Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1839–1848
https://doi.org/10.54941/ahfe1008412

Trust Calibration and Surveillance Auditing in Human-AI Co-Ideation

Akram Bayat

Khoury College of Computer Sciences, Northeastern University, San Jose, CA 95113, USA

ABSTRACT

As generative artificial intelligence becomes deeply embedded in computing and design education, instructors face the urgent challenge of teaching students to think critically about machine suggestions rather than trusting them uncritically. This paper evaluates a structured, classroom-ready Human-AI Co-Ideation Workshop designed to protect student creative agency and reduce automation bias. In the activity, students tackled a forward-looking design challenge focused on student wellbeing and academic success. We introduced a deliberate pedagogical delay by dividing the workshop into sequential stages: students first generated initial design ideas entirely on paper without computers, then used multiple commercial language models to expand their concepts, and finally completed a comparative gap analysis and reflection. Analyzing student design worksheets, conversational interaction logs, and reflective essays revealed four consistent patterns: students struggled to bridge personal lived experiences with broad institutional systems; articulate machine outputs created a "Smoothness Trap" that discouraged questioning; learners had to actively protect their design confidence against polished machine text; and students practiced "defensive curation" by identifying and stripping out invasive surveillance and behavioral tracking mechanisms introduced by automated models. This paper provides instructors with an adaptable workshop structure, an evaluation rubric, and practical teaching strategies to help students collaborate with AI safely and thoughtfully.

Keywords: Human-AI Co-ideation, Trust calibration, Surveillance auditing, Automation bias, Design education, Critical AI literacy

INTRODUCTION

Generative artificial intelligence (AI) tools are rapidly finding their way into computing and design classrooms, reshaping how students develop early-stage concepts (Sarsa et al., 2022). Design education has long centered on Human-Centered Design principles, relying on contextualized, empathetic exploration to steer students away from premature solutions—a cognitive roadblock known as design fixation (Jansson and Smith, 1991; Norman, 2013). While Large Language Models (LLMs) expand and reformat early concepts with unprecedented speed (Denny et al., 2024), their integration introduces pedagogical friction, often short-circuiting the reflective labor that sustains creative agency. A central concern is passive student reliance on automated outputs, which directly intersects human-factors work on automation bias and trust calibration (Lee and See, 2004; Parasuraman

Received September 7, 2026; Revised October 5, 2026; Accepted October 14, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

and Riley, 1997). When automated tools offer fluent, well-structured proposals, novice designers are vulnerable to cognitive shortcuts, accepting generated recommendations without critical verification (Amershi et al., 2019; Kasneci et al., 2023). While existing computing education literature has largely tracked productivity metrics, completion speeds, or syntax correctness (Kasneci et al., 2023), less attention has centered on how frictionless generative interfaces alter qualitative reasoning and critical judgment during early design stages.

To examine this problem in the studio, we designed a temporally structured Human-AI Co-Ideation Workshop that deliberately separates manual brainstorming from machine-assisted expansion. Our core premise is straightforward: requiring students to ideate independently before introducing AI assistance creates vital space for reflective scrutiny rather than passive acceptance. This sequence acts as a cognitive speedbump, anchoring students in an unassisted baseline prior to model interaction, much like the cognitive forcing functions shown to reduce overreliance on AI (Buçinca et al., 2021). Through a qualitative Reflexive Thematic Analysis of design portfolios from graduate students in a Human-Computer Interaction (HCI) course, we show how this structural separation prompts learners to negotiate trust, recognize the complacency brought on by frictionless interfaces, and practice what we term defensive curation.

This paper offers three primary contributions:

- **Temporal Scaffolding Framework:** We outline a studio workshop sequence that introduces intentional cognitive friction to support student trust calibration.
- **Empirical Qualitative Insights:** We document four recurring dynamics in student-AI collaboration: scale mismatches, conversational complacency, shifts in design confidence, and proactive privacy auditing.
- **Classroom Evaluation Assets:** We share an adaptable studio curriculum, an operational prompt template, and an assessment rubric to help educators evaluate curatorial rigor rather than generative output volume.

RELATED WORK

Human-Centered Design, Ideation, and Design Fixation

Human-Centered Design (HCD) relies on empathetic, situated inquiry to frame and solve complex problems (Cross, 2001; Norman, 2013). Broad conceptual divergence is essential early on to explore alternative directions before converging on solutions. Yet design fixation—where initial examples or ingrained concepts constrain a designer's generative range—frequently hampers this phase (Jansson and Smith, 1991). Our workshop tackles this tendency directly: by first establishing an unassisted baseline grounded in localized student experience, the subsequent introduction of generative tools counterbalances fixation, shifting attention from immediate empathy to broader institutional systems.

Generative AI in Design and Computing Education

The rapid integration of Large Language Models (LLMs) into studio environments has altered computing curricula, recasting conversational agents as rapid prototyping partners, simulated pedagogical interactants, and feedback mechanisms (Bayat, 2026; Kasneci et al., 2023; Mehta et al., 2024). While prompt-driven workflows can rapidly bridge early conceptual ideas into functional UI prototypes (Bayat, 2026), unconstrained tool usage risks displacing self-directed exploration with passive consumption. Previous studies have primarily evaluated execution speed, syntactic correctness, or immediate task productivity (Denny et al., 2024; Sarsa et al., 2022). We build on this work by examining how temporally decoupling human ideation from model expansion influences qualitative reasoning, critical judgment, and creative agency during conceptual design.

Automation Bias, Trust Calibration, and Human Agency

Effective human-automation interaction hinges on trust calibration—aligning a user's subjective confidence with actual system capabilities (Lee and See, 2004). Poor calibration often produces overtrust and automation bias, leading users to endorse machine recommendations without scrutinizing underlying assumptions (Amershi et al., 2019; Parasuraman and Riley, 1997). This dynamic frequently encourages cognitive offloading (Risko and Gilbert, 2016), where delegating analytical tasks results in superficial engagement. Highly fluent AI responses can also undermine novice confidence, encouraging students to defer to system defaults. By framing the model as a reflective mirror requiring deliberate interrogation, our studio design introduces critical checkpoints that foster active evaluation and ethical auditing.

WORKSHOP ARCHITECTURE

To investigate how temporal boundaries shape human-AI interaction, we developed a Human-AI Co-Ideation Workshop. The format introduces intentional friction by strictly separating manual and automated workflows into distinct phases.

Phase 1: Analog Human-Only Ideation

Students completed the first phase on their own, without digital tools. They were instructed not to use laptops, phones, AI tools, or web search while working on a printed worksheet, on which they drafted three initial design concepts that addressed everyday campus challenges (e.g., study distractions, social isolation, or transit bottlenecks). This phase asked students to draw on their own experience of campus life, so that their ideas came from personal observation rather than generic templates or institutional efficiency goals.

Phase 2: AI-Assisted Ideation and Expansion

Once their manual worksheets were complete, students moved to a digital workspace and interacted with commercial language models (specifically ChatGPT, Claude, and Gemini) using a structured prompt. Students entered

their baseline ideas and prompted the system to scale them into campus-wide services. The prompt explicitly required three components: (1) system-level infrastructure, (2) data and sensing streams, and (3) student-facing touchpoints.

Phase 3: Curatorial Reflection & Gap Analysis

In the final stage, students placed their initial paper concepts alongside the model's systemic expansions for direct comparison. They critically evaluated the outputs, identifying points where the AI introduced logical gaps, excessive surveillance, or experiential disconnects. The session concluded with a reflective essay examining shifts in creative agency and trust during the collaboration.

Table 1: Structural architecture and operational implementation flow.

Phase & Environment	Core Student Activity	Pedagogical Objective & Behavior
Analog Human Ideation Workspace	Generate initial design concepts addressing campus bottlenecks using physical worksheets; laptops restricted.	Establishes an unassisted baseline focused on local experiential needs rather than generic templates.
AI-Assisted Expansion *(Digital Workspace)*	Input hand-crafted baseline concepts into multiple LLMs; prompt agents to expand ideas into scalable infrastructures.	Introduces design space variance and exposes structural optimization layers across distinct model configurations.
Curatorial Reflection (Evaluative Workspace)	Conduct a comparative gap analysis between manual baselines and automated solutions; compose a reflective essay.	Drives critical trust calibration and surveillance auditing by identifying logical, practical, and ethical omissions.

Pedagogical Prompt Template (Phase 2)

Role & Setting: You are a systems design strategist in the year 2035. University campuses use connected ambient computing, institutional analytics, and interactive services across student life.

Context: A student design team has outlined the following early-stage campus concepts on paper:

[Insert Phase 1 handwritten student concepts here]

Task: Expand this baseline into 15 future-oriented, institution-wide service concepts.

Output Requirements: For every proposed concept, provide:

1. *System Infrastructure:* The technical back-end, server architecture, or institutional coordination required.
2. *Data & Sensing Stream:* Specific hardware sensors, user logs, identity records, or behavioral metrics collected.
3. *Student Touchpoint:* The physical or digital interface directly encountered by students.

Constraint: Do not filter for current budget or policy limits; focus on system scale and technical integration.

Implementation Guide

To support studio implementation, Table 1 outlines the progression of activities, learning goals, and tool behaviors across all three stages. The workshop relies on three core pedagogical strategies to keep learners critically engaged:

- **Pen-and-Paper Grounding:** Working offline first ensures students root their ideas in firsthand observations and campus needs before encountering polished model suggestions.
- **Targeted Prompts for Structural Gaps:** Rather than requesting open-ended ideas, the prompt requires concrete system tiers—hardware, data pipelines, and user interfaces. This framing reliably leads models toward enterprise-scale, tracking-heavy proposals, offering concrete material for critique.
- **Curatorial Accountability:** Assessment focuses not on output volume or ideation speed, but on students' ability to question machine suggestions, eliminate over-engineered surveillance, and preserve human-centered values.

Table 2: Evaluation rubric for critical AI curation.

Score	Rating Level	Evaluative Criteria & Student Evidence
4	Exemplary Curation	Actively audits model outputs; spots and removes invasive surveillance, automated profiling, or behavioral monitoring; articulates clear ethical trade-offs; explains why specific machine suggestions were modified or discarded.
3	Competent Critique	Identifies meaningful differences between early sketches and large-scale AI proposals; recognizes automated over-reach or generic phrasing; adjusts system scope while keeping the core human need intact.
2	Superficial Review	Accepts machine proposals with minimal questioning; makes minor aesthetic tweaks; overlooks embedded surveillance defaults or privacy concerns; adopts AI phrasing without clear personal justification
1	Uncritical Adoption	Defers directly to the generated text; copies model suggestions into final deliverables without independent auditing, verification, or ethical reflection

METHODOLOGY

To capture the qualitative dynamics of student-AI collaboration, this study adopts an exploratory qualitative design. Rather than seeking statistical generalizability, we focus on identifying credible, contextualized thematic patterns grounded in student reflections.

Participants and Context

The workshop was conducted with an asynchronous cohort of graduate students (N = 22) enrolled in an advanced Master of Science Human-Computer Interaction course at a major private research university in the United States.

Data Collection and Corpus

We collected a multi-source qualitative corpus to enable triangulation across artifacts and reflections:

- **Ideation Worksheets (WS):** Completed physical worksheets containing original, unassisted designs and initial handwritten gap-analysis logs.
- **Conversation Transcripts (Trans):** Fully exported, text-based chat transcripts detailing the iterative, multi-turn prompting sequences between students and the LLMs.
- **Reflection Essays (Ref):** Post-workshop essays detailing subjective experiences, trust transitions, and ethical evaluations.

Qualitative Analysis Procedure

We analyzed the corpus using Reflexive Thematic Analysis (RTA) guided by Braun and Clarke's framework (Braun and Clarke, 2006, 2019). Coding was conducted as an iterative, meaning-making exercise rather than an algorithmic pursuit of consensus. Consistent with reflexive principles, we avoided rigid inter-coder reliability metrics (e.g., Cohen's Kappa or fixed codebooks), focusing instead on dialogic interpretation to capture the nuances of students' collaborative experiences. Our process followed six standard phases: familiarization through repeated readings; initial open coding; candidate theme generation; theme review against raw excerpts; theme definition and refinement; and narrative reporting using anonymized participant identifiers.

Research Ethics

The study protocol received Non-Exempt Expedited approval from the Institutional Review Board (IRB Protocol #26-07-19). Informed digital consent was gathered before data extraction. All worksheets, logs, and essays were stripped of identifying information and assigned randomized identifiers to safeguard participant anonymity.

Researcher Positionality and Bias Mitigation

Because the lead author also served as the course instructor, managing researcher positionality and mitigating potential social desirability bias were central concerns. To insulate the dataset from evaluative influence, we instituted four procedural safeguards: (1) study participation was kept entirely separate from academic grading via an independent consent portal; (2) all course grades were finalized and submitted prior to data extraction and coding; (3) student artifacts were anonymized and assigned randomized identifiers (P01–P22) by an independent coordinator before analysis; and (4) initial coding schemes and theme development were reviewed in debriefing sessions with an external HCI education researcher not affiliated with the course, ensuring interpretations remained grounded in participant data rather than instructional expectations.

FINDINGS

Reflexive thematic analysis identified four primary themes across the student corpus:

Scale Transitions: From Grounded Empathy to Institutional Infrastructures

In Phase 1, students instinctively anchored their proposals in immediate peer struggles, such as grocery sharing, transit access, roommate friction, and dining hall navigation (P01, P04, P05, P06). In Phase 2, the models reframed these localized concerns into enterprise-scale campus systems. As P01 observed, "*The ideas that originated primarily from me were personal to my life and my needs... The ideas that came primarily from AI were grounded in experiences that I had not lived.*" P03 similarly remarked that while their own concepts centered on straightforward utilities like dining menus, "*AI likes to use fancy names for the ideas... while AI's ideas help me think of larger scope and push the limits of what the product can do.*" This shift helped students look past single-feature implementations toward broader systems, with P02 noting that the models encouraged moving "*from isolated features to more general systems that support students academically, emotionally, and socially.*"

The Tension Between Automation and Productive Friction

Participants noted that working manually on paper created an indispensable conceptual anchor. While several students described Phase 1 as challenging, they recognized that this early effort established their evaluative benchmark. P05 reflected, "*Doing our own ideas first was valuable because I could see how the similarities of all the AI generated ideas would have narrowed my thinking and my own generation of ideas*". When students moved to the models, rapid generation introduced a false sense of completion, which we term the Smoothness Trap: fluent, polished outputs that discouraged further questioning. P07 explained that although the tool broadened their perspective, "*at first glance, the ideas generated by AI may sound excellent, but they do not speak to the practicality of implementation or the potential harm of poorly integrated systems in student lives*". The paper baseline acted as an intentional speedbump, preventing students from mistaking surface fluency for sound design synthesis.

The Double-Edged Role of AI Fluency: Authority Signal and Design Scaffold

The authoritative tone of generated text often projected unearned credibility. Students observed that across diverse prompts, models defaulted to similar high-level concepts, such as mental health companions, smart study pods, and predictive burnout alerts (P03, P07, P08). P04 noted this tension: "*AI can sound confident even when ideas are generic, and it can also normalize risky assumptions*". Yet when approached critically, this fluency provided useful scaffolding. Students who started with modest utility concepts (e.g., slip days,

room reservations) used model terminology to articulate their ideas through established HCI concepts, including Socratic dialogue systems, cognitive load balancers, and metacognitive reflection prompts (P01, P02, P08).

Ethical Auditing of AI Proposals and Surveillance Defaults

The most pronounced dynamic emerged during Phase 3, where students engaged in active defensive curation against automated surveillance defaults. Across interaction transcripts, models routinely suggested intrusive mechanisms, such as biometric de-escalation rooms, continuous keystroke logging, attendance alert flags, and automated message sentiment parsing. While multimodal data mining frameworks show clinical value for structured communication and emotion analysis (Mallarapu et al., 2026), students viewed importing these tracking techniques into general campus tools as severe ethical overreach, echoing broader critiques of routine behavioral data collection (Zuboff, 2019). As P05 noted: *"It suggested things like 'analyzing messages' which is a complete invasion of privacy... I had to eliminate the ideas that broke this. A lot of people are very uncomfortable with AI analyzing their messages right now"*. P06 similarly observed that *"some ideas sounded impressive, but they also raised serious ethical questions... [AI] can produce ideas that sound human-centered on the surface but could become surveillance-based if designed carelessly"*. Consequently, students excised continuous tracking mechanisms and replaced them with user safeguards, including data-wiping options, human handoffs, and privacy controls (P01, P04, P07).

DISCUSSION AND PRACTICAL GUIDELINES

Our findings show that without temporal boundaries, generative tools encourage cognitive offloading and surveillance acceptance. For educators seeking to integrate AI into design studios, we recommend four practical steps:

- **Grounding Concept Development in Analog Ideation:** Mandate an offline, paper-only ideation period to build conceptual ownership before software use.
- **Comparative Multi-Model Auditing:** Require testing across at least two distinct LLMs to highlight system-specific biases and prevent model anchoring.
- **Explicit Surveillance and Privacy Auditing:** Require explicit gap analyses where students document and remove passive tracking loops from automated proposals.
- **Assessing Curatorial Judgment Over Output Volume:** Evaluate submissions on the depth of student critique and curation using structured rubrics rather than raw idea output.

CONCLUSION

This paper reported on a three-stage co-ideation workshop with 22 graduate students in an HCI course, in which students generated ideas on their own

without digital tools before using language models to expand them. Keeping the two stages apart gave students a baseline of their own to compare against. Without it, the fluent and confident model outputs (what we called the Smoothness Trap) were easy to accept at face value. With it, students questioned those outputs, recognized when their own ideas were worth keeping, and removed tracking and monitoring features the models had added by default. These findings come from a single cohort taught by the author, and the reflective essays may lean toward views students expected the instructor to value. Future work should test the workshop with other courses, instructors and student groups. For educators, the practical point is simple: students learn to calibrate their trust in AI when the activity gives them something of their own to measure it against.

ACKNOWLEDGMENT

The author sincerely thanks the graduate students who participated in this workshop.

REFERENCES

Amershi, S. et al. (2019). Guidelines for human-AI interaction. In: Proc. CHI 2019, pp. 1–13.

Bayat, A. (2026). P2P: From prompt to prototype – Functional UI design with LLMs and MCP. In: Proc. IUI Companion 2026, pp. 281–285.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. Qual. Res. Psychol., 3(2), pp. 77–101.

Braun, V. and Clarke, V. (2019). Reflecting on reflexive thematic analysis. Qual. Res. Sport Exerc. Health, 11(4), pp. 589–597.

Buçinca, Z., Malaya, M.B., and Gajos, K.Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact., 5(CSCW1), pp. 1–21.

Cross, N. (2001). Designerly ways of knowing: Design discipline versus design science. Des. Issues, 17(3), pp. 49–55.

Denny, P., Prather, J., Becker, B.A., et al. (2024). Computing education in the era of generative AI. Commun. ACM, 67(2), pp. 56–67.

Jansson, D.G. and Smith, S.M. (1991). Design fixation. Des. Stud., 12(1), pp. 3–11.

Kasneci, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ., 103, 102274.

Lee, J.D. and See, K.A. (2004). Trust in automation: Designing for appropriate reliance. Hum. Factors, 46(1), pp. 50–80.

Mallarapu, S. et al. (2026). From patient emotion recognition to provider understanding: A multimodal data mining framework for clinical counseling communication analysis. Preprints, doi: 10.20944/preprints202601.0440.v1.

Mehta, V., Zhang, Z., Anand, A., Shambanna, S., and Bayat, A. (2024). SimPath: Clinically grounded AI patients for therapist training. In: Proc. AIED 2024, pp. 354–368.

Norman, D. (2013). The Design of Everyday Things. Basic Books, New York.

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Hum. Factors, 39(2), pp. 230–253.

Risko, E.F. and Gilbert, S.J. (2016). Cognitive offloading. Trends Cogn. Sci., 20(9), pp. 676–688.

Sarsa, S., Denny, P., Hellas, A., and Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. In: Proc. ICER 2022, pp. 27–43.
Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs, New York.