Multi-Dimensional Evaluation of an Augmented Reality Head-Mounted Display User Interface for Controlling Legged Manipulators

RODRIGO CHACÓN QUESADA and YIANNIS DEMIRIS, Personal Robotics Laboratory, Imperial College London, London, United Kingdom

Controlling assistive robots can be challenging for some users, especially those lacking relevant experience. Augmented Reality (AR) User Interfaces (UIs) have the potential to facilitate this task. Although extensive research regarding legged manipulators exists, comparatively little is on their UIs. Most existing UIs leverage traditional control interfaces such as joysticks, Hand-Held (HH) controllers and 2D UIs. These interfaces not only risk being unintuitive, thus discouraging interaction with the robot partner, but also draw the operator's focus away from the task and towards the UI. This shift in attention raises additional safety concerns, particularly in potentially hazardous environments where legged manipulators are frequently deployed. Moreover, traditional interfaces limit the operators' availability to use their hands for other tasks. Towards overcoming these limitations, in this article, we provide a user study comparing an AR Head-Mounted Display (HMD) UI we developed for controlling a legged manipulator against off-the-shelf control methods for such robots. This user study involved 27 participants and 135 trials, from which we gathered over 405 completed questionnaires. These trials involved multiple navigation and manipulation tasks with varying difficulty levels using a Boston Dynamics's Spot, a 7 df Kinova robot arm and a Robotiq 2F-85 gripper that we integrated into a legged manipulator. We made the comparison between UIs across multiple dimensions relevant to a successful human–robot interaction. These dimensions include cognitive workload, technology acceptance, fluency, system usability, immersion and trust. Our study employed a factorial experimental design with participants undergoing five different conditions, generating longitudinal data. Due to potential unknown distributions and outliers in such data, using parametric methods for its analysis is questionable, and while non-parametric alternatives exist, they may lead to reduced statistical power. Therefore, to analyse the data that resulted from our experiment, we chose Bayesian data analysis as an effective alternative to address these limitations. Our results show that AR UIs can outpace HH-based control methods and reduce the cognitive requirements when designers include hands-free interactions and cognitive offloading principles into the UI. Furthermore, the use of the AR UI together with our cognitive offloading feature resulted in higher usability scores and significantly higher fluency and Technology Acceptance Model scores. Regarding immersion, our results revealed that the response values for the AR Immersion questionnaire associated with the AR UI are significantly higher than those associated with the HH UI, regardless of the main interaction method with the former, i.e., hand gestures or cognitive offloading. Derived from the participants' qualitative answers, we believe this is due to a combination of factors, of which the most important is the free use of the hands when using the HMD, as well as the ability to see the real environment without the need to divert their attention to the UI. Regarding trust, our findings did not display discernible differences in reported trust scores across

R. Chacón Quesada was supported by the University of Costa Rica. This research was supported in part by UKRI Grant EP/V026682/1 and Y. Demiris was supported by a Royal Academy of Engineering Chair in Emerging Technologies.

Author's Contact Information: Rodrigo Chacón Quesada (Corresponding author), Personal Robotics Laboratory, Imperial College London, London, United Kingdom; e-mail: r.chacon-quesada17@imperial.ac.uk; Yiannis Demiris, Personal Robotics Laboratory, Imperial College London, London, United Kingdom; e-mail: y.demiris@imperial.ac.uk.

ACM 2573-9522/2024/6-ART30
https://doi.org/10.1145/3660649

UI options. However, during the manipulation phase of our user study, where participants were given the choice to select their preferred UI, they consistently reported higher levels of trust compared to the navigation category. Moreover, there was a drastic change in the percentage of participants that selected the AR UI for completing this manipulation stage after incorporating the cognitive offloading feature. Thus, trust seems to have mediated the use and non-use of the UIs in a dimension different from the ones considered in our study, i.e., delegation and reliance. Therefore, our AR HMD UI for the control of legged manipulators was found to improve human–robot interaction across several relevant dimensions, underscoring the critical role of UI design in the effective and trustworthy utilisation of robotic systems.

CCS Concepts: • **Human-centered computing** → **User studies**; • **Computing methodologies** → **Mixed/augmented reality**; • **Computer systems organization** → **Robotics**;

Additional Key Words and Phrases: Human-robot interaction, trust, cognitive workload, technology acceptance, fluency, system usability, immersion, user interfaces, augmented reality, head-mounted display, legged manipulators, Bayesian data analysis, cognitive offloading

ACM Reference format:
Rodrigo Chacón Quesada and Yiannis Demiris. 2024. Multi-Dimensional Evaluation of an Augmented Reality Head-Mounted Display User Interface for Controlling Legged Manipulators. *ACM Trans. Hum.-Robot Interact.* 13, 2, Article 30 (June 2024), 37 pages.
https://doi.org/10.1145/3660649

1 Introduction

Motivated by the risks associated with hazardous inspection tasks, such as those in construction and nuclear sites, legged manipulators are increasingly deployed in industrial and civil scenarios. Often tasked with repetitive actions, these robots are expected to autonomously execute them, yet human operators must initially define movements and inspection actions, delegate tasks and rely on the robot for future completion. Trust becomes pivotal for operators that rely on legged manipulators for such tasks. Traditional **Hand-Held (HH) User Interfaces (UIs)**, like joysticks and **two-dimensional (2D)** screen-based UIs [6, 44], typically control these robots. However, recent advances in **Augmented Reality (AR) Head-Mounted Display (HMD)** UIs for legged manipulators, as explored in our previous work [12, 13], are expected to revolutionise established UI paradigms. This article builds on our earlier investigations to assess how AR HMD UIs impact relevant **human–robot interaction (HRI)** metrics for successful collaboration. Furthermore, we delve into understanding the transfer of trust levels assigned by operators to legged manipulators across varied tasks and environmental contexts, exploring the influence of both HH and AR HMD UIs on this transfer process.

To this end, we conducted a user study with 27 participants, engaging them in various navigation and manipulation tasks of differing difficulty levels. The study employed a **Boston Dynamics (BD)**'s Spot, a 7 df Kinova robot arm and a Robotiq 2F-85 gripper integrated into a legged manipulator. Participants progressed through five stages (illustrated in Figure 1), involving two distinct UIs (HH and AR), two task types (navigation and manipulation), two task difficulties for navigation (obstacle-free and with obstacles) and two performance outcomes for manipulation (success and failure). Alongside validated metrics for successful HRIs [5, 28, 29, 68, 82] (refer to Section 2.3), we utilised standardised trust questionnaires to gauge two key factors in HRI: delegation and reliance. We considered these factors relevant for inspection applications involving legged manipulators [56, 60]. A video abstract is available at https://youtu.be/I9IfQbMKc4c.

Our experimental design, discussed in Section 4, employs a factorial setting where the same group of participants completed a series of Likert-type instruments after experiencing five

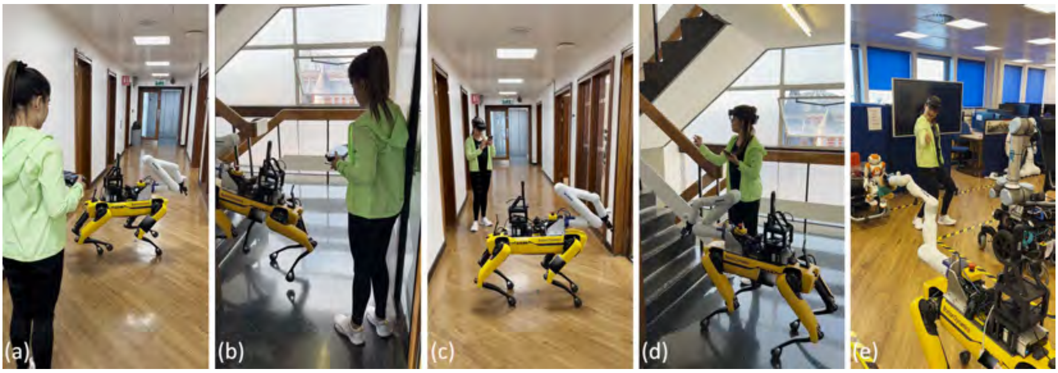


Fig. 1. Our user study involved navigation and manipulation tasks with varying difficulty levels using a BD's Spot, a 7 df Kinova robot arm and a Robotiq 2F-85 gripper that we integrated into a legged manipulator. We compared an AR HMD UI we designed against off-the-shelf HH UIs provided by the manufacturers. For this comparison, we used validated quantitative metrics for successful HRIs. Furthermore, we used standardised trust questionnaires to measure how the level of trust assigned by the participants to the legged manipulator varies across experimental conditions. These experimental conditions were (a) Navigation—Obstacle-Free—HH Controller, (b) Navigation—With Obstacles—HH Controller, (c) Navigation—Obstacle-Free—Relevant Experience (AR) HMD UI, (d) Navigation—With Obstacles—AR HMD UI and (e) Manipulation—User-Selected UI.

different conditions. This design yields longitudinal data [66]. The distribution of such data is often unknown and may include atypical measurements and outliers. Therefore, analysing these data using parametric and semi-parametric procedures imposing restrictive distributional assumptions on observed longitudinal samples becomes questionable [66]. While more suitable non-parametric equivalents exist [66], their use may result in unnecessary statistical power loss and precision compared to continuous alternatives [81]. Bayesian data analysis has proven to be among the most effective alternatives [46, 61]. Consequently, we analysed our data using a Bayesian framework. Aiming at making such analysis useful to the wider community, in Section 5 we show how Bayesian data analysis can be applied to multiple well-established questionnaires.

In our previous study [13], we reported pilot results using our proposed UI. The AR UI enhanced immersion and performed comparably in terms of technology acceptance, fluency and system usability. However, off-the-shelf controls surpassed it in time performance and cognitive workload. Recognising the importance of these metrics, we enhanced the AR UI with a new cognitive offloading functionality, based on insights from the pilot study. With an increased sample size, we present aggregated results in Section 6. To summarise, the primary contributions of our work are twofold:

- First, we emphasise the significance of Bayesian data analysis as opposed to standard parametric or semi-parametric procedures in evaluating HRI experiments. Through careful illustration, we demonstrate the importance of Bayesian analysis, showcasing instances where standard procedures may lead to erroneous conclusions. This analysis is non-trivial but is crucial for the relevant evaluation and interpretation of experimental results.
- Second, our work underscores the critical role of UI design in the effective and trustworthy utilisation of modern technologies, particularly robotic systems. By incorporating cognitive offloading principles into the UI, we made a substantial impact on key parameters crucial for HRI, as highlighted in our thorough investigation of objective and subjective metrics. The study, involving 27 participants utilising an author-designed AR UI and a off-the-shelf HH control method for a legged manipulator robot, provides valuable insights into trust,

immersion, usability, fluency, technology acceptance and cognitive workload, supported by a deep Bayesian statistical analysis.

2 Related Work

2.1 AR-Based Robot Control

Several researchers have previously considered the problems involved in AR-based robot control [57, 79]. Applications in this domain can be classified into two categories: remote interactions (i.e., teleoperation) [73], where the user sends control commands from outside the robot's location, and proximal interactions, where the user and the robot share the same location. Applications in the latter domain include robot programming [2, 22, 26, 45], trajectory planning [23, 40] and assistive robotics [8, 9]. For example, the authors of [22, 23] proposed an AR approach for robot programming and trajectory planning applied to a robotic arm. An HH device was used to input the start and goal points and to modify the automatically generated path. A computer display was used to show a video stream overlaid with virtual objects such as the robot's **three-dimensional (3D)** model and the path to be followed by the end-effector of the robot. A similar application was presented in [45], except that the authors used an HMD to deploy their UI. This UI was designed for skill-based programming on an assembly task. Another example is [2], where the authors performed user trials with a tour-guide robot controlled using hand gestures and AR buttons overlaid on the robot's screen. A video stream of the robot's view shows several virtual elements to the user. Finally, [40] presented an AR HMD UI to control a virtual drone through a 'point and go' method where the user inputs the start and target position using hand gestures. The interface then shows the virtual robot following the trajectory selected by the path-planning algorithm.

Most teleoperation applications overlay a video stream from a remote location with virtual objects, such as a 3D robot model or virtual handles that the user then accesses using either a computer display [63] or a **Virtual Reality (VR)** HMD [52]. Examples in this domain include [34], a teleoperation platform for controlling a robot through a touch-screen-based AR interface. The system allows the user to manipulate each part of the robot using a 3D model of the robot and virtual handles overlaid onto the model. The users specify the desired pose of the robot by touching and dragging the part they want to control. Another example is proposed in [86], where the authors presented a teleoperation system for maintenance robots using an AR interface. The remote scene is reproduced at the user side using a mixture of virtual and real objects using a VR headset. A HH device controls the robot by defining target positions and 'drawing' paths along which it should move its end-effector.

Proximal interactions typically use a computer screen or an AR HMD to overlay the environment with virtual elements such as the planned motion of a robot [40], the torques experienced by the joints of a robotic arm [59] and virtual buttons [2]. Input modalities for remote and proximal interactions include touch-screens [34], hand gestures [2, 40, 55], fiducial markers [51], haptic [89], HH devices [22, 86], electroencephalography signals [83], eye-gaze [88] or a combination of these.

2.2 User Studies with AR UIs for Robot Control

Most existing user studies in the domain of AR UIs are limited to teleoperation and programming applications [2, 9, 22, 26, 34, 54]. Next, we introduce representative examples of user studies outside these applications demonstrating the benefits in HRIs derived from using AR UIs to control single [10, 14, 15, 25, 72] and multirobot platforms [11, 43].

In [10], the authors evaluated an AR HMD UI for controlling a smart wheelchair. Their results show that the task load reported by participants is lower when controlling the smart wheelchair with their UI compared to using the joystick. In [72], the authors evaluated an AR application that disambiguates item references. They compared their AR application against physical actions for reducing robot uncertainty and found their approach to be more accurate, faster and improve usability, trust and workload. In [15], the authors developed a user study simulating a collaborative manufacturing task using a robotic arm. Their results show that the AR interface feels more novel, and a standard joystick interface feels more dependable to users. However, the AR interface was found to reduce **Physical Demand (PD)** and task completion time while increasing robot utilisation. In [14], the authors present a user study leveraging AR HMD UIs and eye-gaze for bidirectional communication between users and a robotic arm for a collaborative task. Their results show that their UI feels more novel to users and reduces PD and task completion time while increasing robot utilisation. In [25], the authors tested an AR-assisted robot learning framework for minimally invasive surgery tasks. Their results show that their proposed framework requires lower workload demand and achieves higher performance and efficiency.

In [11], the authors presented a user study evaluating an AR HMD UI for controlling a multi-robot platform. Their results showed a significant improvement in fluency perception derived from their UI. Finally, in [43], the authors compared an AR interface that allows users to interact with multiple mobile robots simultaneously deployed using an HMD against when deployed on a tablet device. They showed that the performance and preference of the interface depend on the task and the complexity of the required interaction.

2.3 Subjective Metrics for Assessing HRIs

Following standard practice in HRI and interaction design, we assessed our platform using established Likert-type questionnaires, common in HRI research [10, 11, 13, 74, 75, 90, 91]. Next, we introduce these and describe why they are relevant to the HRI field.

2.3.1 Cognitive Workload. Cognitive workload is the level of measurable mental effort put forth by an individual in response to one or more cognitive tasks. Cognitive workload is subjective and can vary between individuals, tasks and situations [29]. The NASA **Task Load Index (TLX)** is a widely used subjective assessment tool designed to measure the relative importance of six factors that determine how much workload a user experiences while performing a task. These six factors are (1) **Mental Demand (MD)**: How much mental and perceptual activity was required? (2) PD: How much physical activity was required? (3) **Temporal Demand (TD)**: How much time pressure did the user feel? (4) Performance (P): How successful do users thinks they were in accomplishing the goals? (5) Effort (E): How hard did the users have to work to accomplish their level of performance? and (6) Frustration level (F): How insecure, discouraged, irritated, stressed and annoyed were the users? [33]. These are rated using six 21-point Likert scales with higher scores indicating higher perceived workload.

In safety-critical environments, such as the ones where legged manipulators are deployed, maintaining an appropriate cognitive workload is crucial to avoid errors and accidents [24]. Managing workload ensures that humans can effectively monitor and control robots without compromising safety. More generally, managing cognitive workload is essential in the HRI field to optimise task performance, enhance user experience, ensure safety and promote the acceptance and adoption of robotic technologies in various applications [29]. Integrating insights from cognitive workload research contributes to the development of more effective and user-friendly HRIs [10].

2.3.2 Technology Acceptance. The **Technology Acceptance Model (TAM)** is used to assess the acceptance of people towards technology products. Originally proposed by Fred Davis in 1989, TAM has since undergone several modifications and extensions [17, 18, 82]. TAM2 is one of such extensions [82]. It includes a total of 10 five-point Likert scale items encompassing the following three dimensions: (1) Perceived usefulness, which is the people's tendency to use or not an application to the extent they believe it will help them perform their tasks better. (2) Perceived ease of use, which refers to the degree to which a person believes that using a particular system would be free from effort. It encompasses factors such as the simplicity of interaction, user-friendly interfaces and the ease of learning. (3) Intention of use, which refers to the likelihood of using new technology in the future.

While TAM was originally developed in the context of information technology, its principles are generalisable to various technological domains, including robotics [78]. Applying TAM in HRI research helps researchers and designers better understand the factors that influence the acceptance of robots by users. It can guide the development of robotic systems that are user-friendly, efficient and align with users' needs and expectations [35]. For instance, in robot interface design, factors such as the simplicity of the robot's interface, intuitive controls and clear communication mechanisms contribute to robot acceptance.

2.3.3 Fluency. In the context of HRI, fluency is defined as a high level of coordination and adaptation between humans and robots, particularly in repetitive tasks [36, 37, 68]. Previous work encompasses various factors influencing HRI fluency, such as the approval of the robot's contribution to team success, optimal work design, user preferences and the impact of job training. Additionally, researchers have examined the interplay between efficiency, worker satisfaction and the control individuals have over their roles in human–robot teams. Their considerations extend to workload, emotion regulation and trust, recognising that fluency plays a role in perceived job satisfaction, emotional experiences, engagement in teamwork and perceptions of the robot's attributes [68].

The subjective assessment of human–robot fluency in HRI includes a questionnaire with 30 items using a five-point Likert scale covering the following seven dimensions: (1) human–robot fluency, which evaluates the overall fluency between the human and the robot; (2) robot relative contribution, which evaluates the robot's contribution to the team; (3) trust in a robot, which evaluates the trust the robot evokes; (4) positive teammate traits, which evaluates the robot's perceived character traits related to it being a team member; (5) improvement, which are measurements applicable to learning and adaptation scenarios; (6) working alliance for human–robot teams, which measures the quality of the human–robot teamwork; and (7) individual measures, which measure the sense of commitment from the users and the robot towards achieving team goals [36].

2.3.4 System Usability. Usability focuses on how well users can learn and use a device to achieve their goals. It encompasses various factors, including the efficiency of the system, ease of learning, user satisfaction and the overall effectiveness of a **UI** [7]. The **Hybrid System Usability Scale (H-SUS)** is a widely used questionnaire for assessing the perceived usability of a system [5]. It consists of 10 items that combine pictorial and verbal information with a five-point Likert scale. The H-SUS questionnaire covers dimensions such as (1) the complexity of the system, (2) ease of use and (3) the user's confidence in using it. The questionnaire provides a standardised and reliable way to assess subjective perceptions of usability, contributing valuable insights into the user experience of a given system.

While the H-SUS was not specifically designed for evaluating HRIs, assessing the usability of interfaces, communication methods and overall user satisfaction remains relevant in HRI to enhance the effectiveness of robotic systems in various applications. In HRI, the design of robot interfaces is paramount to facilitate the interaction between humans and robots, influencing the overall user experience [69]. The usability of these interfaces, determining how easily users can command the robot and comprehend its responses, plays a crucial role in shaping the success of the interaction. Similarly, in scenarios requiring human control or command over robots for specific tasks, the usability of control systems becomes pivotal, impacting the ease of communication and the robot's responsiveness [1].

2.3.5 Immersion. Immersion is a form of cognitive and emotional absorption that promotes enjoyment and engagement in a task or while learning [28]. It is characterised by a sense of presence, involvement and absorption in the virtual or AR environment. For location-aware AR applications, immersion is measured using **AR Immersion (ARI)**, a 21-item seven-point Likert-type instrument based on a multi-levelled immersion model with multidimensionality at each of its three levels [28]. These levels are (1) Engagement: this first level is divided into the constructs of *interest*, which measures the user's interest in the activity, and *usability*, which measures the user's perception of the usability of the application. (2) Engrossment: this second level is divided into the constructs of *Emotional Attachment*, which measures the emotional attachment to the activity, and *Focus of Attention*, which measures the user's focus during the activity. (3) Total Immersion: this third level is divided into the constructs of *Presence*, which measures the user's sense of feeling surrounded by a blended yet realistic physical/virtual environment, and *Flow*, which measures the user's absorption in the activity.

While the ARI questionnaire was not specifically designed for HRI, its principles can be applied when evaluating the immersive aspects of AR UIs used in conjunction with robots. It provides a structured way to gather subjective feedback on users' perceptions and experiences, which can be beneficial for optimising the design and integration of AR elements in various HRI applications. By using the ARI questionnaire, researchers and designers can gain insights into how users perceive and engage with AR content, which is valuable for improving the design and usability of AR-based human–robot interfaces [79].

2.3.6 Trust. Trust in HRI is defined as a person's calculated exposure to the risk of damage from the activities of influential others [32]. Recent efforts for understanding and measuring people's trust in robots have aimed at improving trust measurements by creating and adopting standardised trust questionnaires for easy comparison across experiments [16]. In previous work, multiple researchers have developed survey instruments and methods to quantify trust [20, 21, 31, 39, 41, 53, 84, 85]. For this article, we measure trust in HRI using a 10-item, seven-point Likert-type instrument, which is based on four items from the **Integrative Model of Organisational Trust (IMOT)** [60], which focus on delegation, i.e., assigning tasks or responsibilities to the robot with the expectation that it will carry them out autonomously or with minimal human intervention, and an additional six items from the **Reliance Intention Scale (RIS)** [56], which focus on reliance, i.e., the extent to which humans depend on the robot to perform tasks or provide assistance. We chose these two from the many factors involved in trust in HRI as we deemed them relevant for the inspection applications, typical for legged manipulators, i.e., human's trust in a robot may lead to increased reliance and a greater willingness to delegate tasks [56].

2.4 Trust Transfer

Trust transfer refers to the phenomenon that makes the level of trust assigned by a human to a robot or other devices vary across tasks and environmental contexts as a consequence of the human's prior knowledge of the robot's or device's capabilities and past experiences. Such past experiences include, for example, the observation of success or failure during the execution of a task [77]. In previous work, researchers have explored multi-task and multi-device trust transfer [67, 77].

In [77], the authors presented a user study to investigate how human trust in robot capabilities transfers across multiple tasks across two disparate domains: household tasks and autonomous driving. They showed that inter-task trust transfer depends on perceived task similarity, difficulty

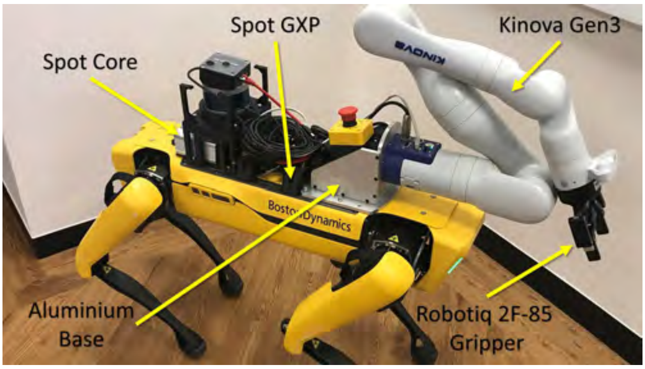


Fig. 2. Hardware setup. We used a BD Spot, a 7 df *Kinova Gen3* robot arm and a *Robotiq 2F-85 gripper* that we integrated into a legged manipulator as the robot platform for our user study. We called the integrated platform SpoK. We powered the Kinova arm using Spot General Expansion Payload (*GXP*), an interface to Spot's main payload port that offers regulated power at 24 V with a maximum power output of 150 W. We provide the onboard computing processing of the robot platform using *Spot Core*, a computer equipped with an Intel i5 CPU, 16 GB of RAM, 512 GB of storage and Ubuntu Desktop 18.04. The *aluminium base* is a bespoke payload adaptor we built to mount the Kinova arm [12].

and observed robot performance. Their results were consistent across domains. More recently, in [67], the authors explored, through a user study, whether trust can be transferred among multiple devices, the effect of the similarity among these devices and the effect of the agency attributed to each of them. Their results showed that trust can be transferred across devices and that clarifying the distinction between devices by attributing different agencies to each prevents trust transfer.

Nonetheless, to our knowledge, no other work has explored how the level of trust assigned by an operator to a legged manipulator transfers across tasks and environmental contexts. Furthermore, no previous work has evaluated how HH and AR HMD UIs affect such transfer.

3 System

3.1 Legged Manipulator

We used a BD Spot, a 7 df Kinova robot arm and a Robotiq 2F-85 gripper that we integrated into a legged manipulator as the robot platform for our user study (see Figure 2). Spot is a battery-powered legged robot capable of mobility on various terrains with built-in obstacle avoidance. The Kinova robot arm has 7 df and an integrated vision module. We mounted the arm on Spot using a bespoke aluminium base, which we screwed onto Spot's mounting rails. We powered the arm using Spot General Expansion Payload, an interface to Spot's main payload port that offers regulated power at 24 V with a maximum power output of 150 W. For the gripper, we used a two-finger 85 mm gripper from Robotiq, attached off-the-shelf to the arm. We provided the on-board computing processing of the robot platform using Spot Core, a computer equipped with an Intel i5 CPU, 16 GB of RAM, 512 GB SSD, Ubuntu Desktop 18.04. We called the integrated platform SpoK [12]. We developed the code to interface with SpoK using BD's Python API 3.0.0, ROS Noetic and Docker.

3.2 AR HMD UI for Controlling Legged Manipulators

To build the AR UI we used in our study, we used the Microsoft HoloLens 2 as HMD and Unity 2019.4.10f1, Mixed Reality Toolkit v2.5.3 and ROS Sharp. The AR UI components provide several visual cues and control commands for the user (see Figure 3). Raising the left palm displays two

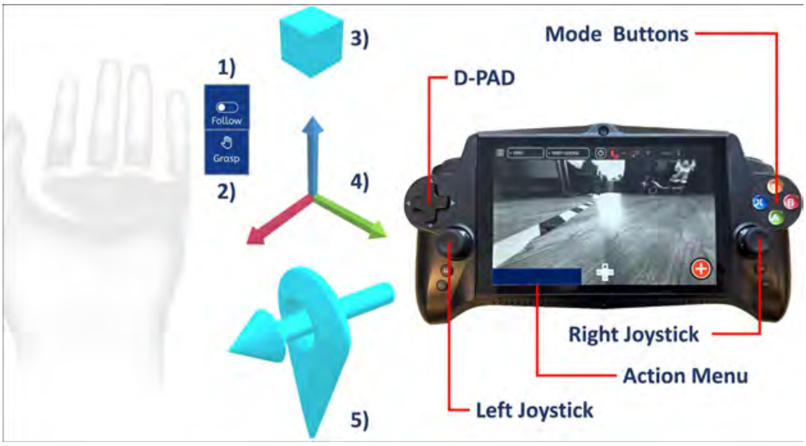


Fig. 3. Left: AR HMD UI elements—Raising the left palm displays two options for interacting with SpoK: (1) Follow and (2) Grasp. The UI incorporates several elements, including (3) Anchor, (4) Spok's odometry frame and (5) GoTo marker icons. Right: Off-the-shelf UI developed by BD. See Section 3.4 for further details. Go to Spot controller configurations for more information.

options for interacting with the robot: Follow and Grasp (1 and 2 in Figure 3). We use a coordinate system symbol to represent the origin of Spot's odometry frame, i.e., the position where Spot was when turned on (4 in Figure 3). We also included a GoTo marker to represent target poses for the robot. The user can move and rotate this marker using hand gestures to place it at the desired location with the desired orientation (5 in Figure 3). Then, the user can press the button on top of the icon to send the desired pose to the robot.

The UI also offers a 'follow' behaviour [12]. When using this behaviour, the users can navigate their environment while the robot maintains the distance between them within a fixed range. Finally, derived from our experience with our previous smaller study [13], we improved the functionality of the AR UI by exploiting the concept of cognitive offloading. Cognitive offloading refers to using physical activity to alter the information processing requirements of a task to reduce cognitive demand [71]. When using the controller, we noticed that some users rotated their bodies to align themselves with the robot and facilitate its positioning. Similarly, we added a new functionality to the AR UI that allows the users to position themselves where they want the robot to go and orient themselves in the direction they want the robot to face. Then, after receiving a voice command ('Come Here'), the robot positions itself to match the required pose [13]. See the supplementary video accompanying this article for more details.

3.3 Colocalisation

Colocalisation is the process that allows localising the robots and the HMD under a global coordinate system. The 3D poses send from the UI are relative to the HMD's frame of reference and cannot be directly used by the robots. To find the transformation between their frames, three main colocalisation approaches exist: marker-based [64], which require instrumenting the environment with fiducial markers that the robot and the HMD must have in their view; map-based [70], in which a robot or an HMD define a reference coordinate system by building a map of the environment first; and vision-based [19].

For our study, we used a vision-based colocalisation approach [12, 13, 19]. This approach extracts sparse visual features from images and 3D poses provided by the simultaneous localisation and

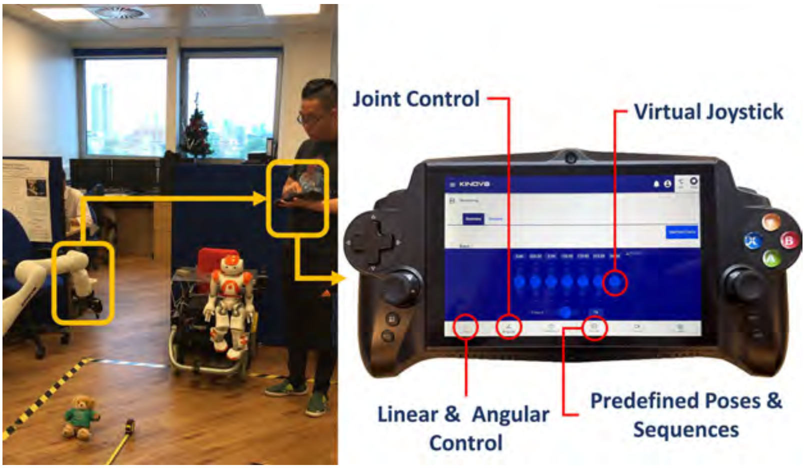


Fig. 4. A participant uses the HH UI to control the robot arm. The Kinova Kortex Web App UI provides multiple virtual joysticks and control modes for the arm and the gripper. Furthermore, the UI allows the user to select between predefined poses or sequences of poses. Go to Kinova Kortex Web App for further details.

mapping system of the HMD and the robot. These features are then matched and used to localise the HMD and the robot against a spatially and temporally stable 6 df pose relative to these features. This pose is known as an anchor and provides a common reference frame between the different map representations built on-line by the robot and the HMD.

Several cloud-based services exist to create these anchors for AR applications [4, 30]. For this work, we use Spatial Anchors [62]. We use a small blue cube to represent an anchor in the AR UI (3 in Figure 3). This cube allows the user to visually verify that the HMD found and placed the anchor correctly.

3.4 Off-the-Shelf Control Methods

For our study, we compared our AR UI against the off-the-shelf UIs developed by the manufacturers to control Spot and the Kinova Gen3 Robot Arm. For this, we used the HH controller provided by BD. We illustrate these UIs in Figures 3 and 4. Spot's off-the-shelf UI allows the user to control the robot using the right and left joysticks for position and orientation. In addition, the UI shows views from the robot's cameras, which the user can change using a D-pad. Using any of these cameras, Spot's UI also offers a 'touch-to-go' option for controlling the robot. Touch-to-go allows the user to drive Spot to a specified location by touching the desired position on the camera image displayed on the controller screen. With this option, the user can also drag a virtual icon while touching the screen for finer location specification, working as a HH AR control method. Using the action menu or mode buttons (Figure 3), the user can access different gait modes available, e.g., stairs mode. The arm's UI provides multiple virtual joysticks and control modes for the arm and the gripper. Furthermore, the UI allows the user to select between predefined poses or sequences of poses (Figure 4).

4 Experiment

Figure 1 illustrates the experimental procedure. It complements the following written descriptions, offering a more comprehensive and accessible representation of the experimental process.

4.1 Participants

Our study received ethics committee approval from Imperial College London. We recruited 27 participants through e-mail advertisements on our university campus. We did not offer financial incentives for participating in the study. We conducted the experiments at the Personal Robotics Laboratory at Imperial College London. The participants had to be at least 18 years and sign a consent form before proceeding with the experiment.

4.2 Experimental Conditions

The participants completed the experiment in five stages (see Figure 1). These stages involved two different UIs, HH and AR, two different task types, navigation and manipulation, two different task difficulties for the navigation task, obstacle-free and with obstacles, and two different performance outcomes for the manipulation task, success and failure. We describe these five stages as follows:

(1) Navigation—obstacle-free—HH: This stage involved navigating with the robot across an obstacle-free corridor and visiting three waypoints. The start position was always the same. We placed the waypoints 5 m apart along a straight line (see top of Figure 5). The participants completed this stage using the HH UI.
(2) Navigation—with obstacles—HH: This stage involved navigating through a 90-cm wide door opening, a 180-cm door opening and taking the robot up a flight of stairs. This stage also involved visiting three waypoints (see bottom section of Figure 5). The participants completed this stage using the HH UI.
(3) Navigation—obstacle free—AR: This stage was identical to navigation—obstacle-free—HH, except for the UI the participants used to complete it. The participants completed this stage using the AR HMD UI.
(4) Navigation—with obstacles—AR: This task was identical to navigation—with obstacles—HH, except for the UI used by the participants to complete it. The participants completed this stage using the AR HMD UI.
(5) Manipulation—user-selected UI: This last stage involved picking an object from the floor while the robot remained still. The participants had a maximum of 5 minutes to successfully grasp an object placed on the floor 130 cm from the robot's base-link[1]. Furthermore, the participants had to select their preferred UI to complete this stage. The participants made this selection after having experienced both UIs while completing the four previous stages. Nonetheless, they had to remove whatever UI they had last. This was to avoid participants basing their decisions on wanting to finish the experiment as fast as possible. In addition, we controlled for whether the participants succeeded or failed at this task. We illustrate this task in Figures 4 and 6.

The order of the experimental stages was pseudo-randomised using either a (1)(2)(3)(4)(5) or a (3)(4)(1)(2)(5) order to account for temporal effects on the participants' performance, such as fatigue, habituation or learning effects. We randomly assigned the participants to the (1)(2)(3)(4)(5) or (3)(4)(1)(2)(5) order, such that half of the participants started using the HH UI to complete the first two stages and then the AR HMD UI for the following two stages. We refer to this group of participants as $G_{\text{HH}\rightarrow\text{AR}}$. The other half started using the AR HMD UI. We refer to this group of participants as $G_{\text{AR}\rightarrow\text{HH}}$.

[1] Any position within the range from 150 cm and 112.5 cm relative to the robot's base-link and along its X-axis (forward) would place the object in-front-of the robot, making it reachable by the arm without moving the base [12]; 130 cm is approximately the middle value for this range.

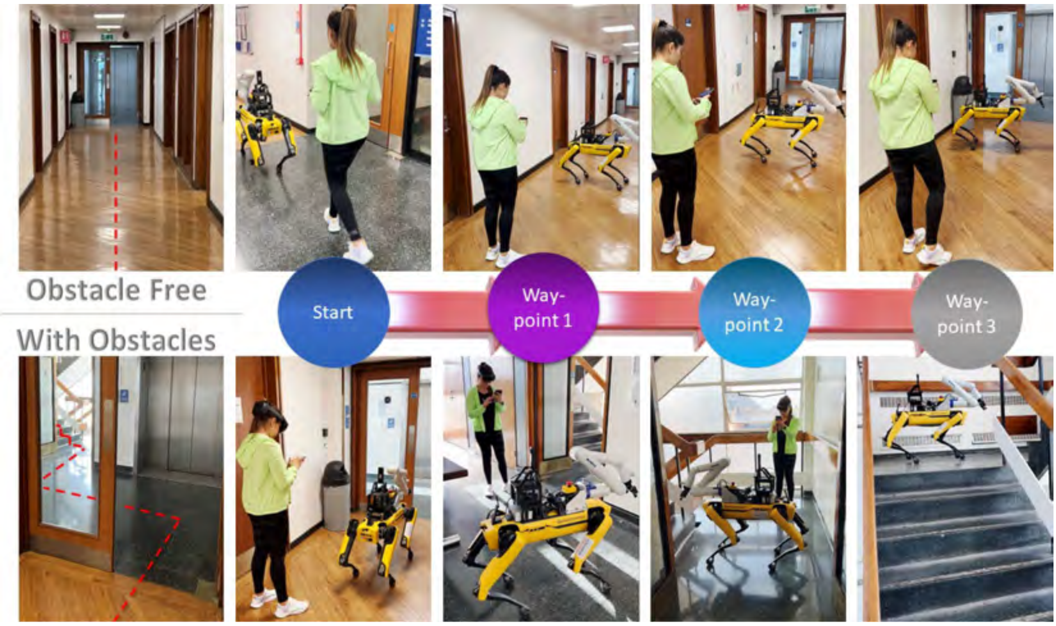


Fig. 5. Experimental conditions—navigation. Top—The obstacle-free stages involved navigating with the robot across an obstacle-free corridor along three waypoints. Bottom—With obstacle stages involved navigating through doors and taking the robot up a flight of stairs along three waypoints. We gave the participants a secondary task where they had to take photos using a 360 camera while the robot was rotated perpendicular to the corridor at each waypoint.

4.2.1 Secondary Task. For the first four stages, we gave the participants a secondary task where they had to take photos using a 360 camera with the robot rotated perpendicular to the corridor at each waypoint. The 360 camera (Ricoh Theta Z1) was onboard the robot, and the participants had to use a mobile application running on a separate device (Apple iPhone 13 Pro) to take the photo. We gave participants this secondary task to simulate an inspection task, typical of legged robots, where they needed to use their hands to take pictures in addition to controlling the robot. While we acknowledge that the data capture process could be seamlessly integrated into the UIs, we deliberately kept it separate. This decision was made to explore whether requiring participants to use their hands for other tasks while using an HMD would yield any reported benefits.

4.3 Procedure

After signing a consent form, completing a safety screening questionnaire and providing standard demographic data, we introduced the participants to the robot's parts and capabilities. In addition, we explained the different stages involved in the study. Next, we continued with a training session with both UIs. Then, participants continued with the five experimental stages. After each stage, the participants completed a trust questionnaire (see Section 2.3.6). In addition, after completing stages (2) and (4) (see Section 4.2), we asked participants to complete another five questionnaires (see Section 5.2). Finally, after stage (5), the participants completed a free text questionnaire elucidating their experience with the study and explaining their reasons for selecting their preferred UI.

An average session took 2.25 hours to complete. Although we acknowledge that 2.25 hours could potentially involve fatigue as a confounding factor, each of the previous stages took no longer than 6 minutes. Participants dedicated the remaining time to the training session and to

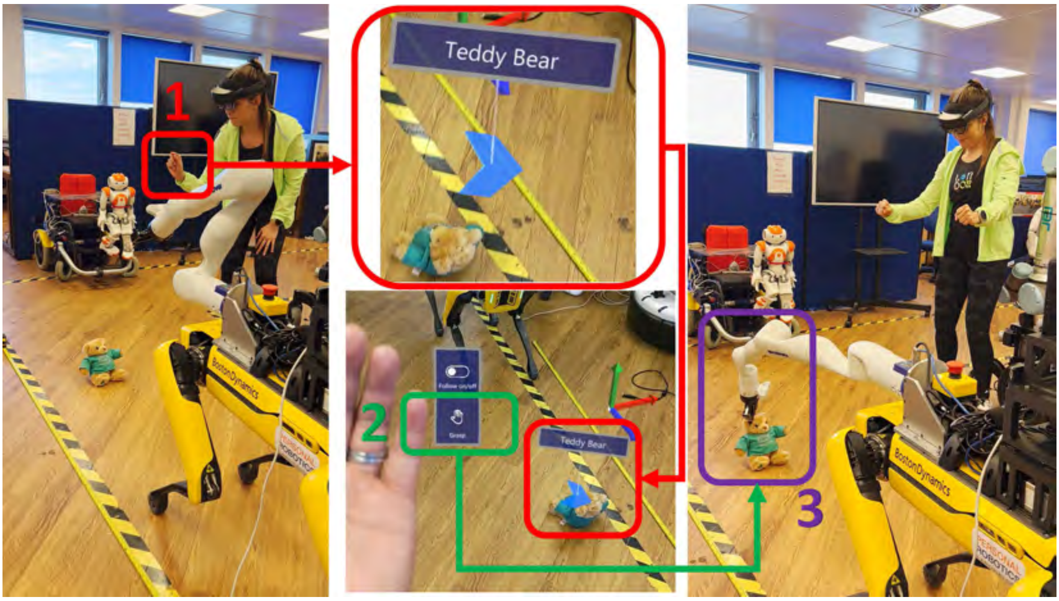


Fig. 6. Experimental conditions—manipulation. This stage involved picking an object from the floor while the robot remained still. The participants had to select their preferred UI to complete this task. Here—A participant using the AR UI to control the robot arm. For this, the participants had to (1) place an AR marker onto the object using hand gestures and (2) select the 'Grasp' option from the hand menu to (3) make the robot execute the grasping behaviour autonomously. See [12] for more details.

complete questionnaires between stages while the participants were seated. Furthermore, we offered participants refreshments and multiple breaks to minimise fatigue. Moreover, none of the participants indicated or expressed discomfort or displayed signs of fatigue.

5 Analysis

Our experimental design, as introduced in Section 4.2, employed a factorial setting where the same group of participants completed a series of Likert-type instruments after experiencing five different conditions. This design yields longitudinal data [66]. In traditional **Null Hypothesis Significance Testing (NHST)**, a parametric test such as a repeated measures ANOVA would be used to analyse the data. This test can assess the main effects of each factor as well as any interactions between factors [38]. However, the choice of the specific statistical test can depend on the assumptions of the data and the nature of the research question. Furthermore, as the number of factors increases, the complexity of interpreting interactions also grows, so researchers rarely apply this test on studies with more than three factors [38].

The distribution of longitudinal data is often unknown and may include atypical measurements and outliers. Therefore, using parametric and semi-parametric procedures imposing restrictive distributional assumptions on observed longitudinal samples becomes questionable [66]. While more suitable non-parametric equivalents exist [66], their use may result in unnecessary statistical power loss and precision compared to continuous alternatives [81]. Bayesian data analysis has proven to be among the most effective alternatives [46, 61]. Consequently, we analysed our data using a Bayesian framework. In the following sections, we detail the Bayesian models applied to analyse the data collected during the user study.

5.1 Objective Measures

5.1.1 Time to Complete Task. We measured the time participants used to complete tasks for each experimental condition and compared their performance under both UIs. In traditional NHST, a t-test would be used for a two-group comparison of metric data such as this one. However, the t-test assumes normality, leading to large estimates of within-group variances in the presence of outliers—data values significantly deviating from expected values [46]. Unlike Bayesian approaches, the t-test only assesses means equality, lacking a test for variances equality. An additional F-test would be necessary, but it inflates p values for both tests and assumes normally distributed data, being sensitive to deviations from this assumption [47].

Bayesian approaches employ *robust estimation* to mitigate the impact of outliers [46, 61]. Robust estimation replaces the standard distribution, usually the normal distribution, in metric models with a thicker-tailed distribution like the t distribution. The t distribution is parameterised with μ, σ and ν. Similar to a normal distribution, μ and σ control the mean and width, while ν (normality parameter) governs tail heaviness. Its value ranges continuously from 1 (very heavy tails) to ∞ (normal distribution). Equation (1) outlines the robust estimation model we used to compare the participants' task completion time for the two UIs.

$$\begin{aligned} T_i &\sim \mathrm{t}\,(\nu, \mu, \sigma) & \sigma &\sim \mathrm{Uniform}(L = 0, H = 70) \\ \mu &= \alpha[\mathrm{UI}] & \nu &= 1 + \nu_{-1}\;. \\ \alpha[\mathrm{UI}] &\sim \mathrm{Normal}(M = 180, S = 30) & \nu_{-1} &\sim \mathrm{Exponential}(1/29) \end{aligned} \tag{1}$$

Here, T denotes the list of time-to-complete-task values for each stage, with index i representing the row number. UI = $\{1, 2\}$ serves as an index variable, where UI = 2 corresponds to participants using the AR UI, and UI = 1 indicates the HH controller. Priors for each α, the mean time for each UI follow a normal distribution with mean $M = 180$ and **Standard Deviation (SD)** $S = 30$. Navigation stages were designed for an average completion time of 180 seconds and a maximum of 360 seconds, resulting in a broad range of plausible times ($180 \pm 2 \times 30$). The σ priors are flat, constraining σ between zero and 70 seconds, indicating that 95% of recorded times would fall within 140 seconds of the average. ν_{-1} follows an exponential distribution with a mean of 29, ranging from 0 to ∞. To allow ν to vary from 1 to ∞, we add 1 to ν_{-1}, changing its mean to 30. The t distribution exhibits significant variation for small ν values; for ν greater than 30, it is practically normal. The ν prior provides equal opportunity for small (<30) and larger (>30) values [46]. Prior predictive simulation [61] indicates that, before seeing the data, the model expects 0.0% of participants to finish the task in less than 0 seconds and 1.0% in more than 360 seconds, aligning with our design choices.

5.2 Subjective Measures

We assessed our platform using the Likert questionnaires discussed in Section 2.3. These questionnaires involve ordinal outcome variables, with responses scaling in one direction ($1 < 2 < 3 < 4 < 5$), though not necessarily in equal steps [81]. Typically, researchers report each item's mean, SD, and the number of observations [81]. However, using means and SDs assumes continuous Likert scales with equal step sizes, which may not accurately estimate distances between items due to potential non-linearities in Likert judgements [81].

Non-parametric tests like Mann-Whitney U test, Wilcoxon signed-rank test and Kruskall-Wallis H test offer alternatives to using means and SDs for estimating distances between items in ordinal outcomes [38]. These tests dichotomise a rated feature to compare the highest- and lowest-rated items, less affected by distortions in Likert ratings' averages, as the comparison remains essentially ordinal. However, such dichotomisation leads to unnecessary statistical power loss and precision compared to continuous alternatives, such as Bayesian approaches [81].

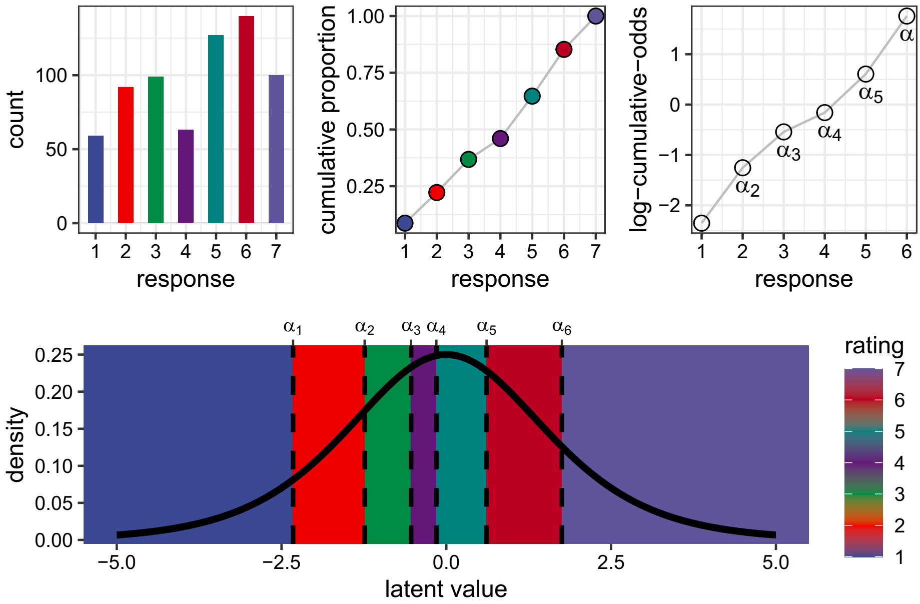


Fig. 7. *Top left*: Histogram representation of a seven-point Likert scale ($K = 7$), displaying response values on the x-axis and their counts on the y-axis. *Top middle*: Likelihood, $Pr\ (r = k)$, for each possible response value $k \in [1, 7]$, represented as cumulative probabilities $Pr\ (r \leq k)$. *Top right*: Cumulative probabilities as log-cummulative-odds using Equation (3), which represent the estimated locations of latent thresholds $\hat{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_6\}$. *Bottom*: Bayesian representation of the Likert scale. Cumulative Link Mixture Models (CLMM) represent the histogram through an underlying continuous distribution, here as a logistic distribution, utilising a logit-link function. In the figure, coloured regions translate from latent values to Likert responses (values between 1 and 7).

5.2.1 *Bayesian Model for Likert-Type Instruments in HRI.*

Likert data can be visualised through histograms, displaying response values on the x-axis and their counts on the y-axis, forming a discrete distribution. Nonetheless, a more effective alternative summarises Likert data by assuming that a latent continuous distribution underlies the discrete distribution [81]. This is illustrated in Figure 7. In such case, the likelihood of encountering a specific Likert response corresponds to the probability of selecting a value from the latent distribution within the defined lower and upper bounds of that particular Likert response value. These bounds are represented by $\hat{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_{K-1}\}$, a vector of intercepts α_k, with one entry for each possible response value $k \in [1, K]$, e.g., $K = 7$ for a seven-point response scale.

This Bayesian representation of Likert data can be implemented using **Cumulative Link Mixture Models (CLMM)** [81], which additionally can accommodate multi-item questionnaires $Q \in [1, M]$, and the participation of multiple individuals $P \in [1, N]$, characteristics common in most HRI studies. Here, M is the number of Likert-items in the questionnaire, and N is the number of participants in the study. For example, the NASA-TLX questionnaire has $M = 6$ Likert items: MD, PD, TD, performance, effort and frustration. Thus, $Q \in [1, 6]$.

CLMM use a cumulative logit-link function and a categorical distribution as likelihood to fit a list of response values (R) for all Likert items in a questionnaire [61]. In our case, we obtained R from the participants' responses to the questionnaires in our user study (see Section 2.3). The

CLMM we employed as the Bayesian representation of our data are outlined as follows:

$$\begin{array}{rr} R_i \sim \text{Ordered-logit}\,(\phi_i, \alpha) & \alpha_{\mathrm{p}}\,[P] \sim \text{Normal}\,\left(0, \sigma_{\mathrm{p}}\right) \\ \phi_i = \alpha_{\mathrm{q}}[Q_i] + \alpha_{\mathrm{p}}[P_i] & \widetilde{\alpha} \sim \text{Normal}\,(0, 1.5) \\ \alpha_k \sim \text{Normal}\,(0, 1) & \sigma_{\mathrm{q}} = \text{Half- Normal}\,(0, 1) \\ \alpha_{\mathrm{q}}\,[Q] \sim \text{Normal}\,\left(\widetilde{\alpha}, \sigma_{\mathrm{q}}\right) & \sigma_{\mathrm{p}} = \text{Half-Normal}\,(0, 1) \end{array}. \tag{2}$$

Here, the index i takes the row numbers in R. To account for the ordered nature of the response values in a Likert scale, each intercept α_k in $\hat{\alpha}$ is defined using a cumulative-logit-link function

$$\log \frac{Pr\,(r_i \leq k)}{1 - Pr\,(r_i \leq k)} = \alpha_k \mid k = \{1, 2, \ldots, K-1\}, \tag{3}$$

where $Pr\,(r_i \leq k)$ is the cummulative probability for a response value k. Notice that the cumulative logit of the highest response value, e.g., 7, is infinity. Therefore, for K possible response values, there are $K - 1\alpha_k$ intercepts in $\hat{\alpha}$. Each intercept α_k requires a prior. When the priors cannot be assigned based on prior scientific knowledge, the standard practice is to assign regularising priors to these parameters [61]. Given that the intercepts α_k are expressed on a logarithmic scale, Normal $(0, 1)$ is considered a suitable prior [61].

In Equation (2), the ordered-logit distribution embeds the link function in the likelihood function by incorporating the linear model ϕ as follows:

$$\log \frac{Pr\,(r_i \leq k)}{1 - Pr\,(r_i \leq k)} = \alpha_k - \phi_i. \tag{4}$$

Here, ϕ models the associations between predictor variables and the outcome R. By incorporating ϕ this way, the model estimates the locations of the ordered intercepts α_k, while other coefficients estimate a constant shift in the distribution's location linked to changes in predictor values [81]. These other coefficients include an effect for each Likert-item, $\alpha_{\mathrm{q}}\,[Q]$, and an effect for each participant in the study, $\alpha_{\mathrm{p}}\,[P]$. Noticeably, only one global mean parameter, $\widetilde{\alpha}$, is necessary for these effects; there is no need for separate mean parameters because they are both added in ϕ.

Next in Equation (2) are the σ parameters, which estimate variation across clusters in the data. Each cluster variable needs its own SD parameter (σ_{q} and σ_{p}) to adapt pooling across units, whether Likert-items or participants. Pooling here means using information from each Likert-item and each participant to improve estimates for others, utilising adaptive regularising priors for better cluster feature estimates [61]. Following standard practice, we use weakly regularising half-normal priors for σ^{q} and σ^{p} as these priors are preferred for non-linear models with logit and log links [61].

Using Equation (2), we independently analysed the Likert questionnaires we used in our study. For this, we adjusted their expressions for ϕ accordingly. In this context, the generalised representation is as follows:

$$\phi_i = \alpha_{\mathrm{q}_X}\,[Q] + \alpha_{\mathrm{p}_X}\,[P] + \alpha_{\mathrm{UI}_X}\,[\mathrm{UI}_i] + \alpha_{\mathrm{G}_X}\,[G_i], \tag{5}$$

where $\alpha_{\mathrm{UI}_X}\,[\mathrm{UI}]$, $\alpha_{\mathrm{q}_X}\,[Q]$ and $\alpha_{\mathrm{p}_X}\,[P]$ denote the UI, Likert-item and participant effects, respectively, for a particular questionnaire denoted by 'X.' The index variables $\mathrm{UI} = \{1, 2\}$ and $G \in \{1, 2\}$ remain consistent across all questionnaires. $\mathrm{UI} = 2$ signifies participants using an AR HMD UI to complete the task, while $\mathrm{UI} = 1$ represents the other UI condition. G has one entry per experimental group. $G = 2$ indicates participants belonging to group $G_{\mathrm{AR} \rightarrow \mathrm{HH}}$, and $G = 1$ signifies participation in group $G_{\mathrm{HH} \rightarrow \mathrm{AR}}$ (refer to Section 4.2). Lastly, the parameter $\alpha_{\mathrm{G}_X}\,[G]$ captures the effect of the group on the questionnaire 'X.'

To analyse how the level of trust assigned by the operator to the legged manipulator transfers across experimental conditions ($\alpha_{\mathrm{D}}, \alpha_{\mathrm{C}}, \alpha_{\mathrm{PE}}$) and how HH and AR HMD UIs affect such transfer

(α_{UI}), we extended Equation (5) as follows:

$$\phi_i = \alpha_{\mathrm{q}}[Q_i] + \alpha_{\mathrm{p}}[P_i] + \alpha_{\mathrm{UI}}[\mathrm{UI}_i] + \alpha_{\mathrm{D}}[D_i] + \alpha_{\mathrm{C}}[C_i] + \alpha_{\mathrm{PE}}[PE_i] + \alpha_{\mathrm{G}}[G_i]. \tag{6}$$

Here, α_{G}, α_{q}, α_{p}, UI and G remain as in Equation (5), and $D = \{1, 2\}$, $C = \{1, 2\}$ and $PE = \{1, 2\}$ are additional index variables. D represents the difficulty of the task. $D = 2$ for navigation with obstacles and $D = 1$ for obstacle-free navigation. C represents the category of the task. $C = 2$ for the manipulation task and $C = 1$ for the navigation tasks. PE represents the performance during the manipulation task. $PE = 2$ when the participants succeeded and $PE = 1$ when they failed to grasp the object from the floor (see Section 4.2).

5.3 Statistical Analysis

We fitted the CLMM from Sections 5.1 and 5.2 using the *ulam* interface in R [61], with one and four Markov chains, respectively. Multiple chains are used to verify that the joint posterior sampling worked correctly. By default, each chain consists of 1,000 iterations, split equally between warmup and sampling [61]. The *posterior* of relevant parameters was reported for all models using a 95% **Highest Density Interval (HDI)**. To aid visualisation, we present a single plot for fixed ($\widetilde{\alpha}$, σ_{q}), parameter (α_{UI} and α_G) and question effects ($\alpha_{\mathrm{q}}[Q]$). For the robust estimation model, pair plots were provided, depicting correlations between parameters in the posterior [61].

In addition to posterior and pair plots, we include plots displaying the distribution of differences between categories for each parameter in the linear model. This difference distribution, known as a *contrast* [61], is visualised in each contrast plot, showing the mean value, the percentage of positive and negative contrasts, the 95% HDI, the **Region of Practical Equivalence (ROPE)** (with two smaller dotted lines) and the percentages within and outside the ROPE. The ROPE represents a range of parameter values equivalent to the null value for practical purposes [49]. To make a dichotomous decision about the null value, we use the rule that if the 95% HDI falls entirely outside the ROPE, we reject the null value; if it falls entirely inside, we accept the null value for practical purposes. Otherwise, we remain undecided [49].

The default ROPE range suggested by [48] is −0.1 to 0.1 for a standardised parameter. A value of 0.2 conventionally represents a small effect size in psychological research [46], and the ROPE limits are set at half that size. For logistic models with parameters expressed in log-odds ratio, the range is −0.18 to 0.18 [58]. Despite this, Bayesian data analysis emphasises the continuous posterior distribution over dichotomous decision-making, and the focus is on the full information provided by the posterior.

6 Results

We show the demographics of the participants, a summary of their reported experience with VR, AR, computer games and robots and the UI they selected for the manipulation stage in Table 1. We divided the demographics into two groups to differentiate the group of participants from our previous smaller user study [13], from the group of participants recruited after. The former group interacted with the AR HMD UI using hand gestures to manipulate the GoTo marker, while the latter group interacted with the UI using the added cognitive offloading feature (see Section 3.2).

We recruited 27 participants (9 female, median age: 26 years, age range: 21–37 years) for our study. From the 27 participants, 24 were right-handed, 16 had experience with AR, 19 had experience with VR, all had experience with computer games and 23 had experience with robots. All participants were healthy without prior history of neurological impairments according to self-report via the safety screening questionnaire. Our user study involved 135 trials across five experimental conditions (Section 4.2) and produced 405 completed questionnaires.

Table 1. Demographics of the Participants, Their Reported Experience with Other Technologies and the UI They Selected to Complete the Manipulation Stage ($N = 27$)

		Hand Gestures		Cognitive Offloading	
		n	%	*n*	%
Gender	Female	7	41	2	20
	Male	10	59	8	80
Age groups	18–24	5	29.4	3	30
	25–34	10	58.8	5	50
	35–44	2	11.8	2	20
Experience	VR	13	76.5	3	30
	AR	13	76.5	6	60
	Computer games	17	100	10	100
	Robots	16	94.1	7	70
UI for stage (5)	HH	8	47	2	20
	AR	9	53	8	80

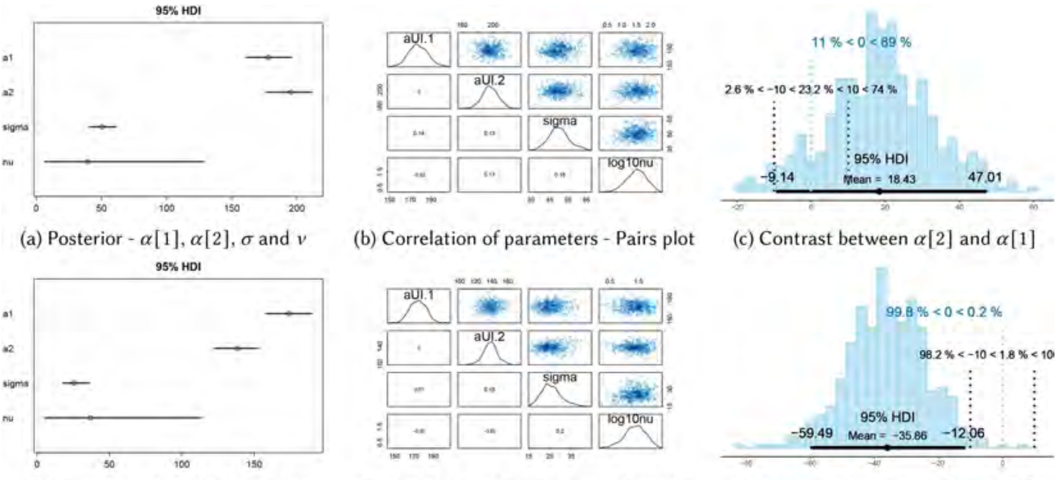
Table 2. Posterior Results (α_{UI}) for Objective and Subjective Metrics in Our User Study, Aggregated and with Cognitive Offloading Samples, across HH and AR UIs

Metric		Aggregated		Cognitive Offloading	
		HH	AR	HH	AR
Time (s) (mean, 95% HDI)	Obstacle-free	177.8, 160.7–196.3	196.2, 175.1–215.7	174.3, 156.4–188.1	138.5, 121.1–157.4
	With Obstacles	178.7, 160.1–200.3	208.6, 183.6–239.3	173.3, 152.2–196.3	143.9, 121.5–165.5
NASA-TLX (mean, SD)		−0.32, 0.64	0.06, 0.64	−0.03, 0.69	−0.44, 0.69
TAM (mean, SD)		0.26, 0.6	0.34, 0.65	−0.12, 0.63	0.63, 0.63
Fluency (mean, SD)		0.06, 0.69	0.24, 0.69	−0.06, 0.65	0.42, 0.66
H-SUS (mean, SD)		0.29, 0.66	0.26, 0.66	0.07, 0.66	0.55, 0.66
ARI (mean, SD)		−0.34, 0.65	0.88, 0.65	−0.35, 0.64	0.87, 0.65

Table 2 presents the α_{UI} posterior results for both objective and subjective metrics obtained from our user study, summarising the effect of the UI on each of these metrics. Time-related metrics include mean and 95% HDI, while results for the NASA-TLX, TAM, fluency, H-SUS and ARI questionnaires are reported in mean and SD on the logit scale. The data are segmented by aggregation status (aggregated and with cognitive offloading) and further categorised for HH and AR UIs. The results for the trust questionnaire are presented separately in Figure 15.

6.1 Objective Measures

In Figure 8, we report the results for the time to complete the navigation—obstacle-free stage with both UIs, and in Figure 9 the results for the navigation—with obstacles stage. Figures 8(a) and 9(a) show the posterior on $\alpha[1]$, $\alpha[2]$, σ and ν for the aggregated sample (see Section 5.1.1 for definitions). Table 2 summarises the most relevant results from the posterior. The contrast between $\alpha[2]$ and $\alpha[1]$ in Figure 8(c) indicates that the difference of means for the navigation—obstacle-free stage is 18.4 seconds, with 11% of the distribution being negative. Similarly, the contrast in Figure 9(c) indicates a positive difference of means for the navigation—with obstacles stage of about 29.8 seconds, with only 4% of probability for the distribution being negative. These results suggest that,



(a) Posterior - $\alpha[1]$, $\alpha[2]$, σ and ν (b) Correlation of parameters - Pairs plot (c) Contrast between $\alpha[2]$ and $\alpha[1]$

(d) Posterior - $\alpha[1]$, $\alpha[2]$, σ and ν (e) Correlation of parameters - Pairs plot (f) Contrast between $\alpha[2]$ and $\alpha[1]$

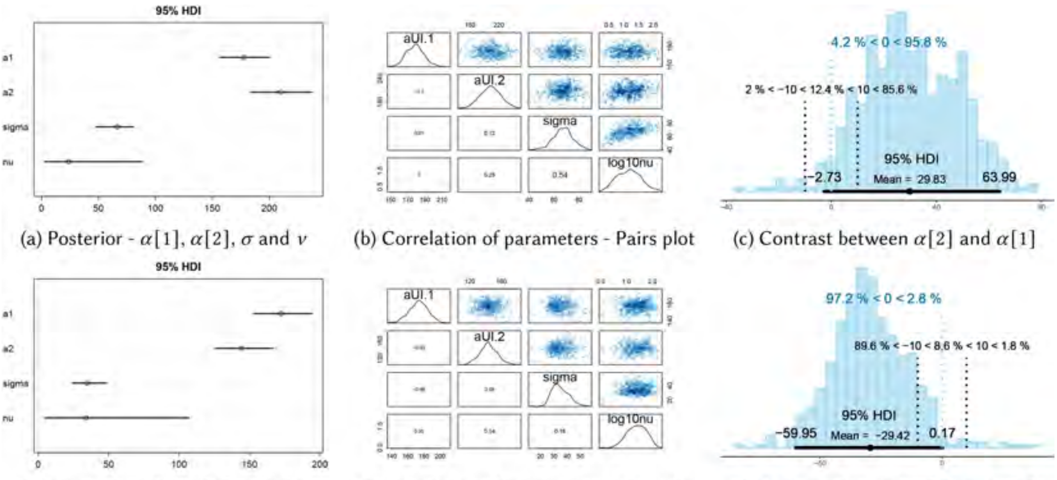
Fig. 8. Results for the time to complete the navigation—obstacle-free stage. (a)–(c) Aggregated results (N = 27). (d)–(f) Cognitive offloading results (N = 10). In (a) and (d), we show the marginal posterior distributions on the α, σ and ν parameters. (b) and (e) Corresponding pair plots for each case (ν's posterior distribution on a logarithmic scale). (c) and (f) The posterior distribution of the difference of means $\alpha[2] - \alpha[1]$. The ROPE was defined from -10 to 10 seconds in this case. These results suggest that, for the aggregated sample, participants took longer to complete the navigation obstacle-free stage when using the AR UI. Nonetheless, when focusing only on the participants that completed the navigation stages using the cognitive offloading feature in AR, this effect was drastically reverted.

for the aggregated sample, participants took longer to complete the navigation stages when using the AR UI.

However, when focusing only on the participants that completed the navigation stages using the cognitive offloading feature in AR, this effect is drastically reverted. From Figure 8(f), notice that the mean contrast is negative and that its 95% HDI does not overlap a ROPE we defined between ±10 seconds. For the navigation — with obstacles stage, this difference is also negative (Figure 9(f)). However, although the difference in means in the effect of each UI was also reverted in this case, the 95% HDI overlaps the ROPE. Nonetheless, such results show that participants were faster using the AR UI with the cognitive offloading feature than with the HH controller.

In the context of the ν parameters, the posterior distribution for the navigation—obstacle-free stage yielded a mean value of 43. Due to its exponential nature (refer to Equation (1)), the distribution of ν is highly skewed in its original scale (see Figure 8(a)). To better visualize this, we present the distribution on a logarithmic scale in Figure 8(b), where the mean is 1.51, surpassing its prior mean of 30 (1.47 on the log10 scale). This indicates a preference for larger ν values (>30), implying a nearly normal distribution and no outliers in the data.

Contrastingly, for the navigation with obstacles stage, shown in Figure 9(a), the mean ν value is 1.18 on the log scale, notably lower than its prior mean of 1.47. This suggests a better fit for smaller (< 30) ν values due to the presence of outliers, likely stemming from the increased difficulty of this stage. Further support is provided by Figure 9(b), illustrating pairwise plots of the parameters in Equation (1)). Notably, σ values exhibit positive correlation with $\log_{10}(\nu)$ values, indicating that a normal distribution with a larger ν corresponds to a wider distribution with a larger σ. This correlation signals the presence of outliers in the data [46].



(a) Posterior - $\alpha[1]$, $\alpha[2]$, σ and ν

(b) Correlation of parameters - Pairs plot

(c) Contrast between $\alpha[2]$ and $\alpha[1]$

(d) Posterior - $\alpha[1]$, $\alpha[2]$, σ and ν

(e) Correlation of parameters - Pairs plot

(f) Contrast between $\alpha[2]$ and $\alpha[1]$

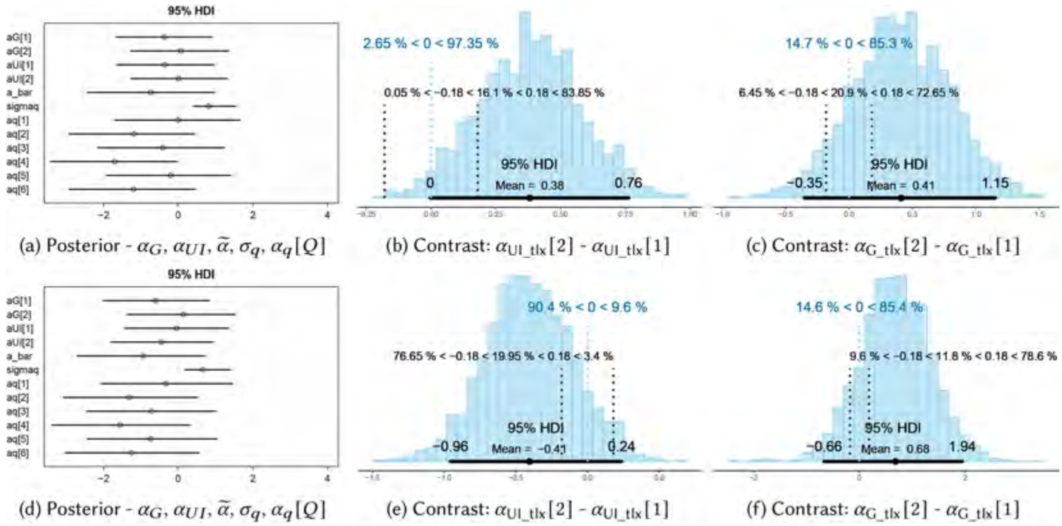
Fig. 9. Results for the time to complete the navigation—with obstacles stage. (a)–(c) Aggregated results (N = 27). (d)–(f) Cognitive offloading results (N = 10). In (a) and (d), we show the marginal posterior distributions on the α, σ and ν parameters. (b) and (e) Corresponding pairs plots for each case (ν's posterior distribution on a logarithmic scale). (c) and (f) The posterior distribution of the difference of means $\alpha[2] - \alpha[1]$. The ROPE was defined from −10 to 10 seconds in this case. These results suggest that, for the aggregated sample, participants took longer to complete the navigation with obstacles stage when using the AR UI. Nonetheless, when focusing only on the participants that completed the navigation stages using the cognitive offloading feature in AR, this effect was drastically reverted. Furthermore, incorporating this feature helped decrease the variability in the data that was derived from the task's increased difficulty.

When focusing solely on participants who completed the navigation stages using the cognitive offloading feature in AR, this correlation diminishes to 0.16 (see Figure 9(e)), with a mean ν value around 1.38 (on the log10 scale) for both stages (refer to Figures 8(d) and 9(d)). These outcomes suggest that integrating the cognitive offloading feature into the AR UI contributed to reducing variability associated with task difficulty.

In an NHST framework, our data would be submitted to a paired t-test. Using R's built-in t-test function, we obtained $p = 0.09735$ ($t = -1.7198, \text{df} = 26$) as a result of such test on the data for the navigation—obstacle-free stage. As the p-value is greater than 0.05, the conventional decision would be not to reject the null hypothesis. This conclusion is in line with Figure 8(c), where the 95% HDI overlaps the ROPE, which we conservatively set to ±10 seconds. However, for the navigation—with obstacles stage, we obtained $p = 0.01178$ ($t = -2.709, \text{df} = 26$). As the p-value is lower than 0.05 in this case, we would reject the null hypothesis. This will conflict with the Bayesian analysis from Figure 9(c), where the 95% HDI would not fall entirely outside the ROPE, regardless of its size. The reason for this conflict is that the t-test assumes normality in the distributions. Therefore, its estimate of the group variances is large when the data has outliers, which we demonstrated is the case for these data.

6.2 Subjective Measures

6.2.1 Cognitive Workload. In Figure 10, we report the results for TLX questionnaire. All the participants completed the questionnaire correctly. Thus, $P = \{1, 2, 3, ..., 27\}$. As the questionnaire is composed of six scales, $Q = \{1, 2, 3, ..., 6\}$. We show the posterior distribution for the aggregated results in Figure 10(a). MD ($\alpha_q[1]$, mean = −0.05, SD = 0.89) and effort ($\alpha_q[5]$, mean = −0.25,



(a) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (b) Contrast: $\alpha_{\text{UI\_tlx}}[2]$ - $\alpha_{\text{UI\_tlx}}[1]$ (c) Contrast: $\alpha_{\text{G\_tlx}}[2]$ - $\alpha_{\text{G\_tlx}}[1]$

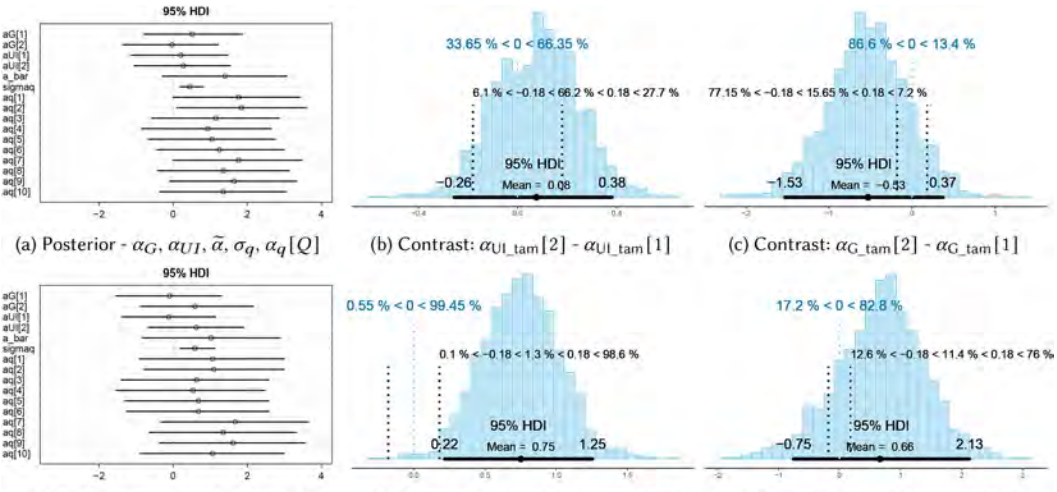
(d) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (e) Contrast: $\alpha_{\text{UI\_tlx}}[2]$ - $\alpha_{\text{UI\_tlx}}[1]$ (f) Contrast: $\alpha_{\text{G\_tlx}}[2]$ - $\alpha_{\text{G\_tlx}}[1]$

Fig. 10. Results for the NASA-TLX questionnaire. (a)–(c) Aggregated results (N = 27). (d)–(f) Cognitive offloading results (N = 10). (a) and (d) Posterior distributions for the parameters in the model ($\alpha_{\text{UI\_tlx}}[UI]$, $\alpha_{\text{G\_tlx}}[G]$, $\widetilde{\alpha}$, σ_q and $\alpha_q[Q]$). In line with the results from our previous smaller study, MD ($\alpha_q[1]$) and effort ($\alpha_q[5]$) are the main contributors to the TLX score, closely followed by TD ($\alpha_q[3]$). Plots (b), (c) and (e), (f) show the posterior distribution of the difference $\alpha_{\text{UI\_tlx}}[2] - \alpha_{\text{UI\_tlx}}[1]$ and $\alpha_{G\_\text{tam}}[2] - \alpha_{G\_\text{tam}}[1]$, respectively. The results presented here suggest that, for the aggregated sample, using the AR UI resulted in a higher cognitive workload. Nonetheless, when focusing only on the participants that interacted with the AR UI using the cognitive offloading feature, this effect is reverted. See Section 6.2.1 for the complete analysis.

SD = 0.89) were the main contributors to higher TLX scores. These are followed by TD ($\alpha_q[3]$, mean = −0.46, SD = 0.88), PD ($\alpha_q[2]$, mean = −1.22, SD = 0.90), frustration ($\alpha_q[6]$, mean = −1.24, SD = 0.89) and performance ($\alpha_q[4]$, mean = −1.74, SD = 0.92).

Higher TLX scores are associated with higher cognitive workload [29]. Therefore, from Table 2, the fact that the mean value for $\alpha_{\text{UI}}[2]$ is larger than for $\alpha_{\text{UI}}[1]$ suggest higher cognitive workload associated with using the AR UI for the aggregated sample. We can visualise this in Figure 10(b), where the contrast between UIs is positive within the 95% HDI and only 2.65% probability of the difference is negative. Such difference would imply a change in the mean response value for a scale in the questionnaire from 10.9 when using AR to 9.6 when using the HH controller, i.e., a difference of over a point per scale in the questionnaire.

Nonetheless, when focusing only on the participants that interacted with the AR UI using the cognitive offloading feature (Figure 10(d)–(f)), this effect is reverted (see Figure 10(e)). Using the AR UI results in a lower TLX score than when using the HH controller (refer to Table 2). Furthermore, the mean effect value for the dimensions of MD ($\alpha_q[1]$, mean = −0.34, SD = 0.98), effort ($\alpha_q[5]$, mean = −0.75, SD = 0.97) and TD ($\alpha_q[3]$, mean = −0.73, SD = 0.95) are lower for these participants, while PD ($\alpha_q[2]$, mean = −1.33, SD = 0.96), frustration ($\alpha_q[6]$, mean = −1.26, SD = 0.96) and performance ($\alpha_q[4]$, mean = −1.59, SD = 0.97) remain relatively similar to the aggregated results (see Figure 10(d)). Thus, the expected response value for a scale in the questionnaire changes its mean value from 9.2 when using AR to 10.6 when using the HH controller. Although the difference in means in the effect of each UI decreased to −0.41 (logit scale), the 95% HDI overlaps the ROPE spanning ±0.18.



(a) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (b) Contrast: $\alpha_{\text{UI\_tam}}[2]$ - $\alpha_{\text{UI\_tam}}[1]$ (c) Contrast: $\alpha_{\text{G\_tam}}[2]$ - $\alpha_{\text{G\_tam}}[1]$

(d) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (e) Contrast: $\alpha_{\text{UI\_tam}}[2]$ - $\alpha_{\text{UI\_tam}}[1]$ (f) Contrast: $\alpha_{\text{G\_tam}}[2]$ - $\alpha_{\text{G\_tam}}[1]$

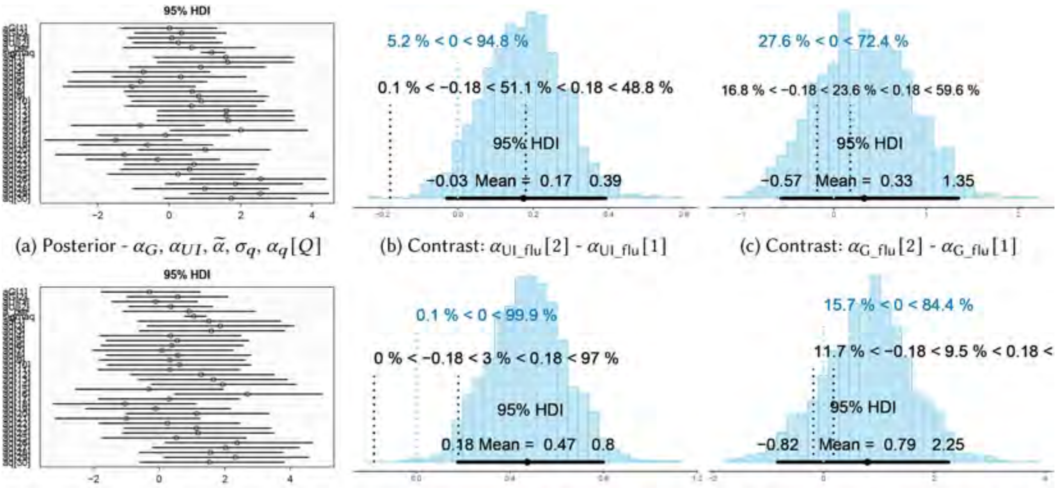
Fig. 11. Results for the TAM questionnaire. (a)–(c) Aggregated results (N = 27). (d)–(f) Cognitive offloading results (N = 10). (a) and (d) Posterior distributions for the parameters in the model ($\alpha_{\text{UI\_tam}}[\text{UI}]$ and $\alpha_{\text{G\_tam}}[G]$, $\widetilde{\alpha}$, σ_q and $\alpha_q[Q]$). Notice how all three dimensions of the questionnaire, i.e., intention of use, perceived usefulness and perceived ease of use, had a positive effect on the response values. Plots (b), (c) and (e), (f) show the posterior distribution of the difference $\alpha_{\text{UI\_tam}}[2] - \alpha_{\text{UI\_tam}}[1]$ and $\alpha_{G\_\text{tam}}[2] - \alpha_{G\_\text{tam}}[1]$, respectively. The results presented here suggest that, for the aggregated sample, both UIs have comparable TAM scores. Nonetheless, when focusing only on the participants that interacted with the AR UI using the cognitive offloading feature, the results suggest that the use of the AR UI is followed by significantly higher TAM scores. See Section 6.2.2 for the complete analysis.

Regarding the effect of the group on the TLX score, in Figure 10(c) we can see how the contrast between groups is mostly positive (mean = 0.41, SD = 0.39), meaning that starting the experiment with the AR UI contributed to higher response values than those who started with the HH controller. Although higher in mean value, this effect is similar even if we only consider the participants who used the cognitive offloading feature (mean = 0.68, SD = 0.67, see Figure 10(f)).

In an NHST framework, the data for the overall population would be submitted to a Wilcoxon signed-rank test. Using R's built-in t-test function, we obtained $p = 0.02785$ $(V = 3879.5)$ as the results for the test. As the p-value is lower than 0.05 in this case, we would reject the null hypothesis, which is consistent with the results from our previous smaller study. Nonetheless, this conflicts with the Bayesian analysis from Figure 10(b), where the 95% HDI does not fall entirely outside the ROPE spanning ±0.18. The reason is that tests such as the Wilcoxon signed-rank test dichotomise a rated feature to compare the highest- and lowest-rated items. Such dichotomisation results in statistical power and precision loss relative to continuous alternatives [81].

6.2.2 *Technology Acceptance.*

In Figure 11, we report the results for TAM questionnaire. All the participants completed the questionnaire correctly. Thus, $P = \{1, 2, 3, ..., 27\}$. This questionnaire is composed of 10 scales. Therefore, $Q = \{1, 2, 3, ..., 10\}$. We show the posterior for the fixed ($\widetilde{\alpha}$, σ_q) and question ($\alpha_q[Q]$) effects in Figure 11(a) for the overall sample, and in Figure 11(d) for the participants that interacted with the AR UI using the cognitive offloading feature. From Figure 11(a) to (d), notice how all three dimensions from the questionnaire, i.e., intention of use ($\alpha_q[1–2]$), perceived usefulness ($\alpha_q[3–6]$) and perceived ease of use ($\alpha_q[7–10]$), had a mean positive effect on the response values.



(a) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (b) Contrast: $\alpha_{\text{UI\_flu}}[2]$ - $\alpha_{\text{UI\_flu}}[1]$ (c) Contrast: $\alpha_{\text{G\_flu}}[2]$ - $\alpha_{\text{G\_flu}}[1]$

(d) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (e) Contrast: $\alpha_{\text{UI\_flu}}[2]$ - $\alpha_{\text{UI\_flu}}[1]$ (f) Contrast: $\alpha_{\text{G\_flu}}[2]$ - $\alpha_{\text{G\_flu}}[1]$

Fig. 12. Results for the fluency questionnaire. (a)–(c) Aggregated results (N = 20). (d)–(f) Cognitive offloading results (N = 9). (a) and (d) Posterior distributions for the parameters in the model ($\alpha_{\text{UI\_fluency}}[UI]$ and $\alpha_{\text{G\_fluency}}[G]$, $\widetilde{\alpha}$, σ_q and $\alpha_q[Q]$). Most dimensions from this questionnaire, i.e., human–robot fluency ($\alpha_q[1–3]$), trust in robot ($\alpha_q[8–9]$), positive teammate traits ($\alpha_q[10–12]$), improvement ($\alpha_q[13–15]$) and individual measures ($\alpha_q[27–30]$), had a positive effect in the response values. The exceptions were robot relative contribution ($\alpha_q[4–7]$), and working alliance for H–R teams ($\alpha_q[16–26]$). Plots (b), (c) and (e), (f) show the posterior distribution of the difference $\alpha_{\text{UI\_flu}}[2] - \alpha_{\text{UI\_flu}}[1]$ and $\alpha_{\text{G\_flu}}[2] - \alpha_{\text{G\_flu}}[1]$, respectively. These results suggest that, for the aggregated sample, the scores are higher for the AR UI. Moreover, when focusing on the participants that interacted with the AR UI using the cognitive offloading feature, the results suggest that the use of the AR UI results in significantly higher fluency scores. See Section 6.2.3 for the complete analysis.

From Table 2, notice that the mean values for $\alpha_{\text{UI}}[2]$ and $\alpha_{\text{UI}}[1]$ are similar for the aggregated sample. This result suggests comparable TAM scores for both UIs, which is in line with our previous pilot study. In fact, in Figure 11(b), we can see that the contrast between $\alpha_{\text{UI\_tam}}[2]$ and $\alpha_{\text{UI\_tam}}[1]$ has a mean value close to zero. Nonetheless, over two-thirds of the distribution is positive, suggesting higher response values associated with the AR UI.

When focusing only on the participants that interacted with the AR UI using the cognitive offloading feature, Figure 11(e) indicates that the difference of means in the effect of each UI is about 0.75 (logit scale), with the 95% HDI excluding the ROPE spanning ±0.18. This suggests that the use of the AR UI results in significantly higher TAM scores.

Regarding the effect of the group on the TAM questionnaire, in Figure 11(c), we can see how the contrast between groups is mostly negative, meaning that participants who started with the AR UI reported relatively lower response values than those who started with the other UI. Nonetheless, when focusing only on the participants that interacted with the AR UI using the cognitive offloading feature, this effect is reverted (see Figure 11(f)).

6.2.3 Fluency. In Figure 12, we report the results for the fluency questionnaire. We removed the data from seven participants for not completing the questionnaire correctly, i.e., $P = \{1, 2, 3, ..., 20\}$). One of these participants was part of the group that interacted with the UI using the cognitive offloading feature. This questionnaire is composed of 30 scales. Therefore, $Q = \{1, 2, 3, ..., 30\}$. We show the posterior for the fixed ($\widetilde{\alpha}$, σ_q) and question ($\alpha_q[Q]$) effects in Figure 12(a) for the overall sample, and in Figure 12(d) for the participants that interacted with the AR UI using cognitive offloading.

From Figure 12(a) to (d), most dimensions from this questionnaire, i.e., human–robot fluency ($\alpha_q[1–3]$), trust in robot ($\alpha_q[8–9]$), positive teammate traits ($\alpha_q[10–12]$), improvement ($\alpha_q[13–15]$) and individual measures ($\alpha_q[27–30]$), had a mean positive effect in the response values. The exceptions were robot relative contribution ($\alpha_q[4–7]$) and working alliance for H–R teams ($\alpha_q[16–26]$).

From Table 2, notice that the mean value for $\alpha_{\text{UI}}[2]$ is greater than for $\alpha_{\text{UI}}[1]$ for the aggregated sample. This suggests higher fluency scores for the AR UI. Indeed, from Figure 12(b), we can see that the contrast between $\alpha_{\text{UI\_fluency}}[2]$ and $\alpha_{\text{UI\_fluency}}[1]$ is mostly positive, i.e., only 5.2% of the distribution is negative. Nonetheless, the effect size for this difference is relatively small for having practical implications. Such difference would imply a change in the mean response value for a scale in the questionnaire from 3.24 when using the AR UI to 3.12 when using the HH controller. This is in line with our previous smaller study.

When focusing on the participants that interacted with the AR UI using the cognitive offloading feature, Figure 12(e) indicates that the difference of means in the effect of each UI is 0.47 (logit scale), with 99.9% of the contrast between them being positive, and the 95% HDI excluding the ROPE spanning ±0.18. This suggests that using the AR UI results in significantly higher fluency scores.

Regarding the effect of the group on the fluency questionnaire, in Figure 12(c), we can see how the contrast between groups is mostly positive, meaning that participants who started with the AR UI reported relatively higher response values than those who started with the other UI. When focusing only on the participants that interacted with the AR UI using the cognitive offloading feature, Figure 12(f), this effect remains similar.

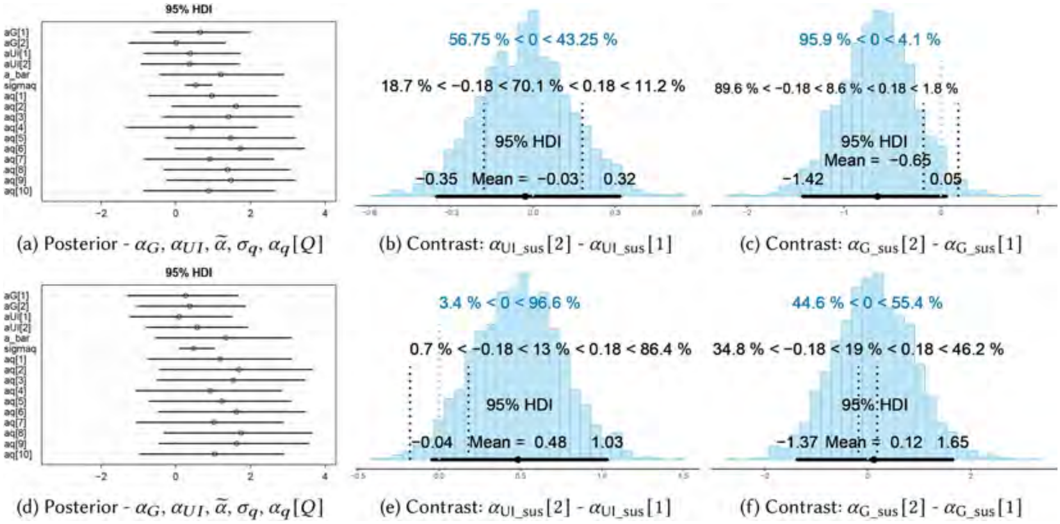
6.2.4 *System Usability.*

In Figure 13, we report the results for the H-SUS questionnaire. We removed the data from one participant for not completing the questionnaire correctly. Thus, $P = \{1, 2, 3, ..., 26\}$. This questionnaire is composed of 10 scales. Therefore, $Q = \{1, 2, 3, ..., 10\}$. We show the posterior for the fix ($\widetilde{\alpha}$, σ_q) and question ($\alpha_q[Q]$) effects in Figure 13(a) for the overall population, and in Figure 13(d) for the participants that interacted with the AR UI using the cognitive offloading feature, respectively. From Figure 13(a) to (d), notice that all of the scales in the questionnaire had a mean positive effect in the response values ($\alpha_q[1–10]$).

From Table 2, notice that the mean values for $\alpha_{\text{UI}}[2]$ and for $\alpha_{\text{UI}}[1]$ are similar for the aggregated sample. This result suggests comparable H-SUS scores for both UIs, which is in line with our previous pilot study. In Figure 13(b), we can see that the contrast between $\alpha_{\text{UI\_sus}}[2]$ and $\alpha_{\text{UI\_su}}[1]$ has a mean value close to zero, with the contrast distribution having a similarly positive and negative probability mass.

When focusing on the participants that interacted with the AR UI using the cognitive offloading feature, Figure 13(e) indicates that the difference of means in the effect of each UI is 0.48 (logit scale), with 96.6% of the contrast between them being positive. However, its 95% HDI overlaps the ROPE spanning ±0.18. Nonetheless, this result suggests that using the AR UI results in higher H-SUS scores.

Regarding the effect of the group on the H-SUS questionnaire, in Figure 13(c), we can see how the contrast between groups is mostly negative, meaning that participants who started with the AR UI reported relatively lower response values than those who started with the other UI. When focusing only on the participants that interacted with the AR UI using cognitive offloading, Figure 13(f), the group contrast has a mean value closer to zero and a similarly positive and negative probability mass.



(a) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (b) Contrast: $\alpha_{\text{UI\_sus}}[2]$ - $\alpha_{\text{UI\_sus}}[1]$ (c) Contrast: $\alpha_{\text{G\_sus}}[2]$ - $\alpha_{\text{G\_sus}}[1]$

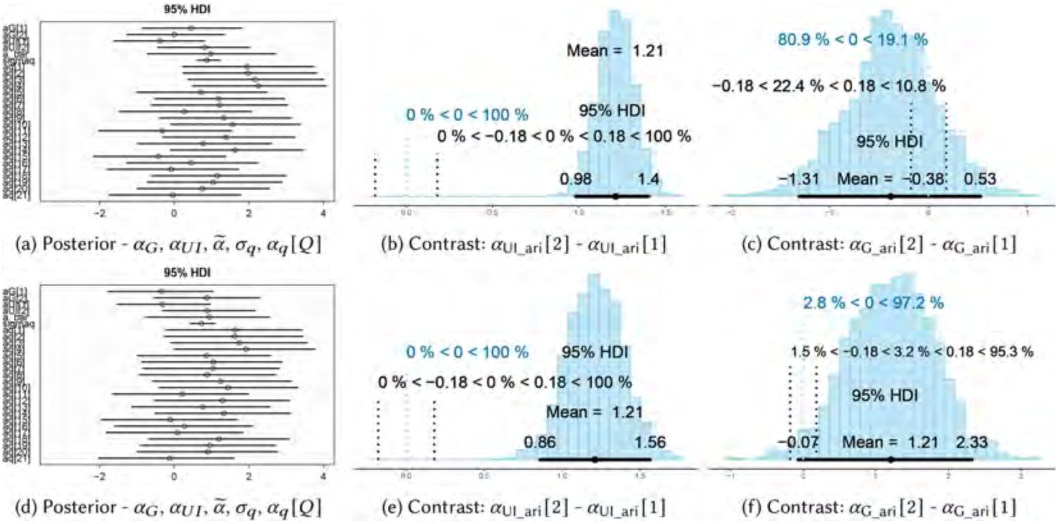
(d) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (e) Contrast: $\alpha_{\text{UI\_sus}}[2]$ - $\alpha_{\text{UI\_sus}}[1]$ (f) Contrast: $\alpha_{\text{G\_sus}}[2]$ - $\alpha_{\text{G\_sus}}[1]$

Fig. 13. Results for the H-SUS questionnaire. (a)–(c) Aggregated results ($N = 26$). (d)–(f) Cognitive offloading results ($N = 10$). (a) and (d) Posterior distributions for $\alpha_{\text{UI\_sus}}[UI]$, $\alpha_{\text{G\_sus}}[G]$, $\widetilde{\alpha}$, σ_q and $\alpha_q[Q]$. Overall, all of the scales in the questionnaire had a positive effect on the response values. Plots (b), (c) and (e), (f) show the posterior distribution of the difference $\alpha_{\text{UI\_sus}}[2] - \alpha_{\text{UI\_sus}}[1]$ and $\alpha_{G\text{\_sus}}[2] - \alpha_{G\text{\_sus}}[1]$, respectively. In line with our previous smaller study, the results presented here suggest that, for the aggregated sample, both UIs have comparable H-SUS scores. However, when focusing on the participants that interacted with the AR UI using the cognitive offloading feature, the results showed that the use of the AR UI results in higher usability scores. See Section 6.2.4 for the complete analysis.

6.2.5 Immersion. In Figure 14, we report the results for ARI questionnaire. All the participants completed the questionnaire correctly. Thus, $P = \{1, 2, 3, ..., 27\}$. This questionnaire is composed of 21 scales. Therefore, $Q = \{1, 2, 3, ..., 21\}$. We show the posterior for the fixed ($\widetilde{\alpha}$, σ_q) and question ($\alpha_q[Q]$) effects in Figure 14(a) for the overall sample, and in Figure 14(d) for the participants that interacted with the AR UI using the cognitive offloading feature.

From Figure 14(a), most dimensions from this questionnaire, i.e., 'interest' ($\alpha_q[1–4]$), 'usability' ($\alpha_q[5–8]$), 'emotional attachment' ($\alpha_q[9–11]$), 'focus of attention' ($\alpha_q[12–14]$) and 'flow' ($\alpha_q[19–21]$), had a mean positive effect in the response values. The exception was 'presence' ($\alpha_q[15–18]$), where half the scales had a mean positive effect and the other half a mean negative effect. This exception does not appear in Figure 14(d), where all the dimensions in the questionnaire had a mean positive effect on the response values.

From Table 2, notice that the mean value for $\alpha_{\text{UI}}[2]$ is substantially larger than for $\alpha_{\text{UI}}[1]$ for the aggregated sample. This result suggests much higher ARI scores associated with using the AR UI. This result is also in line with our previous smaller study. We can visualise the difference in Figure 14(b), where the contrast between $\alpha_{\text{UI\_ari}}[2]$ and $\alpha_{\text{UI\_ari}}[1]$ is reliably positive, with 100% probability for the difference being positive and completely outside the ROPE. Such difference would translate into a change in the mean response value for a scale in the questionnaire from 4.9 when using AR to 3.85 when using the HH controller, a difference of over a point in the Likert scale per question in the questionnaire. The result is similar when we only consider the participants that interacted with the AR UI using the cognitive offloading feature in Figure 14(e). The ARI scores associated with using the AR UI are also significantly higher.



(a) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (b) Contrast: $\alpha_{\text{UI\_ari}}[2]$ - $\alpha_{\text{UI\_ari}}[1]$ (c) Contrast: $\alpha_{\text{G\_ari}}[2]$ - $\alpha_{\text{G\_ari}}[1]$

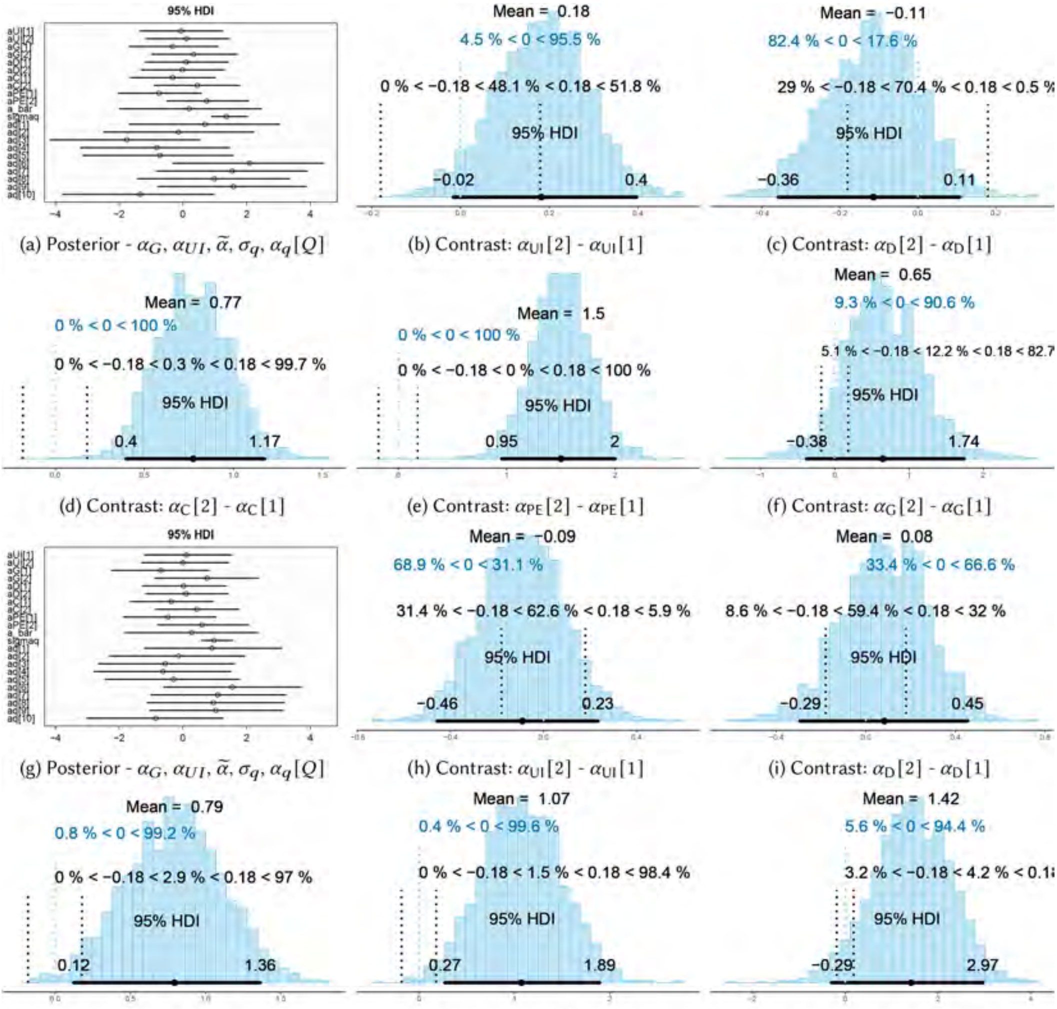
(d) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (e) Contrast: $\alpha_{\text{UI\_ari}}[2]$ - $\alpha_{\text{UI\_ari}}[1]$ (f) Contrast: $\alpha_{\text{G\_ari}}[2]$ - $\alpha_{\text{G\_ari}}[1]$

Fig. 14. Results for the ARI questionnaire. (a)–(c) Aggregated results (N = 26). (d)–(f) Cognitive offloading results (N = 10). Most dimensions from this questionnaire, i.e., interest, usability, emotional attachment, focus of attention and flow, had a positive effect on the response values. The exception was 'presence,' where half the scales had a positive effect, and the other half had a negative effect. The mean value for $\alpha_{\text{UI}}[2]$ is substantially larger than for $\alpha_{\text{UI}}[1]$, suggesting much higher ARI scores associated with using the AR UI. In line with our previous pilot study, the results presented here show that, for the aggregated sample, the ARI scores associated with using the AR UI are significantly higher. The results are similar when we only consider the participants that interacted with the AR UI using the cognitive offloading feature. The ARI scores associated with using the AR UI are also significantly higher. See Section 6.2.5 for the complete analysis.

Regarding the effect of the group on the ARI questionnaire, in Figure 14(c), we can see how the contrast between groups is mostly negative, meaning that participants who started with the AR UI report relatively lower response values than those who started with the other UI. Nonetheless, when focusing only on the participants that interacted with the AR UI using the cognitive offloading feature, this effect is drastically reverted (see Figure 14(f)).

6.2.6 Trust. In Figure 15, we report the results for the trust questionnaire. We removed the data from three participants for not completing the questionnaire correctly. One of these participants was part of the group that interacted with the UI using the cognitive offloading feature. Thus, $P = \{1, 2, 3, ..., 24\}$. This questionnaire is composed of 10 scales. Therefore, $Q = \{1, 2, 3, ..., 10\}$.

From Figure 15(a), notice that the scales $\alpha_q[3–4]$ ('I really wish I had a good way to monitor the decisions of the robot' and 'I would be comfortable allowing the robot to implement its decisions, even if I could not monitor it') in the IMOT dimension of the questionnaire had a negative mean contribution on the response values. Instead, $\alpha_q[1]$ ('If I had it my way, I would NOT let the robot have any influence over issues that are important to the task') had a positive mean contribution, while $\alpha_q[2]$ ('I would be comfortable giving the robot complete responsibility for the task') had a mean contribution close to zero. In contrast, most scales in the RIS dimension ($\alpha_q[5–10]$) had a mean positive effect on the participants' responses. The exceptions were $\alpha_q[5]$ ('I would rely on the robot without hesitation') and $\alpha_q[10]$ ('I would be comfortable allowing this robot to make all decisions'), for which their mean effect was negative. These effects remain similar even if we only consider the participants that interacted with the AR UI using the cognitive offloading feature



(a) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (b) Contrast: $\alpha_{\mathrm{UI}}[2]$ - $\alpha_{\mathrm{UI}}[1]$ (c) Contrast: $\alpha_{\mathrm{D}}[2]$ - $\alpha_{\mathrm{D}}[1]$

(d) Contrast: $\alpha_{\mathrm{C}}[2]$ - $\alpha_{\mathrm{C}}[1]$ (e) Contrast: $\alpha_{\mathrm{PE}}[2]$ - $\alpha_{\mathrm{PE}}[1]$ (f) Contrast: $\alpha_{\mathrm{G}}[2]$ - $\alpha_{\mathrm{G}}[1]$

(g) Posterior - α_G, α_{UI}, $\widetilde{\alpha}$, σ_q, $\alpha_q[Q]$ (h) Contrast: $\alpha_{\mathrm{UI}}[2]$ - $\alpha_{\mathrm{UI}}[1]$ (i) Contrast: $\alpha_{\mathrm{D}}[2]$ - $\alpha_{\mathrm{D}}[1]$

(j) Contrast: $\alpha_{\mathrm{C}}[2]$ - $\alpha_{\mathrm{C}}[1]$ (k) Contrast: $\alpha_{\mathrm{PE}}[2]$ - $\alpha_{\mathrm{PE}}[1]$ (l) Contrast: $\alpha_{\mathrm{G}}[2]$ - $\alpha_{\mathrm{G}}[1]$

Fig. 15. Results for the trust questionnaire. (a)–(f) Aggregated results (N = 24). (g)–(l) Cognitive offloading results (N = 9). The results presented in this section showed that the participants were more willing to rely than to delegate on the robot, i.e., the RIS dimension of the questionnaire contributed more positively to the scores than the IMOT dimension did. Regarding the effect of the UI, our results do not show a difference in the reported trust scores associated with each of them, regardless of whether we consider the aggregated results or the results for the participants using cognitive offloading. When focusing on task difficulty and task performance, our results showed to be in line with previous studies on trust transfer [67, 77]. Easier tasks resulted in higher trust scores while failing to complete a task resulted in significantly lower trust scores. Finally, regarding the effect of the group, starting the experiments with the AR UI resulted in comparatively higher trust scores. This result suggests that trust transfers more easily from the AR UI to the HH UI than in the opposite direction. See Section 6.2.6 for the complete analysis.

(see Figure 15(g)). These results suggest that, overall, the participants were willing to rely more than to delegate on the robot.

In addition, Figure 15(a) show the posterior distributions for all other parameters in the linear model. We show the contrast distributions for the aggregated sample from Figure 15(b) to (f). Starting from the difference in UIs, we can see that the contrast between $\alpha_{\mathrm{UI}}[2]$ and $\alpha_{\mathrm{UI}}[1]$ is mostly positive with a mean value of 0.18 (logit scale). However, the effect size for this difference is relatively small, and almost half of the distribution falls within the ROPE. This contrast would imply a change in the mean response value for a scale in the questionnaire from 4.16 when using the AR UI to 3.46 when using the HH controller. Although the mean value for the contrast between UIs lowers to −0.09 (see Figure 15(h)) when focusing on the participants that interacted with the AR UI using the cognitive offloading feature, a high percentage of the distribution (around 63%) still falls within the ROPE. Therefore, our results do not show a difference in the reported trust scores due to the UIs, regardless of whether we consider the aggregated results or the results for the participants using cognitive offloading.

Next, from Figure 15(c), we can see that the contrast for task difficulties is mostly negative with a mean value of −0.11 (logit scale). This is in line with previous work on trust transfer [77]. Nonetheless, in this work the effect size for this difference is relatively small, i.e., it would imply a change in the mean response value for a scale in the questionnaire from 4.13 when the task is easy to 4.03 when it is difficult. Moreover, around 70% of the distribution falls within the ROPE. Although the mean value for the contrast between difficulties increases to 0.08 for the participants that interacted with the AR UI using the cognitive offloading feature (see Figure 15(i)), a high percentage of the distribution, around 60%, still falls within the ROPE.

Continuing with the effect of the task category, in Figure 15(d), we can see that the contrast between $\alpha_{\mathrm{C}}[2]$ (manipulation) and $\alpha_{\mathrm{C}}[1]$ (navigation) is reliably positive with the 95% HDI completely outside the ROPE. A difference such as this one is in line with previous work on trust transfer [77] and would imply a change in the mean response value for a scale in the questionnaire from 4.43 for the manipulation category to 3.77 for the navigation category.

Regarding task performance, from Figure 15(e), we can see that the contrast between $\alpha_{\mathrm{PE}}[2]$ (success) and $\alpha_{\mathrm{PE}}[1]$ (failure) is reliably positive and completely outside the ROPE as well, thus also in line with previous work on trust transfer [77]. Furthermore, such contrast would imply a change in the mean response value for a scale in the questionnaire from 4.73 when the participants succeed in the task to 3.50 when they fail. The results for task category and task performance are similar when we focus on the participants that interacted with the AR UI using the cognitive offloading feature (see Figure 15(j) and (k)).

Lastly, regarding the effect of the group on the trust questionnaire, in Figure 15(f), we can see that the contrast between groups is mostly positive. This means that participants who started with the AR UI reported relatively higher response values than those who started with the other UI. This effect is similar when focusing on the participants that interacted with the AR UI using the cognitive offloading feature (see Figure 15(l)).

7 Discussion

7.1 Time Performance and Cognitive Workload

Time performance and cognitive workload are crucial metrics for the successful deployment of AR UIs in the type of applications that are typical for legged manipulators. However, when comparing our AR UI against the off-the-shelf HH UI on time performance, we found that, for the overall sample, the time to complete the task in either navigation condition was higher when participants used the AR UI. From our experience with the user study, we believe that this was derived from the use of hand gestures and the GoTo marker as main interaction method, which also had a negative impact on the cognitive workload reported by the participants. Moreover, from Table 1, our intuition is that the level of familiarity all participants had with game controllers, higher than their experience with AR and VR, also played a role.

From the analysis of the video recordings, we argue that the use of hand gestures was not intuitive for most participants, resulting in higher levels of MD and effort, as well as a longer time to complete the tasks. This might be due to problems with hand tacking, which in turn made participants doubt about how to apply the gestures. Aiming at improving these metrics, we modified the functionality of the AR UI by incorporating the use of our proposed cognitive offloading feature. This added feature drastically reduced the time required to complete both navigation tasks and the TLX scores associated with the use of the AR UI. When we focused on the participants that made use of the cognitive offloading feature with the AR UI, we showed that the TLX score was lower for this UI when compared with the HH UI. Furthermore, the participants were faster with the AR UI than they were when they used the HH UI. Thus, through our user study, we provided evidence to demonstrate that AR UIs can outpace HH-based control methods and reduce the cognitive requirements when designers include hands-free interactions and cognitive offloading principles into the UI.

7.2 Fluency, Technology Acceptance, System Usability and Immersion

Technology acceptance, fluency, system usability and immersion are important metrics to assess the quality of HRIs. The statistical analysis we performed in our previous smaller study [13] revealed that the differences in TAM, fluency and H-SUS responses reported by the participants were not significant. Thus, our conclusion was that both UIs performed just as well in these dimensions. Nonetheless, the results presented here showed that incorporating hands-free interactions and cognitive offloading principles into the AR UI led to much better outcomes on these metrics. The use of the AR UI together with our cognitive offloading feature resulted in higher usability scores and significantly higher fluency and TAM scores. These are encouraging results as they show that the cognitive offloading feature has positive effects beyond the dimensions it was intended to improve, i.e., time performance and cognitive workload. Regarding immersion, our results revealed that the response values for the ARI questionnaire associated with the AR UI are significantly higher than those associated with the HH UI, regardless of the main interaction method with the former, i.e., hand gestures or cognitive offloading. Derived from the participants' qualitative answers, we believe this is due to a combination of factors, of which the most important are the free use of the hands when using the HMD, as well as the ability to see the real environment without the need to divert their attention to the UI.

7.3 Trust

Human–robot trust is crucial for the successful collaboration between assistive robots and their users. In this study, we investigated how the level of trust assigned by an operator to a legged manipulator is maintained and transferred across tasks and environmental contexts. Furthermore, we evaluated how HH and AR HMD UIs affect such transfer. From the results, we found that the participants were more willing to rely than to delegate on the robot, i.e., the RIS dimension of the questionnaire contributed more positively to the scores than the IMOT dimension did. Our intuition, based on the participants' comments, is that this happened because the robot platform met or exceeded the participant's expectations, resulting in higher reliability, while the need from some participants to feel in control played against higher delegation scores.

Regarding the effect of the UI, our results do not show a difference in the reported trust scores associated with each of them. Nonetheless, trust is a multi-dimensional construct and in our study we focused on the dimensions of reliance and delegation. Trust is known to mediate the use and non-use of autonomous systems [76]. In fact, for the manipulation category, for which participants had the opportunity to choose their preferred UI, participants reported higher trust scores than those reported for the navigation category. Furthermore, we believe that the drastic change in the percentage of participants that selected the AR UI for completing the manipulation stage after incorporating the cognitive offloading feature (see Table 1), could be due to some other trust dimension not considered in our study.

Regarding the other factors considered in our analysis for trust, i.e., task difficulty, and task performance, our results showed to be in line with previous studies on trust transfer [67, 77]. Easier tasks resulted in higher trust scores, while failing to complete a task results in significantly lower trust scores. Finally, regarding the effect of the group, starting the experiments with the AR UI resulted in comparatively higher trust scores. This result suggest that trust transfers more easily from the AR UI to the HH UI than in the opposite direction.

7.4 Participants' Comments

Regarding the object grasping task, eight of the participants (47%) chose to complete this task using the AR UI. The reasons given for this were obtained from the post-experiment questionnaire and included enjoyment, ease of use, engagement, immersion, feeling focused, involved and interested in the activity, finding it more intuitive, precise, novel, accurate and less mentally demanding. Some participants explicitly mentioned finding their hands free for the secondary task convenient. The reasons for choosing the controller included performance, ease of use, reliability, sense of control, a more gentle learning curve and their level of familiarity with this kind of UI. From this last group of participants, five expressed regretting their decision due to the complexity associated with controlling a robot arm using a virtual joystick. This only happened once with the AR UI.

As we mentioned in the previous section, there was a drastic change in the percentage of participants that selected the AR UI for completing the manipulation stage after incorporating the cognitive offloading feature (see Table 1). Based on the participants' comments, the cognitive offloading feature made participants feel that the interaction with the robot during the manipulation task would be easier using the AR UI than using the HH UI. From the two participants in the group that used the cognitive offloading feature and selected the HH UI for the manipulation task, one mentioned regretting its decision, and both mentioned 'feeling more in control' as the main reason for choosing this UI. This is in contrast to the remaining participants, who instead mentioned the higher degree of autonomy shown by the robot with the AR UI as the main reason for their preference. In addition, some participants also explicitly mentioned finding their hands free for the secondary task convenient.

7.5 The Role of Body Pose in Cognitive Offloading for AR HMD UIs

Traditional HH controllers not only could prove unintuitive, discouraging the interaction with the robot partner; they divert the operator's attention into the UI, bringing additional safety concerns to potentially dangerous applications. AR UIs aim to enhance the interaction between users and assistive robots by leveraging the users' existing abilities. Moreover, AR-based UIs overlay digital information onto the real-world surroundings, allowing users to perceive and interact with the robot in a more contextually relevant manner. By doing so, these UIs facilitate more natural and potentially less demanding interactions, reducing cognitive workload compared to conventional computer systems. AR HMD UIs, in particular, provide a broader range of interaction methods, including voice commands, hand gestures, eye-gaze and head-gaze, thereby expanding accessibility to a more diverse user base [80].

However, the results from our study suggest that hand gestures as the main input method for an AR HMD UI may pose challenges and lead to higher MD and effort for most participants. In designing our AR HMD UI with body pose and voice commands as the primary input method, we incorporated insights from the cognitive offloading principle of 'Thinking with the Body,' a concept rooted in cognitive sciences that emphasises the active use of the body to alleviate cognitive demand [71]. From this, several advantages emerged:

- Reduced MD: Body pose, as compared to hand gestures, may inherently require less mental effort for users. Engaging in natural body movements could align with users' cognitive processes more seamlessly, potentially resulting in lower MD during interaction.
- Intuitive interaction: Leveraging body pose aligns with users' natural ways of moving and interacting with the environment. This intuitiveness contributes to a smoother cognitive integration of actions, facilitating a more natural and less mentally taxing interaction with the AR HMD UI.
- Enhanced reliability: Hand tracking issues associated with gestures could lead to uncertainty and increased cognitive effort. Body pose, encompassing a broader range of movements, might offer a more reliable and consistent means of input, reducing cognitive load related to doubts or hesitations during interaction.
- Streamlined cognitive processing: By relying on body pose, users may engage in cognitive offloading by externalising certain cognitive processes through physical actions. This alignment of physical movements with cognitive tasks can streamline cognitive processing and contribute to a more efficient user experience.

In summary, adopting body pose as the primary input method for an AR HMD UI, as opposed to hand gestures, aligns with the principles of cognitive offloading by providing a more natural, reliable and less mentally demanding interaction for users.

7.6 Cognitive Offloading Strategies—Beyond Body Poses

AR HMD UIs excel in utilising body pose for cognitive offloading, primarily due to their advanced spatial awareness and colocalisation capabilities. Nevertheless, other UIs lacking such capabilities can still derive benefits from cognitive offloading through the approach of 'Putting Cognition Into-The-World' [71]. This strategy of cognitive offloading involves transforming the external environment into a repository of representational information, eliminating the need for an internal mental representation. In the context of UIs lacking spatial understanding capabilities, incorporating cognitive offloading by using this strategy could become valuable, mirroring the natural ways in which individuals externalise information to enhance cognitive efficiency and reduce mental load [27, 42].

7.7 Cognitive Offloading in HRI—Beyond Industrial Legged Manipulators

The distinctive features of legged robots [3] make them ideal research platforms for studying trust transfer in our experiment. Their ability to navigate diverse terrains, including those perceived as riskier such as stair navigation, allows for an exploration of factors influencing trust transfer in HRIs, including the impact of the UI. Furthermore, the distinctive interaction paradigm of legged manipulators in industrial settings involves a collaborative partnership between the human operator and the robot. In this context, the human operator takes an active role in defining the robot's movements and specifying location-based inspection actions.

While existing research has extensively covered legged manipulators [50], our emphasis lies in addressing the unique challenges associated with their UI. Many current UIs for legged manipulators, often relying on traditional interfaces such as joysticks and handheld controllers [6, 44], may not be well-suited for the intricacies of industrial inspections, e.g, obstacle-rich environments, narrow passages, uneven surfaces, among others, potentially hindering their effectiveness and safety in such environments. This work aimed at filling this gap by exploring novel UI designs tailored to the distinctive needs of legged manipulators deployed in industrial inspection scenarios.

However, the application of the cognitive offloading principles applied to our UI design has the potential to extend far beyond AR HMD UIs and legged manipulators in industrial settings. The versatility of cognitive offloading concepts that allows it to reach multiple types of UIs, as discussed in the previous sub-section, also opens up an array of possibilities for various robot platforms and HRI tasks across diverse domains. For instance, in domestic settings, service robots supporting daily tasks like cleaning, cooking, or companionship could also capitalise on cognitive offloading principles. Intuitive and user-friendly interfaces, designed with cognitive offloading in mind, could simplify interactions and enhance the overall user experience. Search and rescue robots operating in complex environments could utilise cognitive offloading strategies to improve operator decision-making and navigation. By incorporating cognitive offloading principles into the UIs of these robots, operators can efficiently manage information and collaborate in high-stress situations. In essence, the application of cognitive offloading principles offers a pathway to create more effective, user-friendly and context-aware interfaces for diverse range of HRI tasks.

7.8 Limitations

Our study examines trust changes over time as a within-subject variable. However, we assessed trust using repeated discrete questionnaires across several short trials. Although such an assessment is adopted by the majority of prior literature, it does not fully acknowledge the dynamic nature of trust, which can strengthen or decline over time. Furthermore, because periodically reporting trust may be cumbersome for users, this approach introduces operational challenges in high-workload and time-critical settings, such as the ones where legged manipulators are required. Additional work is needed to develop computational models of trust capable of measuring or inferring trust on a continuous basis as users repeatedly interact with their robot companions. An additional limitation of our study is the exclusive focus on two UIs, namely the HH UI and the AR HMD UI, and the omission of alternative interaction styles, such as haptic interfaces [65, 87], which could have different implications for operator engagement and robot control.

The arguments and recommendations in Sections 7.6 and 7.7 are rooted in theoretical considerations and conceptual frameworks, lacking empirical testing to assess the practical viability, effectiveness, or user acceptance of the proposed cognitive offloading strategies in different UIs and HRI scenarios. This absence of empirical validation constitutes a limitation, as real-world applications of these strategies may face unforeseen challenges and complexities not accounted for in the theoretical exploration. Factors such as user adaptability, system reliability and the dynamic nature of real-world environments could impact the actual effectiveness of the suggested cognitive offloading approaches, highlighting the need for future research to conduct thorough empirical testing and validation for a more robust understanding of their applicability and limitations.

8 Conclusions and Future Work

In this article, we provided a user study comparing an AR HMD UI we developed for controlling legged manipulator against off-the-shelf control methods for such robots. This user study involved 27 participants and 135 trials, from which we gathered over 405 completed questionnaires. These trials involved multiple navigation and manipulation tasks with varying difficulty levels using a BD Spot, a 7 df Kinova robot arm and a Robotiq 2F-85 gripper that we integrated into a legged manipulator. We made this comparison baseline across multiple metrics relevant to a successful HRI. These metrics included cognitive workload, technology acceptance, fluency, system usability and immersion. Furthermore, in this article, we investigated how the level of trust assigned by an operator to a legged manipulator transfers across tasks and environmental contexts. Moreover, we explored how off-the-shelf control methods and AR HMD UIs affect such transfer. We believe our study is valuable as there is no previous user study involving UIs for legged manipulators in the literature, and very few studies in HRI are focused on trust transfer. Furthermore, to our knowledge, no other study evaluates how different UIs affect such transfer. In addition, our analysis emphasises the significance of Bayesian data analysis over standard parametric or semi-parametric procedures in evaluating HRI experiments. Through careful illustration, we demonstrated the importance of Bayesian analysis, showcasing instances where standard procedures may lead to erroneous conclusions.

Our results demonstrate that AR UIs can outpace HH-based control methods and reduce the cognitive requirements when designers include hands-free interactions and cognitive offloading principles into the UI. Furthermore, the use of the AR UI together with our cognitive offloading feature resulted in higher usability scores and significantly higher fluency and TAM scores. Regarding immersion, our results revealed that the response values for the ARI questionnaire associated with the AR UI are significantly higher than those associated with the HH UI, regardless of the main interaction method with the former, i.e., hand gestures or cognitive offloading. Derived from the participants' qualitative answers, we believe this is due to a combination of factors, of which the most important is the free use of the hands when using the HMD, as well as the ability to see the real environment without the need to divert their attention to the UI. Regarding trust, our findings did not display discernible differences in reported trust scores across UI options. However, during the manipulation phase of our user study, where participants were given the choice to select their preferred UI, they consistently reported higher levels of trust compared to the navigation category. Moreover, there was a drastic change in the percentage of participants that selected the AR UI for completing this manipulation stage after incorporating the cognitive offloading feature. Thus, trust seems to have mediated the use and non-use of the UIs in a dimension different from the ones considered in our study, i.e., delegation and reliance. Therefore, AR HMD UIs for the control of legged manipulators were found to improve HRI across several relevant dimensions, underscoring the critical role of UI design in the effective and trustworthy utilisation of robotic systems. Future work will investigate how eye-gaze patterns can serve as a cue for the online personalisation of trustworthy HRI.

Acknowledgements

We thank Dr Ahmed Al-Hindawi for the interesting discussions on Bayesian data analysis and many other great topics that ultimately helped our research. We also thank our colleagues from the Personal Robotics Laboratory for their help in proofreading this manuscript.

References

[1] George Adamides, Georgios Christou, Christos Katsanos, Michalis Xenos, and Thanasis Hadzilacos. 2014. Usability guidelines for the design of robot teleoperation: A taxonomy. *IEEE Transactions on Human-Machine Systems* 45, 2 (2014), 256–262.

[2] Víctor Alvarez-Santos, Roberto Iglesias, Xose Manuel Pardo, Carlos V Regueiro, and Adrián Canedo-Rodriguez. 2014. Gesture-based interaction with voice feedback for a tour-guide robot. *Journal of Visual Communication and Image Representation* 25, 2 (2014), 499–509.

[3] Anybotics. 2023. How to Hire an Industrial Inspection Robot. Retrieved from https://www.anybotics.com/news/how-to-hire-an-industrial-inspection-robot-part-1/. Accessed: 2024-01.

[4] Apple. 2021. More to Explore with ARKit 5. Retrieved from https://developer.apple.com/augmented-reality/arkit/. Accessed: 2021-09.

[5] Juergen Baumgartner, Nicole Ruettgers, Annigna Hasler, Andreas Sonderegger, and Juergen Sauer. 2021. Questionnaire experience and the hybrid system usability scale: Using a novel concept to evaluate a new instrument. *International Journal of Human-Computer Studies* 147 (2021), 102575.

[6] C. Dario Bellicoso, Koen Krämer, Markus Stäuble, Dhionis Sako, Fabian Jenelten, Marko Bjelonic, and Marco Hutter. 2019. Alma-articulated locomotion and manipulation for a torque-controllable robot. In *Proceedings of the International Conference on Robotics and Automation (ICRA '19)*. IEEE, 8477–8483.

[7] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[8] Daniel Caetano, Fernando Mattioli, Edgard Lamounier, and Alexandre Cardoso. 2014. [DEMO] On the use of augmented reality techniques in a telerehabilitation environment for wheelchair users' training. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR '14)*. 329–330.

[9] Daniel S. D. Caetano, Caroline Valentini, Fernando Mattioli, Paulo Camargos, Thiago Sá, Alexandre Cardoso, Edgard Lamounier, and Eduardo Naves. 2020. The augmented reality telerehabilitation system for powered wheelchair user's training. *Journal of Communication and Information Systems* 35, 1 (2020), 51–60.

[10] Rodrigo Chacon -Quesada and Yiannis Demiris. 2019. Augmented reality controlled smart wheelchair using dynamic signifiers for affordance representation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '19)*. 4812–4818.

[11] Rodrigo Chacon-Quesada and Yiannis Demiris. 2020. Augmented reality user interfaces for heterogeneous Multirobot control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '20)*. 11439–11444. DOI: https://doi.org/10.1109/IROS45743.2020.9341422

[12] Rodrigo Chacon Quesada and Yiannis Demiris. 2022. Holo-SpoK: Affordance-aware augmented reality control of legged manipulators. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '22)*. 856–862. DOI: https://doi.org/10.1109/IROS47612.2022.9981989

[13] Rodrigo Chacon Quesada and Yiannis Demiris. 2023. Design and evaluation of an augmented reality head-mounted display user interface for controlling legged manipulators. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '23)*, 11950–11956.

[14] Wesley P. Chan, Morgan Crouch, Khoa Hoang, Charlie Chen, Nicole Robinson, and Elizabeth Croft. 2022. Design and implementation of a human-robot joint action framework using augmented reality and eye gaze. arXiv:2208.11856 Retrieved from https://doi.org/10.48550/arXiv.2208.11856

[15] Wesley P. Chan, Geoffrey Hanks, Maram Sakr, Haomiao Zhang, Tiger Zuo, Hendrik F. M. Van der Loos, and Elizabeth Croft. 2022. Design and evaluation of an augmented reality head-mounted display interface for human robot teams collaborating in physically shared manufacturing tasks. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 3 (2022), 1–19.

[16] Meia Chita-Tegmark, Theresa Law, Nicholas Rabb, and Matthias Scheutz. 2021. Can you trust your trust measure?. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*. ACM, New York, NY, 92–100. DOI: https://doi.org/10.1145/3434073.3444677

[17] Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly 13, 3* (1989), 319–340.

[18] Fred D. Davis, Richard P. Bagozzi, and Paul R. Warshaw. 1989. User acceptance of computer technology: A comparison of two theoretical models. *Management Science* 35, 8 (1989), 982–1003.

[19] Jeffrey Delmerico, Roi Poranne, Federica Bogo, Helen Oleynikova, Eric Vollenweider, Stelian Coros, Juan Nieto, and Marc Pollefeys. 2022. Spatial computing and intuitive interaction: Bringing mixed reality and robotics together. *IEEE Robotics Automation Magazine* 29, (2022), 45–57. DOI: https://doi.org/10.1109/MRA.2021.3138384

[20] Munjal Desai. 2012. *Modeling Trust to Improve Human-Robot Interaction.* Ph.D. Dissertation. University of Massachusetts Lowell.

[21] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI '13)*. 251–258. DOI: https://doi.org/10.1109/HRI.2013.6483596

[22] Hongchao Fang, Soh K. Ong, and Andrew Y. C. Nee. 2012. Interactive robot trajectory planning and simulation using augmented reality. *Robotics and Computer-Integrated Manufacturing* 28, 2 (2012), 227–237.

[23] Hongchao Fang, Soh K. Ong, and Andrew Y. C. Nee. 2013. Orientation planning of robot end-effector using augmented reality. *The International Journal of Advanced Manufacturing Technology* 67, 9–12 (2013), 2033–2049.

[24] P. C. Anacleto Filho, Lincoln da Silva, Ana Pombeiro, Nelson Costa, Paula Carneiro, and Pedro Arezes. 2023. Assessing mental workload in industrial environments: A review of applied studies. *Occupational and Environmental Safety and Health V* 492, (2023), 677–689.

[25] Junling Fu, Maria C. Palumbo, Elisa Iovene, Liu Qingsheng, Burzo Ilaria, Alberto C. L. Redaelli, Giancarlo Ferrigno, Elena De Momi. 2023. Augmented reality-assisted robot learning framework for minimally invasive surgery task. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '23)*. IEEE, 11647–11653.

[26] Richard Fung, Sunao Hashimoto, Masahiko Inami, and Takeo Igarashi. 2011. An augmented reality system for teaching sequential tasks to a household robot. In *RO-MAN*. IEEE, 282–287.

[27] Paula Gauselmann, Yannick Runge, Christian Jilek, Christian Frings, Heiko Maus, and Tobias Tempel. 2023. A relief from mental overload in a digitalized world: How context-sensitive user interfaces can enhance cognitive performance. *International Journal of Human–Computer Interaction* 39, 1 (2023), 140–150.

[28] Yiannis Georgiou and Eleni A. Kyza. 2017. The development and validation of the ARI questionnaire: An instrument for measuring immersion in location-based augmented reality settings. *International Journal of Human-Computer Studies* 98 (2017), 24–37.

[29] Joshua A. Gomer and Christopher C. Pagano. 2011. NASA task load index for human-robot interaction workload measurement. *International Test and Evaluation Association Journal* 32 (2011), 210–214.

[30] Google. 2021. Cloud Anchors Overview for Android. Retrieved from https://developers.google.com/ar/develop/java/cloud-anchors/overview-android. Accessed: 2021-09.

[31] Robert J. Hall. 1996. Trusting your assistant. In *Proceedings of the 11th Knowledge-Based Software Engineering Conference (KBSE '96)*. 42–51. DOI: https://doi.org/10.1109/KBSE.1996.552822

[32] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[33] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*. Peter. A. Hancock and Najmedin Meshkati (Eds.), Vol. 52. Elsevier, 139–183.

[34] Sunao Hashimoto, Akihiko Ishida, Masahiko Inami, and Takeo Igarashi. 2011. Touchme: An augmented reality based remote robot manipulation. In *Proceedings of the 21st International Conference on Artificial Reality and Telexistence (ICAT '11)*, Vol. 2, 61–66.

[35] Tanja Heuer and Jenny Stein. 2020. From HCI to HRI: About users, acceptance and emotions. In *Proceedings of the 2nd International Conference on Human Systems Engineering and Design : Future Trends and Applications (IHSED '19)*. Universität der Bundeswehr München, Springer, 149–153.

[36] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (June 2019), 209–218. DOI: https://doi.org/10.1109/THMS.2019.2904558

[37] Guy Hoffman and Cynthia Breazeal. 2007. Cost-based anticipatory action selection for human–robot fluency. *IEEE Transactions on Robotics* 23, 5 (October 2007), 952–961. DOI: https://doi.org/10.1109/TRO.2007.907483

[38] Guy Hoffman and Xuan Zhao. 2020. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 1 (2020), 1–31.

[39] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Alvin Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88. DOI: https://doi.org/10.1109/MIS.2013.24

[40] Courtney Hutton, Nicholas Sohre, Bobby Davis, Stephen Guy, and Evan S. Rosenberg. 2019. An Augmented reality motion planning interface for robotics. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR '19)*. IEEE, 1313–1314.

[41] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.

[42] Tao Jin, Jiamin He, Wenrui Wang, Zhengxin Wu, and Haoran Gu. 2022. How mobile touch devices foster cognitive offloading in the elderly: The effects of input and feedback. *International Journal of Human–Computer Interaction* 40, (2022), 1–11.

[43] Florian Kennel-Maushart, Roi Poranne, and Stelian Coros. 2023. Interacting with multi-robot systems via mixed reality. In *Proceedings of the International Conference on Robotics and Automation (ICRA '23)*. 11633–11639

[44] Tobias Klamt, Diego Rodriguez, Max Schwarz, Christian Lenz, Dmytro Pavlichenko, David Droeschel, and Sven Behnke. 2018. Supervised autonomous locomotion and manipulation for disaster response with a centaur-like robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '18)*. IEEE, 1–8.

[45] Jan Krieglstein, Gesche Held, Balazs A. Balint, Frank Nagele, and Werner Kraus. 2023. Skill-based robot programming in mixed reality with ad-hoc validation using a force-enabled digital twin. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '23)*. IEEE, 11612–11618.

[46] John Kruschke. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, Academic Press.

[47] John K. Kruschke. 2013. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General* 142, 2 (2013), 573.

[48] John K. Kruschke. 2018. Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science* 1, 2 (2018), 270–280.

[49] John K. Kruschke and Torrin M. Liddell. 2018. The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review* 25 (2018), 178–206.

[50] G. Satheesh Kumar, S. Aravind, R. Subramanian, Shashank K. Bharadwaj, Ronak R. Muthuraman, R. Steve Mitchell, Aditya Bucha, M. Sriram, K. J. Shri Hari, and Nikhil J. Robin. 2021. Literature survey on four-legged robots. In *Trends in Mechanical and Biomedical Design*. E. Akinlabi, P. Ramkumar, and M. Selvaraj (Eds.), Lecture Notes in Mechanical Engineering. Springer, 691–702.

[51] Ryotaro Kuriya, Takeshi Tsujimura, and Kiyotaka Izumi. 2015. Augmented reality robot navigation using infrared marker. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '15)*. 450–455.

[52] Donghyeon Lee and Young S. Park. 2018. Implementation of augmented teleoperation system based on robot operating system (ROS). In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '18)*. 5497–5502.

[53] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[54] Sam Lee, Nathan P. Lucas, Richard D. Ellis, and Abhilash Pandya. 2012. Development and human factors analysis of an augmented reality interface for multi-robot tele-operation and control. In *Proceedings of the Unmanned Systems Technology XIV*, Vol. 8387. SPIE, 254–261.

[55] Yuan Lin, Shuang Song, and Max Q. H. Meng. 2016. The implementation of augmented reality in a robotic teleoperation system. In *Proceedings of the IEEE International Conference on Real-time Computing and Robotics (RCAR '16)*. 134–139.

[56] Joseph B. Lyons and Svyatoslav Y. Guznov. 2019. Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science* 20, 4 (2019), 440–458.

[57] Zhanat Makhataeva and Huseyin A. Varol. 2020. Augmented reality for robotics: A review. *Robotics* 9, 2 (2020), 21.

[58] Dominique Makowski, Mattan S. Ben-Shachar, and Daniel Lüdecke. 2019. bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software* 4, 40 (2019), 1541.

[59] Levi Manring, John Pederson, Dillon Potts, Beth Boardman, David Mascarenas, Troy Harden, and Alessandro Cattaneo. 2020. Augmented reality for interactive robot control. In *Proceedings of the 37th IMAC, A Conference and Exposition on Structural Dynamics Special Topics in Structural Dynamics & Experimental Techniques*, Vol. 5. Springer, 11–18.

[60] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review* 20, 3 (1995), 709–734.

[61] Richard McElreath. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

[62] Microsoft. 2021. Spatial Anchors. Retrieved from https://azure.microsoft.com/en-us/services/spatial-anchors/#overview. Accessed: 2021-09.

[63] Paul Milgram, Shumin Zhai, David Drascic, and Julius Grodski. 1993. Applications of augmented reality for human-robot communication. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '93)*, Vol. 3. IEEE, 1467–1472.

[64] Faizan Muhammad, Amel Hassan, Andre Cleaver, and Jivko Sinapov. 2019. Creating a shared reality with robots. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI '19)*. IEEE, 614–615.

[65] James F. Mullen, Josh Mosier, Sounak Chakrabarti, Anqi Chen, Tyler White, and Dylan P. Losey. 2021. Communicating inferred goals with passive augmented reality and active haptic feedback. *IEEE Robotics and Automation Letters* 6, 4 (2021), 8522–8529.

[66] Kimihiro Noguchi, Yulia R. Gel, Edgar Brunner, and Frank Konietschke. 2012. nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software* 50, 12 (2012), 1–23.

[67] Kohei Okuoka, Kouichi Enami, Mitsuhiko Kimoto, and Michita Imai. 2022. Multi-device trust transfer: Can trust be transferred among multiple devices? *Frontiers in Psychology* 13 (2022), 920844.

[68] Mateusz Paliga and Anita Pollak. 2021. Development and validation of the fluency in human-robot interaction scale. A two-wave study on three perspectives of fluency. *International Journal of Human-Computer Studies* 155 (2021), 102698.

[69] Sina Radmard, A. Jung Moon, and Elizabeth A. Croft. 2015. Interface design and usability analysis for a robotic telepresence platform. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '15)*. IEEE, 511–516.

[70] Christopher Reardon, Kevin Lee, and Jonathan Fink. 2018. Come see this! augmented reality to enable human-robot cooperative search. In *Proceedings of the IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR '18)*. IEEE, 1–7.

[71] Evan F. Risko and Sam J. Gilbert. 2016. Cognitive offloading. *Trends in Cognitive Sciences* 20, 9 (2016), 676–688.

[72] Eric Rosen, David Whitney, Michael Fishman, Daniel Ullman, and Stefanie Tellex. 2020. Mixed reality as a bidirectional communication interface for human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '20)*. IEEE, 11431–11438.

[73] Kenneth Ruffo and Paul Milgram. 1992. Effect of stereographic + stereovideo 'tether' enhancement for a peg-in-hole task. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 1425–1430.

[74] Nicole Salomons, Kaitlynn T. Pineda, Adérónké Adéjare, and Brian Scassellati. 2022. "We make a great team!": Adults with low prior domain knowledge learn more from a peer robot than a tutor robot. In *Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI '22)*. IEEE, 176–184.

[75] Brian Scassellati, Laura Boccanfuso, Chien-Ming Huang, Marilena Mademtzi, Meiying Qin, Nicole Salomons, Pamela Ventola, and Frederick Shic. 2018. Improving social skills in children with ASD using a long-term, in-home social robot. *Science Robotics* 3, 21 (2018), eaat7544. DOI: https://doi.org/10.1126/scirobotics.aat7544

[76] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and Peter A. Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors* 58, 3 (2016), 377–400.

[77] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. 2020. Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research* 39, 2–3 (2020), 233–249.

[78] Ruth M. Stock and Moritz Merkle. 2017. A service robot acceptance model: User acceptance of humanoid robots during service encounters. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops '17)*. IEEE, 339–344.

[79] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. 2022. Augmented reality and robotics: A survey and taxonomy for AR-enhanced human-robot interaction and robotic interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '22)*. 1–33.

[80] Vildan Tanriverdi and Robert J. K. Jacob. 2000. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '00)*. ACM, 265–272.

[81] Jack E. Taylor, Guillaume A. Rousselet, Christoph Scheepers, and Sara C. Sereno. 2022. Rating norms should be calculated from cumulative link mixed effects models. *Behavior Research Methods* 55, (2022), 1–22.

[82] Viswanath Venkatesh and Fred D. Davis. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science* 46, 2 (2000), 186–204.

[83] Yanxin Wang, Hong Zeng, Aiguo Song, Baoguo Xu, Huijun Li, Lifeng Zhu, Pengcheng Wen, and Jia Liu. 2017. Robotic arm control using hybrid brain-machine interface and augmented reality feedback. In *Proceedings of the 8th International IEEE/EMBS Conference on Neural Engineering (NER '17)*. 411–414.

[84] Anqi Xu and Gregory Dudek. 2016. Towards modeling real-time trust in asymmetric human–robot collaborations. In *Robotics Research*. M. Inaba and P. Corke (Eds.), Springer, 113–129.

[85] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. 408–416.

[86] A. W. W. Yew, S. K. Ong, and A. Y. C. Nee. 2017. Immersive augmented reality environment for the teleoperation of maintenance robots. *Procedia Cirp* 61 (2017), 305–310.

[87] James E. Young, Youichi Kamiyama, Juliane Reichenbach, Takeo Igarashi, and Ehud Sharlin. 2011. How to walk a robot: A dog-leash human-robot interface. In *RO-MAN*. IEEE, 376–382.

[88] Hong Zeng, Yanxin Wang, Changcheng Wu, Aiguo Song, Jia Liu, Peng Ji, Baoguo Xu, Lifeng Zhu, Huijun Li, and Pengcheng Wen. 2017. Closed-loop hybrid gaze brain-machine interface based robotic arm control with augmented reality feedback. *Frontiers in Neurorobotics* 11 (2017), 60.

[89] Zhou Zhao, Panfeng Huang, Zhenyu Lu, and Zhengxiong Liu. 2017. Augmented reality for enhancing tele-robotic system with force feedback. *Robotics and Autonomous Systems* 96 (2017), 93–101.

[90] Mark Zolotas and Yiannis Demiris. 2019. Towards explainable shared control using augmented reality. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '19)*. 3020–3026.

[91] Mark Zolotas, Joshua Elsdon, and Yiannis Demiris. 2018. Head-mounted augmented reality for explainable robotic wheelchair assistance. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '18)*. 1823–1829. DOI: https://doi.org/10.1109/IROS.2018.8594002

Received 29 July 2023; revised 21 February 2024; accepted 15 April 2024