Human-Computer Interaction & Emerging Technologies, Vol. 195, 2025, 462–473
https://doi.org/10.54941/ahfe1006263

Toward Intuitive Interaction: A Cognitive Workflow Analysis of Human-Robot Interaction in Extended Reality Interfaces

Pattaraorn Yu[1], Arisara Jiamsanguanwong[2], and Gim Song Soh[1]

[1]Singapore University of Technology and Design, Singapore
[2]Human-Robot Collaboration and Systems Integration Research Unit, Department of Industrial Engineering, Faculty of Engineering, Chulalongkorn University, Thailand

ABSTRACT

Extended-Reality (XR) technologies promise to enhance human–robot interaction (HRI) by offering intuitive spatial interfaces and immersive input. However, traditional evaluation methods, such as task completion time, error rates, or NASA-TLX often obscure where cognitive and physical demands arise or are reduced within the interaction process. This study conducts a cognitive workflow analysis of XR interfaces by integrating established methodologies: Goal-Directed Task Analysis (GDTA), Norman's Seven Stages of Action, and Applied Cognitive Task Analysis (ACTA). These methods collectively trace how interface design affects user cognition, from mission goals to task-level interactions, revealing specific gulfs of execution and evaluation. We apply this approach to compare two XR interface types: a grid-based menu and spatial affordance-based pop-ups, within an emergency-response scenario using Microsoft HoloLens 2. The analysis uncovers hidden cognitive challenges, such as inefficient visual search and occlusion issues, often missed by conventional metrics. The findings offer XR designers actionable insights into usability challenges and demonstrate how cognitive analysis can guide more intuitive interface development.

Keywords: Human-robot interaction (HRI), Multi-robot system (MRS), Cognitive task analysis (CTA), Extended reality (XR)

INTRODUCTION

Extended-Reality (XR), which encompasses augmented, virtual, and mixed realities, has emerged as a transformative platform for enhancing human–robot interaction (HRI). By spatially aligning virtual information with physical environments, XR interfaces promise significant improvements in situational awareness, spatial reasoning, and reduced cognitive workload compared to traditional 2D displays (Chen et al., 2024; Roldán et al., 2017). These advantages are particularly valuable in complex, time-critical domains such as search-and-rescue, infrastructure maintenance, and defense, where operators must coordinate multiple heterogeneous robots (Roldán et al., 2017).

Despite XR's potential, substantial limitations persist in real-world deployments. Operators often experience cognitive overload even when supervising a relatively small number of robots, typically not more than three, beyond which performance sharply declines due to increased visual complexity and interaction demands (Wang et al., 2025). Although recent advancements such as motion predictive interfaces (Gamage et al., 2021), gazed-based menu interaction (Ahn et al., 2021), multimodal interactions (Shi et al., 2023), and haptic feedback (Kudry & Cohen, 2023) aim to improve usability, they are typically evaluated using outcome-based metrics such as task completion time or NASA-TLX scores (Chang & Hayes, 2024). These metrics, while useful, do not reveal where or how interaction design affects user cognition during task execution. Studies such as AR teleoperation (Walker et al., 2019), data visualization platform (Nafis et al., 2025), and XR for HRI (Szafir, 2019), attempted to clarify these mental and physical challenges using Norman's gulf of execution and evaluation concept, but remain incomplete in providing comprehensive insight.

Thus, the purpose of this study is to conduct a cognitive workflow analysis of XR interfaces by integrating established methodologies: Goal-Directed Task Analysis (GDTA), Norman's Seven Stages of Action, and Applied Cognitive Task Analysis (ACTA) to clarify how interface design influences user cognition and task performance. GDTA decomposes high-level mission goals into sub-goals and decision points (Endsley et al., 2003), Norman's Seven Stages of Action highlights the "gulfs" between user intentions and system feedback (Norman, 1986), and ACTA identifies workload-intensive involved (Militello & Hutton, 1998). This study applied to a multi-robot emergency-response scenario using HoloLens 2 on grid-based and object-affordance-based interfaces.

The implications of present study are for the design and evaluation of XR interfaces in human–robot interaction. By integrating GDTA, ACTA, and Norman's Seven Stages of Action, it demonstrates how cognitive and physical workload can be systematically traced from mission-level goals to task-level actions. This approach enables early-stage interface evaluation without extensive user testing, offering designers a rigorous means to identify usability challenges.

METHODOLOGY

Participants

Two subject matter experts (SMEs) participated in the study. Participant 1 (P1), a 27-year-old male, has five years of experience in designing extended reality interfaces for HRI. Participant 2 (P2), a 37-year-old female, has four years of experience in developing mixed reality systems. Both were members of the development team responsible for the interfaces evaluated in this study.

Task

Participants interacted with two mixed reality interfaces intended for a simulated emergency response mission.

Scenario: A three-room virtual building contained four active fires, two fire extinguishers, three first aid kits, and three survivors (Figure 1a). A legged ground robot could carry and actuate extinguishers, while a drone could transport kits. A mission brief stated: *"Extinguish all fires and deliver first aid kits to all survivors. Each extinguisher may be used only twice."*

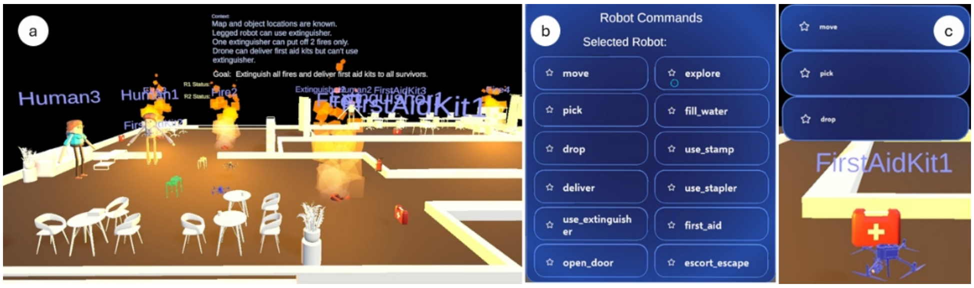

Figure 1: (a) Simulated emergency response scenario in mixed reality; (b) grid-based command interface; (c) object-affordance-based command interface appears when robot waypoint virtual object collides with interactable objects.

Stimulus

Two types of interfaces were used as stimulus in this study: *(1) Grid-based command interface* (Figure 1b): A static floating panel listing twelve robot actions (e.g., Move, Pick, Drop, Extinguish, etc.), requiring operators to select a robot, position a robot waypoint and choose actions from the menu. The panel remained fixed on the left side of the operator's field of view. *(2) Object-affordance-based interface* (Figure 1c): When a robot waypoint contacted an actionable object, an object-affordance menu popped up above it, displaying only the commands valid for that object. This emphasized spatial affordances and reduced menu-search load.

Both control interfaces shared an identical situational-awareness layer that rendered mission brief, accessible at any time during the trial, real-time positional updates and status of robots (e.g. moving to a goal, goal reached) and objects (e.g. being picked up, dropped, or extinguished by a robot) as spatial holograms over a resizable occupancy grid map of the building (Figure 1a).

Apparatus

The application ran on a workstation equipped with an Intel(R) Core(TM) i9-12900 processor and an NVIDIA GeForce RTX 4070 GPU with 12 GB of VRAM. The robot control and simulation stack were hosted on a Hyper-V virtual machine running Ubuntu 20.04 with ROS Noetic. Multi-robot navigation was simulated using the TurtleBot fake node with a preconfigured occupancy grid map to emulate robot localization and path planning. Communication between Unity and ROS platform was established using the ROS–Unity TCP Connector, enabling real-time bidirectional exchange of robot status, goals, and control signals.

All visualization and interaction tasks were conducted using the Microsoft HoloLens 2 head-mounted display, running a custom-built mixed reality (MR) environment (Yu, 2022) developed in Unity (2021.3 LTS). The application was built using the Microsoft Mixed Reality Toolkit 3 (MRTK3), which supported spatial mapping, hand gesture tracking, and interactive holographic UI components. Users interacted with the MR interface using standard HoloLens 2 mid-air gestures, including poke, pinch, and pinch-and-drag, enabling both near and far interactions with virtual elements. The system rendered real-time overlays of environmental geometry, robot positions, and status information.

Procedure

Participants were briefed on mission objectives, operational constraints, robot capabilities, and control mechanisms for both interfaces. The session also covered GDTA task decomposition, symbolic annotation for Norman's Seven Stages of Action, and the cognitive demand table. Each participant then completed a 10-minute practice session with both interfaces to ensure baseline familiarity. During formal trials, they executed the mission scenario three times per interface, presented in a pseudorandom sequence to reduce learning effects. Afterward, participants independently created GDTA and Norman cycle diagrams and analyzed cognitive demands, which were compiled into cognitive demand tables. Individual data were then consolidated to produce the final diagrams and demand tables used for analysis.

Measurement

Cognitive workload was assessed using a three-layer analytic protocol integrating GDTA (Goal-Directed Task Analysis), an annotated application of Norman's Seven Stages of Action, and ACTA's Cognitive Demand Table.

GDTA: Mission-Level Decomposition

Following GDTA methodology (Endsley et al., 2003), the mission-level goal was decomposed into a hierarchy of cognitive subgoals. Each subgoal was annotated with situation-awareness (SA) requirements and decision points. Decomposition proceeded until each leaf node represented an operational goal which defined as a desired system state independent of specific interfaces or robot capabilities. Operational goals were intentionally verb-free to avoid biasing toward particular control actions, ensuring the analysis remains valid across varying levels of automation or interface design.

Norman's Seven Stages of Action Cycle With Symbolic Notation

Each operational goal was realised through one or more interaction cycles model using Norman's Seven Stages of Action (Norman, 1986), annotated with symbolic variables (G, I, M, V, S) representing goal, intention, control mechanisms, interface variables, and system state, respectively. The description of each stage is explained in Table 1. Subscripts distinguished psychological (i.e. M_s, V_s from physical (i.e. M_f, V_f) processes, enabling fine-grained workload tracing.

The number of cycles required to complete an operational goal depended on the system's autonomy and interface design: high-autonomy systems may complete compound commands in a single cycle, while lower-autonomy systems require multiple micro-managed interaction cycles. The GDTA structure remained constant, while the Norman-layer realization adapted accordingly maintaining a clear separation between goal structure and interface-dependent execution.

ACTA: Cognitive Demand Table

The Cognitive Demand Table was adapted from the Applied Cognitive Task Analysis (ACTA) framework (Militello & Hutton, 1998), summarizing GDTA and Norman's seven stages effort perceived as demanding, their contributing factors, common errors, and coping strategies.

Table 1: Norman's seven stages definition.

Stage	Symbolic Expression	Description
N1. Perception	$V_f \rightarrow V_s$ and/or $S_f \rightarrow S_s$	Translation of interface variables V or system state S into psychological description via sensorimotor afferent inputs (such as visual, audio, haptic, etc.).
N2. Interpretation	V_s, S_s (perception) $\rightarrow S_s$ (comprehension)	Translation of the perceived inputs (Situation awareness level 1) into meanings relevant to the goal (Situation awareness level 2) (Endsley et al., 2003).
N3. Evaluation	$S_s = G_s$?	Evaluation if the system state S has met the interaction goal G.
N4. Interaction goal formation	G_s	Interaction goal G is psychological description arises from comparing system state S and a GDTA operational goal If the operational goal is met, exit this interaction cycle and fulfil other operational goals until the mission goal is met. Otherwise, form new interaction goal G to fulfil the operational goal.
N5. Intention formation	$I_s = S_s - G_s$	Intention I, is also a psychological description, surfaced as the mismatch of the current system state S and interaction goal G.

Continued

Table 1: Continued

Stage	Symbolic Expression	Description
N6. Action specification	$I_s \rightarrow M_s$	Translation of the psychological description of intention into psychological description of actions based on interface's control mechanisms M. A single intention may involve interactions with multiple control mechanisms. This stage requires an understanding of the casual relationships between the control mechanisms M, interface variable V, and the resulting system state S, as described by: $V_f = f(M_f)$ and $S_f = h(V_f)$.
N7. Action execution	$M_s \rightarrow M_f$	Translation of psychological description of actions into physical actions on control mechanisms M via sensorimotor efferent outputs (such as gaze, hand, body, speech, etc.).

Note: Subscript s and f denote the psychological and physical aspects of the variable, respectively.

RESULT

The GDTA diagram (Figure 2) structures the ultimate mission goal G0 (stabilize the emergency site) into two principal cognitive goals: G1 (mitigation of fire hazards) and G2 (provision of medical aid). Each branch terminates in operational goals that articulate technology-independent desired system states: G1.1 (a robot is equipped with a functional extinguisher), G1.2 (a fire is extinguished), G2.1 (a robot carries a first-aid kit), and G2.2 (a survivor has received the kit). These operational goals articulate desired outcomes rather than interface-specific commands, ensuring analytical validity across variations in interface design or system autonomy.

To fulfil each operational goal, the operator first observes the system state S either through the interface or directly from the physical environment, corresponding to Norman's gulf of evaluation (see Figure 3, stages N1-N3). The gap between the system state and the desired operational goal defines an interaction goal G (stage N4). For example, if the operational goal is G1.1—"*Robot* is equipped with a functional extinguisher"—but the system state shows that "*Robot* is carrying a depleted extinguisher," then the initial interaction goal G is "*Robot* drops the unusable extinguisher." Once this goal is accomplished through Norman's seven stages cycle, the system state updates to "*Robot* is not equipped with a functional extinguisher and is not near one." Subsequent interaction goals then become "move to a functional extinguisher" followed by "pick it up".

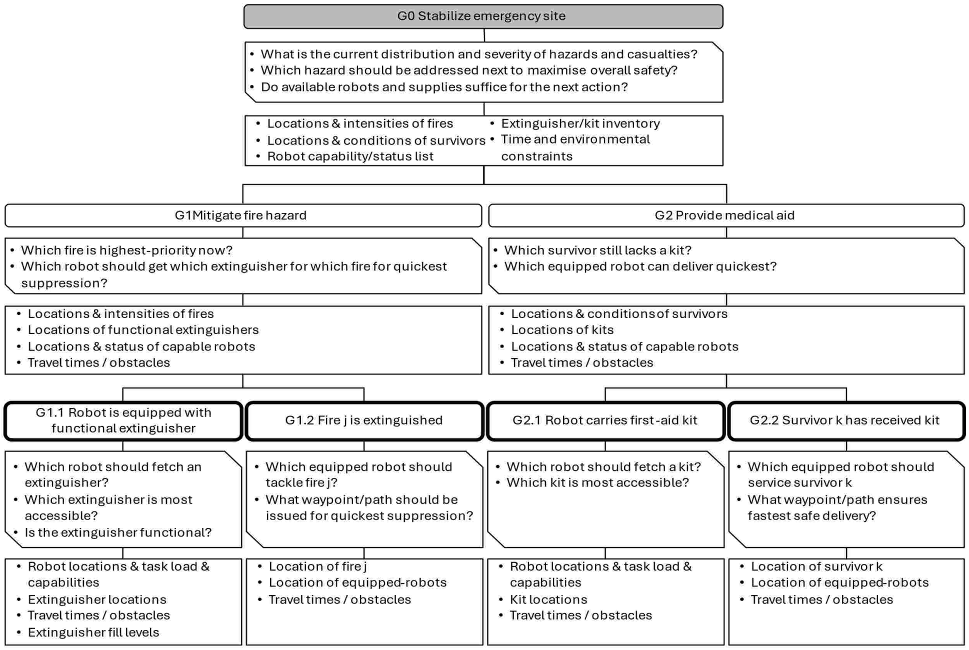

Figure 2: Goal-direct task analysis (GDTA) diagram.

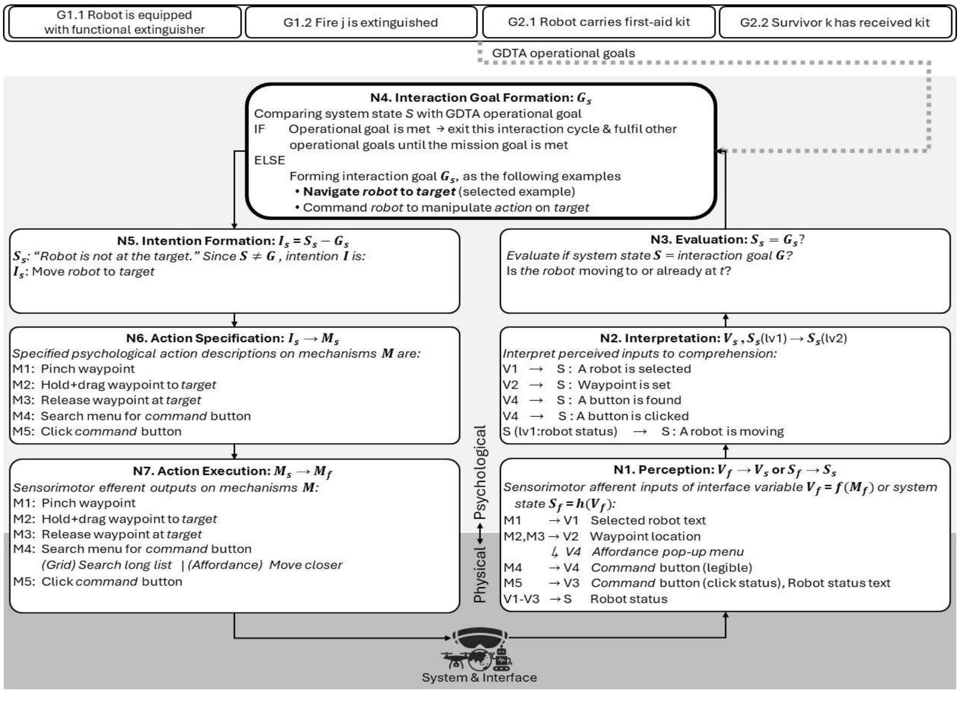

Figure 3: Norman's seven stages of action diagram as applied to fulfilling an operational goal. The example illustrates the formation of the interaction goal "Navigate *robot* to *target*" based on the system state "*Robot* is not at the *target*", where the *target* may be an extinguisher, fire, first aid kit, or survivor—corresponding to G1.1, G1.2, G1.3, and G1.4, respectively.

The number of interaction goals (and thus Norman's cycles) required to fulfil an operational goal depends on the system state and the level of system autonomy. For instance, if the robot system supports composite commands that chain navigation and manipulation into a single interaction, the second and third interaction goals could merge into one.

Each interaction goal G, expressed psychologically as G_s, translates into an intention I_s (stage N5). This intention guides the specification of control mechanisms M_s (N6), physical actions M_f (N7), resulting interface-level changes V_f, and updates in the system state S_f. These changes are perceived as V_s and S_s (N1) and interpreted into a higher-level understanding of the system state S_s (N2). Figure 3 illustrates this analysis for a representative interaction goal across both interface types.

The Norman-based analysis identified five control mechanisms M and four interface feedback variables V, with each V linked to corresponding M. M1 is the pinch gesture used for robot selection (V1); M2 and M3 are drag-and-release gestures for waypoint designation (V2); M4 refers to menu search actions that lead to command button visibility (V4); and M5 is button clicking actions that produce confirmatory feedback (V3). The grid and affordance interfaces differ notably in how the desired command button is found (V4). In the grid interface, V4 results entirely from explicit visual search across an extensive menu panel via M4. In the affordance interface, V4 results from both M2 and M4: waypoint interactions with actionable objects (M2) automatically bring up a concise and object-specific menu for visual search (M4). The dual-purpose functionality of M2—both waypoint designation (V2) and pop-up mechanism for command visibility (V4)—streamlines interaction by using one action to trigger two interface feedback responses.

Table 2: Cognitive demand table.

Difficulty: Mental & Physical	Why Difficult	Common Errors	Cues and Strategies Used
Mission Goal G0 **1.** Remember goals, constraints, contexts **2.** Situation Awareness **3.** Goal decomposition	**1.** High memory demand **2.** Spatial search for all important objects (clutter, small icons, big map area) **3.** Problem solving	**1.** Forget goals, violating constraints **2.** Miss important objects **3.** Fail to capture important subgoals	**1.** Glance at mission brief panel **2.** Moving in MR space to change viewpoint, manipulate the map **3.** Training
Operational Goals Gx.x **1.** Multitasking between subgoals for multiple robots **2.** Multi-robot allocation and prioritization	**1.** Simultaneously tracking multiple robots **2.** Attention switching and recalling next action	**1.** Losing track of next action. **2.** Idle robots overlooked **3.** Forgetting to switch select robot	**1.** Visualizing pending tasks **2.** Keeping track of robot status
(Affordance) **Action Specification N6** of M2 → V4 for menu pop-up	Affordance command interface is not visible by default. Need to remember the menu pop-up mechanism	Forget how to find the menu	Memorizing and training

Continued

Table 2: Continued

Difficulty: Mental & Physical	Why Difficult	Common Errors	Cues and Strategies Used
Action Execution N7 of **M1, M2, M3** for waypoint manipulation	Small waypoint is hard to aim, grab, manipulate, release with hand gesture	Fail to grab, manipulate and release as intended	Moving closer, training hand gesture for better recognition
(Grid) **Action Execution N7** of **M4** for menu search	A list of buttons requires time to search through	Waste time looking for desired button	Memorizing and training
(Affordance) • **Action Execution N7** of **M4** for menu search • **Action Execution N7** of **M5** for button click	**1.** Pop-up menu is small and sometimes occluded by other objects **2.** Pop-up menu disappears	**1.** Fail to find the desired button **2.** Fail to click on the desired button	**1.** Moving closer to change viewpoint **2.** Execute M2 → V4 again to pop the menu
Perception N1 of extinguisher functional status	Lack visualization of extinguisher remaining usage	Mistaken attempt to use a depleted extinguisher	Memorizing

This structured analysis systematically identified all control mechanisms and feedback variables required to realize interaction goals via the interfaces. To complement structural insights derived from the GDTA (Figure 2) and Norman-based analyses (Figure 3), we compiled an ACTA-style Cognitive Demand Table (Table 2). This table captures where cognitive and physical challenges emerged across the mission workflow, detailing the sources of difficulty, common errors, and coping strategies. It provides a granular view of workload distribution, highlighting how an interface influenced user performance and effort, and where current design features failed to adequately support user's needs.

DISCUSSIONS

This study contributes a fine-grained analysis of cognitive workload in mixed reality interfaces for multi-robot supervision and teleoperation, revealing that even seemingly intuitive interfaces impose significant hidden burdens.

At the mission level (G0), the primary challenge was working memory strain. Operators were required to retain mission goals, constraints, and robot capabilities while simultaneously extracting relevant information for decision making. This was further compounded by spatial search demands within a 3D interface, where small or overlapping holographic elements frequently required the user to adjust viewpoint. As the mission decomposed into GDTA-derived operational goals, workload shifted toward the coordination of multiple robots and their respective task assignments. Operators had to continuously monitor the progress of each robot while recalling the next intended actions—a cognitively demanding process of switching attention between multiple concurrent Norman cycles, each corresponding to a different goal and robot state.

Interestingly, despite its seemingly more natural and efficient design, the affordance-based interface introduced several unexpected sources of effort. While it replaced the static, exhaustive grid with a concise, context-aware command list, Table 2 shows that this design led to new challenges. The actions required to execute menu search (stage N7 on M4) shifted from scanning a long panel to managing small-sized, occasionally occluded, and transient pop-up menus. This also increased the difficulty of executing button clicks (stage N7 on M5). To cope, users often had to reposition the viewpoint or re-trigger the pop-up menu to regain visibility, thus redistributed rather than alleviated the effort as intended. Furthermore, the dual function of M2 —serving both waypoint manipulation and triggering the pop-up menu— could create difficulties for novice users unfamiliar with the affordance pop-up mechanism (stage N6, for specifying which M influences V4). Because the command menu is not persistently visible, it could lead to failed command execution. In contrast, the grid interface anchored its menu panel to a consistent position within the user's field of view. While less adaptive to context and non-scalable to more complex system, its fixed placement offered better predictability during command search and ease the button clicking action.

During the gulf of evaluation stages (N1-N3), both interfaces provided timely and sufficient cues for assessing interface and system responses within the simulated MR scenario. However, they lacked feedback on the extinguisher's remaining usage. Furthermore, in real-world remote teleoperation, interface feedback must reliably indicate that the action has been executed, and that the physical environment is responding as intended. This is especially important given that the rendered system state on the interface may be subjected to delay caused by factors such as sensor update frequency, network latency, and computing speed.

The findings from the three-layered analysis suggest three general design guidelines for improving multi-robot interaction in XR-based HRI systems. First, interfaces should better *assist with situation awareness, high-level planning, and multitasking across multiple robots*. Second, they should *reduce the number of Norman's interaction cycles* by supporting compound actions or queued commands, allowing operator to delegate more and focus on situation awareness and decision making. Third, *within-cycle workload should be minimized* by reducing effort for adjusting task and interface variables V, simplifying action sequences M and choosing input mechanisms that provide rapid and low-error control.

One concrete example is to refine the affordance interface by anchoring its pop-up menu to a fixed position relative to the user's view, similar to the grid interface, while preserving its context-aware and concise command list. This hybrid approach could retain advantages of affordance interface while mitigating the undesirable additional coping mechanisms for the small and occluded buttons.

CONCLUSION

This study conducted a cognitive workflow analysis of XR interfaces by integrating established methodologies: GDTA, Norman's Seven Stages of Action, and ACTA. By combining these methods, the study exposed hidden cognitive workload that often escapes detection through traditional usability metrics. This structured approach enables interface designers and researchers to interpret and justify design decisions with cognitively grounded evidence, offering a deeper understanding of how interaction patterns affect mental effort and performance. In addition, the results of this study contribute to the design of XR-HRI systems that better support intuitive interaction and can scale effectively with increasing system complexity and task demands.

However, the study has limitations. The analysis was conducted with a small number of subject matter experts and focused on a single emergency-response scenario, which may limit generalizability. Additionally, the evaluation was performed in a simulated environment using HoloLens 2; thus, real-world operational constraints were not fully captured. Future work will extend this methodology to broader user populations and diverse operational contexts, including real-time field evaluations. Further research is also needed to refine and automate parts of the analytic process, enabling faster iteration in early-stage XR interface development. Expanding the framework to accommodate collaborative HRI and adaptive autonomy could further enhance its applicability to next-generation XR systems.

ACKNOWLEDGMENT

Authors would like to sincerely thank Youngbin Song and researchers from Singapore University of Technology and Design for their invaluable support throughout this work.

REFERENCES

Ahn, S., Santosa, S., Parent, M., Wigdor, D., Grossman, T., & Giordano, M. (2021). *StickyPie: A Gaze-Based, Scale-Invariant Marking Menu Optimized for AR/ VR* Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

Chang, C. T., & Hayes, B. (2024). A Survey of Augmented Reality for Human–Robot Collaboration. *Machines*, *12*(8), 540.

Chen, J., Sun, B., Pollefeys, M., & Blum, H. (2024). A 3d mixed reality interface for human-robot teaming. 2024 IEEE International Conference on Robotics and Automation (ICRA).

Endsley, M. R., Bolté, B., & Jones, D. G. (2003). *Designing for situation awareness: An approach to user-centered design*. CRC press.

Gamage, N. M., Ishtaweera, D., Weigel, M., & Withana, A. (2021). So predictable! Continuous 3D hand trajectory prediction in virtual reality. The 34th Annual ACM Symposium on User Interface Software and Technology.

Kudry, P., & Cohen, M. (2023). Enhanced Wearable Force-Feedback Mechanism for Free-Range Haptic Experience Extended by Pass-Through Mixed Reality. *Electronics*, *12*(17), 3659.

Militello, L. G., & Hutton, R. J. (1998). Applied cognitive task analysis (ACTA): A practitioner's toolkit for understanding cognitive task demands. *Ergonomics*, *41*(11), 1618–1641.

Nafis, F. A., Rose, A., Su, S., Chen, S., & Han, B. (2025). Are We There Yet? Unravelling Usability Challenges and Opportunities in Collaborative Immersive Analytics for Domain Experts. International Conference on Human-Computer Interaction.

Norman, D. A. (1986). Cognitive engineering. *User centered system design*, *31*(61), 2.

Pattaraorn Yu, A. A., Youngbin Song and Gim Song Soh. (2022, 15 – 18 November 2022). A Mixed Reality Platform for Online Motion Planning of Mobile Robots via Hololens 2. Asian Society for Precision Engineering and Nanotechnology (ASPEN2022), Singapore.

Roldán, J. J., Peña-Tapia, E., Martín-Barrio, A., Olivares-Méndez, M. A., Del Cerro, J., & Barrientos, A. (2017). Multi-robot interfaces and operator situational awareness: Study of the impact of immersion and prediction. *Sensors*, *17*(8), 1720.

Shi, R., Wei, Y., Qin, X., Hui, P., & Liang, H.-N. (2023). Exploring gaze-assisted and hand-based region selection in augmented reality. *Proceedings of the ACM on Human-Computer Interaction*, *7*(ETRA), 1–19.

Szafir, D. (2019). Mediating human-robot interactions with virtual, augmented, and mixed reality. Virtual, Augmented and Mixed Reality. Applications and Case Studies: 11th International Conference, VAMR 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21.

Walker, M. E., Hedayati, H., & Szafir, D. (2019). Robot teleoperation with augmented reality virtual surrogates. 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

Wang, X., Shen, L., & Lee, L.-H. (2025). A Systematic Review of XR-Enabled Remote Human-Robot Interaction Systems. *ACM Computing Surveys*.