Toward Mixed-Initiative Assistance for Real-World Procedures

Riku Arakawa
4th-year Ph.D. student, Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, USA
rarakawa@cs.cmu.edu

ACM Reference Format:
Riku Arakawa. 2025. Toward Mixed-Initiative Assistance for Real-World Procedures. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In an era where augmented reality (AR) and artificial intelligence (AI) are rapidly converging, the potential for intelligent systems to support users in performing various tasks is immense. While there has been a desire for an assistant that can guide users through complex procedures as though a human were present [7], achieving effective systems is not without challenges. In this position paper, I argue that task assistance must acknowledge and adapt to the imperfections inherent in context awareness while balancing user agency within interactions.

First, the promise of AI-powered assistive systems hinges on their ability to perceive and interpret a user's context accurately. Yet, sensor limitations, environmental variability, and computational constraints often lead to imperfect context awareness. These imperfections can result in assistance that is either too intrusive or insufficiently adaptive, ultimately diminishing the user experience.

Secondly, balancing user agency becomes critical. Users must feel in control without being overwhelmed by the system's suggestions, while the system should learn and adjust its behavior based on ongoing interactions. Achieving this equilibrium is a key challenge: how can a system be both assertive enough to provide meaningful assistance and subtle enough to respect the user's autonomy?

I begin with introducing our series of efforts in this direction: *PrISM* (Procedural Interaction from Sensing Module) project[1]. This framework[2] offers a foundation for tracking user context during procedural tasks using multimodal sensing while also providing different assistive interactions using the context such as proactive intervention and question-answering, as presented in Figure 1. Then, to tackle the two key challenges, I propose an approach to integrating these interactions into a mixed-initiative experience [2] to achieve *co-adaptation* through interactions. Finally, I share insights from deploying our PrISM assistants in healthcare settings to highlight key challenges that must be addressed for successful real-world implementation. Overall, this paper aims to foster discussions on reliable human-assistant interactions, their implementation, and their effectiveness across diverse task scenarios.

[1] https://rikky0611.github.io/projects/prism.html
[2] https://github.com/cmusmashlab/prism

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 CHALLENGES IN UNDERSTANDING USER CONTEXT DURING TASKS

There are various aspects of context, including user actions, mental states, and more. As a fundamental form of contextual information for procedural interactions, I focus on step context—identifying which step of a procedure the user is currently performing.

To evaluate how state-of-the-art models track steps, we collected a multimodal dataset of kitchen tasks, including making a sandwich and making coffee. Each procedure consists of approximately 10 atomic steps. We recruited 12 participants who performed these tasks freely (*i.e.*, choosing the order of steps themselves) while being recorded by a range of multimodal sensors: a camera, a smartwatch equipped with audio and IMU sensors, and privacy-preserving ambient sensors, including a 2D LiDAR, doppler radar, and a low-resolution thermal camera. These sensors were chosen to capture diverse environmental signals for step context inference.

We manually annotated step transitions and processed the collected data using state-of-the-art machine-learning models. For example, we used GPT-4V [1] to analyze camera data and *PrISM-Tracker* [5] to process audio and IMU data. Despite leveraging multiple modalities, the best-performing model achieved approximately 80% frame-level tracking accuracy across the four tasks on average. We observed that certain actions were difficult to capture due to occlusions or being out of the camera's field of view, while other sensors were susceptible to noise and interference (*e.g.*, a passenger making a loud sound). While further improvements in individual modality processing could be explored, these results highlight the limitations of sensing technology in reliably tracking user context.

This has significant implications for developing robust human-assistant interactions. If an assistant misidentifies the user's current step, it may provide incorrect guidance, leading to confusion or errors. Thus, designing adaptive interactions that can recover from errors and handle uncertainty is crucial for making procedural assistance more reliable.

3 IMPORTANCE OF BALANCING USER AGENCY AND SYSTEM CONTROL

Using the above step tracking, we have developed assistive interactions with varying levels of user agency: question answering [3] and proactive intervention [4]. In the Q&A interaction, the assistant responds to user queries (*e.g.*, *"What should I do next?"*), where Large Language Models (LLMs) are enhanced by step context to generate more accurate and context-aware responses. In the Observer interaction, the assistant proactively intervenes when it detects that the user is likely to make a mistake (*e.g.*, *"Have you wiped the pan?"*) by probabilistically modeling and forecasting user behavior.

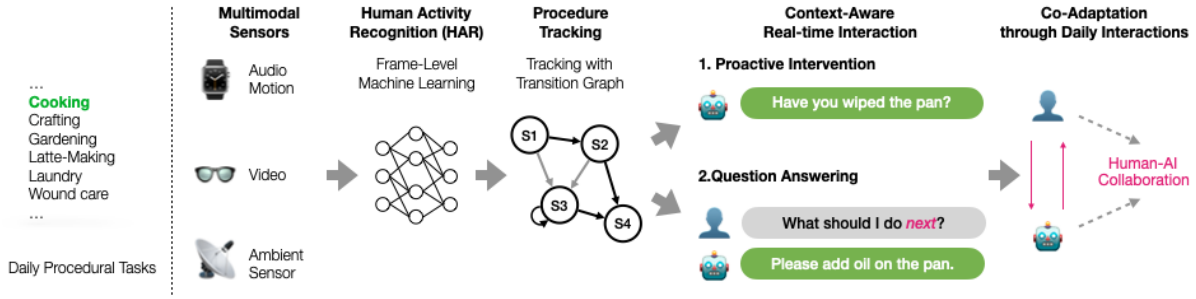

Figure 1: PrISM Framework for task assistants for procedural tasks.

Integrating these interaction modalities into a single assistant system presents a remarkable challenge, as determining the appropriate level of agency shared between humans and intelligent agents in real-time is crucial. A system that exerts too much control may leave users feeling disempowered, whereas one that provides too little assistance may fail to offer adequate support. Moreover, it is essential to design a system that allows users to navigate tasks at their own pace, override the assistant when necessary, and express disagreement when appropriate. Additionally, user needs and interaction preferences evolve over time (*e.g.*, from a beginner to a proficient user), making a fixed balance of interaction suboptimal. Thus, assistant systems should support adaptive personalization, such as adjusting the frequency of proactive interventions based on user preferences and situational demands.

4 ENABLING MIXED-INITIATIVE INTERACTION WITH FEEDBACK LOOP

To address these challenges (*i.e.*, assistant's imperfection in understanding user context and needs for adaptation to each user and scene), we enable the assistant to leverage user-assistant interactions as feedback to refine its support strategies. Specifically, we propose a method for extracting contextual information from dialogue interactions and dynamically updating the context model. For example, if a user asks, *"What should I do after washing my dish?"*, this utterance serves not only as a query for the Q&A module but also as implicit feedback about the user's current step. This approach allows the assistant to continuously refine its context understanding throughout a task session, improving tracking accuracy. Furthermore, it supports flexible initiative balancing, enabling the system to handle diverse forms of language interactions, such as Q&A, reminders, confirmations, and self-narration. During the workshop, I will introduce the framework overview and current results from multiple daily tasks such as latte-making and skin care.

5 APPLICATION EXAMPLES

A particularly promising application area for PrISM is healthcare, where the stakes are especially high. We have been collaborating with medical professionals to support post-operative care for skin cancer patients. Adhering to specific care routines is crucial, yet variations in healing processes and individual behaviors often complicate standardized protocols. Another critical application is assisting individuals with dementia in their daily routines at home. By continuously monitoring and adapting to user interactions, the system provides tailored support that helps maintain independence while ensuring safety. Our deployment of the PrISM assistant in these settings has revealed several challenges, including privacy concerns and the difficulty of responding to spontaneous patient questions due to differences between everyday language and formal medical knowledge. I will share these insights to explore potential pathways for effectively deploying task-support systems in real-world healthcare environments.

6 EXPECTED DISCUSSION

I anticipate that this position paper will spark a rich discussion at the workshop, particularly on several key issues:

Diverse Computing Platforms. : As AI systems are deployed on a range of sensor and display platforms—from wearable devices to stationary setups—the implications for feedback mechanisms and context accuracy differ significantly. How might our approach adapt across these different platforms?

Scenario-Dependent Stakes. : The balance between assistance and user control may shift dramatically based on the stakes involved. In high-stress or safety-critical situations, should the system take a more assertive role? Conversely, in lower-stakes environments, might a more reserved approach be preferable?

Cost of Training Models. : Data-driven approaches inherently come with the cost of data collection. How can we minimize it? What will the ideal end-user experience of creating a new task support?

Mixed-Initiative Dynamics. : Designing systems where control is genuinely shared between human and machine remains a formidable challenge. In this regard, Mackay [6] introduced the idea of a *human-computer partnership*, where humans and intelligent agents *collaborate* to achieve superior performance compared to working independently. What principles should guide the design of mixed-initiative interactions to ensure such partnership?

Through this discussion, I hope to inspire new research directions and collaborative efforts that will refine the proposed approach and expand its applicability, ultimately transforming how technology supports human endeavors in complex, real-world settings.

REFERENCES

[1] Open AI. 2022. ChatGPT. https://openai.com/index/chatgpt/

[2] James E Allen, Curry I Guinn, and Eric Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14, 5 (1999), 14–23.

[3] Riku Arakawa, Jill Fain Lehman, and Mayank Goel. 2024. PrISM-Q&A: Step-Aware Voice Assistant on a Smartwatch enabled by Multimodal Procedure Tracking and Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (2024), 180:1–180:26. https://doi.org/10.1145/3699759

[4] Riku Arakawa, Hiromu Yakura, and Mayank Goel. 2024. PrISM-Observer: Intervention Agent to Help Users Perform Everyday Procedures Sensed using a Smartwatch. In *[conditionally accepted] Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST 2024, Pittsburgh, PA, USA, October 13-16 2024*. ACM. https://doi.org/10.1145/3654777.3676350

[5] Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. 2022. PrISM-Tracker: A Framework for Multimodal Procedure Tracking Using Wearable Sensors and State Transition Information with User-Driven Handling of Errors and Uncertainty. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4 (2022), 156:1–156:27. https://doi.org/10.1145/3569504

[6] Wendy E. Mackay. 2023. Creating Human-Computer Partnerships. In *Computer-Human Interaction Research and Applications - 7th International Conference, CHIRA 2023, Rome, Italy, November 16-17, 2023, Proceedings, Part I (Communications in Computer and Information Science, Vol. 1996)*. Springer, 3–17. https://doi.org/10.1007/978-3-031-49425-3_1

[7] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 2023. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 20213–20224. https://doi.org/10.1109/ICCV51070.2023.01854