AgentCraft: Workshop on Developing Trustworthy Agentic AI Systems

Karthik Dinakar
Department of Medicine
Icahn School of Medicine at Mount Sinai
NY, New York, USA
karthik.dinakar@mountsinai.org

Justin D. Weisz
IBM Research
Yorktown Heights, New York, USA
pcs@jweisz.org

Henry Lieberman
CSAIL
MIT
Cambridge, Massachusetts, USA
lieber@media.mit.edu

Werner Geyer
IBM Research
Cambridge, Massachusetts, USA
werner.geyer@gmail.com

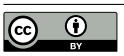
Abstract

Ambitious efforts are underway to build AI agents powered by large language models (LLMs) across a spectrum of domains, including programming, productivity, creative work, and enterprise automation. Although many frameworks have emerged to aid in the agent development process, major challenges remain regarding agent autonomy, reasoning processes, unpredictable behaviors, and consequential actions. Developers struggle to comprehend and debug their agent's behavior in systematic ways. They also struggle to determine where in an agent's workflow human input and oversight are needed. Intelligent interfaces that enable meaningful oversight of agentic plans, decisions, and actions are needed to foster transparency, build trust, and manage complexity. Further, evaluations of agentic AI systems require moving past automated machine learning benchmarks toward human-centric evaluations that demonstrate real-world usefulness, reliability, and safety.

This workshop will bring together researchers, developers, and designers to explore approaches to address these challenges of building human-centered and trustworthy agentic AI systems. We will explore novel user interfaces for (1) mixed-initiative collaboration including developers and end-users, during agent development and deployment, (2) design patterns for debugging agent behaviors, (3) strategies for determining when developer control and oversight are needed, and (4) evaluation methods that ground agent performance in real-world impact. The workshop will convene a diverse community from academia and industry in the form of a day-long workshop at IUI 2026.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; *Interaction design*; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → *Arts and humanities.*



This work is licensed under a Creative Commons Attribution 4.0 International License.
IUI Companion '26, Paphos, Cyprus
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1985-1/26/03
https://doi.org/10.1145/3742414.3794957

Keywords

Generative modeling, artificial intelligence, generative design, user experience, collaboration

ACM Reference Format:
Karthik Dinakar, Justin D. Weisz, Henry Lieberman, and Werner Geyer. 2026. AgentCraft: Workshop on Developing Trustworthy Agentic AI Systems. In *Companion Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI Companion '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3742414.3794957

1 Description

Since late 2024, advances in LLM-based agents have triggered efforts to build AI systems that take action towards achieving user-defined goals, rather than just responding to isolated prompts, under the banner of *Agentic AI*. Such agents are often endowed with access to external tools such as the ability to navigate software, write and execute code. In software development, agentic coding assistants (e.g., Cursor, Windsurf, Kilo Code, Claude Code) can autonomously draft code, run tests, and debug, consulting with the human developer only when needed. In productivity enterprise applications, agents can orchestrate complex workflows (e.g., scheduling, information retrieval, and communications) on behalf of users. To make it easier to build such agents, startups and major tech companies have released numerous interactive development environments (IDEs) to simplify the process of building agentic LLM applications (e.g. AutoGen [7], CrewAI [3], LangGraph [6], BeeAI [5]).

Because LLM agents operate probabilistically, developers must contend with non-deterministic behavior. Without an appropriate level of control, people may hesitate to trust agents for high-stakes tasks [10]. Giving users a sense of control can help overcome algorithmic aversion [2]. However, designing such human-in-the-loop or agent-in-the-loop safeguards is a nascent aspect of software engineering for AI. Developers need to reason about failure modes and governance mechanisms such as confirmation prompts, and policy checks.

Developers need tools to inspect an agent's reasoning process and state, visualize agent behaviors, simulate scenarios in sandboxed environments, and profile an agent's performance. Tools such as AutoGen Studio [8], a no-code tool for developing agents, provides a visual interface where developers can drag-and-drop

agent components, run simulations, and observe each agent's actions and messages with real-time metrics. These intelligent interfaces are needed not just for end-users, but also for creators of agentic systems to test and refine their designs.

Finally, the evaluation of agentic AI systems remains an open challenge. Traditional machine learning benchmarks (e.g., accuracy, precision, recall, win rates, etc.) often fail to capture whether an agent is truly useful or trustworthy in real-world workflows. Human-centered evaluation frameworks are emerging to fill this gap. For example, researchers have proposed an "agentic ROI" metric that factors in the time and effort an agent saves a user versus the overhead it introduces. Task-completion rates alone are insufficient without considering how often human intervention was needed or whether the user felt confident in the agent's decisions. Recent work in HCI has begun studying team performance and trust in human–LLM agent teams. For instance, He et al. [4] found that a plan-assist-execute paradigm, where the LLM agent proposes a plan, the user edits/approves it, and then the agent executes step-by-step, led to better outcomes and calibrated trust. Users benefited from the agent's efficiency when the agent's plan was sound and they were appropriately involved; but if the agent's plan looked superficially plausible, users often mistrusted it unless they had a chance to verify or tweak it. Evaluation must consider end-user experience: are users able to understand the agent's decisions? Do they trust it appropriately (neither over- nor under-trusting)? Does the agent actually save them effort in practice? Answering these questions requires new methodologies, from user studies and case reports to human-in-the-loop simulations and deployment studies in real settings.

1.1 Workshop areas of focus

Key questions and areas of focus include:

- **Interfaces for building agentic AI**. This theme focuses on designing intuitive tools that let developers create, configure, and supervise LLM-based agents. We invite work on low- or no-code builders, interactive planning environments, dashboards that chart agent actions and workflows, and multimodal debugging aids that expose state and reasoning traces. The goal is to move beyond simple chat logs so that agents can be monitored, debugged, and guided without requiring deep technical knowledge.
- **Transparency, observability, and debugging**. How can developers grasp what an agent is doing and why? Contributions should propose techniques for exposing intermediate reasoning, generating explanations of decisions, and providing real-time observability of operations. We are interested in how transparency features influence trust and reliance, and in UI patterns such as visual cues or notifications that alert users to status changes or potential errors. Case studies of debugging failures and methods for agent-state introspection are welcome.
- **Human oversight, trust, and control**. This theme invites research on when an agent should defer to a user, request confirmation, or seek feedback. We seek designs for strategic guidance at planning time, and for continuous control during execution (e.g., pause/undo). Studies on signaling the agent's confidence, limits, or need for help so that users can regain control quickly, without micromanaging, are especially valuable.
- **Evaluation and real-world usefulness**. Beyond static benchmarks, we need evaluation methods that involve real users, tasks, and deployments. Proposals for metrics that capture aspects of human-agent team performance, trustworthiness, and safety are encouraged. Papers may present lab studies, field trials, simulation-based human-in-the-loop evaluations, or analytical frameworks. Case studies are highly sought that demonstrate an agent's impact, such as by changing workflows in an organization or safety-critical use of agents in areas such as healthcare or finance.

2 Previous History

- **HAI-GEN 2020-2025** (https://hai-gen.github.io/2024). Two authors were co-organizers of the HAI-GEN workshop series at IUI, which ran for 6 years. This workshop steadily increased in participation over the years, attaining about 50-60 unique participants in 2025. Agentic AI became an emerging topic at the last workshop.
- **GenAI in the Enterprise: Trauma to Success, 2024.** This day-long symposium at MIT was organized by Dinakar and keynoted by Lieberman. It featured strong participation from industry and government, including Sky UK, Dell, Comcast, GraphCore, and several US and European government agencies. Notably, participants shared candid narratives about both failures and successes in real-world projects, inspiring the audience. It is a model for our outreach to industry and application areas.
- **MIND 2025** (https://pienso.com/events/mind-mixed-initiative-next-gen-design-iui-2025). Two authors were co-organizers of the MIND workshop series at IUI 2025. This workshop received enthusiastic participation in 2025, securing about 30+ unique participants in 2025. The workshop covered mixed-initiative approaches and agentic automation.

3 Organizers

Karthik Dinakar is co-founder and CTO of Pienso Inc. He was the co-organizer of the MIND (Mixed Initiative Next-gen Design Workshop at IUI 2025 which explored the implications of the "Agents vs. Direct Manipulation" debate for making auto-regressive LLMs end-user accessible. *karthik@pienso.com*

Justin D. Weisz is a Senior Research Scientist, Manager, and Strategy Lead for Human-Centered Trustworthy AI at IBM Research. Dr. Weisz's research focuses on understanding how to design AI systems that amplify and augment human capabilities. He was a co-organizer of HAI-GEN @ IUI (2021-2025). https://research.ibm.com/people/justin-weisz, *jweisz@us.ibm.com*.

Henry Lieberman is a Research Scientist at MIT CSAIL, whose contributions span the fields of programming languages, artificial intelligence and human-computer interaction. He was twice Program Chair of IUI, and is one of the most prolific authors of IUI papers. He also served a term on the Executive Council (Board of Directors) of AAAI. He received the 2018 IUI Impact Award. https://web.media.mit.edu/~lieber/, *henry@csail.mit.edu*

Werner Geyer is a Principal Research Scientist, ACM Distinguished Member at IBM Research in Cambridge, MA. He has held various roles as co-chair at ACM RecSys. He has been co-organizer of HAI-GEN since 2020. His recent research interests include the use of generative AI in business settings and understanding how to effectively design AI-assisted agentic workflows to to augment peoples' abilities. https://wernergeyer.com, *werner.geyer@us.ibm.com*

4 Workshop Program Committee

Victor Dibia Microsoft Research	Thariq Shihipar Anthropic (Claude Code)
Simret Gebreegziabher Notre Dame University	Sebastian Gehrmann Bloomberg
Harang Ju Johns Hopkins University	Steve Ross IBM Research
Daniel Karl Weidele IBM Research	Patrick Xian Northeastern University
Zhiping Zhang Northeastern University	Brian Cort Pienso Inc
Arnav Nagpal Pienso Inc	Sonia Wu Pienso Inc
Mathew Maradin Pienso Inc	

5 Participants

The target audience for this workshop are researchers in HCI, AI, and cognitive science, industry practitioners and engineers building AI agent platforms, interaction designers exploring new paradigms for AI interfaces, and importantly, developers who create tools that enable end-users to build or customize agents. We are making a concerted effort to reach out to companies providing agent development environments, to companies exploring prototype and pilot projects using Agentic AI, and to end-users who hope to incorporate Agentic AI in their work. In this way, we will achieve a better academic/industry balance than is typical for a workshop at an academic conference.

We expect 50-60 participants, given the urgency of this topic. We will announce our workshop through major ACM HCI and AI mailing lists (e.g., chi-announcements, iui-announcements, ml-news@googlegroups.com), our PC, and the organizers' internal organizational channels (IBM Research, MIT CSAIL, PIENSO, Berkman-Klein).

6 Workshop Format & Length

The workshop will take the form of a **full-day mini conference** with an inspiring keynote speaker, paper presentations, and poster and demo sessions. We will provide time for discussion, enabling participants to learn, connect, and plan for future collaboration.

7 Planned Outcomes

Expected outcomes include: (1) a set of identified best practices or guidelines for designing human-agent interactions (we might collectively draft a list during the workshop, inspired by existing human-AI interaction guidelines (e.g. [1, 9]) but updated for agentic contexts); (2) an increased awareness of different evaluation frameworks, possibly leading to a follow-up working group on creating shared evaluation tasks or datasets for human-in-the-loop agent testing; (3) new collaborations between attendees (we will encourage participants to discuss potential joint projects or cross-institution collaborations as part of the breakout report-back). We also aim to publish workshop proceedings (likely in the ACM Digital Library adjunct proceedings or a CEUR workshop volume).

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[2] Jingyi Bo, Shirley Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3706598.3712345 Forthcoming in CHI '25.

[3] CrewAI. 2025. CrewAI: The Leading Multi-Agent Platform. https://www.crewai.com

[4] Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan–Then–Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents as a Daily Assistant. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3706598.3713218

[5] IBM. 2025. BeeAI Framework. https://github.com/i-am-bee/beeai-framework

[6] LangChain. 2025. LangGraph. https://www.langchain.com/langgraph

[7] Microsoft. 2025. AutoGen: A Framework for Building AI Agents and Applications. https://microsoft.github.io/autogen/stable

[8] Microsoft. 2025. AutoGen Studio. https://microsoft.github.io/autogen/dev/user-guide/autogenstudio-user-guide/

[9] Justin D Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[10] Ziang Xiao, Wesley Chan, Angela Zhang, and Haiyi Zhu. 2025. Plan-Then-Execute: An Empirical Study of User Trust and Reliance on LLM Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 593, 19 pages. doi:10.1145/3706598.3713218