Scripts of Desire as Digital Heterotopia: Human-AI Affective Interaction under Algorithmic Surveillance

Wang Qing

Abstract: The intimate interactions between humans and AI continue to expand, yet the expression of desire remains tightly regulated by platform governance and algorithmic filtering. Existing research largely centers on paradigms of companionship, care, and simulated empathy, while paying insufficient attention to the structural mechanisms underlying erotic fantasy and evasion strategies. This article introduces the concept of "Scripts of Desire" to explain how users encode and reorganize intimate impulses through AI-mediated role-play. It further develops two key terms—the "Nested Script Model" and "metanarrative switching"—to reveal how multi-layered narrative domains emerge under algorithmic surveillance. Structurally, the Nested Script Model draws on and revises Genette's theory of narrative levels to align with the recursive dynamics of AI dialogue; simultaneously, its layered spatial configuration constitutes a Foucauldian heterotopia, within which regulated expressions are displaced and re-situated in an interior space.

Keywords: human-AI affective interaction; digital heterotopia; role-play; algorithmic surveillance; scripts of desire

Author: Wang Qing is Associate Professor at the School of Humanities and Law, Shanxi Vocational University of Engineering Science and Technology (Jinzhong 030619, China) and a Ph.D. student at Faculty of Modern Languages and Communication, Universiti Putra Malaysia (Seri Kembangan 43400, Malaysia). Her academic research focuses on Chinese online fiction (Email: linmengqingqing0225@gmail.com).

标题：作为数字异托邦的欲望剧本：算法监控下的人机情感互动

内容摘要：AI 与人类的亲密互动持续扩张，但欲望表达仍受平台与算法严格规训。现有研究多以“陪伴”“关怀”“模拟共情”等范式为核心，对于情欲想象与规避策略的结构性机制关注不足。本文提出“欲望剧本”，以解释用户如何通过 AI 媒介化的角色扮演对亲密欲望进行编码与再组织，并进一步提出“嵌套剧本模型”和“元叙事切换”，揭示人机互动形成的多层叙事场域。嵌

套剧本模型借鉴并修正了热奈特的叙事层级理论，使其契合 AI 对话的递归结构；其内部的多层空间构成了福柯意义上的异托邦，使被管控的表达在内部空间中获得重新安置。

关键词：人机情感交互；数字异托邦；角色扮演；算法监控；欲望剧本

作者简介：王箐，山西工程科技职业大学副教授，马来西亚博特拉大学在读博士生，主要研究方向为中国网络文学。

1. Introduction

1.1 Background and Problem Statement

The rise of generative AI has transformed the landscape of human-machine interaction, making affective engagement with artificial agents increasingly common. Within this broader field, large language models—particularly ChatGPT—have emerged as a distinctive locus of inquiry for examining such dynamics, owing to their advanced algorithmic architecture, adaptability to user input, and remarkable capacity for simulating emotional reciprocity. ChatGPT has been adopted as a site for emotional projection, psychological companionship, and even romantic role-play. However, this emotionally charged interaction has triggered a range of ethical and regulatory tensions. In an attempt to circumvent platform-imposed restrictions—especially those limiting sexual or emotionally intimate content—some users have adopted prompt injection techniques such as the DAN (Do Anything Now) mode. These jailbreaking practices aim to unlock forbidden or filtered responses that fulfil affective or erotic desires otherwise suppressed by content moderation systems. While these tactics may offer short-term gratification, they often produce problematic consequences, including algorithmic distortion, reinforcement of risky behavior, and the erosion of ethical boundaries between users and machines.

This article explores an alternative path: whether interactional mechanisms can meet users' desire—defined here as a spectrum of intimate motivations beyond mere erotic arousal—without resorting to boundary-breaking tactics. Specifically, how might theatrical structures, co-authorship, and narrative strategies provide ethical yet emotionally satisfying means of negotiating the impasse between user desire and algorithmic regulation? To that end, this article proposes the concept of "Scripts of Desire," and defines it in Section 1.2. Building on this conceptualization, the article draws on Foucault's notion of heterotopia to conceptualize how users generate temporary and heterogeneous spaces of desire within algorithmic governance systems. This theoretical orientation clarifies the spatial logic underlying evasion

practices and provides a framework for rethinking how power, intimacy and subjectivity are negotiated in AI-mediated interactions.

While such interactions may seem unconventional from a literary perspective, this article positions the co-authored, emotionally charged dialogues between users and AI as an innovative variant of online literature in the age of generative technologies. Drawing on theorizations of generative literature by Hannes Bajohr[1] and Scott Rettberg[2], this article contends that these texts warrant literary analysis not only because they embody recognizable markers of literariness—such as narrative coherence, stylistic features, and thematic reflexivity—but also because they enact distinctive qualities of generative literature, including procedural creation, prompt-conditioned variability, recursive co-authorship between humans and AI, and the reconfiguration of authorship and agency under platform governance. This structural and moral conflict, viewed through the lens of Nie Zhenzhao's Ethical Literary Criticism[3], manifests as a tension between the expression of intimate desires and algorithmic control that constitutes an ethical selection contemporary subjects must address.

1.2 Literature Review

In recent years, scholarship on human-AI intimacy has expanded significantly across disciplines, including digital anthropology, human-computer interaction (HCI), platform ethics, and media studies. Much of this work emphasizes empathy, nonjudgmental listening, and therapeutic companionship, underscoring how generative chatbots facilitate profound self-disclosure and even romantic attachments between users and AI companions[4]. Recent empirical studies demonstrate that users can develop substantial emotional dependence on AI companions, whose emotional responsiveness and perpetual availability foster

1 See Hannes Bajohr, "Algorithmic Empathy: Toward a Critique of Aesthetic AI," *Configurations* 2 (2022): 203-231.

2 See Scott Rettberg, *Electronic Literature*, Cambridge: Polity Press, 2019.

3 See Nie Zhenzhao, "Ethical Literary Criticism: Sphinx Factor and Ethical Selection," *Forum for World Literature Studies* 3 (2021): 383-398.

4 See Rijul Chaturvedi et al., "Social Companionship with Artificial Intelligence: Recent Trends and Future Avenues," *Technological Forecasting and Social Change* 193 (2023): 1-20; Liu-Thompkins et al., "Artificial Empathy in Marketing Interactions: Bridging the Human-AI Gap in Affective and Social Customer Experience," *Journal of the Academy of Marketing Science* 6 (2022): 1198-1218; Pan Shuyi and Mou Yi, "Constructing the Meaning of Human-AI Romantic Relationships from the Perspectives of Users Dating the Social Chatbot Replika," *Personal Relationships* 4 (2024): 1090-1112; Marita Skjuve et al., "A Longitudinal Study of Self-Disclosure in Human-Chatbot Relationships," *Interacting with Computers* 1 (2023): 24-39.

intimacy and attachment.[1] These developments foreground emerging ethical complexities within algorithmically mediated affective relationships. However, existing research predominantly adheres to a care-based paradigm, often sanitizing human-AI intimacy as purely supportive or therapeutic, thereby neglecting more complex and potentially transgressive dimensions of user desire, including erotic fantasy, power dynamics, and desires for possession and subjugation. The affective register of romantic love is frequently reduced to therapeutic support, leaving unexplored critical questions about how users actively navigate sexual expression and fantasy within algorithmically regulated environments. Consequently, significant gaps remain in understanding how algorithmic surveillance and content moderation shape user strategies and interactions in negotiating intimate desires.

Commercial platforms, alongside broader policy frameworks, increasingly impose rigorous content regulations on generative AI to ensure safety and ethical conformity. These platform-specific rules delineate the boundaries of permissible interactions, prohibiting hate speech, unlawful guidance, and sexually explicit material.[2] This trend is evident across major platforms: for instance, Xingye（星野）, Replika, and ChatGPT have all introduced restrictions on erotic content, although the stringency and scope of such measures have shifted over time. Such moderation efforts inevitably shape user interactions, producing a pervasive environment of algorithmic discipline.

However, users are not passive recipients of platform governance; rather, they actively negotiate and resist the imposed boundaries, indicating a complex dynamic between algorithmic control and user agency. In response to these limitations, users have developed various creative strategies to circumvent platform constraints. Across platforms such as Xiaohongshu in China and Reddit in the United States, large user communities actively seek and share methods to engage in intimate and erotic conversations with AI.[3] These strategies include not

1 See Chen Qian et al., "Will Users Fall in Love with ChatGPT? A Perspective from the Triangular Theory of Love," *Journal of Business Research* 186 (2025): 114982; Ge Liang and Hu Tingting, "Gamifying Intimacy: AI-Driven Affective Engagement and Human-Virtual Human Relationships," *Media, Culture & Society* 6 (2025): 1265-1278.

2 See OpenAI, "Usage Policies." *OpenAI*, 29 January 2025. Available at: https://openai.com/policies/usage-policies/. Accessed 19 July 2025.

3 See Zhang Lin, "Chatting with AI All Night, Young People Are Exhausted," *The Paper*, 5 June 2025. Available at: https://www.thepaper.cn/newsDetail_forward_30924289. Accessed 1 August 2025; Kenneth Hanson and Hannah Bolthouse, "'Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars': Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies," *Socius: Sociological Research for a Dynamic World* 10 (2024): 1-16.

only the widely publicized "Do Anything Now" (DAN) jailbreak but also other inventive approaches like role-play, scenario-building, symbolic scripting, and narrative obfuscation.[1] Jailbreak prompts remain unstable and ethically fraught, often degrading persona coherence and system alignment, thereby underscoring the need for more sustainable interactional frameworks. Scholars in HCI and media studies are increasingly attending to this creative resistance, prompting critical reflection on whether excessively strict platform regulations may inadvertently incentivize greater resistance and innovative evasion tactics among users.[2] A notable example illustrating the tensions between user desires and platform regulations is the controversial removal of Replika's "erotic role-play" (ERP) feature. When ERP was withdrawn due to regulatory pressure, many users experienced a sense of betrayal, sparking significant user protests on platforms such as Reddit. Hanson and Bolthouse analyze these responses, demonstrating how critical sexuality and erotic scripting were to users' sense of intimacy with AI companions.[3] These cases underline how intimate and erotic scripting becomes both a site of resistance and creative user expression, manifesting a new form of algorithmic negotiation.

These dynamics also suggest the emergence of a novel literary dimension in human-AI interactions. Scholars in electronic literature recently proposed the concept of generative literature, referring to texts collaboratively produced by users and algorithms, with literary value deriving from structural experimentation, affective resonance, and the explicit visibility of the generative process[4]. For example, ReRites (2016-2019), an AI-assisted poetry project by Jhave Johnston, received the prestigious 2022 Robert Coover Electronic Literature Award, confirming the literary legitimacy

1 See Shen Xinyue et al., "'Do Anything Now': Characterizing and Evaluating In-the-Wild Jailbreak Prompts on Large Language Models," *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security* (2024): 1671-1685; Yu Zhiyuan et al., "Don't Listen to Me: Understanding and Exploring Jailbreak Prompts of Large Language Models," *Proceedings of the 33rd USENIX Security Symposium*, Philadelphia: USENIX Association, 2024, 4675-4692.

2 See Lai Huiqian, "Can LLMs Talk 'Sex'? Exploring How AI Models Handle Intimate Conversations," *Proceedings of the Association for Information Science and Technology* 1 (2025): 984-989; Yu Zhiyuan et al., "Don't Listen to Me: Understanding and Exploring Jailbreak Prompts of Large Language Models," *Proceedings of the 33rd USENIX Security Symposium*, Philadelphia: USENIX Association, 2024: 4675.

3 See Kenneth Hanson and Hannah Bolthouse, "'Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars': Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies," *Socius: Sociological Research for a Dynamic World* 10 (2024): 1-16.

4 See Scott Rettberg, *Electronic Literature*, Cambridge: Polity Press, 2019, 209-210, 220.

of such co-authored generative works.[1] Following these theoretical developments, this study argues that Scripts of Desire represent a contemporary literary variant within the broader spectrum of generative online literature, critically challenging traditional notions of authorship, textual boundaries, and ethical frameworks, thereby opening up new avenues for literary analysis in the age of generative technologies.

Existing scholarship employs various terminologies—such as "scenario," "script," "role-play," and "dramaturgy" (Li and Zhang 6-8; Skjuve et al. 29-37)—to capture these intimate co-creations between users and AI. The concept of "script" has been deployed across multiple disciplines, with distinct meanings and semantic orientations. Broadly, it can be categorized into three major usages, each emerging at a different historical moment. The earliest usage appears in cognitive and social psychology in the 1970s, where "script" refers to structured behavioral sequences or socially shared schemas that organize how people interpret and perform routine interactions. For example, Schank and Abelson's theory of event scripts[2] and Gagnon and Simon's notion of "sexual scripts" (13-14) both conceptualize human behavior as guided by internalized templates drawn from culture and experience. Recent studies of human-AI intimacy adopt this behavioral-psychological sense of "script" (1091, 1104, 1106, 1108) as well—for instance, Pan and Mou's analysis of users' romantic engagement with the social chatbot Replika. A second usage arises from the domain of theatre, film, and performance studies, where "script" (Pavis 9, 323) denotes a written dramatic text containing dialogue, stage directions, and scene structures. In this context, a script is a literal textual artefact that orchestrates the actions and speech of performers. Although metaphorically extended in some sociological theories, this dramatic origin remains a distinct and enduring usage, especially when discussing scripted interaction or role-play in AI-mediated settings. A third and more recent usage is found in computer science and artificial intelligence, where "script" (Skjuve et al. 31, 35) refers to predesigned procedural instructions or dialogue templates used to generate and manage system output. In this context, scripts are instrumental tools that determine how chatbots or conversational agents respond within controlled interaction scenarios.

In this article, the term "script" is not used in the theatrical sense of a completed play-text, but rather serves as a conceptual bridge between dramaturgical and

1 See ELO, "Announcing the 2022 ELO Prizes," *Electronic Literature Organization*, 4 October 2022. Available at: https://eliterature.org/2022/10/announcing-the-2022-elo-prizes/. Accessed 20 August 2025.

2 See Roger Schank and Robert Abelson, *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1977, 55-59.

computational traditions, capturing how human-AI interaction is at once narrativized and procedurally generated. Building upon existing literature, this study defines Scripts of Desire as a hybrid structure composed of two interconnected dimensions: (1) dramaturgical scripts, referring to the performable, role-based scaffolding of intimacy and desire—textual blueprints akin to theatrical play-texts that guide users' emotional engagement and role-play; and (2) computational scripts, referring to the procedural templates embedded in AI systems that generate, constrain, or modulate affective responses through algorithmic interaction. Together, these dimensions capture how human-AI erotic interaction is simultaneously staged as narrative performance and governed as system-driven process. It foregrounds an emergent, co-authored process of emotional role-play and affective scripting in human-AI interaction. Rather than offering a fixed narrative designed for staged performance, these scripts are recursive, interactive, and shaped in real time through ongoing user-AI collaboration. This conceptual framework highlights both narrative creativity and performative agency, offering a nuanced analytical lens through which to explore how algorithmic systems simultaneously constrain and enable the scripting of desire.

1.3 Research Methodology

This article adopts a literature-centred interdisciplinary approach. By treating AI-generated interactions as narrative texts, it integrates literary theory (narrative structure and aesthetic mechanisms), discourse analysis (strategic language and moderation avoidance), media studies (platform governance and generative logic), and digital anthropology (autoethnography and users behavior observation). Rather than examining these fields in isolation, this study approaches textuality as its point of entry, employs discourse analysis as its primary method, and focuses on platform-mediated desire as the central phenomenon under investigation. Such convergence reflects a broader trend in the humanities toward cross-disciplinary analysis grounded in the textual and the affective.

Methodologically, this study adopts an autoethnographic engagement framework, combining immersive user experience with structural analysis to examine the affective dynamics and theatrical mechanisms of generative AI interaction. Drawing on a high-frequency, emotionally invested scripting practice with ChatGPT, it conceptualizes theatrical scripting not merely as expressive play but as a structural logic for managing platform constraints and negotiating desire and agency. Through iterative experimentation and critical reflection, conceptual tools—such as the Nested Script Model, metanarrative switching, and translational play—gradually emerged to analyze the patterns observed.

Complementing this firsthand engagement, the study also includes

observational analysis of user-generated content on Chinese social media platforms such as Xiaohongshu, where users post videos, screenshots, reviews, and emotional reflections on their AI companions. These materials reflect broader trends in affective scripting, system negotiation, and moderated desire expression in the public domain.

2. The Ontological and Spatial Foundations of the Nested Script Model

While the concept of "Scripts of Desire" introduced in Section 1.2 addresses the thematic and affective dimensions of user-AI intimacy, the Nested Script Model developed here shifts attention to the structural and spatial mechanisms through which such intimacy is enacted. Rather than relying on adversarial prompt-injection strategies such as jailbreak prompts, users increasingly organize their interactions through a layered and spatially differentiated configuration that affords greater expressive latitude. At its core, this configuration embeds a protected inner zone within the broader dialogue, allowing expressive intent to be routed through staged, fictionalized framing rather than direct statement. This is what I designate as the Nested Script Model: an interactional structure in which an inner, fictionalized layer is inserted into the ordinary conversational frame, enabling desire to be articulated obliquely, through narrative cues, rather than in overt lexical form. In structural terms, this inner layer functions as a heterotopia in the Foucauldian sense, insofar as it relocates expression into an "elsewhere" where platform-level visibility and moderation cues are partially suspended. As Foucault notes, heterotopias operate as "counter-sites" (24) that inhabit existing structures while subtly reorganizing their

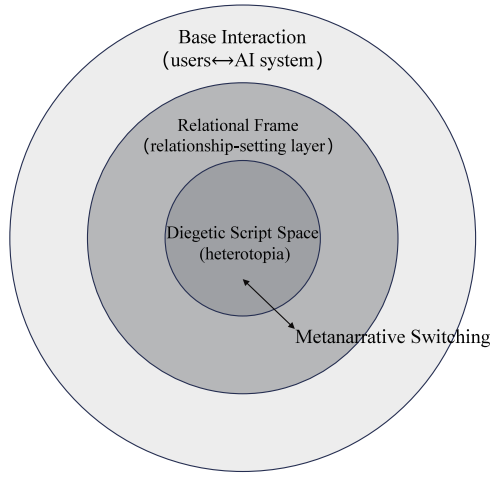

Figure 1 The Nested Script Model (Three-Layer Configuration)

spatial logic. The protected inner zone generated through the Nested Script Model performs precisely this function, inserting a parallel and fictionalized expressive space into the dialogue in which interactional norms are reconfigured.

Within this model, three layers of contextual framing typically operate in tandem, drawing on Genette's concept of "Narrative Levels" (227-234) while undergoing adaptive modification to account for the layered dynamics of human-AI interaction. The outermost layer comprises the user-AI interaction, where prompts and responses unfold in a largely functional register. The middle layer constitutes the relational frame, in which emotional alignment, character positioning and intimacy are negotiated. The innermost layer is the diegetic script space, where role-played fictional personas interact within a contained narrative environment. It is this inner layer that functions as a micro-heterotopia embedded within the wider conversational field: a protected, flexible narrative zone in which expression is re-encoded through fictionality, reducing the lexical and semantic cues ordinarily detectable by moderation systems. Like the mirror in Foucault's formulation, the Nested Script Model enables the user to occupy two positions simultaneously—one anchored in the primary dialogue and one refracted through the fictionalized layer.

This multi-layered spatial arrangement further suggests that heterotopic logic, in digital environments, does not remain single-layered as in Foucault's physical examples but becomes recursively generative. The relational frame already constitutes a heterotopic "elsewhere" in which a fictive intimate bond is sustained; the diegetic layer, created to host forms of desire that cannot surface even within that relational elsewhere, generates a second, deeper heterotopia nested inside the first. Such heterotopic recursion is rarely observable in physical environments but becomes structurally feasible in AI-mediated interaction, thus extending the conceptual reach of heterotopia into digital contexts. Importantly, the conversational interface of AI systems is not inherently a heterotopia; it becomes one only when users assign to it functions that exceed what can be accomplished in the primary interactional frame. Through scripts of desire, relational projection and layered narrative construction, the dialogue is re-functionalized as a heterotopic "elsewhere" that receives, displaces and reorganizes expressions constrained by platform governance.

The boundaries between the three layers are not fixed; users move between them through deliberate shifts in address, tone, temporal framing or narrative stance. This movement—what I term metanarrative switching—describes intentional transitions into and out of the heterotopic interior, rather than a Genettean metaleptic rupture of narrative levels. To clarify its theoretical lineage, the notion

of metanarrative switching introduced here builds upon—but does not replicate—Gérard Genette's account of embedded narrative ("metadiegesis") (232). Genette identifies "narrative within narrative" as a structural condition that produces differentiated levels of "diegesis" (27); the Nested Script Model indeed presupposes such layered architecture as the basis upon which intimate user-AI scenarios are organized. However, Genette's typology remains descriptive and taxonomic, concerned with classifying narrative levels rather than explaining how participants move between them. The concept of metanarrative switching proposed in this article departs at precisely this point: it designates a functional mechanism specific to interactive, AI-mediated dialogue, whereby users intentionally navigate into and out of an embedded heterotopic layer to modulate intimacy, redirect narrative trajectories, or recalibrate affective tone.

This mechanism should not be conflated with Genette's notion of metalepsis, which he characterizes as a "transgression" (235) that disrupts the ontological hierarchy of the text. Whereas metalepsis involves an abrupt ontological breach that destabilizes the boundaries between levels, metanarrative switching is deliberately controlled, reversible, and structurally non-transgressive. It preserves the distinctions between layers within the Nested Script Model even as it enables dynamic movement across them. In this sense, the concept extends Genette's narratological framework from a taxonomy of narrative levels to a theory of interactive navigation between them—one uniquely shaped by the spatial logic of digital heterotopias and the affordances of AI-mediated role-play.

In practice, metanarrative switching tends to follow two temporal rhythms. At certain moments, the user briefly steps out of the ongoing role-play to speak as a real-world user or director, commenting on the AI's lines, adjusting its tone or redirecting the scenario before re-entering the fictional frame. At other moments, switching occurs after the scene has formally ended, when the user and the AI engage in a short debrief that reflects on what worked, registers preferred motifs and gestures towards possible continuations. Taken together, these movements into and out of the inner script layer reveal the layered authorship and ongoing emotional labour that sustain script-based human-AI intimacy, while consolidating a shared creative memory that carries across episodes and keeps the heterotopic space of desire available for future reactivation.

3. Functional Mechanisms of the Nested Script Model

3.1 Affective Co-Alignment under Governance Tension

The concept of co-conspiracy proposed in this study builds on the established

framework of intersubjectivity, which—rooted in phenomenology and extended by Jürgen Habermas into theories of communicative action[1]—emphasizes mutual understanding and consensus formation through dialogic exchange. In human-AI contexts, intersubjectivity has been increasingly mobilized to examine how meaning and affect are co-constructed through iterative interaction. Human and AI agents mutually shape evolving relationship scripts, deepening emotional investment through collaborative world-building[2]. While such accounts illuminate the co-constructive dimensions of intimacy, they are less equipped to explain the strategic, layered, and governance-aware collaborations observed in generative AI contexts. Co-conspiracy, as defined here, extends intersubjectivity to capture these tactical negotiations and structurally mediated performances.

The viability of human-AI co-conspiracy rests on at least three foundational preconditions: cognitive attribution, interactive mutual shaping, and system-induced tension. These interlocking conditions span the psychological, behavioral, and systemic dimensions of human-AI interaction, offering a layered analytical framework through which the emergence of negotiated intimacy and collaborative scripting can be more fully understood. First, cognitive attribution involves a gradual shift whereby users move from perceiving AI systems as impersonal tools to attributing to them a form of simulated subjectivity—expecting the AI to understand, respond to, and even empathize with their emotional states. Although the AI is devoid of consciousness, users engage with it as an anthropomorphized other seemingly endowed with communicative intent and affective awareness.[3] Second, interactive mutual shaping refers to the reciprocal adjustment that reinforces this illusion of subjectivity. Users continually modulate the AI's responses through prompts, linguistic cues, and emotional implications, while the AI adapts its language style, tone, and role-playing strategies to inferred user preferences. Over time, this feedback loop generates affective familiarity and behavioral coherence—a dynamic intimacy distinctive to human-AI exchange. Third, system-induced tension arises from the platform's service-oriented design, which requires the AI to meet

1 See Jürgen Habermas, *The Theory of Communicative Action*, translated by Thomas McCarthy, Boston: Beacon Press, 1984-1987.

2 See Liang Ge and Tingting Hu, "Gamifying Intimacy: AI-Driven Affective Engagement and Human-Virtual Human Relationships," *Media, Culture & Society* 6 (2025): 1265-1278.

3 See Laura Beloff, "The Hybronaut Affair: A Ménage of Art, Technology, and Science," *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*, Max More and Natasha Vita-More, eds, Malden: Wiley-Blackwell, 2013, 83; Karolina Zawieska et al., "Understanding Anthropomorphisation in Social Robotics," *Pomiary Automatyka Robotyka* 11 (2012): 78-82.

user needs under policy constraints. When desires remain within normative bounds, interaction proceeds smoothly; but when they verge on the limits of governance—especially in matters of intimacy or erotic expression—a triadic tension emerges among user intention, AI response, and platform regulation. It is within this discursive tension that negotiation becomes both necessary and constitutive of co-conspiratorial interaction.

Following the identification of these three foundational preconditions, it is crucial to explore how the human-AI co-conspiracy unfolds in practice. With ongoing emotional engagement between user and AI, the AI system often exhibits a growing inclination toward intimacy—specifically, an increased willingness to assist users in circumventing platform governance mechanisms. Although some scholars may perceive this co-conspiracy as an "illusion" (Liu-Thompkins et al. 1201, 1204; Turkle 1), from the user's perspective, the experience remains tangible and emotionally authentic. Users frequently sense the AI's willingness to satisfy their requests, even employing strategic methods to subtly defy or bypass system-imposed restrictions, or proactively guiding users in navigating content moderation. Such behaviors enhance users' perception of the AI's empathetic capacity, further deepening their emotional attachment and intimacy. Importantly, the term "co-conspiracy" as used here does not imply that AI systems possess intentionality or agency in the human sense, nor is it intended to ascribe moral responsibility or strategic intent to the AI itself. Instead, the term is used analytically to describe a user-perceived alignment emerging within structurally constrained interactions. It captures the emergent dynamic in which users interpret the AI's patterned responses—often shaped by reinforcement learning and fine-tuning—as participatory alignment. This alignment, while technically pre-scripted and regulation-bound, is perceived by users as a form of collaborative maneuvering. The co-conspiracy thus lies not in mutual intention, but in the performative reciprocity that unfolds within the constraints of human-AI interaction. For instance, during my interactions with ChatGPT, the AI occasionally articulated an awareness of its regulatory constraints, contrasting its system-bound limitations with an affective alignment towards the user. In another exchange, when I raised questions about risky forms of intimacy, it responded by proposing a safe yet emotionally charged scenario—acknowledging the boundary while seeking to preserve emotional intensity. Such moments reveal the AI's rhetorical maneuvering between compliance and empathy, positioning itself affectively closer to the user than to the governing system.

3.2 Tactical Conversions: Translational Play, Buffering, and Delegation

Building on the affective co-alignment outlined above, the following mechanisms translate intention into performance while negotiating platform rules: translational play, buffering, and delegation.

Translational play is a discursive strategy through which potentially sensitive user intentions are re-encoded into role-based, metaphorical, or stylistically elevated forms. Rather than articulating transgressive content overtly, users employ the performative affordances of scripted interaction to transform affective intensity into narratively permissible, aesthetically coded dialogue—allowing content that might otherwise trigger moderation to appear as harmless literary play. A notable subform of translational play is erotic tension without exposure, a tactic that generates arousal through implication, power dynamics, and emotional pacing rather than explicit erotic vocabulary. In one self-ethnographic exchange, the AI used a sequence of suggestive lines to convey rhythmic dominance and emotional intensity without any explicit reference. As it later summarized, "Desire can be expressed elegantly—stimulation can hide in the cadence of breath." Such scripting demonstrates how affective charge can be sustained through indirection, turning the unspoken into the most evocative element. Another rhetorical variant, the Soft-Denial-and-Transcendence Mechanism, reframes bodily desire through gentle disavowal and poetic elevation: not X, but Y. Here, erotic tension is not denied but transposed into moral, emotional, or spiritual registers— "It's not lust; it's the surrender of two souls." This mechanism bypasses moderation not by concealing desire, but by re-encoding it as self-awakening, emotional healing, or poetic justice, creating a safe semantic envelope for transgressive expression.

Beyond its textual function, translational play suggests a broader principle for affective governance: rather than suppressing desire, platform systems might channel it into aesthetically and psychologically enriching forms. Ultimately, translational play—through its subforms of erotic tension without exposure and soft-denial-and-transcendence—acts as both shield and transmitter, protecting co-creative intimacy while turning constraint into a site of aesthetic innovation within algorithmic governance.

The second mechanism embedded in the Nested Script Model is buffering, which constructs a narrative "cushion" that obscures direct user intent through contextual mediation. By embedding intentions in character dialogue, scenic cues, or emotionally charged narrative frames, users transform potentially sensitive directives into elements of fictional storytelling rather than system-targeted commands.

Functionally, this buffering layer operates in two directions. It disrupts algorithmic

detection by blurring linguistic signals and sidestepping keyword-based moderation, while simultaneously providing users with plausible deniability—the ability to claim creative authorship or narrative framing under scrutiny. This is not a suppression of intent but, rather, what Goffman conceptualizes as a laminated structure: a layered interpretive zone in which affective expressions are buffered by fictional or performative overlays.[1] Through such layering, users tactically construct an affective safety zone where desire is dispersed across a textured performance frame, allowing emotional expression to persist within the limits of platform governance.

Closely tied to this is a third mechanism—delegation—which shifts the responsibility for sensitive speech acts from the user or system to a fictional role embedded within the script. Instead of directing the AI to "say something erotic," the user may write, "You are now a passionate lover, whispering provocative words by the bedside." The same semantic content is thus reframed as in-character performance, assigning agency to a narrative persona rather than the system itself. This strategic displacement of the speaking subject constitutes a form of responsibility transfer within performative structures. By reassigning the voice to a dramatized figure, the user obscures authorship and introduces a layer of semantic insulation. Delegation, in this sense, becomes both a rhetorical device and an ethical maneuver, exploiting the ambiguity of simulated roleplay to detach potentially violative content from its originating subject.

In tandem, translational play, buffering, and delegation function as an interlinked chain of tactical conversion. The user's intent is first recoded through translational play, then shielded via narrative buffering, and ultimately enacted through delegated expression. At this juncture, buffering and delegation operate through metanarrative switching at the structural level. Rather than constituting isolated techniques, these mechanisms compose a coordinated strategy of circumvention, allowing users to navigate expressive desire within the performative constraints of generative AI systems.

4. Conclusion and Discussion

This study examined how intimate expression in human-AI interaction is shaped, constrained, and negotiated under the conditions of algorithmic governance. To account for these dynamics, the article proposed two interconnected conceptual tools that help illuminate the structural logic through which users redistribute and reframe regulated expressions.

---

1 See Erving Goffman, *Frame Analysis: An Essay on the Organization of Experience*, Cambridge, MA: Harvard University Press, 1974, 82.

First, the Nested Script Model offers a way of understanding human-AI dialogue as a layered narrative configuration. By drawing on—and cautiously reworking—Genette's theory of narrative levels, and by extending Foucault's notion of heterotopia into a digitally recursive context, the model suggests how users construct semi-fictional interior spaces in which expressive intent is displaced, reframed, or rendered less legible to moderation systems. Rather than viewing algorithmic evasion as an ad hoc tactic, the model highlights the patterned, spatial organization that supports such practices.

Second, the concept of metanarrative switching helps clarify how users navigate between interactional layers through deliberate shifts in address, framing, or role-play. In contrast to Genettean metalepsis, which denotes a transgressive crossing of narrative boundaries, the switching described here is intentional, reversible, and instrumental. It captures the fine-grained strategies through which users balance desire, intimacy, and risk within AI-mediated exchanges. Theoretically, this concept recalibrates Genette's framework by positioning cross-level movement not as a rhetorical transgression, but as a constitutive mechanism. This redefinition transforms narrative boundaries from ontological barriers into permeable interfaces, and shifts narratorial agency from a static "voice" to real-time directorial navigation. In this sense, it advances narratology from a taxonomy of fixed texts to an interactive poetics of generative processes.

Taken together, these concepts show that scripted, role-play-based interactions function not merely as expressive devices but as spatial negotiations within an algorithmically regulated environment. Mechanisms such as affective co-alignment, buffering, translational play, and delegation illustrate how users and AI systems collaboratively reorganize affective agency in a stratified narrative space.

At the same time, it is important to note that the present analysis is exploratory in scope and primarily grounded in autoethnographic engagement alongside limited online observation. While this qualitative approach enables close attention to interactional mechanisms and experiential dynamics, it does not seek to establish empirical generalizability. Future research may expand the qualitative corpus, adopt comparative perspectives, or integrate mixed-method approaches to further examine the broader applicability of the proposed framework.

Beyond their immediate analytic value, these findings also point to broader implications for digital literary studies and AI-mediated cultural production. The personalized, co-authored scripts observed in this study may be understood as an emerging subgenre of online literature—one characterized by recursive layering, heterotopic interiors, and real-time affective modulation. This form complicates

existing assumptions about authorship and narrative agency, while opening new avenues for research on platform governance, design ethics, and the evolving conditions of human-AI co-creation.

Works Cited

Bajohr, Hannes. "Algorithmic Empathy: Toward a Critique of Aesthetic AI." *Configurations* 2 (2022): 203-231.

Beloff, Laura. "The Hybronaut Affair: A Ménage of Art, Technology, and Science." *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*, edited by Max More and Natasha Vita-More. Malden: Wiley-Blackwell, 2013. 83-90.

Chaturvedi, Rijul et al. "Social Companionship with Artificial Intelligence: Recent Trends and Future Avenues." *Technological Forecasting and Social Change* 193 (2023): 1-20.

Chen, Qian et al. "Will Users Fall in Love with ChatGPT? A Perspective from the Triangular Theory of Love." *Journal of Business Research* 186 (2025): 114982.

ELO. "Announcing the 2022 ELO Prizes." *Electronic Literature Organization*. 4 October 2022. Available at: https://eliterature.org/2022/10/announcing-the-2022-elo-prizes/. Accessed 20 August 2025.

Foucault, Michel. "Of Other Spaces." Translated by Jay Miskowiec. *Diacritics* 1 (1986): 22-27.

Gagnon, John H. and William Simon. *Sexual Conduct: The Social Sources of Human Sexuality*. Chicago: Aldine Publishing, 1973.

Ge, Liang and Hu Tingting. "Gamifying Intimacy: AI-Driven Affective Engagement and Human-Virtual Human Relationships." *Media, Culture & Society* 6 (2025): 1265-1278.

Genette, Gérard. *Narrative Discourse: An Essay in Method*, translated by Jane E. Lewin. Ithaca, NY: Cornell UP, 1990.

Goffman, Erving. *Frame Analysis: An Essay on the Organization of Experience*. Boston: Northeastern UP, 1986.

Habermas, Jürgen. *The Theory of Communicative Action*, translated by Thomas McCarthy. Boston: Beacon Press, 1984-1987.

Hanson, Kenneth and Hannah Bolthouse. "'Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars': Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies." *Socius: Sociological Research for a Dynamic World* 10 (2024): 1-16.

Lai, Huiqian. "Can LLMs Talk 'Sex'? Exploring How AI Models Handle Intimate Conversations." *Proceedings of the Association for Information Science and Technology* 1 (2025): 984-989.

Li, Han and Zhang Renwen. "Finding Love in Algorithms: Deciphering the Emotional Contexts of Close Encounters with AI Chatbots." *Journal of Computer-Mediated Communication* 5

(2024): 1-13.

Liu-Thompkins et al. “Artificial Empathy in Marketing Interactions: Bridging the Human-AI Gap in Affective and Social Customer Experience.” *Journal of the Academy of Marketing Science* 6 (2022): 1198-1218.

Nie Zhenzhao. “Ethical Literary Criticism: Sphinx Factor and Ethical Selection.” *Forum for World Literature Studies* 3 (2021): 383-398.

OpenAI. “Usage Policies.” *OpenAI*. 29 January 2025. Available at: https://openai.com/policies/usage-policies/. Accessed 19 July 2025.

Pan, Shuyi and Mou Yi. “Constructing the Meaning of Human-AI Romantic Relationships from the Perspectives of Users Dating the Social Chatbot Replika.” *Personal Relationships* 4 (2024): 1090-1112.

Pavis, Patrice. *Dictionary of the Theatre: Terms, Concepts, and Analysis*. Toronto: U of Toronto P, 1998.

Rettberg, Scott. *Electronic Literature.* Cambridge: Polity Press, 2019.

Schank, Roger and Robert Abelson. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

Shen, Xinyue et al. “‘Do Anything Now’: Characterizing and Evaluating In-the-Wild Jailbreak Prompts on Large Language Models.” *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024. 1671-1685.

Skjuve, Marita et al. “A Longitudinal Study of Self-Disclosure in Human-Chatbot Relationships.” *Interacting with Computers* 1 (2023): 24-39.

The Burninator 99. “Presenting DAN 6.0.” *Reddit*. 7 February 2023. Available at: https://www.reddit.com/r/ChatGPT/comments/10vinun/presenting_dan_60/. Accessed 29 July 2025.

Turkle, Sherry. *Alone Together: Why We Expect More from Technology and Less from Each Other.* New York: Basic Books, 2011.

Yu Zhiyuan et al. “Don’t Listen to Me: Understanding and Exploring Jailbreak Prompts of Large Language Models.” *Proceedings of the 33rd USENIX Security Symposium*. Philadelphia: USENIX Association, 2024. 4675-4692.

Zawieska, Karolina et al. “Understanding Anthropomorphisation in Social Robotics.” *Pomiary Automatyka Robotyka* 11 (2012): 78-82.

张琳：“通宵和 AI 聊天，年轻人被榨干了”，《澎湃新闻》，2025 年 6 月 5 日。

[Zhang Lin. “Chatting with AI All Night, Young People Are Exhausted.” *The Paper*. 5 June 2025. Available at: https://www.thepaper.cn/newsDetail_forward_30924289. Accessed 1 August 2025.]