What Happened and Why? Trace-Guided Micro-Episodes with Elicited User Explanations for Product Iteration

SIRUI TAO*, UC San Diego, USA
WILLIAM P. MCCARTHY, Autodesk AI Lab, USA
STEVEN P. DOW†, UC San Diego, USA

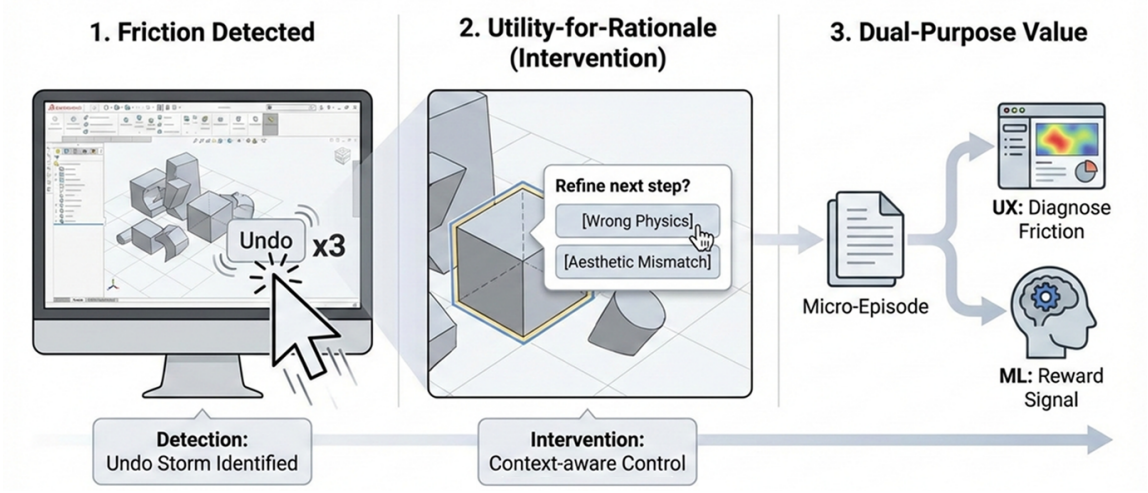

Fig. 1. The Micro-Episode Lifecycle. (1) High-friction moments (e.g., "undo storms") trigger (2) a **Utility-for-Rationale** intervention that aids the user while capturing intent. (3) The resulting micro-episodes provide vulnerable insights for both UX and ML teams.

ABSTRACT

Teams shipping AI workflows in design tools can measure usage yet often struggle to explain *why* features fail. In creative work, standard metrics are ambiguous: a long session could imply productive exploration or frustrating struggle with stochastic outputs. We argue for *trace-guided micro-episodes*, a unit of analysis binding interaction logs—what users did—to their intent. Rather than relying on disruptive surveys, we propose a "utility-for-rationale" paradigm: systems offer optional, context-aware controls at likely friction points, capturing user explanations as a byproduct of real-time error recovery. This approach converts ambiguous telemetry into causal evidence without breaking flow. We posit this methodology serves a dual purpose: equipping teams with diagnostic clarity to iterate on vague failure modes (e.g., controllability vs. quality) while generating the grounded alignment data required to train future agents.

CCS Concepts: • **Human-centered computing** → **User models**; **Activity centered design**; *Interactive systems and tools*; *Contextual design*; • **Computing methodologies** → *Artificial intelligence.*

*Corresponding author.
†Principle Investigator.

Authors' Contact Information: Sirui Tao, UC San Diego, La Jolla, CA, USA, s1tao@ucsd.edu; William P. McCarthy, Autodesk AI Lab, San Francisco, CA, USA, william.mccarthy@autodesk.com; Steven P. Dow, UC San Diego, La Jolla, CA, USA, spdow@ucsd.edu.

2026. Manuscript submitted to ACM

Additional Key Words and Phrases: creativity support tools, causal evaluation, interaction traces, generative AI, design workflows

ACM Reference Format:
Sirui Tao, William P. McCarthy, and Steven P. Dow. 2026. What Happened and Why? Trace-Guided Micro-Episodes with Elicited User Explanations for Product Iteration. In *CHI 26 Workshop "Herding CATs: Making Sense of Creative Activity Traces", April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

When teams deploy AI in complex creative software—such as Computer-Aided Design (CAD)—usage logs often mislead. High acceptance rates for a feature like "smart auto-connect" suggest success, yet users may accept the connection only to spend minutes manually repairing the resulting geometry errors. This ambiguity grows in GenAI workflows where outputs are non-deterministic: a generation that appears "successful" in logs often shifts user effort from productive modeling to tedious verification. This challenge intensifies as tools shift toward *generative interfaces* [7, 9]. Because these interfaces are dynamically composed per session, static logs lose the "semantic anchor" required to explain user actions.

This gap in evidence is sharpest during early pilots and user studies, where data is scarce. Industry pilots and HCI tool evaluations both operate with constrained cohorts and heterogeneous tasks [10, 11]. While benchmarks attempt to standardize tasks [8], teams observing open-ended usage face an interpretation bottleneck: *why* did behavior change? We argue for a standard evaluation package that (i) captures scalable traces, (ii) strategically elicits brief clarifications *only* when ambiguity is high, and (iii) synthesizes evidence for iteration [6, 15]. We instantiate this as *trace-guided micro-episodes*: short trace windows anchored to consequential moments, paired with minimal, user-controlled clarification.

2 Related Work

Evaluating creative and AI workflows. Evaluations of creativity support tools involve multiple objectives and resist single metrics [10]. Recent work cautions that studies often overemphasize short-term productivity while under-measuring outcomes like learning [11]. In production AI, adoption hinges on user-facing costs—controllability, transparency, verification—often missed by aggregate metrics [3]. Within GenAI workflows, process-level effects (explore, verify) are hard to interpret from outcomes [17], motivating the need for diagnostic evidence to guide iteration, rather than just aggregate scores for reporting [2].

Connecting traces to intent. While telemetry scales, traces show *what* happened without clarifying *why* [12]. Prior work improves interpretability through visualization [12, 14] and evolution graphs [16]. Approaches like Think-Aloud Computing [6] and ClearFairy [15] capture intent near action. Our focus differs in mechanism: rather than supporting the ongoing task, we frame brief clarification as a low-friction *recovery mechanism* anchored to breakdowns. We *structure* the resulting micro-episodes to specifically inform product iteration, helping teams disambiguate failure modes.

3 Position: An Evaluation Stack for Messy Creative Work

We propose an evaluation stack producing *micro-episodes*: compact records connecting *what happened* to *why* through minimal, user-controlled clarification. The stack operates in three stages:

Observation. The stack records traces and assistance touchpoints (e.g., generations, undo/redo). In GenAI tools, this includes prompts, re-ranking, and branching [16] alongside outcomes [17]. Crucially, the system segments traces into

candidate moments signaling friction: *undo storms* (rapid reversals), *oscillations* (toggling), or *abandonment*. For each, it logs the interface state to reconstruct the user's view.

Clarification. Moments often remain ambiguous: undoing a suggestion may reflect low quality, hidden constraints, or goal changes. Instead of disruptive surveys, the stack anchors a non-intrusive "Clarify / Fix" utility to the friction point. For instance, after a rapid undo, a transient tooltip might appear asking: *"Wrong shape or wrong style?"* Selecting an option immediately adjusts the next generation. By framing clarification as a *mechanism for control*, the system captures a diagnostic micro-episode (trace + state + rationale) as a byproduct of error recovery.

Synthesis. Micro-episodes reveal recurring failure modes. The stack aggregates patterns, using AI to cluster traces and clarifications into UX-facing workflow maps and system-facing breakdown sets. This complements visualization [12, 14] by adding causal evidence only where traces are ambiguous, helping teams prioritize fixes for issues like controllability or verification burden.

4 Discussion

4.1 Compatibility with causal evaluation designs

The stack improves interpretability but does not alone establish causality. However, it complements rigorous designs: whether randomizing UI variants [5] or using micro-randomized trials to estimate proximal effects [4], micro-episodes help explain observed differences and surface actionable breakdowns to inform what to change next.

4.2 Generalizability beyond CAD

While CAD offers precise anchoring, the evaluation problem is universal. Across writing, editing, and coding, people backtrack and verify; GenAI amplifies this via stochasticity. The stack applies wherever tools expose identifiable touchpoints. Domain adaptation involves defining episode schemas matching the unit of work and selecting conservative moments where clarification reduces ambiguity.

4.3 From Evaluation to Agent Training

Beyond product iteration, micro-episodes address the "credit assignment" problem in agent training. As current models exhaust static web data and move into the "era of experience" [13], they require grounded interaction signals. However, SOTA methods like SWEET-RL [18] and DigiRL [1] currently rely on synthetic proxies—such as unit tests or VLM evaluators—to approximate these signals. Micro-episodes bridge this gap by providing the *ground-truth* rationale (e.g., "rejected due to physics") and state snapshots required to train robust Reward Models and Critics, replacing noisy heuristics with explicit human reasoning.

4.4 Limitations and future directions

Poorly chosen candidate moments can distract users, and clarification may bias behavior. Next steps include exploring diverse heuristics for identifying candidate moments to minimize disruption, validating whether micro-episodes reduce analysis time without degrading insight quality, and building analyst tools to translate episode clusters into concrete product changes.

Acknowledgments

We thank Shm Garanganao Almeda, Howard Ziyu Han, Prof. Nikolas Martelaro, and Frederic Gmeiner for insightful discussions that helped shape this work in its early stages. We also thank Tony Li, Matthew Beaudouin-Lafon, Prof. William G. Griswold, and Devon Tao for their helpful feedback during the internal talk presentation.

We produced all original writing in this manuscript and used ChatGPT to assist with shortening and revising the text, as well as supporting literature searches. Nano-Banana was used for visualization.

References

[1] Hao Bai, Yifei Zhou, Jiayi Pan, Mert Cemri, Alane Suhr, Sergey Levine, and Aviral Kumar. 2024. Digirl: Training in-the-wild device-control agents with autonomous reinforcement learning. *Advances in Neural Information Processing Systems* 37 (2024), 12461–12495.

[2] Frederic Gmeiner, Jamie Lynn Conlin, Eric Handa Tang, Nikolas Martelaro, and Kenneth Holstein. 2024. An Evidence-based Workflow for Studying and Designing Learning Supports for Human-AI Co-creation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 42, 15 pages. doi:10.1145/3613905.3650763

[3] Aman Khullar, Nikhil Nalin, Abhishek Prasad, Ann John Mampilli, and Neha Kumar. 2025. Nurturing Capabilities: Unpacking the Gap in Human-Centered Evaluations of AI-Based Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 101, 18 pages. https://doi.org/10.1145/3706598.3713278

[4] Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. 2015. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology* 34, S (2015), 1220.

[5] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.* 18, 1 (Feb. 2009), 140–181. doi:10.1007/s10618-008-0114-1

[6] Rebecca Krosnick, Fraser Anderson, Justin Matejka, Steve Oney, Walter S. Lasecki, Tovi Grossman, and George Fitzmaurice. 2021. Think-Aloud Computing: Supporting Rich and Low-Effort Knowledge Capture. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 199, 13 pages. doi:10.1145/3411764.3445066

[7] Yaniv Leviathan, Dani Valevski Matan Kalman, Danny Lumen, Eyal Segalis Eyal Molad, Shlomi Pasternak, Vishnu Natchu, Valerie Nygaard, and Srinivasan Cheenu Venkatachary James Manyika Yossi Matias. 2025. Generative UI: LLMs are Effective UI Generators. *github* (2025).

[8] William P McCarthy, Saujas Vaduguru, Karl Dd Willis, Justin Matejka, Judith E Fan, Daniel Fried, and Yewen Pu. 2025. mrCAD: Multimodal Communication to Refine Computer-aided Designs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 22905–22921.

[9] Bryan Min, Allen Chen, Yining Cao, and Haijun Xia. 2025. Malleable Overview-Detail Interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.

[10] Christian Remy, Lindsay MacDonald Vermeulen, Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. 2020. Evaluating Creativity Support Tools in HCI Research. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 457–476. doi:10.1145/3357236.3395474

[11] Samuel Rhys Cox, Helena Bøjer Djernæs, and Niels van Berkel. 2025. Beyond Productivity: Rethinking the Impact of Creativity Support Tools. In *Proceedings of the 2025 Conference on Creativity and Cognition (C&C '25)*. Association for Computing Machinery, New York, NY, USA, 735–749. doi:10.1145/3698061.3726924

[12] Jeffrey Rzeszotarski and Aniket Kittur. 2012. CrowdScape: interactively visualizing user behavior and output. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 55–62. doi:10.1145/2380116.2380125

[13] David Silver and Richard S Sutton. 2025. Welcome to the era of experience. *Google AI* 1 (2025).

[14] Amy Smith, Barrett R Anderson, Jasmine Tan Otto, Isaac Karth, Yuqian Sun, John Joon Young Chung, Melissa Roemmele, and Max Kreminski. 2025. Fuzzy Linkography: Automatic Graphical Summarization of Creative Activity Traces. In *Proceedings of the 2025 Conference on Creativity and Cognition (C&C '25)*. Association for Computing Machinery, New York, NY, USA, 637–650. doi:10.1145/3698061.3726915

[15] Kihoon Son, DaEun Choi, Tae Soo Kim, Young-Ho Kim, Sangdoo Yun, and Juho Kim. 2025. ClearFairy: Capturing Creative Workflows through Decision Structuring, In-Situ Questioning, and Rationale Inference. *arXiv preprint arXiv:2509.14537* (2025).

[16] Sarah Sterman, Molly Jane Nicholas, and Eric Paulos. 2022. Towards Creative Version Control. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 336 (Nov. 2022), 25 pages. doi:10.1145/3555756

[17] Sirui Tao, Ivan Liang, Cindy Peng, Zhiqing Wang, Srishti Palani, and Steven P. Dow. 2025. DesignWeaver: Dimensional Scaffolding for Text-to-Image Product Design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 425, 26 pages. doi:10.1145/3706598.3714211

[18] Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. 2025. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478* (2025).