

Research questions **(RQ2)** was formulated as: *Is the 3D environment in virtual reality influencing performance positively or negatively, if at all?* This question can be answered using multiple assessed metrics. Subjectively, users perceived higher performance using the screen interface (indicated by significantly lower values on the NASA TLX performance sub-scale) than compared to VR. The SAIT questionnaire, in contrast, indicated that the visualization in VR helped them to better understand the terrain than the screen interfaces. Objectively, we measured the number of science targets that were detected as an indication for coverage and performance. Overall, VR conditions yielded significantly less coverage when compared to the screen modalities. however, interaction effects reveal that SFA ($\mu \approx 39\%$) and VRFA ($\mu \approx 32\%$) perform at similar levels with no significant differences. Thus neither interface is influencing the performance positively or negatively when using FA. Highly significant differences are observed for the waypoint inputs, which could be attributed to differences in the used peripherals and input methods, as well as familiarity with the equipment. Robots might explore completely separate physical spaces in an environment with large distances between them; similar drawbacks were found in [170]. Further, the robots might be out of view in the VR environment as it is true to scale. This would require locations switching to observe either robot at a given time in VR. Despite having the option to jump to the location of each robot immediately (in addition to walking or teleporting), there seems to be an overhead for navigating the virtual world in VR. Overall, the VRWP condition performed the worst at $\mu \approx 21\%$ coverage, which advocates for higher levels of autonomy and automation when it comes to selecting where to explore next.

6.5.3 Influence of Autonomy and Interfaces on Situational Awareness

The study yields no significant differences for situational awareness and autonomy levels. Thus, the answer to **(RQ3)**: *Is higher autonomy reducing or increasing situational awareness, if at all?* is that there is no evidence of autonomy influencing SA. However, a significant difference in SA has been observed for the interfaces used where average accuracy for the

screen interface is 94% compared to 87% for VR. While statistically significant, the accuracy can be rounded to about 90% for either interface and would be within a 5% margin of error, indicating similarly good SA for both interfaces. Answer times on screen are lower, indicating that information needed to provide SA might be available more readily than in VR. Absolute differences are limited to a couple of seconds on average. It could be argued that SA is another measurement of performance, which is shown not to be different for the level of autonomy used and, at best, similar when looking at the interfaces.

6.5.4 Impact of Workload and Autonomy Levels on Trust

We reduced the notion of trust to a single question within the SAIT questionnaire and did not observe any evidence of different levels of trust between conditions. Thus, the findings regarding **(RQ4)**: *Do higher levels of trust correlate with higher autonomy and/or physiological measurements?* do not provide sufficient evidence to reject the null hypothesis (H_0 = there is no correlation between trust and the variables of autonomy or physiological measures). This indicates that there is no significant correlation between trust and the variables of autonomy or physiological responses within the limitations of this study. The question of whether high workload and stress yield differences in the level of trust remains unanswered in this study.

6.5.5 Limitations

Participants were predominantly students pursuing higher engineering degrees. Most were experienced computer users, however more than half had no prior exposure to using VR devices. Despite others having used VR headsets before, limited exposure compared to computer screen interfaces might pose a limitation of this study. This novel technology might have required additional workload to complete the exploration tasks for some users (novelty effect). While designing interfaces for both experts and non-expert users simultaneously is desirable [136], including both might be limiting generalizability to specific populations. A large percentage of participants declared to have expertise in robotics, which does not necessitate familiarity with autonomous systems. This poses another limitation, as not knowing how the autonomy was implemented users might have used additional mental resources to reason about autonomous decisions, leading to unnecessary interventions, thus limiting performance. Gender balance was not achieved, as only six female participants took part in the study. Another limitation might be fatigue as each complete experiment cycle lasted between three and four hours, however, counterbalancing conditions should have minimized resulting order effects.

6.6 Conclusion

This study focused on developing and validating interfaces for multi-robot planetary exploration missions. We assess the influence of two interfaces and two autonomy levels on workload, situational awareness, and performance. Specifically, we contributed a fielded VR interface with real-time rendering capabilities enabling large-scale exploration of cyber-physical spaces. A human can explore the resulting world model freely, while interacting with multiple robots. We conducted a user study recreating and resembling a fielded system with n=38 participants. We assessed objective workload with a low-cost wearable HRV sensor and compared results to subjective NASA TLX scores. Further, we introduced a short questionnaire for Situation Awareness, Immersion, and Trust (SAIT) that shows that users feel like they understand the terrain better using the VR representation over the screen interface with the simple point cloud.

The user study finds continuous physiological measurements match the objective equivalent of NASA TLX measurements when comparing interface modalities. Using the VR interface over the screen interface resulted in higher workloads, yet detection performance is similar; except for the VRWP condtition. Posthoc SFA, VRFA and SWP perform at similar levels, while VRWP yields the lowest number of science targets detected. Compared to the meta-analysis of NASA TLX scores conducted by Grier et al. [169], our screen interface demonstrates workloads within the 25th percentile, indicating low workload, while our VR interface is categorized in the low to moderate workload range.

Lower autonomy caused significantly higher self-reported workload, while objective HRV measurements do not provide evidence that the level of autonomy is a dominant workload contributor. Lower autonomy, however, required approximately three times more work in the form of human inputs.

No evidence is found that autonomy levels influence situational awareness positively or negatively. Instead, a significant difference in SA can be observed for the interfaces used. Be that as it may, looking at the absolute values of correctness has both interfaces perform at an accuracy of approximately 90% (rounded to the nearest ten percent). In absolute terms, this means we introduced a VR interface provides high levels of situational awareness that are on par with the screen interface. More training and thus higher levels of expertise and familiarity could likely reduce differences even further.

Looking at trust measures, breaking it down to a single question did not yield significant differences between conditions. This motivates future work with a dedicated focus on trust and trustworthiness in autonomous systems. We must include that the lack of expertise in

autonomous systems or novelty effects might have influenced the results. However, knowledge autonomy implementations is not required if interfaces deploy more insights on decision-making.

The results of this work contribute to the understanding of objective and subjective workload measurements for operating multi-robot teams in realistic planetary exploration scenarios. Future work comparing interface modalities can safely deploy online HRV measures to assess workload under similar conditions as presented in this body of work. Other non-invasive workload measurements, such as pupillometry or combinations of objective measurements, are an interesting research area. Adaptive autonomy systems that employ online assessment are still in the early stages and could benefit from larger-scale fielded experiments [44, 164, 171, 172]. This presents exciting avenues for future contributions.

Acknowledgments We acknowledge the support of the Canadian Space Agency (CSA) (19FAPOLA32). The authors thank the NASA BRAILLE Project funded under the NASA PSTAR Program (NNH16ZD A001N), especially Jennifer G. Blank, and the US National Park Service for facilitating permission to work at the Lava Beds National Monument. We would like to thank the whole NeBula team for developing many parts of the autonomy system that we used in our study. Further, we thank the team for the field support that we received. Thank you Benjamin Morrell, Maira Saboia da Silva, Christopher Patterson, Jose Uribe, Xianmei Lei, Sangwoo Moon, and Taeyon Kim. The authors further thank Emily Coffey from the psychology department of Concordia University, Canada, for the discussions around the experimental design. We want to thank the pilot participants that helped to improve the experience for the laboratory study and finally thank you to all participants that provided their time and data to inform this study.

CHAPTER 7 DISCUSSION AND CONCLUSION

7.1 Discussion of Works

The research presented in this thesis aims at improving the fields of human-robot interaction and multi-robot systems, particularly in the context of peaceful applications like exploration and disaster relief. This work endeavored to achieve the four research objectives outlined in Section 1.3 by following the five work packages introduced in Section 3.1. The work packages have been addressed by the main body of this work, the three articles presented in Chapters 4, 5, and 6.

Specifically, **Objective I** – the definition of collaboration protocols between humans and robots targeting the exploration of unknown environments with multi-robot teams – lead to the creation of an autonomy assistant and integrates the human and their base station as entities of a human-robot system, thus extending a traditional interaction graph by a base or interface agent (compared to Figure 1.3).

Objective II – the implementation of a software architecture and the tools needed to reduce workload in human and multi-robot systems – is incrementally realized by all presented articles. Starting from the initial iteration of the autonomy assistant Copilot MIKE (Chapter 4), and the evolution of Copilot to an improved game-inspired screen interface with higher autonomy and resource-constraint task planning. Finally, the integration of biophysiological measurements and the implementation of a processing pipeline for a novel VR interface (that in itself addresses the next objective).

Objective III – the creation of an intuitive interface for the robotic system that allows effective control and interaction with minimal training and little cognitive load overhead for a single human supervisor – is achieved in both Chapters 5 and 6, presenting a screen and virtual reality based method, respectively. The game-inspired screen interface subjectively reduced workload and task switching, while its full-screen mode was deployed and validated as part of the user study in Chapter 6. The work presented in Chapter 6 additionally presents a methodology to create a performant and immersive VR interface for HMRS, which is validated in comparison to the screen interface. Both have been used by expert and non-expert users and allowed multi-robot operations by a single human supervisor with minimal prior training.

Objective IV – the validation of our approaches through realistic exploration and disaster-relief exercises – has been demonstrated and validated during the DARPA Subterranean

Challenge in a search and rescue context, exploring various underground structures, as well as through the field experiments in the caves of the Northern California Lava Beds National Monument. The user study presented in Chapter 6 adds to this validation by deploying an identical processing pipeline (Figure 6.2) and interfaces (Figure 6.3) compared to the real-world experiments. At the same time, the concept of operations remained the same and users, expert or not, successfully deployed multi-robot systems for planetary exploration.

To the best of the authors knowledge, there has not been any work that investigated the influence of autonomy and interfaces on a single-human multi-robot system investigating both subjective and objective ergonomics, while utilizing real-time LiDAR point clouds of large-scale environment in the visualization pipeline; and while creating a cyber-physical explorable space in VR.

Many works that designed VR interfaces, or more general VAM HRI systems, did not use 3D command sequencing as an input modality in their applications, and the most popular way of creating situational awareness was by live streaming video data [67]; potentially because of how VAM allow better depth perception for stereoscopic views. Such views, however, are limited to the view point of the camera used and do not allow a robot-exocentric perspective. While some works utilize 3D environments, they often rely on reconstructed digital twins that are not produced online. Few works deploy LiDAR only solutions, instead RGB-D cameras are often used [21] and fields of view are limited to smaller objects, or live-streams of raw-data limited by the field of view again. One work has been looking at solely the reconstruction aspect of larger environments using circular surface elements (splats) with varying radii achieving 60 fps renderings of static scenes in VR [173]. Specifically, they are presenting the reconstruction of architectural landmarks, i.e., the front view of the Bruxelles city hall, which is a small environment compared to the large natural caves that we are exploring. Their work, like ours, takes advantage of frustum culling, that is, the limitation of rendering to the current field of view within the headset (plus a small margin for movements); frustum culling has become a standard feature as part of many VAM software development kits. A variation of splatting called 4D Gaussian splats is introduced in the work of Wu et al. [174], which is taking into account dynamic scenes. A downside of this technique is the large training time that is needed to create a model; it is in the range of several minutes. With the presented work in Chapter 6, we integrate rendering of high-volume, and large-scale real-time dynamic environments into a cyber-physical space that allows robot-exocentric exploration of the environment – the human user can walk around freely in the ad-hoc created environment to gather situational awareness and understand limitation imposed by the terrain. After all, users reported that the VR representation helped them to better understand the environment subjectively, while task performance was mostly on par with screen interfaces, except in the

case of the combined usage of VR and the waypoint-based low autonomy modality.

7.2 Limitations

While each of the presented articles has its own limitations, it is important to note that some have been addressed incrementally. For instance, the original Copilot MIKE did have scalability issues that needed to be addressed to become a viable system for more than three robots. A purely linear task schedule would easily exceed the time horizon, especially when tasks kept being deferred. Further, some timing differences were observed when comparing simulation results and real-world robot operations, because human and machine errors were not modeled in simulation. This shows the importance of going beyond simulation results when validating HMRS — it is crucial to address the sim-to-real gap. These shortcomings were identified early and addressed in the updated Copilot version that is presented in Chapter 5. The follow-up system allowed for parallel task execution and was tested with up to 11 robots during field tests in preparation for the DARPA SubT Challenge. Hence, scalability issues might exist beyond that number of robots. One immediate limitation comes to mind and that is available screen estate for the miniature robot status cards. However, with minor modifications, it is the authors' belief that heterogeneous human-robot teams with sizes between 20-30 robots could be handled by the system, given the robotic hardware is reliable. Yet, with larger groups of robots, the human might become the limiting factor again.

The validation results we presented might be biased by different demographic factors such as gender, age, and levels of expertise in robotics, particularly in autonomous multi-robot systems, and familiarity with interface technologies. This is applicable to all presented works of this thesis. As autonomy systems become more and more autonomous, decision making and reasoning might require additional transparency considerations, to prevent unnecessary interventions that could limit performance. Further, despite being real-time capable, the visualization is limited in resolution.

Finally, human trust in the system and trustworthiness of the system have only been explored tangentially in this work. While we presented the Situational Awareness, Immersiveness, and Trust (SAIT) questionnaire in this work, the single question addressing trust might have been too limiting for our study. This was a risk taken to limit the number of questions asked to the participants, as the total duration of an experimental session lasted between 3-4 hours including informed consent and training. Nonetheless, fatigue could present an additional limitation, however, counterbalancing conditions should have minimized resulting order effects.

7.3 Potential Impact

Academia: We bridged works between psychophysiology and engineering, informing the design of future systems. The presented methodologies and obtained validation results are a promising contribution to interface design and real-time heart rate variability assessment with low-cost, wearable devices in HMRS. The demonstrated approach could be used as input for adaptive automation systems in various domains beyond search and rescue and exploration, e.g. self-driving cars and fleets, which are assumed to have some kind of human oversight at higher levels. The work could be used to inspire further exchanges between human sciences and engineering in applied settings even outside of the multi-robot realm.

Industry and Economy: Reliable human-autonomy teaming paradigms and effective interfaces, as presented in this work, increase situational awareness and performance, which serves a a widespread enabler for HMRS. This, will have a positive impact on industry and the economy. Outside of disaster relief and exploration, the chemical or mining industries could apply the presented works in their inspection or extraction tasks. A single human supervisor could then oversee a large team of robots and perform hazardous or tedious tasks. In addition to potential performance gains, the humans involved would experience increased comfort and reduce risks by deploying machines; wearing uncomfortable personal protective equipment and squeezing through tight inaccessible spaces will then not be necessary for nominal operations. Especially inspection tasks could be elevated by real-time cyber-physical co-location.

Society: Involving humans and integrating human supervision into automation processes, as demonstrated in this work, leads to enhanced acceptance by society. The presented work has demonstrated that non-expert users can manage multi-robot teams in complex environments with minimal training. Considering a disaster relief scenario, this capability opens access to a vast pool of potential volunteers who would otherwise not be able to contribute to operations. Overall, the realization of autonomous robot teams will present society with transformational opportunities, such as increasing safety and efficiency.

Commercial: Consumer electronics could be used to port the presented methods onto mobile devices. Recent smartphones with integrated LiDAR sensors could then be used for large-scale distributed mapping. Further, applications could be computer games. Specifically, biofeedback gaming which makes use of psychophysiological measurements to adopt the difficulty levels in games. To challenge players even more, an anti-Copilot system that

plans and schedules task with the goal to overwhelm the user to a certain extend could be implemented.

Space Exploration: The motivation for this work originates in answering big existential and philosophical questions, broken down to smaller tasks that can only be achieved by multi-robot systems that collaborate. One example is the CADRE mission where robots autonomously and collaboratively take ground penetrating radar measurements while driving in formation. A single robot would only be able to obtain a 2D slice, whereas via triangulation between agents one can measure a 3D volume. The published works of this thesis have already inspired small elements of the CADRE mission operations and ground data system visualization tools on screen interfaces. In the near-term future, the planned Artemis missions that aim to explore the lunar south pole could deploy HMRS with cyber-physical, real-time interfaces. This includes Copilot capabilities for integrated task planning and scheduling capabilities, which could make missions safer and more effective.

7.4 Future Work

The presented work surrounding Copilot could be expanded by replacing static, task-based estimates of workload and a fixed upper limit with dynamic and adaptive constraints that are derived or learned in real-time. Additionally, multiple sources of non-invasive workload measures could be combined for more precise and redundant workload estimates. Further, the presented approaches could be implemented for novel VAM devices such as the Apple Vision Pro or other future devices. The extension towards human-swarm interfaces, especially when going beyond 20 robots, opens up opportunities to integrate novel interaction modalities. These could include grouping robots into sub-teams, implementing a leader-follower based approach, or other emerging behaviors. Further, it would be great to overcome rendering resolution limitations. One research avenue could be to improve Gaussian splatting techniques like [174] and enable faster (real-time) model creation. Should this not be possible, however, a dynamic loading strategy could be implemented that post-processes the data and updates already visited areas as higher level of detail renderings become available. Future experiment designs and validation studies should investigate, trust and trustworthiness in HMRS in higher levels of detail. Moreover, evaluating whether shorter experiment sessions could provide sufficient validation data would be of interest. However, depending on the task at hand or the team composition, shorter experiment times might be too restrictive to capture several interaction effects. Human-autonomy teaming (HAT) remains an active research topic in the human factors and ergonomics domain, and future work should address the lack

of detailed use-cases that go beyond the boundaries of this thesis. The integration of new AI and machine learning techniques, including large language and foundation models will create novel HAT use cases that the community can leverage and learn from [175]. While we utilized several elements of HAT to create our symbiotic HMRS, [175] describes four critical factors: security, privacy, confidentiality and trust – all of which require future work on both the human and machine teaming side.

A specific future space mission involving a multi-robot system with human supervisors is the CADRE mission. CADRE is planned to operate for approximately one lunar day, the equivalent to 14 consecutive Earth days. This scenario requires consideration of a broader set of human factors, even though the robots will operate fully autonomously.

7.5 Conclusion

In conclusion, the body of works (i) defines collaboration protocols between humans and robots targeting the exploration of unknown environments with multi-robot teams, (ii) implements the software architecture and the tools needed to reduce workload in human and multi-robot systems, (iii) creates multiple intuitive interfaces for a robotic system enabling effective control and interaction with minimal training, while maintaining low cognitive loads, and (iv) validates the approaches through realistic exploration and disaster-relief exercises.

The results indicate that we have created a human and multi-robot planetary exploration system that leads to several symbiotic effects for both the human, and the robots. Most frequently the effects are mutually symbiotic, however some might only benefit one side. Human workload is decreased, while performance and active operation times increase. Longer continuous exploration times are achieved by higher levels of autonomy, which in turn leads to higher area coverage. High levels of human situational awareness can reduce risks of operational robot failures. Alongside this, the human benefits from not being exposed to hazardous environments, while the robot might not be affected by the risk or impacted negatively. Furthermore, by validating the relationship between subjective and objective workload measurements in interface design, we create a strong feeling of immersiveness, nurturing a symbiotic bond between humans and the HMRS's machine-interface. Finally, this work, which enhances human-multi-robot system capabilities for planetary exploration, has been validated and fielded in realistic analog space exploration and disaster relief missions. These efforts not only contribute towards breaking the 'one mission, one rover paradigm' and foster human-autonomy teaming, but also pave the way for upcoming missions like Artemis, and potentially, a key to answering fundamental existential questions.

REFERENCES

[1] D. Morrison, "The nasa astrobiology program," *Astrobiology*, vol. 1, no. 1, pp. 3–13, 2001, pMID: 12448992. [Online]. Available: https://doi.org/10.1089/153110701750137378

[2] N. Aeronautics and S. A. (NASA), "Nasa mars website," 2024, last accessed: July 11, 2024. [Online]. Available: https://science.nasa.gov/mars/

[3] J.-P. de la Croix *et al.*, "Multi-agent autonomy for space exploration on the cadre lunar technology demonstration," in *2024 IEEE Aerospace Conference*, 2024, pp. 1–14.

[4] E. C. Rigobelo, *Symbiosis in Nature*. Rijeka: IntechOpen, Jun 2023. [Online]. Available: https://doi.org/10.5772/intechopen.105293

[5] P. N. Nguyen and S. M. Rehan, "Wild bee and pollen microbiomes across an urban–rural divide," *FEMS Microbiology Ecology*, vol. 99, no. 12, p. fiad158, 11 2023. [Online]. Available: https://doi.org/10.1093/femsec/fiad158

[6] J. M. Carrillo *et al.*, "Living on the edge: Settlement patterns by the symbiotic barnacle xenobalanus globicipitis on small cetaceans," *PLoS ONE*, vol. 10, no. 6, 2015.

[7] S. Moutailler *et al.*, "Co-infection of ticks: The rule rather than the exception," *PLOS Neglected Tropical Diseases*, vol. 10, no. 3, pp. 1–17, 03 2016. [Online]. Available: https://doi.org/10.1371/journal.pntd.0004539

[8] J. C. R. Licklider, "Man-computer symbiosis," *IRE Transactions on Human Factors in Electronics*, vol. HFE-1, no. 1, pp. 4–11, 1960.

[9] F. Barravecchia *et al.*, "Redefining human–robot symbiosis: a bio-inspired approach to collaborative assembly," *The International Journal of Advanced Manufacturing Technology*, vol. 128, no. 5, pp. 2043–2058, September 2023. [Online]. Available: https://doi.org/10.1007/s00170-023-11920-1

[10] A. Dahiya *et al.*, "A survey of multi-agent human–robot interaction systems," *Robotics and Autonomous Systems*, vol. 161, p. 104335, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092188902200224X

[11] R. R. Murphy, "Human-robot interaction in rescue robotics," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 138–153, 2004.

[12] J. J. Marquez, V. Riley, and P. C. Schutte, "Chapter 10 - human automation interaction," in *Space Safety and Human Performance*, T. Sgobba *et al.*, Eds. Butterworth-Heinemann, 2018, pp. 429 – 467. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780081018699000108

[13] A. Valero, P. De La Puente, and D. Rodriguez-Losada, "Exploratory analysis of operator:robot ratio in search and rescue missions," in *IFAC Proceedings Volumes (IFAC-PapersOnline)*, Timisoara, Romania, 2010, pp. 101 – 108.

[14] J. Dul *et al.*, "A strategy for human factors/ergonomics: developing the discipline and profession," *Ergonomics*, vol. 55, no. 4, pp. 377–395, 2012, pMID: 22332611. [Online]. Available: https://doi.org/10.1080/00140139.2012.661087

[15] K. van de Merwe, S. Mallam, and S. Nazir, "Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review," *Human Factors*, vol. 66, no. 1, pp. 180–208, 2024, pMID: 35274577. [Online]. Available: https://doi.org/10.1177/00187208221077804

[16] J. Heard, C. E. Harriott, and J. A. Adams, "A Survey of Workload Assessment Algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, 2018.

[17] J. McLurkin *et al.*, "Speaking Swarmish: Human-Robot Interface Design for Large Swarms of Autonomous Mobile Robots," *AAAI Spring Symposium*, pp. 72–75, 2006.

[18] D. St-Onge *et al.*, "Engaging with robotic swarms: Commands from expressive motion," *Transaction on Human-Robot Interactions*, 2019.

[19] R. Sakagami *et al.*, "Rosmc: A high-level mission operation framework for heterogeneous robotic teams," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5473–5479.

[20] T. H. Chung, V. Orekhov, and A. Maio, "Into the robotic depths: Analysis and insights from the darpa subterranean challenge," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, no. Volume 6, 2023, pp. 477–502, 2023. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-control-062722-100728

[21] H. Hofer, "Real-time visualization pipeline for dynamic point cloud data," Master's thesis, Institut für Softwaretechnik und interaktive Systeme, 2018. [Online]. Available: https://publik.tuwien.ac.at/files/publik_272811.pdf

[22] T. B. Sheridan, *Telerobotics, Automation and Human Supervisory Control*. The MIT Press, 1992.

[23] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[24] R. W. Proud, J. J. Hart, and R. B. Mrozinski, "Methods for determining the level of autonomy to design into a human spaceflight vehicle: a function specific approach," NASA Johnson Space Center, Tech. Rep., 2003.

[25] A. Agha *et al.*, "Nebula: Team costar's robotic autonomy solution that won phase ii of darpa subterranean challenge," *Field Robotics*, vol. 2, pp. 1432–1506, 2022.

[26] W. Jo *et al.*, "Toward a wearable biosensor ecosystem on ros 2 for real-time human-robot interaction systems," *arXiv*, 10 2021. [Online]. Available: http://arxiv.org/abs/2110.03840

[27] T. B. Sheridan, "Human–robot interaction: Status and challenges," *Human Factors*, vol. 58, no. 4, pp. 525–532, 2016, pMID: 27098262. [Online]. Available: https://doi.org/10.1177/0018720816644364

[28] M. Lichtenstern *et al.*, "A prototyping environment for interaction between a human and a robotic multi-agent system," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '12. New York, NY, USA: ACM, 2012, pp. 185–186. [Online]. Available: http://doi.acm.org/10.1145/2157689.2157747

[29] M. J. Schuster *et al.*, "The arches space-analogue demonstration mission: Towards heterogeneous teams of autonomous robots for collaborative scientific sampling in planetary exploration," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5315–5322, 2020.

[30] M. Kaufmann, J. Panerati, and G. Beltrame, "Towards a Symbiotic Human and Multi-Robot Planetary Exploration System: Resilient Topologies for Space Exploration," in *Robotics Science & Systems 2018: Autonomous Space Robotics Workshop*, 2018.

[31] L. Burkhard *et al.*, "Collaborative multi-rover crater exploration: Concept and results from the arches analog mission," in *2024 IEEE Aerospace Conference*, 2024, pp. 1–14.

[32] B. P. Vagvolgyi *et al.*, "Scene Modeling and Augmented Virtuality Interface for Telerobotic Satellite Servicing," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, p. 4241, 2018.

[33] R. K. Panda and S. Saxena, "An insight to multi-tasking in cognitive robotics," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, March 2016, pp. 1018–1023.

[34] A. Rosenfeld *et al.*, "Intelligent agent supporting human–multi-robot team collaboration," *Artificial Intelligence*, vol. 252, pp. 211 – 231, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370217301029

[35] J. Nagi *et al.*, "Wisdom of the swarm for cooperative decision-making in human-swarm interaction," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015-June, no. June, pp. 1802–1808, 2015.

[36] J. Gale, J. Karasinski, and S. Hillenius, "Playbook for uas: Ux of goal-oriented planning and execution," in *Engineering Psychology and Cognitive Ergonomics*, D. Harris, Ed. Cham: Springer International Publishing, 2018, pp. 545–557.

[37] F. D. Pace *et al.*, "An augmented interface to display industrial robot faults," in *AVR*, 2018.

[38] C. Ouali *et al.*, "Voice controlled multi-robot system for collaborative task achievement," in *Advances in Intelligent Systems and Computing*, vol. 751, 2019, pp. 345–360.

[39] H. R. Kam *et al.*, "Rviz: a toolkit for real domain data visualization," *Telecommunication Systems*, vol. 60, no. 2, pp. 337–345, 2015. [Online]. Available: https://doi.org/10.1007/s11235-015-0034-5

[40] B. Ikeda and D. Szafir, "Advancing the design of visual debugging tools for roboticists," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022, pp. 195–204.

[41] A. Kolling *et al.*, "Human Interaction with Robot Swarms: A Survey," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 9–26, 2016.

[42] G. Coppin and F. Legras, "Autonomy spectrum and performance perception issues in swarm supervisory control," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 590–603, March 2012.

[43] S. Music and S. Hirche, "Control sharing in human-robot team interaction," in *Annual Reviews in Control*, vol. 44, 2017, pp. 342–354.

[44] D. St-Onge *et al.*, "Planetary exploration with robot teams: Implementing higher autonomy with swarm intelligence," *IEEE Robotics & Automation Magazine*, vol. 27, no. 2, pp. 159–168, 2019.

[45] D. A. Norman and S. W. Draper, *User Centered System Design; New Perspectives on Human-Computer Interaction*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1986.

[46] J. Y. C. Chen and M. J. Barnes, "Human–agent teaming for multirobot control: A review of human factors issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, Feb 2014.

[47] A. Rajavenkatanarayanan *et al.*, "Monitoring task engagement using facial expressions and body postures," in *Proceedings of the 3rd International Workshop on Interactive and Spatial Computing - IWISC '18*, 2018, pp. 103–108. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3191801.3191816

[48] G. Podevijn *et al.*, "Investigating the effect of increasing robot group sizes on the human psychophysiological state in the context of human-swarm interaction," *Swarm Intelligence*, vol. 10, no. 3, pp. 1–18, 2016.

[49] G. Podevijn *et al.*, *Human Responses to Stimuli Produced by Robot Swarms - the Effect of the Reality-Gap on Psychological State*. Cham: Springer International Publishing, 2018, pp. 531–543. [Online]. Available: https://doi.org/10.1007/978-3-319-73008-0_37

[50] S. Mathot, "Pupillometry: Psychology, physiology, and function." *Journal of Cognition*, vol. 1, no. 1, 2018.

[51] V. Peysakhovich, F. Dehais, and M. Causse, "Pupil Diameter as a Measure of Cognitive Load during Auditory-visual Interference in a Simple Piloting Task," *Procedia Manufacturing*, vol. 3, pp. 5199–5205, 2015.

[52] A. H. Memar and E. T. Esfahani, "Physiological Measures for Human Performance Analysis in Human-Robot Teamwork: Case of Tele-Exploration," *IEEE Access*, vol. 6, pp. 3694–3705, 2018.

[53] M. Micire *et al.*, "Smart spheres: a telerobotic free-flyer for intravehicular activities in space," in *AIAA Space 2013 Conference and Exposition*, 2013, p. 5338.

[54] M. Bualat *et al.*, "Astrobee: Developing a free-flying robot for the international space station," in *AIAA SPACE 2015 conference and exposition*, 2015, p. 4643.

[55] R. Gomes *et al.*, "Human spaceflight robotic medical first responder," in *Proceedings of the 68th International Astronautical Congress (IAC)*, International Astronautical Federation (IAF). International Astronautical Congress, 2017.

[56] S. Kang *et al.*, "Astrobee iss free-flyer datasets for space intra-vehicular robot navigation research," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3307–3314, 2024.

[57] S. Aukstakalnis, *Practical Augmented Reality: A Guide to the Technologies, Applications, and Human Factors for AR and VR*, 2016. [Online]. Available: http://evi.sagepub.com/cgi/doi/10.1177/1356389011400889

[58] B. Sheeran *et al.*, *Robot Guided Emergency Evacuation from a Simulated Space Station*. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2023-0156

[59] C. McGhan and E. Atkins, "A Virtual Rover Interface for Collaborative Human-Robot Exploration Teams," no. May, pp. 1–15, 2007.

[60] R. Mueller *et al.*, "Collaboration in a hybrid team of humans and robot for improving working conditions in an aircraft riveting process," *SAE International Journal of Advances and Current Practices in Mobility*, vol. 1, no. 2019-01-1372, pp. 396–403, 2019.

[61] S. Kohn *et al.*, "Towards a real-time environment reconstruction for vr-based teleoperation through model segmentation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 1–9.

[62] M. Walker *et al.*, "Communicating robot motion intent with augmented reality," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: ACM, 2018, pp. 316–324. [Online]. Available: http://doi.acm.org/10.1145/3171221.3171253

[63] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[64] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*, 2001. [Online]. Available: http://www.amazon.com/Probabilistic-Robotics-Intelligent-Autonomous-Agents/dp/0262201623/ref=sr_11_1/105-3361811-4085215?ie=UTF8&qid=1190743235&sr=11-1

[65] J. O. Burns *et al.*, "Science on the lunar surface facilitated by low latency telerobotics from a lunar orbital platform - gateway," *Acta Astronautica*, vol. 154, pp. 195–203, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0094576517317307

[66] H. Ji *et al.*, "On-site human-robot collaboration for lunar exploration based on shared mixed reality," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 18 235–18 260, 2024. [Online]. Available: https://doi.org/10.1007/s11042-023-16178-z

[67] M. Walker *et al.*, "Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy," *J. Hum.-Robot Interact.*, vol. 12, no. 4, 2023. [Online]. Available: https://doi.org/10.1145/3597623

[68] K. Merckaert *et al.*, "Real-time constraint-based planning and control of robotic manipulators for safe human–robot collaboration," *Robotics and Computer-Integrated Manufacturing*, vol. 87, p. 102711, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584523001862

[69] E. Rosen *et al.*, "Communicating robot arm motion intent through mixed reality head-mounted displays," in *Robotics Research*, N. M. Amato *et al.*, Eds. Cham: Springer International Publishing, 2020, pp. 301–316.

[70] A. V. Taylor *et al.*, "Diminished reality for close quarters robotic telemanipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 11 531–11 538.

[71] T. Kot and P. Novák, "Utilization of the oculus rift hmd in mobile robot teleoperation," *Applied Mechanics and Materials*, vol. 555, pp. 199 – 208, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:109842947

[72] C. Leger *et al.*, "Mars exploration rover surface operations: driving spirit at gusev crater," *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 1815–1822 Vol. 2, 2005.

[73] S. Maxwell *et al.*, "The best of both worlds: Integrating textual and visual command interfaces for mars rover operations," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 2. IEEE, 2005, pp. 1384–1388.

[74] B. Pendleton and M. Goodrich, "Scalable Human Interaction with Robotic Swarms," *AIAA Infotech@Aerospace (I@A) Conference*, pp. 1–13, 2013. [Online]. Available: http://arc.aiaa.org/doi/abs/10.2514/6.2013-4731

[75] L. Chen *et al.*, "Sensor-based activity recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, Nov 2012.

[76] C. D. Wickens, J. G. Hollands, and S. Banbury, *Engineering Psychology and Human Performance*, 4th ed. London: Taylor and Francis, 2012. [Online]. Available: https://polymtl.on.worldcat.org/oclc/929512672

[77] C. D. Wickens *et al.*, *Introduction to Human Factors Engineering (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2003.

[78] M. S. Young and N. A. Stanton, "Malleable attentional resources theory: A new explanation for the effects of mental underload on performance," *Human Factors*, vol. 44, no. 3, pp. 365–375, 2002, pMID: 12502155. [Online]. Available: https://doi.org/10.1518/0018720024497709

[79] M. R. Endsley and D. G. Jones, *Designing for Situation Awareness: An Approach to User-Centered Design*, 2nd ed. Boca Raton, FL: CRC Press, 2012. [Online]. Available: http://www.crcnetbase.com/isbn/9781420063585

[80] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 32–64, 1995. [Online]. Available: https://doi.org/10.1518/001872095779049543

[81] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Human mental workload*, vol. 1, no. 3, pp. 139–183, 1988.

[82] G. Teo *et al.*, "Augmenting robot behaviors using physiological measures of workload state," in *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*, D. D. Schmorrow and C. M. Fidopiastis, Eds. Cham: Springer International Publishing, 2016, pp. 404–415.

[83] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[84] C. M. Bishop, *Pattern recognition and machine learning, 5th Edition*, ser. Information science and statistics. Springer, 2007. [Online]. Available: http://www.worldcat.org/oclc/71008143

[85] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 1569–1576. [Online]. Available: http://papers.nips.cc/paper/2547-two-dimensional-linear-discriminant-analysis.pdf

[86] G. F. Wilson and C. A. Russell, "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," *Human Factors*, vol. 45, no. 4, pp. 635–644, 2003, pMID: 15055460. [Online]. Available: https://doi.org/10.1518/hfes.45.4.635.27088

[87] J. C. Christensen *et al.*, "The effects of day-to-day variability of physiological data on operator functional state classification," *NeuroImage*, vol. 59, no. 1, pp. 57–63, 2012. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2011.07.091

[88] H. A. Abbass *et al.*, "Augmented cognition using real-time eeg-based adaptive strategies for air traffic control," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1, pp. 230–234, 2014. [Online]. Available: https://doi.org/10.1177/1541931214581048

[89] J. Crandall, C. Nielsen, and M. Goodrich, "Towards predicting robot team performance," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, vol. 1, 2003, pp. 906–911 vol.1.

[90] M. Fujino *et al.*, "Comparison of sagat and spam for seeking effective way to evaluate situation awareness and workload during air traffic control task," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 64, no. 1, pp. 1836–1840, 2020. [Online]. Available: https://doi.org/10.1177/1071181320641442

[91] R. S. Pierce, "The effect of spam administration during a dynamic simulation," *Human Factors*, vol. 54, no. 5, pp. 838–848, 2012, pMID: 23156627. [Online]. Available: https://doi.org/10.1177/0018720812439206

[92] B. Mekdeci and M. Cummings, "Modeling multiple human operators in the supervisory control of heterogeneous unmanned vehicles," in *Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems*. ACM, 2009, pp. 1–8.

[93] J. D. L. Croix and M. Egerstedt, "Controllability Characterizations of Leader-Based Swarm Interactions," in *2012 AAAI Fall Symposium Series*, 2012. [Online]. Available: http://www.aaai.org/ocs/index.php/FSS/FSS12/paper/download/5543/5832

[94] E. Peters *et al.*, "Design for collaboration in mixed reality: Technical challenges and solutions," in *2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, Sep. 2016, pp. 1–7.

[95] NASA. (2020) NASA Technology Taxonomy. Last visited 2020-10-11. [Online]. Available: https://www.nasa.gov/offices/oct/taxonomy/index.html

[96] NASA. (2020) NASA's Plan for Sustained Lunar Exploration and Development. Last visited 2020-10-11. [Online]. Available: https://www.nasa.gov/sites/default/files/atoms/files/a_sustained_lunar_presence_nspc_report4220final.pdf

[97] NASA/JPL. (2020) DARPA Subterranean Challenge Team CoSTAR. Last visited 2020-10-11. [Online]. Available: https://costar.jpl.nasa.gov

[98] DARPA. (2020) Subterranean Challenge. Last visited 2020-10-11. [Online]. Available: https://www.subtchallenge.com

[99] D. St-Onge *et al.*, "Planetary exploration with robot teams: Implementing higher autonomy with swarm intelligence," *IEEE Robotics Automation Magazine*, vol. 27, no. 2, pp. 159–168, 2020.

[100] B. Trouvain and H. L. Wolf, "Evaluation of multi-robot control and monitoring performance," in *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*, 2002, pp. 111–116.

[101] J. Delmerico *et al.*, "The current state and future outlook of rescue robotics," *Journal of Field Robotics*, vol. 36, no. 7, pp. 1171–1191, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21887

[102] W. Chi *et al.*, "Embedding a scheduler in execution for a planetary rover," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 28, 2018, pp. 312–320.

[103] K. Otsu *et al.*, "Supervised autonomy for communication-degraded subterranean exploration by a robot team," in *2020 IEEE Aerospace Conference*, 2020, pp. 1–9.

[104] T. S. Vaquero *et al.*, "Traversability-aware signal coverage planning for communication node deployment in planetary cave exploration," in *The International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)*, 2020, *in press.*

[105] S. Mitchell, M. OSullivan, and I. Dunning, "Pulp: a linear programming toolkit for python," *The University of Auckland, Auckland, New Zealand*, 2011.

[106] M. Asada *et al.*, "RoboCup: A treasure trove of rich diversity for research issues and interdisciplinary connections [TC Spotlight]," *IEEE Robotics Automation Magazine*, vol. 26, no. 3, pp. 99–102, 2019.

[107] K. A. Hambuchen *et al.*, *NASA's Space Robotics Challenge: Advancing Robotics for Future Exploration Missions.* American Institute of Aeronautics and Astronautics, 2017. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2017-5120

[108] M. Link and B. Lamboray, *European Space Resources Innovation Centre – ESRIC.* American Institute of Aeronautics and Astronautics, 2021. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2021-4012

[109] A. Agha *et al.*, "Nebula: Quest for robotic autonomy in challenging environments; TEAM costar at the DARPA subterranean challenge," *Submitted to the Journal of Field Robotics*, 2021. [Online]. Available: https://arxiv.org/abs/2103.11470

[110] T. N. Titus *et al.*, "A roadmap for planetary caves science and exploration," *Nature Astronomy*, vol. 5, no. 6, pp. 524—525, 2021. [Online]. Available: https://doi.org/10.1038/s41550-021-01385-1

[111] J. Y. Chen and M. J. Barnes, "Human–robot interaction," *Handbook of human factors and ergonomics*, pp. 1121–1142, 2021.

[112] M. Kaufmann *et al.*, "Copilot mike: An autonomous assistant for multi-robot operations in cave exploration," in *IEEE Aerospace Conference*, 2021.

[113] S. Kohlbrecher *et al.*, "Human-robot teaming for rescue missions: Team vigir's approach to the 2013 darpa robotics challenge trials," *Journal of Field Robotics*, vol. 32, no. 3, pp. 352–377, 2015.

[114] M. Tranzatto *et al.*, "Cerberus: Autonomous legged and aerial robotic exploration in the tunnel and urban circuits of the darpa subterranean challenge," *Field Robotics*, 2022.

[115] D. Szafir and D. A. Szafir, "Connecting human-robot interaction and data visualization," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: Association for Computing Machinery, 2021, p. 281–292. [Online]. Available: https://doi.org/10.1145/3434073.3444683

[116] A. Rahmani *et al.*, "Space vehicle swarm exploration missions: A study of key enabling technologies and gaps," *Proceedings of the 70th International Astronautical Congress*, 2019.

[117] M. Johnson *et al.*, "Coactive design: Designing support for interdependence in joint activity," *J. Hum.-Robot Interact.*, vol. 3, no. 1, p. 43–69, feb 2014. [Online]. Available: https://doi.org/10.5898/JHRI.3.1.Johnson

[118] K. A. Roundtree *et al.*, "Visualization design for human-collective teams," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 417–421, 2019. [Online]. Available: https://doi.org/10.1177/1071181319631028

[119] G. Wallner, N. Halabi, and P. Mirza-Babaei, "Aggregated visualization of playtesting data," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3290605.3300593

[120] M. C. Medlock *et al.*, "Using the RITE method to improve products: A definition and a case study," *Usability Professionals Association*, vol. 51, 2002.

[121] N. Hudson *et al.*, "Heterogeneous ground and air platforms, homogeneous sensing: Team CSIRO data61's approach to the DARPA subterranean challenge," *CoRR*, 2021. [Online]. Available: https://arxiv.org/abs/2104.09053

[122] M. T. Ohradzansky *et al.*, "Multi-agent autonomy: Advancements and challenges in subterranean exploration," *CoRR*, 2021. [Online]. Available: https://arxiv.org/abs/2110.04390

[123] S. Scherer *et al.*, "Resilient and modular subterranean exploration with a team of roving and flying robots," *Submitted to the Journal of Field Robotics*, 2021.

[124] T. Roucek *et al.*, "System for multi-robotic exploration of underground environments CTU-CRAS-NORLAB in the DARPA subterranean challenge," *CoRR*, 2021. [Online]. Available: https://arxiv.org/abs/2110.05911

[125] K. Otsu *et al.*, "Supervised autonomy for communication-degraded subterranean exploration by a robot team," in *IEEE Aerospace Conference*, 2020.

[126] S. S. Bae *et al.*, "A visual analytics approach to debugging cooperative, autonomous multi-robot systems' worldviews," in *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2020.

[127] J. Benton, A. Coles, and A. Coles, "Temporal planning with preferences and time-dependent continuous costs," in *Proceedings of the Twenty-Second International Conference on International Conference on Automated Planning and Scheduling*. AAAI Press, 2012.

[128] A. Khan *et al.*, "A competitive combat strategy and tactics in rts games ai and starcraft," in *Advances in Multimedia Information Processing – PCM 2017*, B. Zeng *et al.*, Eds. Cham: Springer International Publishing, 2018.

[129] S. Kim *et al.*, "PLGRIM: hierarchical value learning for large-scale exploration in unknown environments," *CoRR*, 2021. [Online]. Available: https://arxiv.org/abs/2102.05633

[130] E. Terry *et al.*, "Object and gas source detection with robotic platforms in perceptually-degraded environments," in *RSS Workshop: Robots in the Wild: Challenges in Deploying Robust Autonomy for Robotic Exploration*, 2020.

[131] J. Huang, R. White, and G. Buscher, "User see, user point: Gaze and cursor alignment in web search," in *Proceedings of the 2012 Conference on Human Factors in Computing Systems (CHI)*, 2012. [Online]. Available: https://www.microsoft.com/en-us/research/publication/user-see-user-point-gaze-and-cursor-alignment-in-web-search/

[132] D. Miranda, "2020 nasa technology taxonomy," NASA, Technical Report, 2020.

[133] T. Fong, "Nasa autonomous systems & robotics: Roadmap and investments," in *Lunar surface innovation consortium fall 2021 meeting*, 2021.

[134] National Academies of Sciences, Engineering, and Medicine, *Origins, worlds, and life: a decadal strategy for planetary science and astrobiology 2023-2032*. Washington, DC: The National Academies Press,

2023. [Online]. Available: https://nap.nationalacademies.org/catalog/26522/origins-worlds-and-life-a-decadal-strategy-for-planetary-science

[135] G.-Z. Yang *et al.*, "The grand challenges of science robotics," *Science Robotics*, vol. 3, no. 14, p. eaar7650, 2018.

[136] D. J. Rea and S. H. Seo, "Still not solved: A call for renewed focus on user-centered teleoperation interfaces," *Frontiers in Robotics and AI*, vol. 9, p. 704225, 2022.

[137] M. Kaufmann, K. Sheridan, and G. Beltrame, "Towards human-in-the-loop autonomous multi-robot operations," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, ser. ICMI '21 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 341–343. [Online]. Available: https://doi.org/10.1145/3461615.3486573

[138] R. Hetrick *et al.*, "Comparing virtual reality interfaces for the teleoperation of robots," in *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 2020, pp. 1–7.

[139] J. G. Blank *et al.*, "Testing operational designs for a future robotic mission to a martian lava tube," in *Proceedings of the 2023 74th International Astronautical Congress (IAC 2023)*, Baku, Azerbaijan, 2023.

[140] M. Kaufmann *et al.*, "Human-autonomy teaming for supervised scientific exploration: A pilot study in the wild," Presented at the Human-Multi-Robot-Systems Workshop, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022. [Online]. Available: https://sites.google.com/view/hmrs-iros2022/

[141] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. North-Holland, 1988, vol. 52, pp. 139–183. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166411508623869

[142] N. A. R. Center, "Nasa task load index (tlx) ios application," Available on the Apple App Store, 2023. [Online]. Available: https://humansystems.arc.nasa.gov/groups/tlx/

[143] M. Selvaggio *et al.*, "Autonomy in physical human-robot interaction: A brief survey," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7989–7996, 2021.

[144] F. Ropero *et al.*, "A virtual reality mission planner for mars rovers," in *Proceedings - 6th IEEE International Conference on Space Mission Challenges for Information Technology, SMC-IT 2017*, vol. 2017-December. Institute of Electrical and Electronics Engineers Inc., 2017, pp. 142–146.

[145] D. G. Riley and E. W. Frew, "Fielded human-robot interaction for a heterogeneous team in the darpa subterranean challenge," *J. Hum.-Robot Interact.*, vol. 12, no. 3, 2023. [Online]. Available: https://doi.org/10.1145/3588325

[146] A. Hornung *et al.*, "Octomap: an efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013. [Online]. Available: https://doi.org/10.1007/s10514-012-9321-0

[147] C. Reardon *et al.*, "Augmented reality visualization of autonomous mobile robot change detection in uninstrumented environments," *J. Hum.-Robot Interact.*, 2023, just Accepted. [Online]. Available: https://doi.org/10.1145/3611654

[148] P. Wang *et al.*, "A novel human-robot interaction system based on 3d mapping and virtual reality," in *2017 Chinese Automation Congress (CAC)*, 2017, pp. 5888–5894.

[149] C. J. S. Patterson *et al.*, "Fusing lidar and scientific data to create a multipurpose virtual realty tool for planetary cave mission operations," in *72nd International Astronautical Congress (IAC)*, Dubai, United Arab Emirates, 2021, iAC-21.D4.1.13. Copyright ©2021 by the International Astronautical Federation (IAF). All rights reserved.

[150] B. SAYKRS, "Analysis of heart rate variability," *Ergonomics*, vol. 16, no. 1, pp. 17–32, 1973, pMID: 4702060. [Online]. Available: https://doi.org/10.1080/00140137308924479

[151] M. Malik, "Heart rate variability," *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 1996. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1542-474X.1996.tb00275.x

[152] N. Urrestilla and D. St-Onge, "Measuring cognitive load: Heart-rate variability and pupillometry assessment," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 405–410. [Online]. Available: https://doi.org/10.1145/3395035.3425203

[153] T. Kosch *et al.*, "A survey on measuring cognitive workload in human-computer interaction," *ACM Comput. Surv.*, vol. 55, 7 2023. [Online]. Available: https://doi.org/10.1145/3582272

[154] M. Chiou, N. Hawes, and R. Stolkin, "Mixed-initiative variable autonomy for remotely operated mobile robots," *J. Hum.-Robot Interact.*, vol. 10, no. 4, 2021. [Online]. Available: https://doi.org/10.1145/3472206

[155] M. Quigley *et al.*, "roslibjs: The standard ros javascript library," 2024, gitHub repository, last accessed: July 3, 2024. [Online]. Available: https://github.com/RobotWebTools/roslibjs

[156] M. Bischoff, V. Röhrl, and the ROS# Development Team, "Ros# - a set of software libraries and tools in c# for communicating with ros from .net applications," 2024, gitHub repository, last accessed: July 3, 2024. [Online]. Available: https://github.com/siemens/ros-sharp

[157] T. S. Vaquero *et al.*, "Traversability-aware signal coverage planning for communication node deployment in planetary cave exploration," in *International Symposium on Artificial Intelligence, Robotics and Automation in Space (I-SAIRAS)*, Jet Propulsion Laboratory, National Aeronautics and Space Administration (NASA). Virtual Conference: Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration (NASA), 2020.

[158] M. Saboia *et al.*, "Achord: Communication-aware multi-robot coordination with intermittent connectivity," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 184–10 191, 2022.

[159] M. Kaufmann *et al.*, "Copiloting autonomous multi-robot missions: A game-inspired supervisory control interface," 2022. [Online]. Available: https://arxiv.org/abs/2204.06647

[160] S.-K. Kim *et al.*, "Plgrim: Hierarchical value learning for large-scale autonomous exploration in unknown environments," 2021. [Online]. Available: https://arxiv.org/abs/2102.05633

[161] M. Schaffarczyk *et al.*, "Validity of the polar h10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women," *Sensors (Basel)*, vol. 22, no. 17, p. 6536, 2022.

[162] A. Alaimo *et al.*, "Aircraft pilots workload analysis: Heart rate variability objective measures and nasa-task load index subjective evaluation," *Aerospace*, vol. 7, no. 9, 2020. [Online]. Available: https://www.mdpi.com/2226-4310/7/9/137

[163] R. Castaldo *et al.*, "Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis," pp. 370–377, 2015.

[164] V. Villani *et al.*, "A framework for affect-based natural human-robot interaction," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018, pp. 1038–1044.

[165] Texas Instruments, "Generic attribute profile (gatt)," 2024, accessed: 2024-07-03. [Online]. Available: https://software-dl.ti.com/lprf/sdg-latest/html/ble-stack-3.x/gatt.html

[166] M. Körber, "Theoretical considerations and development of a questionnaire to measure trust in automation," in *Advances in Intelligent Systems and Computing*, vol. 823. Springer Verlag, 2019, pp. 13–30.

[167] M. J. B. Mena *et al.*, "Non-normal data in repeated measures anova: impact on type i error and power," *Psicothema*, 2023.

[168] D. Whitney *et al.*, "Comparing robot grasping teleoperation across desktop and virtual reality with ros reality," in *Robotics Research*, N. M. Amato *et al.*, Eds. Cham: Springer International Publishing, 2020, pp. 335–350.

[169] R. A. Grier, "How high is high? a meta-analysis of nasa-tlx global workload scores," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1, pp. 1727–1731, 2015. [Online]. Available: https://doi.org/10.1177/1541931215591373

[170] J. J. Roldán *et al.*, *Multi-robot Systems, Virtual Reality and ROS: Developing a New Generation of Operator Interfaces*. Cham: Springer International Publishing, 2019, pp. 29–64. [Online]. Available: https://doi.org/10.1007/978-3-319-91590-6_2

[171] A. Duval *et al.*, "The eyes and hearts of uav pilots: observations of physiological responses in real-life scenarios," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 2352–2358.

[172] J. Heard, P. Baskaran, and J. A. Adams, "Predicting task performance for intelligent human-machine interactions," *Frontiers in Neurorobotics*, vol. 16, 2022. [Online]. Available: https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2022.973967

[173] D. Bonatto *et al.*, "Explorations for real-time point cloud rendering of natural scenes in virtual reality," in *2016 International Conference on 3D Imaging (IC3D)*, 2016, pp. 1–7.

[174] G. Wu *et al.*, "4d gaussian splatting for real-time dynamic scene rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 310–20 320.

[175] H. El Alami, M. Nwosu, and D. B. Rawat, "Joint human and autonomy teaming for defense: status, challenges, and perspectives," *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V*, vol. 12538, pp. 144–158, 2023.

APPENDIX A IROS 2022 HMRS BEST POSTER

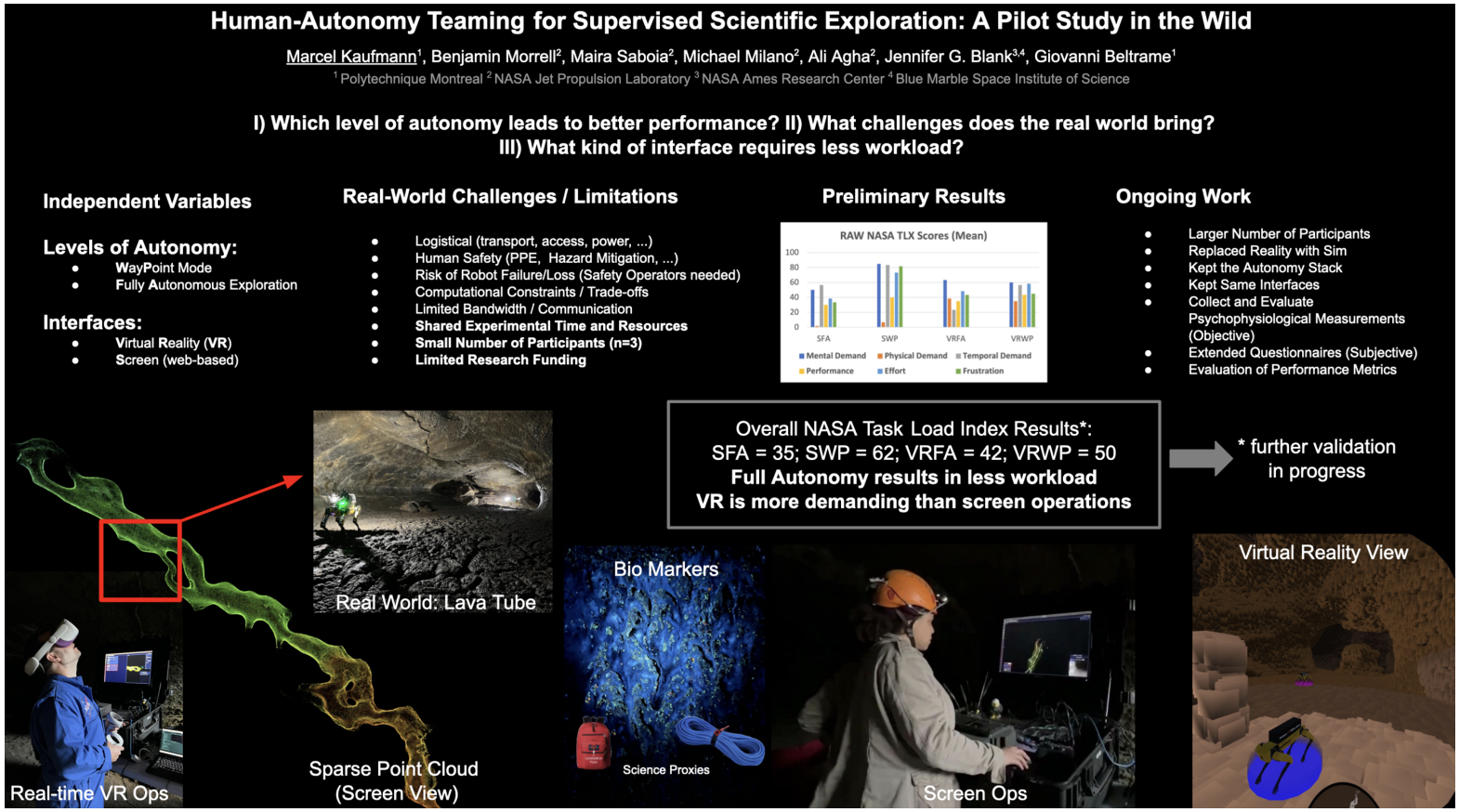

Figure A.1 IROS 2022 HMRS Workshop Best Poster Award

APPENDIX B SPAM QUESTIONS

Table B.1 Questions and Instructions Used to Assess Situational Awareness with SPAM

#	**Instruction/Question**
1	Operator, press the corresponding thumb stick or robot number for the robot that is furthest inside the cave.
2	Operator, press the corresponding thumb stick or robot number for the robot that is closest to the starting point.
3	Operator press 1 or the left thumb stick if you found less than 5 objects, press 4 or right thumbs stick otherwise.
4	Operator press 1 or the left thumb stick if you found more than 5 objects, press 4 or right thumbs stick otherwise.
5	Operator, press the corresponding thumb sticks or robot number for the robots that are currently not moving. Press both options, if needed.
6	Operator, press the corresponding thumb sticks or robot number for the robots that are currently moving. Press both options, if needed.
7	Operator, is robot number 1 (blue) currently stuck? Press 1 or right thumb stick if yes, 4 or left otherwise.
8	Operator, is robot number 4 (purple) currently stuck? Press 1 or right thumb stick if yes, 4 or left otherwise.
9	Operator, is robot number 1 (blue) currently moving? Press 1 or right thumb stick if yes, 4 or left otherwise.
10	Operator, is robot number 4 (purple) currently moving? Press 1 or right thumb stick if yes, 4 or left otherwise.

APPENDIX C REPEATED MEASURES ANOVA RESULTS

Table C.1 Questionnaire results obtained via repeated measures ANOVA for n=38 included participants. SS - the sum of squares, indicating the variance explained by each factor. F is the F-statistic, a ratio of variance between groups to variance within groups. p-unc is the uncorrected p-value, p-GG-corr is the p-value corrected using Greenhouse-Geisser, which adjusts for violations of sphericity. Epsilon (eps) is a measure of sphericity (1.0 indicates no violation of the sphericity assumption).

Source	SS	F	p-unc	p-GG-corr	eps
The robots' level of autonomy made it easy to explore with multiple robots.					
Interface	1273.68	4.44	0.042	**0.042**	1.0
Autonomy	6063.16	7.41	0.00983	**0.00983**	1.0
Interface * Autonomy	190.13	0.67	0.419	0.419	1.0
I felt like I was exploring the cave together with the robots.					
Interface	16737.01	19.25	**<0.001**	**<0.001**	1.0
Autonomy	1813.32	6.09	0.0183	**0.0183**	1.0
Interface * Autonomy	304.11	0.78	0.384	0.384	1.0
The visualization helped me understand the terrain.					
Interface	2569.90	4.64	0.0378	**0.0378**	1.0
Autonomy	59.38	0.18	0.674	0.674	1.0
Interface * Autonomy	738.32	2.66	0.111	0.111	1.0
I knew what the robots were doing approx. x% of the time.					
Interface	65.79	0.14	0.713	0.713	1.0
Autonomy	5813.16	18.25	**<0.001**	**<0.001**	1.0
Interface * Autonomy	515.79	2.26	0.141	0.141	1.0
I can rely on the system.					
Interface	138.32	0.70	0.407	0.407	1.0
Autonomy	304.11	1.09	0.302	0.302	1.0
Interface * Autonomy	72.53	0.31	0.579	0.579	1.0

Table C.2 Repeated Measure ANOVA for NASA TLX results. SS denotes the sum of squares, reflecting the variance explained by each source. F is the F-statistic, measuring the ratio of variance between groups to variance within groups. p-unc represents the uncorrected p-value, indicating the initial significance level, while p-GG-corr is the p-value corrected for violations of sphericity. eps (epsilon) measures sphericity; an eps value of 1.0 indicates no violation of the assumption. N=38 included participants.

Source	**SS**	**F**	**p-unc**	**p-GG-corr**	**eps**
Mental Demand					
Interface	7673.68	18.77	**<0.001**	**<0.001**	1.0
Autonomy	4423.68	7.97	**0.008**	**0.008**	1.0
Interface * Autonomy	716.45	2.93	0.096	0.096	1.0
Physical Demand					
Interface	8400.66	10.78	**0.002**	**0.002**	1.0
Autonomy	3042.11	10.39	**0.003**	**0.003**	1.0
Interface * Autonomy	0.66	0.0036	0.953	0.953	1.0
Temporal Demand					
Interface	1954.11	5.50	**0.024**	**0.024**	1.0
Autonomy	2569.90	4.37	**0.044**	**0.044**	1.0
Interface * Autonomy	102.80	0.32	0.573	0.573	1.0
Performance					
Interface	4697.53	13.81	**<0.001**	**<0.001**	1.0
Autonomy	13.32	0.03	0.855	0.855	1.0
Interface * Autonomy	394.90	1.66	0.205	0.205	1.0
Effort					
Interface	6912.01	22.57	**<0.001**	**<0.001**	1.0
Autonomy	7047.53	11.13	**0.002**	**0.002**	1.0
Interface * Autonomy	4.11	0.02	0.885	0.885	1.0
Frustration					
Interface	1160.53	2.01	0.165	0.165	1.0
Autonomy	290.13	0.48	0.495	0.495	1.0
Interface * Autonomy	852.63	3.49	**0.070**	**0.070**	1.0
Raw TLX Score					
Interface	4651.32	26.68	**<0.001**	**<0.001**	1.0
Autonomy	1542.22	5.22	**0.028**	**0.028**	1.0
Interface * Autonomy	34.58	0.44	0.512	0.512	1.0
Weighted TLX Score					
Interface	4463.33	21.63	**<0.001**	**<0.001**	1.0
Autonomy	1096.95	3.13	0.085	0.085	1.0
Interface * Autonomy	18.02	0.13	0.718	0.718	1.0

Table C.3 Repeated ANOVA for Average SPAM Answer Times. N=38 included participants.

Source	**SS**	**F**	**p-unc**	**p-GG-corr**	**eps**
Interface	180.45	12.42	**0.001**	**0.001**	1.0
Autonomy	3.59	0.37	0.546	0.546	1.0
Interface * Autonomy	14.93	1.61	0.213	0.213	1.0

Table C.4 Repeated ANOVA Results for SPAM Correctly Answered Questions. N=38 included participants.

Source	**SS**	**F**	**p-unc**	**p-GG-corr**	**eps**
Interface	1871.35	6.25	**0.017**	**0.017**	1.0
Autonomy	29.24	0.12	0.730	0.730	1.0
Interface * Autonomy	308.85	1.98	0.168	0.168	1.0

Table C.5 Average Detection Performance repeated measures ANOVA. N=38 included participants.

Source	**SS**	**F**	**p-unc**	**p-GG-corr**	**eps**
Interface	0.785	25.84	**<0.001**	**<0.001**	1.0
Autonomy	0.076	2.29	0.139	0.139	1.0
Interface * Autonomy	0.169	7.03	**0.012**	**0.012**	1.0