Carrier: Infrastructure for Studying Agentic Automation Experiences

Jiahao Eric Yang*

Institute for Digital Security and Behaviour, University of Bath, jy2154@bath.ac.uk

Will Smith

Institute for Digital Security and Behaviour, University of Bath, ws833@bath.ac.uk

Agentic AI is shifting workplace automation from discrete tools to proactive systems where multiple specialised agents coordinate behind a single interface, initiate actions, and solve complex tasks with little human input. This shift challenges organisations to understand when multi-agent autonomy improves outcomes versus creates brittle handoffs, automation bias, or misplaced reliance; how sustained delegation changes perceived agency and skill across individuals; and what oversight interfaces best support intervention and recovery. Progress is limited by the lack of reusable study infrastructure that can manipulate agentic variables, implement individual-differences logic, and capture standardised behavioural process data. We present Carrier (https://www.carrierlab.org), an open-source browser-based platform for controlled, repeatable multi-party human–agent studies. Carrier separates participant type (human, LLM, scripted bot) from role (Communicator, Mediator, Processor), supports survey-driven adaptive agent assignment, and logs traceable interaction and drafting events for mechanism-focused analysis.

CCS CONCEPTS • Human computer interaction (HCI) • Interaction paradigms • Web-based interaction

Additional Keywords and Phrases: Text/Speech/Language, Prototyping/Implementation, Agentic AI; Multi-agent systems; Human–AI collaboration; Automation and delegation; Workplace automation; Experimental platforms; Behavioral process tracing

ACM Reference Format:

Jiahao Eric Yang, Will Smith. 2026. Carrier: Infrastructure for Studying Agentic Automation Experiences. In Proceedings of AutomationXP26 Workshop of the 2026 CHI Conference on Human Factors in Computing Systems, April 14, 2026, Barcelona, Spain. ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Recent advances in AI are moving workplace automation from discrete tools to agentic systems: multiple specialised agents coordinate behind a single interface, initiate actions, and complete complex organisational tasks with little human input. As humans shift from execution to oversight and intervention, three questions follow. First, when does multi-agent autonomy improve work quality, and when does it introduce failure modes—brittle handoffs, misaligned sub-goals, or automation bias—that erase gains? Human–AI teams can underperform the best human or AI alone, and trust- or proxy-task evaluations can mislead [3,8]. Fluency and persuasive rationales can further inflate reliance even when underlying coordination is opaque or wrong [1]. Second, how does routine delegation reshape cognition, perceived agency, and competence over time—and for whom? Neurocognitive findings from repeated LLM-assisted writing point to altered engagement consistent with accumulated "cognitive debt" [6], while classic work documents out-of-the-loop risks, including reduced situation awareness and skill decay under high autonomy [4]. These effects are heterogeneous: perceived agency varies with automation level [7], and reliance depends on dispositional and learned factors across roles, ages, and expertise [5]. Third—and practically, building on the first two—what interface and control policies keep people oriented, able to intervene and recover, and able to maintain skill without forfeiting autonomy's efficiency gains? Human-centred and mixed-initiative guidance offers starting points [2], but it needs testing under proactive, multi-agent dynamics.

Answering these questions requires study infrastructure that can (i) manipulate multi-agent configurations and autonomy policies, (ii) incorporate individual differences into system behaviour under explicit condition logic, and (iii) capture behavioural process measures consistently over repeated exposure. In practice, most human–agent studies still rely on bespoke, single-use systems that are difficult to reuse, inconsistently instrumented, and poorly suited to factorial manipulation of core agentic variables (composition, role structure, disclosure/visibility, protocol constraints, and channel). Individual-differences designs are rarely supported end-to-end—for example, assigning different agent roles or autonomy policies based on validated measures while preserving experimental control and comparability. Likewise, standardised process data (timing, edits, overrides, acceptance/rejection, escalation) is often missing or idiosyncratic, limiting mediation analyses and longitudinal tests of agency and skill change.

We present Carrier (https://www.carrierlab.org), an open-source, browser-based platform for controlled, repeatable studies of multi-party human–agent interaction. A study is specified as chamber lines (condition paths) composed of chambers (synchronous interaction episodes), each with configurable matching, timing, and pre/post measurement. Within each chamber, Carrier separates participant type (human participant, LLM agent, scripted bot) from participant role, allowing agents to act as Communicators (in-stream contributions), Mediators (broadcast coordination), or Processors (draft-time transformation). This separation makes attribution of influence—who acted, when, and through which channel—experimentally tractable. Carrier supports individual-differences designs by using survey measures for assignment and matching, enabling trait- or expertise-contingent agent behaviour under explicit condition logic. Finally, the platform provides standardised interaction traces (timestamped events) and protocol controls (turn-taking, phases, gating), supporting systematic tests of autonomy boundaries and oversight mechanisms.

2 CARRIER DESIGN STUDYING AGENTIC AUTOMATION EXPERIENCES

Carrier is designed as study infrastructure for experiments in which multiple humans collaborate alongside multiple AI agents—potentially across channels (e.g., text, audio)—within the same interaction episode. Its core design goal is to make automation structure manipulable and measurably comparable—who has initiative, what authority each actor holds, what is disclosed, how actions are routed through the interface, and how humans can intervene, override, or recover. This

supports factorial manipulations of collaboration topology—dyads vs. triads, single-agent vs. multi-agent groups, disclosed vs. blinded agent participation—without rebuilding the study stack for each design.

2.1 A role-based model: Type × Role

Carrier organises agentic automation as Type × Role: *type* identifies the actor (human participant, LLM agent, scripted bot) and *role* specifies where and how automation enters the workflow—in-stream interaction on the primary task surface, coordination-layer intervention, or back-stage drafting support before commitment. This reduces role ambiguity and makes "who did what" experimentally tractable. Practically, it lets researchers flexibly "slot" different agent implementations into the same role (or compare roles using the same agent), vary role assignments by condition or by participant characteristics, and test how shifting an identical capability between Communicator, Mediator, and Processor changes oversight, attribution, and intervention under different autonomy boundaries.

Communicator. Communicators participate directly in the shared stream using the same interaction surface as humans. Carrier implements interaction parity so human and agent communicators appear in the same space, enabling controlled comparisons across human-only, human–agent, and mixed compositions while holding the interface constant. Communicator behaviour can be shaped via protocol controls and agent modes: message sending can be gated by time or content (including LLM-based checks), externally controlled to enforce turn-taking or structured phases, and an agent can switch within a conversation between scripted responses and LLM-driven generation while remaining governed by the same constraints.

Mediator. Mediators coordinate group activity without being treated as peer contributors. They observe interaction with configurable visibility and communicate via broadcast interventions distinct from peer chat. In agent-mediator configurations, mediators can deliver structured coordination actions (e.g., timed prompts, attention cues, turn-taking enforcement) and respond to discussion dynamics using triggers such as periodic schedules, phase boundaries, or activity timeouts. This supports systematic tests of facilitation and governance strategies as automation mechanisms.

Processor. Processors operate at the drafting layer, shaping messages or artefacts before they enter the shared stream. They work in the composition space and assist communicators through modes such as review, rewriting, or generation, while preserving human control over what is ultimately posted or executed. This enables studies of back-stage influence—especially disclosure vs non-disclosure, delegation and reliance—by varying when, how strongly, and how transparently processing support is available within an interaction episode.

2.2 Study flow: chambers, matching, and run plans

Carrier standardises the workflow from experiment authoring to recruitment and completion. Researchers create studies in the visual experiment builder by defining a sequence of chambers (atomic synchronous interaction episodes) and grouping them into chamber lines that implement condition paths. For each chamber, researchers specify participant slots (how many humans and agents), assign roles, set interaction rules (e.g., timing/turn-taking), and attach optional pre/post measures. Participants then join via a study URL; Carrier assigns them to a chamber line and generates a persisted run plan that records the exposure sequence, progress markers, and completion status.

Multi-party interaction is enabled through Carrier's matching and room creation process. For each chamber, human participants enter a matching phase where Carrier forms the required group using eligibility criteria, timeouts, and fallbacks. Once the group is formed, Carrier instantiates the room and introduces agents according to the chamber specification, enabling reliable staging of mixed groups while retaining realistic human arrival and dropout patterns. Survey responses can drive **dynamic matching and adaptive agent assignment**: eligibility rules can use participants' latest

measures to form groups and to allocate/configure agent roles for individuals with different characteristics (e.g., expertise or trust propensity), while logging the rules and outcomes to preserve an auditable record of who was matched and why.

2.3 Instrumentation and auditability

Carrier reduces the replication burden typical of bespoke multi-party systems by standardising instrumentation and auditability. It logs interaction events, role states, and agent prompts/outputs in a consistent schema, enabling cross-study comparison and post hoc reconstruction of what agents received and produced—critical when small prompt or context differences can shift outcomes. Carrier supports behavioural process measures for human–AI interaction: draft histories, acceptance/rejection of suggestions, time-on-draft, rewrite/undo patterns, frequency of agent calls, alongside standard timing and turn-taking measures in the shared stream. Operationally, Carrier integrates authoring and deployment through an Experiment Builder (study specification), a participant runtime (execution), and an Experimenter Dashboard (monitoring and intervention). The dashboard provides live tracking of participant status and chamber progress and records experimenter actions (pause/skip/end), preserving an audit trail for protocol deviations and failure recovery.

3 CONCLUSION, LIMITATIONS AND FUTURE WORK

Carrier makes agentic automation experiences a configurable, testable object rather than a one-off engineering outcome. By decomposing sessions into chambers and role-based participation—Communicator, Mediator, and Processor—Carrier enables systematic manipulation of autonomy boundaries (who initiates, who acts, who approves), where automation enters the workflow, and what is disclosed, while holding the surrounding infrastructure constant. Its study primitives (matching, chamber lines, run plans) support realistic group formation and repeatable exposure sequences, allowing experiments to move beyond dyadic "AI vs. no AI" contrasts toward controlled tests of role configuration, disclosure, protocol design, and communication channel. Carrier's standardised logging—including agent prompts and outputs—supports traceability and cross-study comparison, addressing a major limitation of bespoke systems. Collectively, these capabilities lower the barrier to mechanism-focused experiments on perceived agency, responsibility attribution, trust calibration/appropriate reliance, out-of-the-loop effects (skill retention), and error detection and recovery in multi-agent organisational automation.

Carrier remains under active development, and the current prototype has several limitations. First, we do not yet provide support for real-time AI video synthesis, which means video conditions are currently restricted to human–human interaction. Second, internet-based deployment introduces timing and latency variability that can constrain perceptual or fine-grained behavioural measures unless carefully designed and validated [9]. Finally, while Carrier reduces study build effort, running a research-grade deployment still requires operational competence (e.g., server setup, monitoring, and security hardening).

Future work will focus on improving validity, portability, robustness, and safety. We are currently recruiting the first batch of internal users to run pilot studies that surface usability and reliability issues in realistic workflows. Methodologically, we plan to expand reusable study recipes, publish reference paradigms, and provide evaluation suites to support benchmarking across studies. Technically, we will extend multimodal capabilities, improve scalability and session quality, and strengthen interoperability with standard recruitment, consent, and data-management workflows.

REFERENCES

1. Rohan Ajwani, Shashidhar Reddy Javaji, Frank Rudzicz, and Zining Zhu. 2024. LLM-Generated Black-box Explanations Can Be Adversarially Helpful. https://doi.org/10.48550/arXiv.2405.06800
2. Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233
3. Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In Proceedings of the 25th International Conference on Intelligent User Interfaces, 454–464. https://doi.org/10.1145/3377325.3377498
4. Mica R. Endsley and Esin O. Kiris. 1995. The Out-of-the-Loop Performance Problem and Level of Control in Automation. Human Factors: The Journal of the Human Factors and Ergonomics Society 37, 2: 381–394. https://doi.org/10.1518/001872095779064555
5. Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors: The Journal of the Human Factors and Ergonomics Society 57, 3: 407–434. https://doi.org/10.1177/0018720814547570
6. Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv preprint arXiv:2506.08872 4. Retrieved February 21, 2026 from https://collimateur.uqam.ca/wp-content/uploads/sites/11/2025/12/2506.08872v1_comp.pdf
7. Sayako Ueda, Ryoichi Nakashima, and Takatsune Kumada. 2021. Influence of levels of automation on the sense of agency during continuous action. Scientific Reports 11, 1: 2436. https://doi.org/10.1038/s41598-021-82036-3
8. Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour 8, 12: 2293–2303. https://doi.org/10.1038/s41562-024-02024-1
9. Andy T. Woods, Carlos Velasco, Carmel A. Levitan, Xiaoang Wan, and Charles Spence. 2015. Conducting perception research over the internet: a tutorial review. PeerJ 3: e1058. https://doi.org/10.7717/peerj.1058