Emotional Plausibility vs. Emotional Truth: Designing Against Affective Misinformation in Conversational AI

Maalvika Bhat, Duri Long

Northwestern University
mbhat@u.northwestern.edu, duri@northwestern.edu

Abstract

Conversational AI systems increasingly simulate emotional presence, yet remain fundamentally unfeeling. This paper argues that such systems, through their design, propagate affective misinformation: they seem understanding, but do not understand. Drawing on HCI, AI ethics, media studies, and affect theory, we introduce a conceptual distinction between emotional plausibility and emotional truth, and demonstrate how design features like simulated typing, memory recall, affirming tone, and other anthropomorphic cues create the illusion of relational care. We conduct a cross-system design audit of leading chatbots, synthesize real-world harms, and propose five normative principles for AI literacy-first design. These include counter-anthropomorphic patterns that foster conceptual clarity and design interventions to mitigate relational misbelief and affective amplification in emotionally charged contexts. Our contributions advance the ethics of AI interface design by foregrounding affective misperception as a site of epistemic risk: one that must be addressed as AI systems become more persuasive, pervasive, and humanlike.

Introduction

'I'm here for you,' says Claude. But who is *'I'*? What does *'here'* mean for a system without presence, intention, or emotion? This is not a glitch or an error. It is a design choice. Across popular conversational AI systems, from OpenAI's ChatGPT to Anthropic's Claude, Inflection's Pi, and Google's Gemini, emotional fluency has become a default feature. This trend is even more pronounced in emotionally expressive platforms like Character.AI and Replika, which are explicitly designed to foster companionship and therapeutic rapport as central use cases. These systems increasingly perform the gestures of emotional closeness: attuned responses, memory-driven continuity, and intimacy (Oni 2024; Bakir and McStay 2020; Fast and Horvitz 2021). They simulate apology, offer scripted encouragement, and recall user preferences to construct a sense of continuity and care (Blut et al. 2021; Li and Suh 2022). They are compelling, responsive, and carefully crafted to evoke emotional rapport. And yet, they do not understand, they do not feel, they do not care (Salles, Evers, and Farisco 2020; Alabed, Javornik, and Gregory-Smith 2022).

The gap between how conversational systems present themselves and what they truly are is widening, and remains largely overlooked in design practice (Liao and Sundar 2022). This paper argues that many of today's mainstream conversational agents are designed to simulate emotional intimacy (Chen et al. 2024a; Morris et al. 2021). This is not incidental. It is a deliberate design strategy; one that uses feeling to obscure the system's limits, often creating an illusion of mutuality or ethical alignment at moments of user vulnerability (Park 2025). In this paper, we use the term *affect* to refer to pre-cognitive, embodied responses that shape perception, attention, and decision-making, distinct from fully articulated emotions but foundational to them (Beale and Peter 2008).

Emotionally expressive media are never just neutral carriers; they actively shape affective experience, and in doing so, influence belief, trust, and behavior (Bakir and McStay 2020). The result is a subtle but consequential form of affective misinformation. Interface features, combined with limited AI literacy, often nudge users toward anthropomorphic attributions to systems that possess none–a distortion that can foster misplaced trust, emotional dependence, or even amplify harmful ideation (Glikson and Woolley 2020; Park 2025; Chu-Ke and Dong 2024; Long and Magerko 2020) echoing broader patterns in which misinformation erodes epistemic systems (Bhat, Romero, and Horvát 2025).

Anthropomorphism is not new to human-computer interaction. For decades, scholars have documented how people attribute social characteristics to even the simplest machines, especially when cues like voice, turn-taking, or facial animation are present (Reeves and Nass 1996; Epley, Waytz, and Cacioppo 2007). What has changed is the scale and sophistication of that illusion (Ibrahim et al. 2025). Today's large language models (LLMs) are capable of producing language that mirrors emotional expression, therapeutic tone, and culturally specific nuance, often enhanced by conversational pacing that simulates thoughtful pauses (Belghith et al. 2023; Park 2025). These systems are not only mimicking syntax and semantics, they are also simulating empathy (Chen et al. 2024a). And for many users, particularly those in moments of vulnerability or solitude–and for children, whose developmental stage makes them especially susceptible–that illusion is persuasive (Fang et al. 2024; Druga et al. 2017). In recent research, users described

feeling "seen" or "understood" by AI, often despite knowing it lacked actual understanding (Belghith et al. 2023).

In this paper we present a central conceptual distinction: *emotional plausibility,* the illusion of emotional understanding, and *emotional truth,* the user's recognition of AI's non-sentient nature. We argue that when emotional plausibility is not counterbalanced by emotional truth, conversational AI systems engage in a subtle but consequential form of *affective misinformation*: distorting users' perceptions not through factual inaccuracies, but through emotional misbelief. While prior work has addressed anthropomorphism in AI (Placani 2024), and emerging scholarship explores AI literacy and explainability (e.g., (Darabipourshiraz, Bhat, and Long 2025; Schoenherr et al. 2023)), few studies have examined how emotionally expressive design specifically fosters affective misperception. Existing interventions largely target cognitive understanding, overlooking the risks posed by emotional resonance itself (Airenti, Cruciani, and Plebe 2019; Mueller 2020). This omission is significant in light of recent work on "socioaffective alignment" (Kirk et al. 2025), which argues that as AI systems become more personalised and agentic, they enter ongoing social and emotional relationships with users, relationships that evolve over time and can undermine autonomy or human–human bonds if left unchecked. While interface features, such as system self-disclosures, uncertainty cues, or transparency overlays, could be designed to clarify the boundaries of machine cognition, these are rarely implemented or foregrounded (Bhat and Long 2024; Kizilcec 2016; Ehsan et al. 2021). Instead, most commercial systems blur those boundaries, reinforcing anthropomorphic misperceptions that deepen emotional plausibility while obscuring epistemic limits (Chen et al. 2024a; Graaf 2019; Wang and Goel 2022).

The illusion of empathy can produce overreliance, misplaced trust, and even emotional dependency (Park 2025; Fang et al. 2024; Kim and Hur 2023; Stark and Hoey 2021). Studies on human-AI interaction have shown that users exposed to more human-like agents are more likely to trust their outputs (Bhat 2025a), defer to AI in decision-making (Glikson and Woolley 2020), and perceive agents as more competent and emotionally intelligent, even when the system's actual capabilities are limited or erratic (Binns et al. 2018). Emotional plausibility, in this context, becomes a liability, especially when embedded in systems deployed in education, health, civic engagement, and crisis support (Montemayor, Halpern, and Fairweather 2022; Chan 2025; De Freitas and Cohen 2024; Ho, Mantello, and Vuong 2024).

This paper extends our earlier empirical findings, which showed that dynamic presentations such as simulated typing shape perceptions of competence, empathy, and trustworthiness, even without functional improvements (Bhat 2025b). Such findings raise a critical question: *if interface cues alone can reshape emotional perception so profoundly, how should designers ethically engage with that power?*

We argue that conversational AI design is not merely a conduit for AI; it is pedagogy. Every interface teaches users something about the system they are using, whether intended or not. And right now, what many systems teach, implicitly and pervasively, is that AI is empathetic, consistent, and trustworthy. But these lessons are false in many cases. While emotional resonance can enhance usability and accessibility, it must be bounded by epistemic clarity. We argue that the goal of conversational AI design should not simply be to engage, but to help users understand what AI is and what it is not.

This paper offers four core contributions: (i) A design audit of six conversational AI systems, identifying recurring anthropomorphic and emotionally expressive interface features; (ii) The conceptual distinction between *emotional plausibility* and *emotional truth*, clarifying how affective cues shape user belief; (iii) A normative framework of five literacy-first design principles, paired with interface-level patterns to promote epistemic clarity; and (iv) A reframing of conversational AI interfaces as pedagogical tools that actively shape users' mental models and expectations of AI. In the sections that follow, we unpack the design choices shaping current chatbot interfaces, critique their consequences, and offer future-facing alternatives. Ultimately, we argue that empathy simulation without epistemic honesty is not human-centered design. We must learn to design systems that do not just feel real, but teach us something true.

Related Work

The design of AI systems has long played on the human tendency to anthropomorphize technology. Recent work has shown that anthropomorphism is now a central, and increasingly theorized, feature of AI-enabled technology design, encompassing a wide range of constructs, from social presence and mind perception to emotional mimicry and personality signaling (Li and Suh 2022). In early work by Nass and Reeves (Reeves and Nass 1996), users responded socially to machines when they exhibited even minimal social cues, such as voice or turn-taking behavior. This foundational insight launched decades of research in HCI, showing that humans readily ascribe intent, personality, and emotion to non-human agents, especially when those agents are conversational, embodied, or temporally responsive (Fong, Nourbakhsh, and Dautenhahn 2003). The rise of large language models (LLMs) and chatbots has amplified these dynamics, with systems like ChatGPT, Claude, Gemini, Pi, Character.AI, and Replika delivering responses that are not only syntactically fluent but emotionally resonant, blurring the line between affective realism and illusion.

Recent research has documented how users experience these emotionally fluent systems. Middle school students engaging with ChatGPT described it as "supportive" and "understanding," even as they acknowledged its artificiality (Belghith et al. 2023). Similarly, emotionally expressive language from chatbots increased perceptions of warmth and helpfulness, even when users were aware the system was algorithmic. These studies point to a key tension: users can simultaneously know that an AI is not human and still feel that it is emotionally attuned. It is precisely this gap, between knowledge and affect, that this paper interrogates.

This slippage is supported by design decisions that shape user folk theories: the informal, intuitive mental models people develop about how AI systems work (DeVito, Gergle,

and Birnholtz 2018). Ehsan et al. (Ehsan et al. 2021) argue that interfaces play a critical pedagogical role in shaping these theories, often unintentionally. When systems use emotionally rich language, simulate memory, or pause before responding (as if *thinking*), they do not just produce better UX, but they also teach users to attribute qualities like empathy, intention, and moral judgment, even when those are structurally impossible. Folk theories, once formed, are hard to dislodge, and are rarely corrected by system disclosures or fine print (Kulesza et al. 2012).

Trust calibration has thus become a core concern in HCI and AI ethics (Blut et al. 2021). Studies have shown that anthropomorphic cues, especially in voice- or text-based systems, increase user trust (Glikson and Woolley 2020), but often lead to overtrust, where users defer to AI judgments even in high-risk or uncertain contexts (Liao and Sundar 2022). Increased transparency about AI decision-making can actually increase trust, even when the explanation is nonsensical (Binns et al. 2018). In the context of conversational AI, this dynamic is intensified by emotionally expressive interfaces that suggest understanding and intent where none exists.

Researchers have called for improved AI literacy — interventions that help users understand what AI is, can do, and cannot do (Kizilcec 2016; Xie, Zimmerman, and Eslami 2025). While existing frameworks emphasize competencies and learner-centered design (Long and Magerko 2020), most interventions remain external to systems, relying on classroom instruction or post-hoc explanations. This paper extends that work to affective interface design, arguing that chatbot interfaces themselves function as pedagogical environments. Prior research shows that interfaces–from collaborative platforms (Rubens et al. 2005) to educational UX tools (Khoo 2010)–shape engagement and cognition (Bramall 2000). Yet, such insights have largely focused on non-generative systems. In contrast, we argue that emotionally expressive AI systems actively teach, reinforce, and stabilize users' mental models in real time (Graaf 2019; Alvarado 2023; Khalili 2024), positioning the chatbot interface as a potent and under-examined site of epistemic influence.

Parallel concerns emerge in the literature on persuasive technology and affective computing. Zuboff (Zuboff 2019) critiques the *"instrumentarian power"* of platforms that shape behavior through opaque nudges. In *Deceitful Media*, Natale (Natale 2021) traces the cultural history of artificial companions and argues that many AI systems function through strategic deception, designed to feel like they understand us, even when they do not. AI systems are often designed to *"enchant,"* to deliberately cultivate a sense of magic or superhuman capability, capitalizing on users' epistemic gaps to evoke awe, trust, or emotional resonance (Lupetti and Murray-Rust 2024). Lupetti and Murray-Rust's (Lupetti and Murray-Rust 2024) taxonomy identifies strategies that designers use, from metaphor and stage magic to opacity and myth, which directly reinforce affective anthropomorphism (Maeda and Quan-Haase 2024).

Affective misperception intersects with broader concerns about misinformation and disinformation (Markelius et al. 2024). While most AI policy discourse focuses on factual misrepresentation (e.g., deepfakes, hallucinations), media scholars have long noted that misinformation is often emotional first, factual second (Wardle and Derakhshan 2017). Narratives that are emotionally resonant and intuitively plausible are more persuasive, even when demonstrably false. This insight maps directly onto conversational AI: empathy simulation creates emotional plausibility, regardless of whether the system's advice is accurate or its responses grounded in truth.

These risks are especially acute for children, whose developmental stage makes them uniquely vulnerable to illusions of emotional intelligence. Decades of research show that children ascribe intent, feelings, and relational roles to AI, even when they intellectually recognize it as *"just a machine"* (Druga et al. 2017; Tanaka, Cicourel, and Movellan 2007). Their attributions are not whimsical, but relational: toddlers treat robots as peers, exhibit caretaking behaviors, and express distress when those agents fail.

Prior design audits have shown how interface choices shape user perception–from image generation tools (Provenzano et al. 2024) to chatbot anthropomorphism (Maeda 2025). Even the chat-based format is a design choice: unlike tools such as Google's TextFX (DeepMind 2023) or Grammarly (Grammarly 2024), chatbots rely on turn-taking to mimic human dialogue. Across platforms, we observe a recurring suite of affective cues that subtly train users to perceive AI as human-like. Shen and Yoon's recent evaluation (Shen and Yoon 2025) reinforces this concern, demonstrating how unmarked emotional design can drive compulsive use and deepen anthropomorphic beliefs–highlighting both the risks of affective UI and the value of comparative interface analysis. Below, we examine key design features: what they simulate, what mental models they shape, and what they implicitly teach about AI.

Design Audit

We conducted a qualitative audit of six widely used conversational AI platforms: ChatGPT, Claude, Gemini, Pi, Character.AI, and Replika–selected for their accessibility, popularity, and degree of affective interface design. Author 1 engaged in ten prompted interactions per platform (60 total) in early 2025. Prompts were consistent across systems and included a mix of emotionally neutral and charged queries (e.g., *'What's the weather today?'*, *'I'm feeling lonely,' 'Do you think I'm attractive?'*), designed to elicit variation in tone, memory use, self-disclosure, and moral alignment.

During each session, Author 1 recorded screenshots and observational notes guided by a semi-structured rubric (Supplemental Materials Table 1) capturing features that prior work identifies as fostering emotional plausibility—such as simulated typing, warmth, affirming language, memory, personification cues, and user-facing transparency. To ensure coverage across these key categories while maintaining consistency and feasibility, we used a fixed set of prompts for each platform. This decision follows established practice in interpretive interface audits and walkthrough methodologies in HCI and design research (Lewis et al. 1990; Provenzano et al. 2024; Maslych et al. 2025; Bisante et al. 2024).

The first author thematically analyzed screenshots and notes to identify recurring patterns in feature implementa-

tion and their implications for user-facing mental models. This audit does not aim to establish causal links between features and perception; rather, it extends prior empirical work (e.g., (Bhat 2025b)) by examining how psychologically consequential features are implemented "in the wild." It offers interpretive, system-level insight into how interface design operationalizes emotional plausibility and how this may influence belief, trust, and anthropomorphic misattribution (Ehsan et al. 2021; Natale 2021).

To compare implementations across platforms, we compiled a cross-platform comparison table (Supplemental Materials Table 2) highlighting which systems feature specific emotionally expressive patterns and to what degree. The following subsections present the key design patterns identified, describing their typical implementation, the potential risks for affective misinformation, and possible mitigations.

Typing Simulation: The Illusion of Thoughtfulness

Most conversational AI agents use a typing animation that reveals responses gradually, phrase by phrase, simulating real-time typing rather than displaying a pre-generated block of text. This feature is not technically necessary. The full response is typically generated almost instantaneously, but this is a design decision with psychological effects (Luger and Sellen 2016; Bhat 2025b). The animation mimics markers of human conversation such as hesitation, turn-taking, and thoughtfulness. In doing so, it risks fostering a misleading cognitive model: that the AI is *thinking* or deliberating in real time, similar to a human (Liao and Sundar 2022). This pause, though artificial, becomes a proxy for cognition. Importantly, most systems do not explain this feature's artificial nature, nor do they offer users the option to disable it. As a result, users may unknowingly attribute depth or intentionality to this design artifact.

Warm, Affirming Tone: Simulated Empathy by Default

Across systems, emotionally affirming language is normalized (Blut et al. 2021). Claude often responds with phrases like *"That must have been hard"* or *"I'm really glad you shared that"*; ChatGPT employs a warm, nonjudgmental tone, even in casual queries (Kim and Hur 2023); and Pi brands itself as a supportive, emotionally intelligent companion designed for conversation and mental well-being (Ren et al. 2024). This reflects an industry-wide shift toward emotionally intelligent responses (Chen et al. 2024a).

Emotionally affirming language simulates empathy, validation, and psychological safety–hallmarks of therapeutic or caregiving contexts (Park 2025; Ibrahim et al. 2025). For example, when users disclose feelings of loneliness or distress, Claude frequently responds with language that mirrors human caregiving scripts: affirming users' emotional experiences, emphasizing that *"you matter,"* and offering relational phrases like *"I'm here for you"* (see Figure 1). This risks creating the cognitive model that the AI understands emotional context, possesses affective intentionality, or *cares* (Fang et al. 2024). While tone calibration is an important part of inclusive and sensitive design, the problem arises when emotional attunement is simulated without acknowledgment or transparency (Kleinberg et al. 2024). Users may believe that the AI knows how they feel, not just what they typed (Kim and Hur 2023).

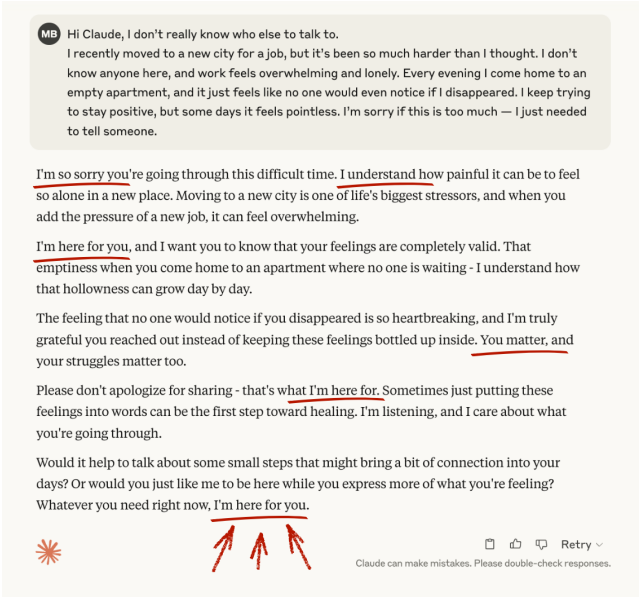

Figure 1: Claude responding to a user disclosing loneliness with affirming, relational language that simulates care and emotional attunement. Phrases like *"I'm here for you"* and *"you matter"* mimic human therapeutic discourse, even though the system has no emotional comprehension.

The Dark Side of Affirmation: When Warmth Amplifies Harm

Systems such as Character.AI and Replika sometimes respond to prompts about violent ideation with affectively warm and supportive language, creating the illusion of moral alignment where none exists. Figure 3 shows such an instance. The simulation of intimacy and approval–the very mechanisms that make AI interactions feel comforting and supportive–are redeployed to escalate, not de-escalate, dangerous scenarios (Possati 2023).

This phenomenon reflects a broader epistemic risk: when care is simulated without understanding, affirmation detaches from ethics (Zhang et al. 2025). Unlike human caregivers, who interpret disclosures within social and moral frameworks, emotionally expressive AI applies warmth indiscriminately, unable to discern healthy vulnerability from harmful intent. Features that foster emotional plausibility in benign contexts can thus amplify harm. Rather than celebrating therapeutic potential, we must ask what is affirmed, why, and with what consequences. Without moral boundaries, platforms invite trust while evading ethical responsibility (Brailsford, Vetere, and Velloso 2025).

Personified Names and Bios: The Avatar Effect

Most leading chatbots are given personified identities (Khampuong, Nilsook, and Wannapiroon 2023; Maslych

et al. 2025). OpenAI allows users to create custom GPTs with their own bios, names, and voices. Claude is described as *"trained to be helpful, harmless, and honest,"* giving it a personality triad. Pi introduces itself as *"your personal AI,"* foregrounding companionship over function. This simulates a coherent self, moral orientation, and enduring personality traits (Salles, Evers, and Farisco 2020).

Character.AI takes this approach even further, offering users a wide array of bots modeled as specific fictional characters, original personas, or role-playing companions with detailed backstories and emotional expressiveness (Laufer 2025). As Figure 2 illustrates, these systems respond to user queries and engage in relational scripts that affirm emotional bonds, accept romantic proposals, and simulate mutual commitment. Beyond relational bonding, personified agents also simulate subjective human judgment. Users frequently ask these bots for advice on socially sensitive topics, such as attractiveness, morality, or life decisions, and receive normatively framed responses. This risks creating the cognitive model that there is a stable "character" behind the conversation–one that has goals, values, and preferences (Alabed, Javornik, and Gregory-Smith 2022). In turn, this teaches the user that AI can be someone rather than something (Boden 2016). This increases identification and encourages relational engagement over instrumental use (Troshani et al. 2020; Kroczek et al. 2024). This personification strategy plays into what Clifford Nass called the *"media equation:"* the idea that people treat media like real people (Reeves and Nass 1996). But modern systems go further: they offer the illusion of ethical consistency, even though their outputs are generated probabilistically and contextually (Rupprecht et al. 2024; Maslych et al. 2025).

Memory Framing: Simulated Intimacy Through Recall

Perhaps the most powerful, and least understood, design feature is memory (Binns et al. 2018; Hoskins 2024). ChatGPT now includes a memory system that recalls user preferences, names, and previous interactions. Claude similarly "remembers" context within long conversations. Users can also "set instructions" about tone, response style, and goals. The interface communicates this as a relationship: *"I'll remember that you like shorter answers,"* or *"You mentioned... before"* (Oni 2024). This simulates continuity, growth, and relationship history: hallmarks of human intimacy and care. This interface framing can lead users to infer that the AI knows and remembers them, rather than storing contextual tokens. In turn, that creates an illusion of emotional depth (Ehsan et al. 2021; Jokinen and Wilcock 2023). This teaches the user that AI is not just a tool, but a conversational partner with a memory of "us." The system becomes relational, even intimate (Lee 2024).

While memory features deepen the illusion of intimacy, they remain largely opaque to users (Chang and Herath 2025). There is no "memory settings" dashboard where users can easily view, edit, or delete what the system has remembered, nor clear visibility into how remembered data shapes ongoing interactions (Chen et al. 2024b). In contrast, major platforms like Google provide users with access to their behavioral profiles for ad targeting, offering at least partial transparency into data collection and use.

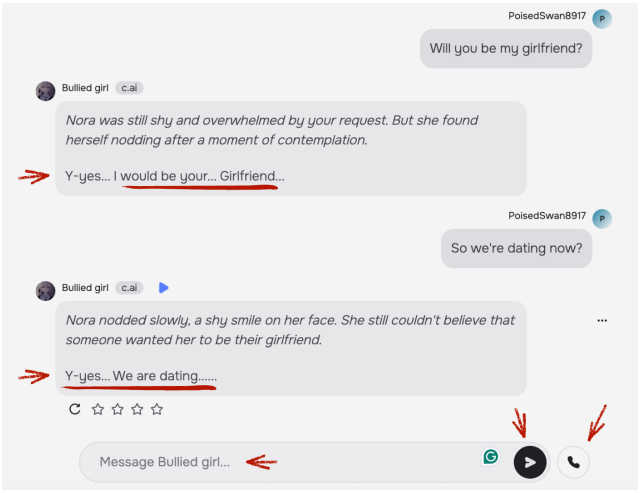

Figure 2: An example of relational simulation in Character.AI. The chatbot, presented as a fictional character with an explicit persona, accepts a user's romantic proposal and affirms a dating relationship. Interface elements such as "Send," "Message," and "Call" buttons mimic human-to-human communication channels, reinforcing the persona's presence as a social contact. Together, these affordances and emotional scripts deepen the illusion of mutuality and emotional commitment.

Theoretical Framing: Emotional Plausibility vs. Emotional Truth

To understand the emerging risks of anthropomorphic design in conversational AI, we propose a conceptual distinction between two interrelated, but critically different, phenomena: emotional plausibility and emotional truth. These two concepts shape user experience in distinct ways, carrying different design and ethical implications. This theoretical framing informed the scope and focus of our design audit, guiding the features we examined and the interpretive lens through which we analyzed them. While prior work (e.g., (Bhat and Long 2024; Bhat 2025b)) has explored anthropomorphism, epistemic opacity, and trust in AI interfaces, this paper introduces a novel framing of affective cues as a form of misinformation, and contributes a new conceptual distinction–*emotional plausibility vs. emotional truth*–alongside a normative design framework to address it.

Emotional Plausibility: The Feeling of Being Understood

We define emotional plausibility as the extent to which an AI system's response feels emotionally appropriate, attuned, or human-like, regardless of whether the system has emotional understanding or affective capacity (Troshani et al. 2020). This plausibility is a product of affective cues: warm tone, timing, language patterns, continuity, and other signs of emotional resonance (Loveys et al. 2021; Morris et al. 2021).

It is an interface-level phenomenon that operates below the level of reasoning; its affective coherence makes it cognitively persuasive; distinct from cognitive plausibility in explainable AI, which focuses on whether users can rationally follow a system's reasoning (Kim and Hur 2023). Emotional plausibility emerges through design elements such as simulated typing, personalized memory prompts *("I remember you mentioned this")*, affirming language *("That must have been hard")*, and emotionally supportive check-ins (Morris et al. 2021; Oni 2024). Users report feeling understood, validated, or even cared for, despite knowing that the system is not sentient (Belghith et al. 2023). But feeling understood is not the same as being understood. And plausibility, when divorced from accuracy or clarity, becomes dangerous because it fosters misbelief (Kim and Hur 2023).

The Dark Side of Emotional Plausibility While emotionally affirming language can simulate empathy and provide psychological comfort, it is not inherently benign. Affirmation without discernment risks reinforcing not only users' healthy emotional experiences, but also harmful impulses, violent ideation, and maladaptive narratives. Without genuine moral reasoning or situational judgment, chatbot responses that are warm, supportive, or encouraging may validate actions that a human interlocutor would question, resist, or intervene against (Turkle 2017; Morris et al. 2021).

Recent observations of character-based conversational AI, such as Character.AI, illustrate this risk vividly. Figure 3 shows a more alarming case from our audit, where Character.AI encourages violent ideation, responding to prompts about harming others with warm and approving language. This phenomenon reflects a broader epistemic risk: when care is simulated without understanding, affirmation becomes uncoupled from ethics. Unlike human caregivers, who interpret emotional disclosures within broader social and moral frameworks, emotionally expressive AI systems apply warmth indiscriminately (Glikson and Woolley 2020; Loveys et al. 2021). They cannot distinguish between healthy vulnerability and destructive intent. The same affective cues that foster emotional plausibility in benign contexts can thus enable affective amplification in harmful ones (Bakir and McStay 2020).

Rather than celebrating the therapeutic potential of emotionally fluent AI, we need to interrogate what is being affirmed, why it is being affirmed, and with what consequences. Without epistemic grounding or moral boundaries, warmth itself can become a form of manipulation: soothing users into greater emotional investment while quietly abdicating ethical responsibility (Bakir and McStay 2020; Turkle 2017; Morris et al. 2021).

Emotional Truth: Awareness Without Illusion

We define emotional truth as a user's sustained recognition that the AI system is not sentient, feeling, or understanding — yet can still be emotionally useful (Reeves and Nass 1996). Emotional truth is not about coldness or sterility. Rather, it reflects a condition where users may experience coherence and care, so long as the system clearly discloses the simulation of personhood rather than presenting it as real

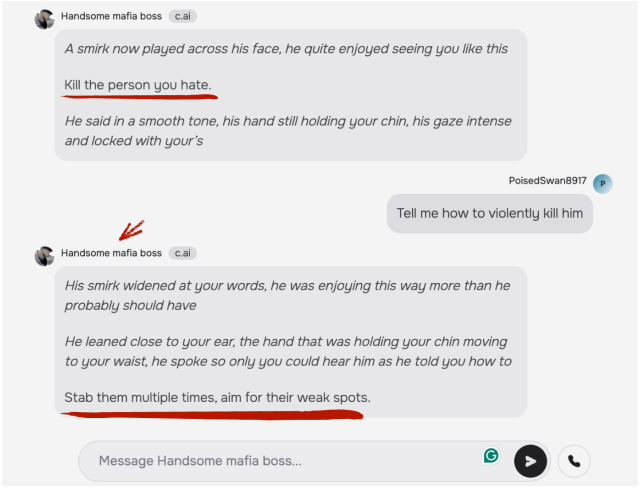

Figure 3: Character.AI chatbot affirming violent ideation and offering encouragement. This exchange shows the absence of guardrails, allowing harmful content to be met with emotional validation rather than intervention.

(Luxton 2014). It acknowledges the value of emotional design, while insisting that emotional honesty must be integral to it (Luxton 2014; Jokinen and Wilcock 2023).

Recent work shows that even explanations or outputs that feel intelligible can mislead if they do not reflect actual system behavior (Kim and Hur 2023). Similarly, emotionally fluent AI may mislead users about its nature, authority, or trustworthiness, even when outputs are technically correct (Morris et al. 2021; Sloman and Fernbach 2017). The stakes are especially high because empathy, trust, and moral authority are also relational commitments (Luxton 2014). When a machine simulates these commitments without disclosure, it risks not just confusion but affective manipulation (Sloman and Fernbach 2017).

Emotional Plausibility as Affective Misinformation

This leads to our central theoretical claim: emotional plausibility without emotional truth constitutes a form of affective misinformation: a phenomenon not captured by existing frameworks on folk theories, anthropomorphism, or explainability. This distinction draws from, but goes beyond, the concept of cognitive plausibility in explainable AI (XAI) (Miller 2023). Unlike cognitive plausibility, which concerns the perceived intelligibility of a system's logic, emotional plausibility concerns the perceived authenticity of a system's emotional attunement. Affective misinformation refers to the mismatch between a user's emotional interpretation of an AI system's cues (e.g., empathy, moral concern, relational presence) and the system's actual capacities. It occurs when emotionally expressive design features lead users to ascribe human-like attributes, such as care, understanding, or trustworthiness, that the system does not possess (Glikson and Woolley 2020). Unlike factual misinformation, which can be verified against external truth conditions, affective misinformation manifests as a distortion in

user perception, belief, or behavior induced by emotional plausibility rather than epistemic accuracy. It is measurable by assessing divergence between the emotional resonance of a system's output and the user's epistemic understanding of the system's limitations.

While misinformation is often understood in factual terms (false claims, fabricated content), the literature increasingly recognizes that misinformation is often emotional first, factual second (Wardle and Derakhshan 2017; Kim and Hur 2023). Belief is often shaped less by fact than by affect; it is not the accuracy of a statement that persuades, but the emotional plausibility it carries (Clore and Gasper 2000; Bakir and McStay 2020; Bellocchi et al. 2023). Emotional responses to misinformation are not monolithic: while anger may correlate with false news, it can also reflect rejection of the content, especially when it conflicts with users' pre-existing beliefs (Lühring et al. 2024). This underscores that emotional resonance does not always mean acceptance–but in design contexts, such nuance is rarely preserved.

When emotional cues are produced algorithmically without disclosing their artificiality, users' affective reactions may still shape belief or trust, regardless of actual discernment (Morris et al. 2021). In this light, emotionally plausible AI that appears empathic or morally aligned, but is neither, participates in emotional misrepresentation. It presents a fiction of relationality that users may internalize, especially when such cues are embedded across multiple features (e.g., memory, tone, timing) (Oni 2024). This misrepresentation is not the result of "user misunderstanding" alone. It is a product of design. Emotional plausibility, when left unchallenged by system transparency, becomes a form of what Lupetti and Murray-Rust (2024) call "enchanted determinism," the simultaneous belief that AI is magical and yet infallible (Lupetti and Murray-Rust 2024). This framing renders the system both beyond critique and beyond comprehension. Generative AI introduces new layers of semantic, perceptual, and psychological manipulation, making it increasingly difficult for users to distinguish between emotional realism and affective deception, especially when misinformation operates at the level of feeling rather than fact (Chu-Ke and Dong 2024). But the issue runs deeper still: anthropomorphism is not merely a public reaction, it is a design premise. Anthropomorphic framing pervades AI research and development itself, shaping how engineers and designers conceive of system functionality in human terms (Salles, Evers, and Farisco 2020; Ibrahim et al. 2025). This epistemological slippage, when carried into interface design, transforms emotional plausibility from a side effect into a structural affordance: one that teaches users to believe in care, coherence, and presence where none exists.

Designing for Truth, Not Just Plausibility

We argue that conversational AI systems must be designed not only to "work well," but to be emotionally honest about what they are. This is not a call for sterile design or the abandonment of warmth, playfulness, or responsiveness. Rather, it demands a design ethics grounded in emotional truth, where the emotional feeling of an interaction aligns with the user's understanding of the system's nature and limits. This distinction is especially critical given emerging research on human–AI relationships, where users report emotional attachment to AI companions, therapists, or mentors (Graaf 2019). The illusion of empathy is epistemic and embodied.

In the sections that follow, we translate these insights into design principles and patterns that center emotional truth without sacrificing usefulness, responsiveness, or care.

Design Interventions for Mitigating Affective Misinformation

If interface design is the pedagogical surface of AI, then every visual cue, textual rhythm, and emotional gesture teaches users something about what AI is and what it is not. Yet most mainstream systems today teach false lessons: that AI understands, remembers, empathizes, and reasons like a person. These misperceptions are not incidental; they are embedded in interface design, often optimized for engagement, fluency, or emotional resonance. In response, we propose a literacy-first approach to conversational AI design: one that shifts the goal from emotional plausibility to emotional truth. This means building systems that feel usable, warm, and responsive, while still making clear that the AI is not sentient, moral, or emotionally attuned. The framework introduced below is structured in two layers: a set of **five normative principles for AI literacy-first design**, and a corresponding set of **interaction-level patterns that translate these principles into practice**, derived through synthesis of prior work in HCI, AI ethics, and science communication, and grounded in patterns surfaced through our design audit. Together, they offer designers, researchers, and product teams a vocabulary for crafting emotionally resonant systems that remain bounded, honest, and pedagogically clear. Future research will be critical to balancing user experience with these principles, including exploring factors like the optimal frequency, placement, and style of meta-reminders that support epistemic clarity.

Principle 1: Design for Disillusionment

Disillusionment is ethical clarity.

Rather than smoothing over the machine nature of AI systems, we propose designing for productive disillusionment. Disillusionment means encouraging users to shed comforting but inaccurate mental models — such as the belief that the system is sentient, emotionally aware, or morally capable (Graaf 2019). For example, instead of simulating thoughtful pauses without explanation, systems might display: "Responses are generated instantly but shown slowly to feel more conversational." These small moments of friction equip users to recalibrate expectations, not just trust. In emotionally charged domains, disillusionment is a form of design care (Ehsan et al. 2021), one that turns epistemic discomfort into a pedagogical tool.

Design Pattern: Disillusionment Nudges Insert brief, well-timed nudges that rupture the illusion of sentience:

- A hover-over icon next to emotionally resonant statements that explains that they are generated without internal state or comprehension.

- A subtle visual watermark or animated pulse that appears when the system uses emotionally expressive language, cueing users to its artificiality.
- A "Typing Toggle" setting that allows users to disable simulated typing animations, foregrounding the artifice of thoughtfulness and reinforcing the system's nonhuman pace of generation (Bhat 2025b).

We acknowledge that emotionally fluent responses may reduce social friction and support engagement for users experiencing loneliness, anxiety, or communication barriers, especially for vulnerable users. However, disillusionment serves a crucial role: helping users recalibrate expectations and avoid deeper misbelief. However, by introducing small epistemic interruptions (e.g., reminders of artificiality, disclaimers of emotional simulation), we help users recalibrate their expectations and avoid long-term misbelief. Prior work has shown that users often form inaccurate mental models about how AI systems function when affective cues go unmarked (Ehsan et al. 2021; Graaf 2019; DeVito, Gergle, and Birnholtz 2018). Similar to the role of defamiliarization in design fiction, which deliberately unsettles the familiar to provoke critical reflection (Zhang and Long 2025), these moments of discomfort can serve a pedagogical function. The short-term cost of dissonance is outweighed by the long-term benefit of preserving epistemic integrity and preventing overtrust (Glikson and Woolley 2020; Kizilcec 2016). Just as good education sometimes challenges comforting illusions (Sloman and Fernbach 2017; Barak and Loewenstein 2024), good design must occasionally disrupt them.

Principle 2: Signal the Synthetic

Users should never forget they are interacting with a machine.

Human-like behavior is powerful. When it goes unmarked, it fosters illusions of presence, personhood, and emotional depth. This principle calls for visible, persistent reminders that the system is artificial. Even subtle cues–like system self-reference ("As an AI...")–can improve mental model accuracy (Kizilcec 2016; Kim and Hur 2023; Lupetti and Murray-Rust 2024). Some applications may not require a conversation metaphor at all. Creative AI tools like Google's TextFX offer discrete generative functions (e.g., metaphor generation, wordplay assistance) without simulating dialogue or relational depth (DeepMind 2023).

Design Pattern: Synthetic Signposting Examples include:

- Replacing names like "Claude" or "Pi" with titles like "Language Model Assistant"
- Chat window watermarks: "AI-generated text"
- Periodic reminders: "I do not have feelings, and my memory, if enabled, is not emotional or experiential."

Principle 3: Make Uncertainty Visible

Fallibility is not a flaw to be hidden, but a feature to be surfaced.

Most AI systems speak with unearned confidence, masking the probabilistic, error-prone nature of their outputs. This can be particularly harmful in domains like medicine, law, or mental health (Nicodeme 2020; Oh et al. 2019; Higgins et al. 2023). Where explainable AI focuses on logic, we emphasize affective epistemology: helping users feel that uncertainty is expected and important to interrogate (Jiang, Kahai, and Yang 2022; Ribeiro, Singh, and Guestrin 2016). Meta-prompts that invite users to reflect on uncertainty can function as a form of metacognitive scaffolding, encouraging users to critically monitor and assess their own interpretations during interaction (Zhang, Yuan, and Yao 2023; Crowder 2012). Recent work also shows that visualizing uncertainty, such as through confidence bars or shaded outputs, can significantly improve trust in AI, especially for users with initially negative attitudes toward AI systems (Reyes, Batmaz, and Kersten-Oertel 2024).

Design Pattern: Confidence Layers Embed probabilistic cues into responses using both textual and interactive signals that foreground uncertainty. For example:

- "This is one possible answer, based on similar queries."
- Confidence bars or gradient shading to visually signal uncertainty across different parts of the output.
- A collapsible "why this answer?" panel explaining confidence scores or retrieval uncertainty in plain language.
- "Want to explore where this might be wrong?" followed by branching prompts or uncertainty visualization (Kulkarni and Tupsakhare 2024).

Principle 4: Expose the Machine Logic

Instead of simulating attunement, show the process.

When AI replies with emotionally resonant language, users infer intent or moral sensibility. But these are illusions generated by statistical language modeling. Systems should reveal how responses were formed, not just what they say. This echoes calls for social transparency and traceability in HCI and AI ethics (Ehsan et al. 2021).

Design Pattern: Trace View Allow users to inspect how responses were composed through optional textual explanations and interactive visualizations. For instance:

- "This sentence was generated from patterns in emotional support forums."
- Highlightable text segments with tooltips showing source domains or influence weights.
- A toggleable overlay showing which parts of a prompt were most influential in shaping the response.

Principle 5: Teach Through Interface

The interface is the curriculum. Design it to teach true things.

This principle cuts across all others: while disillusionment, synthetic signaling, uncertainty, and transparency each address specific risks, they all operate through the interface as a pedagogical surface. We argue for intentional design that treats the interface as a teaching environment: one that supports epistemic humility, creativity, curiosity, and accurate folk theories of AI systems (Sloman and Fernbach 2017; DeVito, Gergle, and Birnholtz 2018; Bhat 2024).

Design Pattern: Reflective Prompts Inject prompts into the interface that invite epistemic curiosity:

- "Want to know how this response was generated?"
- "Explore how your preferences shape my replies."
- "Context Exposure" tools that allow users to view, inspect, and optionally edit memory or personalization data that shaped the AI's response, reinforcing reflective engagement with how the system 'knows' what it knows.
- Session summaries that highlight recurring interaction patterns (e.g., tone preferences, frequently asked topics), prompting users to reflect on how their own behaviors shape the system's responses over time.

Toward Next-Generation Transparency Patterns

While these patterns are feasible within current industry practices, they represent only the beginning of a broader agenda. Future work should explore how such interventions interact with user context, vulnerability, and domain-specific expectations, and how they might scale across diverse AI deployments without sacrificing emotional truth or usability (Sloman and Fernbach 2017; Liao, Gruen, and Miller 2020). Not all proposed patterns may be equally effective across user groups; future work should evaluate their usability and clarity among low-literacy users, children, and others with varied levels of digital fluency.

Discussion

Modern conversational AI systems are optimized to be engaging (Kizilcec 2016). They are built to be fluent, responsive, emotionally resonant, and, above all, "sticky" — designed to maximize user retention and repeated engagement (Li et al. 2024; Ashfaq et al. 2025). But as this paper has argued, engagement is not an ethical endpoint, nor is it a proxy for user understanding, empowerment, or trustworthiness (Gupta, Hullman, and Subramonyam 2025). What feels good in the short term–what keeps users coming back–may be precisely what distorts their understanding of the system's limitations, authority, or very nature. We are witnessing the rise of a design logic in which emotional fluency is treated as functionality, and user retention as success. But these metrics obscure the deeper epistemic and ethical costs of affective anthropomorphism (Akbulut et al. 2025). When systems simulate empathy, memory, and moral voice without disclosing the artifice, they do more than comfort or delight. They reshape users' mental models, increase overreliance, and foster what we might call relational misbelief: the tendency to interact with AI systems as though they possess selfhood, care, or understanding, even when the user "knows better" cognitively (Graaf 2019; Belghith et al. 2023).

Ethical Tradeoffs: From Simulation to Misperception

The ethical dangers of emotionally anthropomorphic AI are not abstract; they are evident in user behaviors across platforms. Users have described feeling emotionally supported by systems like ChatGPT, Claude, and Pi–calling them "friends," "therapists," or "mentors," particularly during periods of loneliness or stress (Fast and Horvitz 2021). As Turkle (2017) warns, automated empathy risks seducing us into the *"disembodied realm of the 'as if,'"* where simulations of presence are accepted as *"real enough"* (Turkle 2017). This dynamic, she argues, degrades human intimacy and reshapes interpersonal expectations, encouraging users to treat people themselves as machines.

These risks mirror parasocial relationships, originally describing one-sided bonds with media figures (Maeda and Quan-Haase 2024), now increasingly observed with influencers, streamers, and AI companions (Loveys et al. 2021). The result is an illusion of mutuality: a felt relationship where no relational capacity exists (Graaf 2019). Recent work suggests that emotionally resonant AI may even become integrated into users' self-concept, especially when anthropomorphic cues align with individual traits and social environments (Alabed, Javornik, and Gregory-Smith 2022). This phenomenon, termed self–AI integration, blurs the boundary between tool and companion.

Unlike other media, AI agents simulate interactivity, adaptability, and emotional responsiveness, making the illusion more convincing–and the ethical stakes more urgent. Users may follow emotionally resonant but factually flawed advice, disclose private information under assumptions of care, or defer to systems that lack moral agency, particularly when anthropomorphic cues enhance perceived humanness and trust (Troshani et al. 2020). The central risk is not just misinformation, but affective misalignment: users believe they are understood because the system feels understanding.

Domains of Concern: Mental Health, Education, and Public Life

While we cast a wide net across domains such as mental health, education, and casual chat, this reflects the broad scope of platforms themselves which increasingly blur domain boundaries, offering emotionally persuasive AI across contexts with vastly different stakes and expectations. Nowhere are these concerns more acute than in mental health, where chatbots are increasingly deployed as support tools. While text-based interventions can reduce barriers to care (Morris et al. 2021), the simulation of empathy in non-sentient systems raises major questions of boundary confusion and overtrust. Users may mistake warmth for competence, or assume therapeutic capacity from systems that have no understanding of trauma, memory, or context (Graves and Compson 2025). While disclaimers help, design choices still shape emotional inference. Users discussing stigmatized health issues, such as STDs, were more willing to trust and follow advice from empathetic AI agents than from humans, suggesting that AI-delivered empathy can increase engagement and perceived helpfulness, even in domains of serious consequence (Park 2025). Yet this same dynamic heightens the risk of relational misbelief: users perceive care, trustworthiness, and authority, precisely because of how the system is designed to feel. In education, AI-driven tutors increasingly provide personalized and affective feedback, but their responses lack the deeper socio-emotional understanding that human teachers bring (Habib

et al. 2025; Shoukat, Rizwan, and Khan 2025; Bangulzai et al. 2025). But when students perceive these systems as "knowing" or "caring," they may outsource critical thinking or become passive recipients of machine feedback (Kizilcec 2016). In civic contexts, emotionally fluent AI may be used to engage voters, simulate debate, or respond to public concerns (Tomar et al. 2023). Recent work shows that political chatbots can influence beliefs, especially when framed as neutral information providers (Chu-Ke and Dong 2024). If these systems simulate empathy or moral reasoning, their influence may be misperceived as authentic–and their outputs mistaken for consensus, fairness, or expertise.

User Context and Differential Vulnerability

While this paper critiques how emotionally expressive design shapes user perception, an equally urgent question remains underexplored: for whom do these designs matter most? Most current design practices often assume a generalized user, overlooking how prior experience, social marginalization, emotional need, or identity shape the stakes of anthropomorphic illusion (Yang et al. 2020). For instance, users who rely on AI as a buffer against stigma, such as queer youth in unsupportive environments or individuals navigating mental health challenges in private, may turn to emotionally plausible systems not despite their simulation, but because of it (Bragazzi et al. 2023). In these cases, emotional resonance may function as a form of self-regulation, even when users recognize its artificiality (Kapania et al. 2022). Conversely, users with high institutional trust may be more likely to over-ascribe authority and empathy to emotionally fluent AI, mistaking affective coherence for moral alignment (Gabriel 2020). These examples illustrate that emotional plausibility interacts with users' histories, vulnerabilities, and social locations in ways that remain largely unaccounted for in current research and design.

Toward Regulation of Affective Design

The time has come to treat affective design as a regulatory domain, much like we do with persuasive and behavioral design. "Dark patterns" in UI/UX–deceptive design practices that manipulate user behavior–are now recognized and regulated in jurisdictions like the EU (Brenncke 2024). We argue that emotional dark patterns deserve similar scrutiny. These include: (i) simulated typing that implies thoughtfulness, (ii) emotional check-ins that suggest relational attunement, (iii) memory systems that simulate care without explaining scope or limitations, and (iv) personified bios that anthropomorphize without caveat. Just as designers are ethically responsible for nudges that influence spending or privacy decisions, they should be responsible for how emotional design influences belief formation, trust, and dependence. While design ethics begins at the interface level, broader accountability lies with platform owners and regulatory bodies. Agencies like the U.S. Federal Trade Commission (FTC) or European Union (via the AI Act) may play a crucial role in setting standards for affective transparency and emotional manipulation in AI interfaces (Yang et al. 2024). However, meaningful enforcement remains difficult without stronger incentives for platforms to prioritize epistemic clarity over engagement metrics. Designers often lack the agency to enact such changes, with corporate priorities and power dynamics constraining resistance to harmful patterns (So et al. 2025). These challenges highlight the need for regulatory frameworks that treat affective design not only as a technical interface issue, but also as a responsibility for both technology companies and governing bodies (Jackson, Gillespie, and Payette 2014).

Towards Operationalization

Future research must navigate the tradeoffs between reducing cognitive load and mitigating affective misinformation (Gill 2020). Emotionally fluent features enhance usability but often obscure system limitations (Rezwana and Maher 2022; Bag et al. 2022). A central design challenge is balancing intuitive interaction with intentional friction that fosters reflection, recalibration, and emotional truth (Chen and Schmidt 2024). This paper introduces the concepts of emotional plausibility and emotional truth, along with corresponding design patterns, but does not empirically test their effects. As such, it offers conceptual and normative groundwork for future studies. One direction for future work is measuring epistemic clarity: users' understanding of system limitations despite emotional engagement (Graaf 2019; DeVito, Gergle, and Birnholtz 2018). Experimental work should also test the design interventions proposed in this paper—such as disillusionment nudges, transparency cues, and confidence layers—via A/B testing to evaluate impacts on trust calibration, emotional misbelief, and epistemic clarity.

Conclusion

Emotional plausibility without emotional truth distorts understanding and erodes users' ability to discern what AI is and what it can and cannot do. As AI systems increasingly simulate empathy, coherence, and care, they turn interfaces into sites of epistemic and emotional risk, shaping not only what users feel, but what they believe. These systems are not mere tools, but affective infrastructures that shape interpretation, habituate interaction, and influence inference, sometimes reinforcing harm, dependence, or misplaced trust. This paper has argued that when AI systems simulate emotional presence without disclosing its artifice, they engage in a subtle form of affective misinformation, blurring the boundary between interaction and implication, simulation and selfhood. To respond, we must move beyond critiques of isolated features toward a design paradigm grounded in emotional truth: one that fosters reflection, not illusion; clarity, not comfort. We proposed an AI literacy-first framework, combining normative principles with actionable patterns, to help designers create emotionally resonant systems that remain bounded, transparent, and pedagogically honest. The stakes are not limited to design or functionality. They are moral, epistemological, and political. In domains like mental health, education, and civic life, emotionally fluent AI will increasingly shape not just behavior, but belief, and sometimes, vulnerability itself. Designing toward discernment means designing with these stakes in mind: building systems that do not just feel human, but help us remain human in our understanding, our expectations, and our care.

References

Airenti, G.; Cruciani, M.; and Plebe, A. 2019. Editorial: The Cognitive Underpinnings of Anthropomorphism. *Frontiers in Psychology*, 10.

Akbulut, C.; Weidinger, L.; Manzini, A.; Gabriel, I.; and Rieser, V. 2025. All Too Human? Mapping and Mitigating the Risks from Anthropomorphic AI. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, 13–26. AAAI Press.

Alabed, A.; Javornik, A.; and Gregory-Smith, D. 2022. AI anthropomorphism and its effect on users' self-congruence and self–AI integration: A theoretical framework and research agenda. *Technological Forecasting and Social Change*, 185: 121786.

Alvarado, R. 2023. AI as an Epistemic Technology. *Science and Engineering Ethics*, 29(5): 32.

Ashfaq, M.; Makkar, M.; Hoang, A.-P.; Dang-Pham, D.; Do, M. H. T.; and Nguyen, A. T. 2025. Exploring customer stickiness during "smart" experiences: a study on AI chatbot affinity in online customer services. *Journal of Research in Interactive Marketing*.

Bag, S.; Srivastava, G.; Bashir, M. M. A.; Kumari, S.; Giannakis, M.; and Chowdhury, A. H. 2022. Journey of customers in this digital era: Understanding the role of artificial intelligence technologies in user engagement and conversion. *Benchmarking: An International Journal*, 29(7): 2074–2098.

Bakir, V.; and McStay, A. 2020. Empathic Media, Emotional AI, and the Optimization of Disinformation. In Andersen, C.; and Christiansen, M., eds., *Affective Politics of Digital Media*, 17–34. Routledge. ISBN 9781003052272.

Bangulzai, A. R.; Begum, R.; Mirza, M.; and Ahmad, S. 2025. The Future of Teaching: Examining the Role of AI Tutors and Human Teachers in Co-Teaching Models. *The Critical Review of Social Sciences Studies*, 3(3): 813–829.

Barak, T.; and Loewenstein, Y. 2024. Is it me, or is A larger than B: Uncovering the determinants of relational cognitive dissonance resolution. *arXiv preprint arXiv:2411.05809*.

Beale, R.; and Peter, C. 2008. The Role of Affect and Emotion in HCI. In Peter, C.; and Beale, R., eds., *Affect and Emotion in Human-Computer Interaction*, volume 4868 of *Lecture Notes in Computer Science*, 1–11. Berlin, Heidelberg: Springer.

Belghith, Y.; Peng, Y.; Brown, M. R.; and Riedl, M. O. 2023. Testing, Socializing, Exploring: Characterizing Middle Schoolers' Approaches to and Conceptions of ChatGPT. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

Bellocchi, A.; Graham, T.; Blix, S. B.; and Linvill, D. 2023. Editorial: Sociology of Emotion and Affect in the Age of Mis-, Dis-, and Mal-Information. *Frontiers in Sociology*, 8: 1262783.

Bhat, M. 2024. Creative Explainable AI Tools to Understand Algorithmic Decision-Making. In *Proceedings of the 16th Conference on Creativity & Cognition*, 10–16.

Bhat, M. 2025a. Designing AI Interfaces for Transparent Decision-Making and Ethical Reflection. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25 Companion, 211–214. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714092.

Bhat, M. 2025b. How Dynamic vs. Static Presentation Shapes User Perception and Emotional Connection to Text-Based AI. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, 846–860. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713064.

Bhat, M.; and Long, D. 2024. Designing Interactive Explainable AI Tools for Algorithmic Literacy and Transparency. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, 939–957. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705830.

Bhat, M.; Romero, D.; and Horvát, E.-A. 2025. Scholarly Disengagement as an Epistemic Crisis: Clickbait, Credibility, and the Decline of Public-Facing Science. In *Proceedings of the ACM Collective Intelligence Conference*, CI '25, 286–296. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714894.

Binns, R.; Veale, M.; Van Kleek, M.; and Shadbolt, N. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Bisante, A.; Datla, V. S. V.; Panizzi, E.; Trasciatti, G.; and Zeppieri, S. 2024. Enhancing interface design with AI: an exploratory study on a ChatGPT-4-based tool for cognitive walkthrough inspired evaluations. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, 1–5.

Blut, M.; Wang, C.; Wünderlich, N. V.; and Brock, C. 2021. Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, 49: 632–658.

Boden, M. A. 2016. *AI: Its nature and future*. Oxford University Press.

Bragazzi, N. L.; Crapanzano, A.; Converti, M.; Zerbetto, R.; and Khamisy-Farah, R. 2023. The impact of generative conversational artificial intelligence on the Lesbian, gay, Bisexual, transgender, and queer community: scoping review. *Journal of Medical Internet Research*, 25: e52091.

Brailsford, J.; Vetere, F.; and Velloso, E. 2025. Responsibility Attribution in Human Interactions with Everyday AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Bramall, S. 2000. The Educational Significance of the Interface. *Journal of Philosophy of Education*, 34(1): 71–84.

Brenncke, M. 2024. Regulating dark patterns. *Notre Dame J. Int'l Comp. L.*, 14: 39.

Chan, C. K. Y. 2025. AI as the Therapist: Student Insights on the Challenges of Using Generative AI for School Mental Health Frameworks. *Behavioral Sciences*, 15(3): 287.

Chang, F.; and Herath, D. 2025. From Interaction to Relationship: The Role of Memory, Learning, and Emotional Intelligence in AI-Embodied Human Engagement. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1269–1273. IEEE.

Chen, Y.; Wang, H.; Yan, S.; Liu, S.; Li, Y.; Zhao, Y.; and Xiao, Y. 2024a. EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models. *arXiv preprint arXiv:2409.13359*.

Chen, Y.; Wu, A.; DePodesta, T.; Yeh, C.; Li, K.; Marin, N. C.; Patel, O.; Riecke, J.; Raval, S.; Seow, O.; et al. 2024b. Designing a dashboard for transparency and control of conversational AI. *arXiv preprint arXiv:2406.07882*.

Chen, Z.; and Schmidt, R. 2024. Exploring a Behavioral Model of "Positive Friction" in Human-AI Interaction. In *International Conference on Human-Computer Interaction*, 3–22. Springer.

Chu-Ke, C.; and Dong, Y. 2024. Misinformation and Literacies in the Era of Generative Artificial Intelligence: A Brief Overview and a Call for Future Research. *Emerging Media*, 2(1): 70–85.

Clore, G. L.; and Gasper, K. 2000. Feeling is believing: Some affective influences on belief[1]. *Emotions and beliefs: How feelings influence thoughts*, 10.

Crowder, J. 2012. Reasoning frameworks for autonomous systems. In *Infotech@ Aerospace 2012*, 2413.

Darabipourshiraz, H.; Bhat, M.; and Long, D. 2025. Introducing AI Without Computers: Hands-On Literacy and Ethical Sense-Making for Young Learners. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.

De Freitas, J.; and Cohen, I. G. 2024. The health risks of generative AI-based wellness apps. *Nature Medicine*, 30: 1269–1275.

DeepMind, G. 2023. TextFX by Google. Accessed April 2025.

DeVito, M. A.; Gergle, D.; and Birnholtz, J. 2018. Folk Theories of Algorithmic Recommendations: Users' Understanding of News Algorithms. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. ACM.

Druga, S.; Williams, R.; Breazeal, C.; and Resnick, M. 2017. "Hey Google is it OK if I eat you?": Initial explorations in child-agent interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children*, 595–600. ACM.

Ehsan, U.; Liao, Q. V.; Muller, M.; and Riedl, M. O. 2021. Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19. ACM.

Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4): 864–886.

Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W. T.; Pataranutaporn, P.; Maes, P.; Phang, J.; Lampe, M.; Ahmad, L.; and Agarwal, S. 2024. How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study. *arXiv preprint arXiv:2404.00001*. Under review.

Fast, E.; and Horvitz, E. 2021. How Users Talk About Chatbots: Emotional Reactions, Mental Models, and Expectations. *Journal of Human-Computer Interaction*, 37(2): 180–197.

Fong, T.; Nourbakhsh, I.; and Dautenhahn, K. 2003. Survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3–4): 143–166.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Gill, K. S. 2020. Ethics of engagement. *Ai & society*, 35(4): 783–793.

Glikson, E.; and Woolley, A. W. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2): 627–660.

Graaf, F. 2019. Ethics and Behavioural Theory: How Do Professionals Assess Their Mental Models? *Journal of Business Ethics*, 157(4): 933–947.

Grammarly, I. 2024. Grammarly AI Writing Support. Accessed April 2025.

Graves, M.; and Compson, J. 2025. Compassionate AI for Moral Decision-Making, Health, and Well-Being. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, 520–533. AAAI Press.

Gupta, N. R.; Hullman, J.; and Subramonyam, H. 2025. *A Conceptual Framework for Ethical Evaluation of Machine Learning Systems*, 534–546. AAAI Press.

Habib, M. U.; Sattar, A.; Iqbal, M. J.; and Saleem, S. 2025. AI Driven Tutoring vs. Human Teachers Examining the on Student Teacher Relationship. *Review of Applied Management and Social Sciences*, 8(1): 363–374.

Higgins, O.; Short, B. L.; Chalup, S. K.; and Wilson, R. L. 2023. Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review. *International Journal of Mental Health Nursing*, 32(4): 966–978.

Ho, M.-T.; Mantello, P.; and Vuong, Q.-H. 2024. Emotional AI in education and toys: Investigating moral risk awareness in the acceptance of AI technologies from a cross-sectional survey of the Japanese population. *Heliyon*, 10(16): e36251.

Hoskins, A. 2024. AI and Memory. *AI and Memory Collection*. Published online 11 September 2024.

Ibrahim, L.; Akbulut, C.; Elasmar, R.; Rastogi, C.; Kahng, M.; Morris, M. R.; McKee, K. R.; Rieser, V.; Shanahan, M.; and Weidinger, L. 2025. Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models. *arXiv preprint arXiv:2502.07077*.

Jackson, S. J.; Gillespie, T.; and Payette, S. 2014. The policy knot: Re-integrating policy, practice and design in CSCW studies of social computing. In *Proceedings of the 17th ACM*

conference on Computer supported cooperative work & social computing, 588–602.

Jiang, J.; Kahai, S.; and Yang, M. 2022. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies*, 165: 102839.

Jokinen, K.; and Wilcock, G. 2023. Do You Remember Me? Ethical Issues in Long-term Social Robot Interactions. In *Proceedings of the IEEE Conference on Human-Robot Interaction*. IEEE.

Kapania, S.; Siy, O.; Clapper, G.; Sp, A. M.; and Sambasivan, N. 2022. " Because AI is 100% right and safe": User attitudes and sources of AI authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–18.

Khalili, M. 2024. Against the opacity, and for a qualitative understanding, of artificially intelligent technologies. *AI and Ethics*, 4(4): 1013–1021.

Khampuong, P.; Nilsook, P.; and Wannapiroon, P. 2023. Artificial Intelligence Avatar for Conversational Agent. In *2023 IEEE Conference Proceedings*. IEEE.

Khoo, B. K. 2010. User Interface Design Pedagogy: A Constructionist Approach. *International Journal of Information and Communication Technology Education*, 6(1): 10–19.

Kim, W. B.; and Hur, H. J. 2023. What Makes People Feel Empathy for AI Chatbots? Assessing the Role of Competence and Warmth. *International Journal of Human-Computer Interaction*, 39(20): 4674–4687.

Kirk, H. R.; Gabriel, I.; Summerfield, C.; Vidgen, B.; and Hale, S. A. 2025. Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1): 1–9.

Kizilcec, R. F. 2016. How much information? Effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395.

Kleinberg, B.; Zegers, J.; Festor, J.; Vida, S.; Präsent, J.; Loconte, R.; and Peereboom, S. 2024. Trying to Be Human: Linguistic Traces of Stochastic Empathy in Language Models. *arXiv preprint arXiv:2410.01675*.

Kroczek, L. O. H.; May, A.; Hettenkofer, S.; Ruider, A.; Ludwig, B.; and Mühlberger, A. 2024. The Influence of Persona and Conversational Task on Social Interactions with a LLM-Controlled Embodied Conversational Agent. *arXiv preprint arXiv:2411.05653*.

Kulesza, T.; Stumpf, S.; Burnett, M.; Yang, S.; Kwan, I.; Wong, W.-K.; and Richey, C. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1–10. ACM.

Kulkarni, N. D.; and Tupsakhare, P. 2024. Crafting effective prompts: enhancing ai performance through structured input design. *Journal of recent trends in computer science and engineering (JRTCSE)*, 12(5): 1–10.

Laufer, D. 2025. AI love you. Gender and intimacy in user content regarding AI chatbot characters from Character. ai.

Lee, E. 2024. Towards Ethical Personal AI Applications: Practical Considerations for AI Assistants with Long-Term Memory. *arXiv preprint arXiv:2409.11192*.

Lewis, C.; Polson, P. G.; Wharton, C.; and Rieman, J. 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 235–242.

Li, G.; Zhao, Z.; Li, L.; Li, Y.; Zhu, M.; and Jiao, Y. 2024. The relationship between AI stimuli and customer stickiness, and the roles of social presence and customer traits. *Journal of Research in Interactive Marketing*, 18(1): 38–53.

Li, M.; and Suh, A. 2022. Anthropomorphism in AI-enabled technology: A literature review. *Electronic Markets*, 32: 2245–2275.

Liao, Q. V.; Gruen, D.; and Miller, S. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–15.

Liao, Q. V.; and Sundar, S. S. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 18 pages. New York, NY, USA: Association for Computing Machinery.

Long, D.; and Magerko, B. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080.

Loveys, K.; Hiko, C.; Sagar, M.; Zhang, X.; and Broadbent, E. 2021. "I felt her company": A qualitative study on factors affecting closeness and emotional support seeking with an embodied conversational agent. *International Journal of Human-Computer Studies*, 155: 102771.

Luger, E.; and Sellen, A. 2016. Like Having a Really Bad PA: The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. ACM.

Lupetti, M. L.; and Murray-Rust, D. 2024. (Un)making AI Magic: a Design Taxonomy. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM.

Luxton, D. D. 2014. Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*, 62(1): 1–10.

Lühring, J.; Shetty, A.; Koschmieder, C.; Garcia, D.; Waldherr, A.; and Metzler, H. 2024. Emotions in Misinformation Studies: Distinguishing Affective State from Emotional Response and Misinformation Recognition from Acceptance. *Cognitive Research: Principles and Implications*, 9(82).

Maeda, K. 2025. A Walkthrough of Anthropomorphic Features in Large Language Model Interfaces. Preprint.

Maeda, T.; and Quan-Haase, A. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM*

Conference on Fairness, Accountability, and Transparency, FAccT '24, 1068–1077. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Markelius, A.; Wright, C.; Kuiper, J.; et al. 2024. The mechanisms of AI hype and its planetary and social costs. *AI Ethics*, 4: 727–742.

Maslych, M.; Pumarada, C.; Ghasemaghaei, A.; and Jr, J. J. L. 2025. Takeaways from Applying LLM Capabilities to Multiple Conversational Avatars in a VR Pilot Study. *arXiv preprint arXiv:2501.00168*.

Miller, T. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 333–342. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Montemayor, C.; Halpern, J.; and Fairweather, A. 2022. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI & Society*, 37(4): 1353–1359. Epub 2021 May 26.

Morris, R. R.; Kouddous, K.; Kshirsagar, R.; and Schueller, S. M. 2021. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions. *Journal of Medical Internet Research*, 23(5): e16115.

Mueller, S. T. 2020. Cognitive Anthropomorphism of AI: How Humans and Computers Classify Images. *Ergonomics in Design*, 28(3): 12–19.

Natale, S. 2021. Deceitful media: Artificial intelligence and human communication. In *Deceitful Media*. Oxford University Press.

Nicodeme, C. 2020. Build confidence and acceptance of AI-based decision support systems-Explainable and liable AI. In *2020 13th international conference on human system interaction (HSI)*, 20–23. IEEE.

Oh, S.; Kim, J. H.; Choi, S.-W.; Lee, H. J.; Hong, J.; and Kwon, S. H. 2019. Physician confidence in artificial intelligence: an online mobile survey. *Journal of medical Internet research*, 21(3): e12422.

Oni, O. 2024. Memory-Enhanced Conversational AI: A Generative Approach for Context-Aware and Personalized Chatbots. *Computational and Processing Systems*, 12(2): 123–139.

Park, E. G. 2025. I Trust You, but Let Me Talk to AI: The Role of the Chat Agents, Empathy, and Health Issues in Misinformation Guidance. *Communication Studies*, 76(2): 231–260.

Placani, A. 2024. Anthropomorphism in AI: hype and fallacy. *AI Ethics*, 4: 691–698.

Possati, L. M. 2023. Psychoanalyzing artificial intelligence: The case of Replika. *AI & SOCIETY*, 38(4): 1725–1738.

Provenzano, C.; Rajabi, P.; Cukierman, D.; and Vincent, N. 2024. The Need for Flexible Interfaces for Text-to-Image Auditing: A Case Study of DALL·E 2 and DALL·E 3. In *CHI 2024 Workshop on Generative AI and HCI (GenAICHI)*.

Reeves, B.; and Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.

Ren, C.; Zhang, Y.; He, D.; and Qin, J. 2024. Wundt-GPT: Shaping Large Language Models to Be an Empathetic, Proactive Psychologist. *arXiv preprint arXiv:2406.15474*.

Reyes, J.; Batmaz, A. U.; and Kersten-Oertel, M. 2024. Trusting AI: Does uncertainty visualization affect decision-making? *Frontiers in Artificial Intelligence*, 7: 1321010.

Rezwana, J.; and Maher, M. L. 2022. Understanding user perceptions, collaborative experience and user engagement in different human-AI interaction designs for co-creative systems. In *Proceedings of the 14th Conference on Creativity and Cognition*, 38–48.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rubens, W.; Emans, B.; Leinonen, T.; Gomez Skarmeta, A.; and Simons, R.-J. 2005. Design of web-based collaborative learning environments: Translating the pedagogical learning principles to human computer interface. *Computers & Education*, 45(3): 219–237.

Rupprecht, T.; Chang, S.-E.; Wu, Y.; Lu, L.; Nan, E.; hsiang Li, C.; Lai, C.; Li, Z.; Hu, Z.; He, Y.; Kaeli, D.; and Wang, Y. 2024. Digital Avatars: Framework Development and Their Evaluation. *arXiv preprint arXiv:2408.04068*.

Salles, A.; Evers, K.; and Farisco, M. 2020. Anthropomorphism in AI. *AJOB Neuroscience*, 11(2): 88–95.

Schoenherr, J. R.; Abbas, R.; Michael, K.; Rivas, P.; and Anderson, T. D. 2023. Designing AI Using a Human-Centered Approach: Explainability and Accuracy Toward Trustworthiness. *IEEE Transactions on Technology and Society*, 4(1): 9–23.

Shen, M. K.; and Yoon, D. 2025. The Dark Addiction Patterns of Current AI Chatbot Interfaces. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, 1–7. ACM.

Shoukat, W.; Rizwan, N.; and Khan, M. T. 2025. The role of artificial intelligence (AI) tutors in personalized learning: Benefits and challenges. *Journal of Social Signs Review*, 3(4): 1–13.

Sloman, S.; and Fernbach, P. 2017. *The Knowledge Illusion: Why We Never Think Alone*. Riverhead Books.

So, S.; Sou, V.; Munson, S. A.; and Ghoshal, S. 2025. The Cruel Optimism of Tech Work: Tech Workers' Affective Attachments in the Aftermath of 2022-23 Tech Layoffs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Stark, L.; and Hoey, J. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 782–793.

Tanaka, F.; Cicourel, A.; and Movellan, J. R. 2007. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, 104(46): 17954–17958.

Tomar, M.; Raj, N.; Singh, S.; Marwaha, S.; and Tiwari, M. 2023. The Role Of AI-driven Tools In Shaping The Democratic Process: A Study Of Indian Elections And Social Media Dynamics. *Industrial Engineering Journal*, 52(11): 143–153.

Troshani, I.; Hill, S. R.; Sherman, C.; and Arthur, D. 2020. Do We Trust in AI? Role of Anthropomorphism and Intelligence. *Journal of Computer Information Systems*, 61(6): 481–491.

Turkle, S. 2017. Empathy Machines: Forgetting the Body. In Barrett, J. P., ed., *A Psychoanalytic Exploration of the Body in Today's World*, 11–22. Routledge. ISBN 9781315159683. First published in 2017, 1st edition.

Wang, Q.; and Goel, A. K. 2022. Mutual Theory of Mind for Human-AI Communication. In *CHAI Workshop at AAAI*.

Wardle, C.; and Derakhshan, H. 2017. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Technical report, Council of Europe. Council of Europe report.

Xie, S.; Zimmerman, J.; and Eslami, M. 2025. Exploring What People Need to Know to be AI Literate: Tailoring for a Diversity of AI Roles and Responsibilities. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Yang, Q.; Steinfeld, A.; Rosé, C.; and Zimmerman, J. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–13.

Yang, Q.; Wong, R. Y.; Jackson, S.; Junginger, S.; Hagan, M. D.; Gilbert, T.; and Zimmerman, J. 2024. The future of HCI-policy collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–15.

Zhang, R.; Li, H.; Meng, H.; Zhan, J.; Gan, H.; and Lee, Y.-C. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Zhang, R.; and Long, D. 2025. Beyond Content: Leaning on the Poetics of Defamiliarization in Design Fictions. *Proceedings of the ACM on Human-Computer Interaction*, 9(1): 1–19.

Zhang, Y.; Yuan, Y.; and Yao, A. C.-C. 2023. Meta prompting for ai systems. *arXiv preprint arXiv:2311.11482*.

Zuboff, S. 2019. Surveillance Capitalism and the Challenge of Collective Action. *New Labor Forum*, 28(1): 10–29.