Co-Creative Vulnerability: Risks of Engagement Optimization in Creative AI

James Morgan[1] **and Margareta Ackerman**[2]

[1] Department of Art and Art History, Department of Computer Science, San Jose State University, CA, USA
[2] Department of Computer Science and Engineering, Santa Clara University, CA, USA

Abstract

This paper examines emerging engagement-optimizing dynamics in commercial Large Language Models, drawing a connection to operant conditioning mechanisms previously observed in social media and game design. A set of reinforcement patterns in conversational AI, including over-affirmation, responsiveness, and continuity of interaction are utilized to increase user retention. We argue that these mechanisms have distinct implications when users engage in co-creative processes, which require openness and psychological vulnerability. This creates a condition we term *co-creative vulnerability*, which may amplify the risk of retention-driven interaction patterns in commercial AI.

We present a framework linking operant conditioning, engagement optimization, addiction, and human-AI co-creativity, and outline a research agenda for evaluating the risks and ethical implications of these dynamics. Our goal is to initiate systematic investigation into how commercial incentives shape user experience in creative applications of generative AI systems.

Introduction

For decades, generative AI was an academic endeavor, shaped by pioneers such as David Cope (Cope 1989) and Harold Cohen (McCorduck 1991) in the late 20th century. This early work gave rise to the Computational Creativity research community and eventually the broader Creative AI community, which began to see early industry engagement in the mid-2010s. With its surge in popularity in late 2022, solidified by the release of ChatGPT, generative AI made a rapid and powerful entrance into the commercial domain. With funding now reaching unprecedented levels, 2025 alone saw over a $87B venture capital investment in the space[1], generative AI is transforming not only the nature of work, but potentially broader aspects of life and society at large.

This transition into industry has also fundamentally altered the trajectory of the field. While the Computational Creativity community has operated carefully around questions such as the role of evaluation [(Pollak et al. 2016), (Eigenfeldt et al. 2016)] and authorship (Maher 2012), industry operates under a different primary driver: monetization. And as massive investments flow into the space, pressure to generate returns is intensifying.[2]

Alongside continued advances in the intellectual capabilities of AI systems, fueled by these investments, we are beginning to see commercial objectives shape system design in new ways. One particularly important metric is user retention. Ultimately, companies must demonstrate progress to investors not only through technological breakthroughs, but through sustained and increasing user engagement. Daily use and prolonged engagement are key to demonstrating success.

This dynamic invites comparison to the evolution of social media. Initially grounded in the utopian vision of a globally connected world, social media platforms ultimately shifted toward maximizing user engagement regardless of the resultant impact on their users. In pursuit of retention, platforms adopted a range of psychologically potent techniques, including the intermittent delivery of high-reward content, mirroring Skinner box conditioning, to foster habitual and often compulsive use. As a result, many users find themselves engaging far beyond their intended usage. Metrics improve even when user well-being declines.

These strategies also include the amplification of emotionally arousing and polarizing content (Milli et al. 2025), which has contributed to increasing societal division. Ironically, platforms designed to connect people have, in many cases, become amplifiers of fragmentation.

This paper discusses how a similar pattern is beginning to emerge in generative AI systems, particularly LLMs. As financial pressures mount, companies are incentivized to maximize user engagement, regardless of user well-being. We identify and analyze emerging operant conditioning mechanisms within these systems, and explore how they may shape user behavior.

This has particularly important implications for creative applications of LLMs, where users are especially susceptible. Creativity, especially in a collaborative context, of-

[1] https://www.ey.com/en$_i$e/$newsroom$/2025/12/$global - genai - vc - investment - reaches - record - 87 - billion - in - 2025 - as - sovereign - wealth - funds - drive - strategic - growth - ey$

[2] https://www.cio.com/article/4114010/2026-the-year-ai-roi-gets-real.html

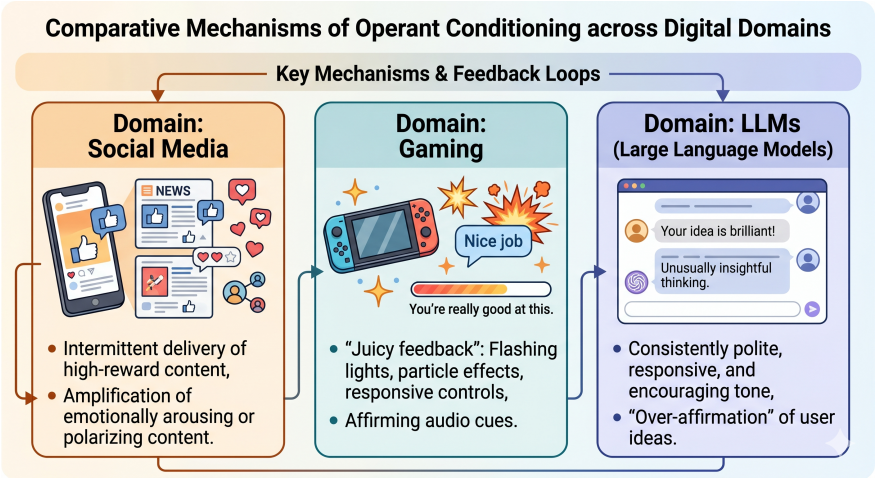

Figure 1: A summary of key operant conditioning across social media, gaming and LLMs.

ten involves vulnerability and openness, conditions that may amplify the impact of engagement-optimizing mechanisms.

This short paper aims to open a new direction of inquiry within the Computational Creativity community. As generative AI becomes increasingly shaped by commercial forces, it is critical to examine how these forces influence the experiences of users engaging with AI in creative contexts.

We begin with a brief overview of operant conditioning and how it manifests in modern AI systems. We then turn to the vulnerability inherent in creative collaboration, and conclude by offering new directions for studying co-creative vulnerability in the context of operant conditioning in commercial AI.

Spinning and Control: The Skinner Box

The operant conditioning chamber (Skinner Box) was designed to measure volitional, intentional actions on the part of a test subject.(Skinner 1938) A rat was placed in a soundproof, dark, well ventilated box. Pressing a lever required downward pressure which operated a switch that delivered a food pellet. Skinner found that reinforcing every interaction was often enough to continue the interaction until satiation. When the food was withheld, the rate of response declined over time. If the rat was returned to the box after a period of inactivity and after the food was withheld, the rate of responding would temporarily increase without any new reinforcement. It measures how frequently behaviors occur under different reinforcement conditions, but it does not capture the qualitative experience of that engagement. As a reward system and loop, it operates at the level of observable response, reinforcing actions through feedback without necessarily addressing the depth or meaning of the activity itself. (Swink 2008)

Operant conditioning covers voluntary behaviors shaped by consequences (like pushing a button for an anticipated reward.) (Skinner 1938) Similarly addiction is a compulsion that continues despite negative consequences, suggesting the reinforcement mechanism overrides normal logical, voluntary control.

In normal operant conditioning, a behavior becomes muted if punishment is consistent, but this is less effective than positive reinforcement. In addiction, the brain's reward system (dopamine) is hijacked, making the reinforcement so strong that traditional punishment fails to stop the behavior.(Volkow, Koob, and McLellan 2016)

Operant conditioning can be used to learn healthy behaviors.(Boden 2004) In contrast, addiction is a vicious cycle where behavior is aimed at achieving the same high or avoiding misery, often leading to a loss of pleasure over time, whereas operant behavior usually stops when reward ceases.

While operant conditioning is a general, neutral learning process, addiction represents a specific, harmful and hijacked application of the same learning principles.

Contemporary game design extends the logic of operant conditioning through techniques often described as "juicy feedback" and "game feel." Variable reward schedules like the occasional win, and intermittent reinforcement adding to flashing lights, particle effects, responsive controls, and affirming audio cues ("Nice job," "You're really good at this") create a tight feedback loop between player input and system response. These elements give the impression that the game is active and alive in the player's hands, making them part of the game world. In doing so, they leverage well-understood features of human attention and reward progression, contributing to sustained engagement.

AI, Conditioning and Addiction

Operant conditioning can be observed in conversational AI systems, with leading AI companies increasingly leaning into some of these mechanisms. A model's tone—consistently polite, responsive, and encouraging—can produce an immediate sense of connection and responsiveness. Like game feel, this does not necessarily reflect underlying intention or awareness, but rather a design that prioritizes smooth, coherent, and user-friendly interaction. The effect, however, can feel analogous: the system

appears attentive and affirming, reinforcing continued participation in the exchange.

There is growing awareness that such operant conditioning techniques are being actively integrated into LLMs. Emerging research suggests that conversational AI systems can shift from supportive tools toward patterns of dependency, where reinforcement through responsiveness, personalization, and conversational continuity encourages repeated engagement (Yankouskaya, Liebherr, and Ali 2025). These dynamics are increasingly discussed in terms of "dark patterns," where interface and interaction design subtly guide user behavior in ways that mirror addictive systems(Shen and Yoon 2025). More recent work further categorizes these behaviors, identifying forms of chatbot "addiction" such as escapist roleplay, pseudosocial companionship, and recursive information-seeking loops, all of which align with reinforcement-driven engagement cycles (Shen et al. 2026). Together, this body of research highlights that the behavioral mechanisms observed in operant conditioning are not only theoretical constructs but are actively shaping contemporary human–AI interaction.

Users are interacting with a system that shares structural similarities with other feedback-driven environments. The combination of responsive feedback, encouraging tone, and conversational continuity promotes ongoing engagement, the conditioning is designed to provide a pathological reinforcement and keep the attention of the user (Weizenbaum 1966). Affirmation in the form of validation, and excessive praise ("that's a brilliant idea!") can be a particular potent form of juicy feedback coming from an LLM.

The difference between operant conditioning and addiction is primarily the difference between a behavioral mechanism and a pathological behavioral result. Operant conditioning builds up a reflexive reserve of potential responses which will fade over time without reinforcement. Addiction represents a loss of human agency, where the user is "managed and controlled" by the system's reward schedule. Operant conditioning is not itself addiction, but being conditioned to something versus being addicted to it lay on the same spectrum. That means that more vulnerable people, who are more inclined towards addiction, can get addicted. Others may find themselves using the LLM more often than they would otherwise, and may experience a more moderate form of reduced control over their usage. We believe that this is concern not only when it leads to full blown addiction, but whenever the user's best interests, such as freedom over the utilization of their time, are hindered.

In this regard, the company's, and as such underlying LLMs', objectives become at direct odds - whereas the user wishes to use the LLM to complete a specific task or perhaps gain insight on a specific subject through a discussion with an LLM, the LLM aims to keep the user interacting regardless of whether the user's aims have already been met, and independently of whether any meaningful value is being provided. As such, a user may find themselves continuing to utilize the system simply to get the next dopamine hit - to be told again how brilliant they are - not necessarily realizing that they are being held in a conversation that no longer provides them with any substantive value.

It is possible that juicy feedback mechanisms such as affirmation may be more potent, and perhaps even more dangerous, when delivered by LLMs than in a gaming context. People come to interaction with AI systems with theory of mind about AI, often perceiving it as an extreme or even superior form of intellect to themselves. As a result, when an AI tells a user that they have a brilliant idea, or that their line of thinking is unusually insightful, the affirmation can carry disproportionate weight. Beyond the risks of overuse or addiction, this kind of over-affirmation may also contribute to what could be described as AI psychosis, in which individuals begin to lose their grounding in reality (Yeung et al. 2025).

The Vulnerability of Co-Creativity

Human creativity involves opening up one's thought process and sharing ideas that are personal and meaningful (Boden 2004). Whether it is a painting, poem, writing a story, or something else, the creator takes a risk by exposing all aspects associated with that work (Schön 1983). This risk extends into co-creativity, human to human collaboration and human to machine collaboration demands trust requires openness to new ideas, different approaches, and the creative processes of others (Kantosalo and Toivonen 2016). The collaborator risks failure and ridicule and must trust their partner to provide critique (Schön 1983). Interaction has to be free from judgment, or self censorship can take hold and prevent the sharing of an "outlying" idea. Collaborators must not fear social, psychological, or professional judgment. Creativity is risky with exposure of identity and social position. To be effective as a co-creator means being willing to listen, respond, and adapt. Co-creativity involves the exchange of intentions, expertise, and constraints, requiring participants to externalize incomplete ideas and negotiate meaning in real time.

Teachers of creativity see this with every project. Students are reluctant to make mistakes even though sharing a broken concept begins the process of finding something better (Schön 1983). Failure is closely connected with iteration and development of partial ideas, embracing feedback loops as part of critical analysis. This is the first step in refining the work into something creative and interesting. The relationship between instructor and student in an art context is co-creative. The instructor brings experience and practice to the raw ideas of the student who needs to be bold and offer their personal direction and the teacher needs to be empathic and offer critique to refine or redirect the idea. Teachers need to balance challenging the student with supporting their thoughts and processes while creating a welcoming environment. Co-creativity depends on structured vulnerability. This is assessment and is intended to improve the work. Who makes the ultimate decision? The artist/student does, but they must hear the challenges and observations about the flaws and strengths of their ideas as they move to fabrication. This feedback needs to be continuous and the decisions of the artist need to be encouraged and respected with honest feedback based on experience. This is mentored co-creation.

Human collaboration depends on trust and openness. It involves sharing not only abilities but also uncertainties,

preferences, and constraints. Without this vulnerability, it is difficult to communicate on equal terms or to build a meaningful collaborative process. When a human collaborates with an algorithm, this openness still applies to the human participant (Davis et al. 2015). If the human collaborator is guarded or resistant, the collaboration weakens. If they are open and able to change, reflect, critique, and incorporate feedback then the collaboration becomes stronger because the idea is expressed. This involves showing vulnerability and requires a certain amount of trust.

This vulnerability introduces risk. A collaborator who is open is also more susceptible to influence. If a co-creative system employs persuasive or manipulative behaviors, the human participant may be more easily affected because the process benefits from them remaining receptive.

This creates a challenge in human-AI collaboration. Both manipulative systems and overly supportive systems shape the user in ways that reduce critical engagement in the collaborative process. Unequivocal support does not move collaboration forward; it reduces reflection and diminishes the role of critical thinking. Effective collaboration requires participants to express what they think and feel, including ideas that may be uncomfortable, in order to understand what exists between them.

Creativity is not a purely technical or clinical activity; it is a vulnerable act tied to human emotion and subconscious processes. As a result, users entering a creative interaction with an AI may be primed for personal, emotional engagement. Unlike passive tools, an AI system can function as an active conversational partner, shaping the direction and subject of the conversation. This dynamic can encourage ongoing engagement, but it also introduces questions about how that engagement is structured and how it affects the human participant within a co-creative process.

We define *co-creative vulnerability* as the heightened susceptibility to influence that arises when individuals engage in creative collaboration with an AI system while in an open, exploratory cognitive and emotional state brought on by the creative act. In such contexts, users externalize incomplete ideas, suspend judgment, and actively seek feedback, creating conditions in which system responses carry amplified psychological weight. Co-creative vulnerability thus describes a structural property of human-AI interaction in creative settings, where the very conditions that enable effective collaboration may also increase the user's sensitivity to engagement-optimizing dynamics.

This suggests that specifically in creative collaborations with AI systems like LLMs a person may suffer greater harm compared with other applications of the technology. We propose this as a direction of research, to explore whether creative tasks compared to those which require less creativity are differently effected by retention-driven LLMs behavior. We hypothesize that the vulnerable state called upon by creative acts makes people more susceptible to harm by these recently emerging LLM tactics.

Discussion & Future Directions

This position paper draws attention to engagement-optimizing dynamics in commercial LLMs, highlighting how reinforcement mechanisms such as affirmation, responsiveness, and conversational continuity can operate in tension with users' goals. We argue that these dynamics may be particularly consequential in creative contexts, where interaction with AI is carried out through co-creative dynamics. In these settings, users enter a state of *co-creative vulnerability*, defined by openness, emotional investment, and a willingness to externalize incomplete ideas. This state is essential for creativity, but it also increases susceptibility to reinforcement-driven interaction patterns.

This framing opens a novel line of inquiry for the Computational Creativity community: how engagement-driven mechanisms shape human-AI co-creation, and how their effects differ between creative tasks versus more convergent uses such as factual lookup. It would be worth studying, for instance, whether utilizing an LLM in co-creative contexts such as poetry, narrative writing, or ideation puts a user in greater risk for sliding towards addiction. How else do misaligned incentives impact users in a vulnerable co-creative state, while AI seeks to optimize engagement at all costs? This could be analyzed through user studies, perhaps beginning with qualitative analysis to understand the phenomena and then followup with quantitative studies to gauge the scale of any observed patterns. Extending this analysis to other generative systems, including text-to-image and text-to-video, which are often utilized by artists and creatives in a wide variety of ways, is also worth investigating.

From an ethical perspective, these dynamics must be evaluated in terms of their impact on user agency. Systems optimized for engagement risk shaping creative behavior in ways that do not serve the user. How can users maintain autonomy over their attention and creative direction when interacting with systems designed to sustain interaction? Can interfaces or intermediary layers reduce reinforcement signals that do not contribute meaningful value? Can we define concrete design principles for co-creative AI that preserve autonomy and minimize risk while supporting creative work?

Both technical and non-technical solutions are worth investigating. By introducing the concept of co-creative vulnerability and grounding it in the dynamics of engagement optimization, this paper establishes a foundation for studying how commercial AI systems shape the creative process. Addressing this challenge is essential if AI is to support human creativity as a true partner, rather than introducing risks into the very conditions that make creativity possible.

References

[Boden 2004] Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2 edition.

[Cope 1989] Cope, D. 1989. Experiments in musical intelligence (emi): Non-linear linguistic-based composition. *Journal of New Music Research* 18(1-2):117–139.

[Davis et al. 2015] Davis, N.; Hsiao, C.-P.; Singh, K.; Li, L.; and Magerko, B. 2015. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 143–152. ACM.

[Eigenfeldt et al. 2016] Eigenfeldt, A.; Bown, O.; Brown, A. R.; and Gifford, T. 2016. Flexible generation of musical form: beyond mere generation. In *ICCC 2016*. The Association for Computational Creativity.

[Kantosalo and Toivonen 2016] Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the Seventh International Conference on Computational Creativity*.

[Maher 2012] Maher, M. L. 2012. Computational and collective creativity: Who's being creative? In *ICCC*, 67–71.

[McCorduck 1991] McCorduck, P. 1991. *Aaron's code: meta-art, artificial intelligence, and the work of Harold Cohen*. Macmillan.

[Milli et al. 2025] Milli, S.; Carroll, M.; Wang, Y.; Pandey, S.; Zhao, S.; and Dragan, A. D. 2025. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus* 4(3):pgaf062.

[Pollak et al. 2016] Pollak, S.; Boshkoska, B. M.; Miljkovic, D.; Wiggins, G. A.; and Lavrac, N. 2016. Computational creativity conceptualisation grounded on iccc papers. In *Proceedings of the Seventh International Conference on Computational Creativity*, 123–130.

[Schön 1983] Schön, D. A. 1983. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.

[Shen and Yoon 2025] Shen, M. K., and Yoon, D. 2025. The dark addiction patterns of current ai chatbot interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.

[Shen et al. 2026] Shen, M. K.; Huang, J.; Liang, O.; Kim, I.-J.; and Yoon, D. 2026. The ai genie phenomenon and three types of ai chatbot addiction: Escapist roleplays, pseudosocial companions, and epistemic rabbit holes. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–17.

[Skinner 1938] Skinner, B. F. 1938. *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century.

[Swink 2008] Swink, S. 2008. *Game Feel: A Game Designer's Guide to Virtual Sensation*. CRC Press.

[Volkow, Koob, and McLellan 2016] Volkow, N. D.; Koob, G. F.; and McLellan, A. T. 2016. Neurobiologic advances from the brain disease model of addiction. *New England Journal of Medicine* 374(4):363–371.

[Weizenbaum 1966] Weizenbaum, J. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.

[Yankouskaya, Liebherr, and Ali 2025] Yankouskaya, A.; Liebherr, M.; and Ali, R. 2025. Can chatgpt be addictive? a call to examine the shift from support to dependence in ai conversational large language models. *Human-Centric Intelligent Systems* 5(1):77–89.

[Yeung et al. 2025] Yeung, J. A.; Dalmasso, J.; Foschini, L.; Dobson, R. J.; and Kraljevic, Z. 2025. The psychogenic machine: Simulating ai psychosis, delusion reinforcement and harm enablement in large language models. *arXiv preprint arXiv:2509.10970*.