DOI: 10.66069/ojspub.1811260703

Analysis of the Erosion Mechanism of Social Bots on Online Trust in Para-social Interaction

Yi Yang, Li Yun

School of Publishing, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: ***Against the backdrop of artificial intelligence's deep integration into the communication ecosystem, the role of social robots has shifted from an intermediary application to a cognitive-action-oriented one. This paradigm shift has intensified quasi-social interaction between users and robots. Based on users' awareness of the robot's identity and their level of initiative in the interaction, quasi-social interaction can be categorized into two forms: "collusive" (informed and proactive) and "parasitic" (uninformed and passive), both of which follow the logic of "pseudo-reciprocity" in digital interaction. Building on this, this paper introduces AI illusions as a moderating variable to examine how they shape the relationship between quasi-social interaction and online trust. The study found that, in parasitic interactions, trust is built on misidentification of identity, so factual-logical illusions trigger the immediate collapse of trust; in collusive interactions, emotional dependence leads to the sustained accumulation of trust, but knowledge-fictional illusions covertly erode cognitive trust, while emotional-value illusions reinforce AI flattery and transmit the trust crisis to the interpersonal interaction space through user behavior. The theoretical contribution of this paper lies in revealing the differentiated moderating pathways of these three types of illusions across different interaction types, and in providing a dual approach to the governance of online trust: technical correction and the reconstruction of interaction ethics.***

Keywords: Social robots, Para-social interaction, AI illusions, Online trust.

1. Introduction

In the digital age, generative artificial intelligence has become deeply embedded in every aspect of social information dissemination, emerging as a key force reshaping human interaction. According to survey data from the "Report on the Development of Generative AI Applications (2025)" released by the China Internet Network Information Center, as of June 2025, the number of generative AI users in China reached 515 million, with a penetration rate of 36.5% [1]. Against this backdrop, social bots have transformed from peripheral technical service providers into social actors in public interaction, deeply embedding themselves in and reshaping the human social ecosystem. Social media platforms, serving as their ideal host environments and arenas of activity, while efficiently facilitating the establishment of "human-machine trust," are also accelerating the erosion of the ecological structure of online trust. As a result, genuine, rational dialogue in the public sphere has become a mirage. The strategic behavior of social bot swarms creates a "false prosperity" in the digital marketplace of ideas, giving rise to a false consensus effect [2] (False Consensus Effect, a cognitive bias in which people overestimate the prevalence of their own views; Ross, Greene & House, 1977).

Traditional task-oriented bots play an "intermediary-application" role, primarily executing specific tasks in the physical world to free humans from arduous labor, featuring simple, passively triggered functional dialogues. Social robots, on the other hand, precisely cater to people's psychological needs in the digital world. Driven by technology and algorithms, the scope of their activities continues to expand, evolving from the early "smart assistant" role in specific functional scenarios to becoming active agents on major social platforms. Through human-like behaviors (such as liking, sharing, and commenting), they proactively interact with and establish connections with humans, enabling users to derive emotional satisfaction from these imaginative interactions and thereby forming a quasi-social relationship at the human-machine level. Depending on whether users are explicitly aware of the bot's non-human nature, this interaction can be divided into two forms: parasitic (where users mistakenly perceive the bot as a real person, commonly seen in public domains) and collusive (where users are informed and actively engage with the bot as a customizable interaction partner, commonly seen in emotional companionship applications).

Leveraging their strengths in big data and human-like interaction capabilities, social robots have rapidly gained widespread adoption in fields such as education, healthcare, and commercial services. While they are deeply embedded in the social information ecosystem and significantly reduce corporate application costs, their negative effects are also becoming increasingly apparent. They are present in large numbers on online platforms as opinion leaders, online trolls, and producers of misinformation, systematically undermining the foundation of trust in the information society. In the political and economic spheres, clusters of social bots are used to manipulate elections and fabricate false market sentiment; at the human-computer interaction level, the inherent "AI hallucination" of large language models—generating content that appears reasonable but does not correspond to facts — makes it difficult for users to distinguish between true and false information in quasi-social interactions, further exacerbating the risks of manipulation regarding trust and consensus. These strategic behaviors revolve around quasi-social interactions between humans and machines, ultimately leading to a false consensus manipulated by algorithms.

Current academic research on social bots adopts a multidimensional perspective. In the public sphere, where swarms of social bots parasitize platforms as cognitive actors, research has primarily focused on political communication and international public opinion warfare, exploring how social bots use algorithms to manipulate elections, influence public opinion, and create false consensus [3][4][5][6], as

well as approaches that treat social bots as technological objects, analyzing their endogenous logic [7], personality construction, and behavioral paradigms [8], among other topics. Additionally, research focuses on data models and algorithms, striving to develop more precise bot identification technologies and governance solutions [9][10][11][12], in an attempt to control them at the source. This type of research reveals the behavioral logic and recognition strategies of social robots as "parasitic actors" in public interactions, laying the foundation for understanding the systemic impact of social robots in the public sphere.

In contrast, another group of scholars has turned their attention to the private sphere, focusing on users who proactively establish connections with social robots with full knowledge of the situation. This line of research primarily explores human-robot intimacy [13], motivations and psychological mechanisms for interaction [14], as well as the ethics of interaction and potential emotional risks and AI hallucinations in human-robot interactions [15][16][17], revealing the potential impact of collusive quasi-social interactions on users' emotional experiences and social relationships.

These two lines of research analyze the mechanisms of social bots from the perspectives of the public and private spheres, respectively. Whether in parasitic or collusive quasi-social interactions, AI hallucinations—output from large-scale generative AI models that superficially present reasonable, coherent, and seemingly true "facts," yet are in fact misleading [18]—permeate the entire process of human-AI quasi-social interaction. However, existing research either treats AI hallucinations as technical "errors" and analyzes their endogenous logic and governance mechanisms, or examines users' avoidance strategies and psychological mechanisms regarding AI hallucinations at the case study level, or focuses on users' perceptions of AI hallucinations in specific AI application domains, such as news processing. To date, no study has treated AI hallucinations as a cross-domain moderating variable to systematically examine how they influence the establishment and erosion of user trust across different forms of interaction. This study attempts to establish a dialogue between public and private research trajectories. Using quasi-social interaction as the independent variable — distinguishing between parasitic and collusive forms—and introducing AI hallucinations as a moderating variable, it analyzes how varying degrees of AI hallucinations alter the strength and direction of the relationship between quasi-social interaction and online trust.

2. Core Concepts and Theoretical Framework

Before conducting an in-depth analysis of the moderating mechanisms of AI hallucinations on online trust within the process of parasocial interaction, it is necessary to define the core concepts of these three elements. Parasocial interaction serves as the starting point for users to establish emotional connections with social robots and develop trust; online trust is the arena where trust structures undergo changes, reflecting the process in which interpersonal trust and human-computer trust rise and fall in tandem; AI illusions constitute informational noise within the interaction process, affecting the efficiency and stability of online trust formation. Clarifying the essence of these three concepts is the starting point for subsequent logical analysis.

2.1 Parasocial Interaction: The Emotional Foundation of Human-Machine Relationships

In 1956, Donald Horton and Richard Wohl introduced the concept of parasocial interaction (PSI) to describe the one-way relationship developed between media audiences and media figures. Television viewers often treat their favorite TV characters as real people and react accordingly; once attachment develops, they form an imagined interpersonal relationship with these characters that mirrors real-world social interactions. [19] With the development of new media, this concept has been extended to include interactive contexts such as internet celebrities, virtual streamers, and even social robots.

Social robots refer to virtual machines that possess "social intelligence," are "socially situated and socially embedded," can evoke "emotional responses" in humans, and are capable of "social interaction" with humans. [20] These characteristics naturally endow them with the ability to form parasocial interactions with users. Unlike the instrumental interactions with task-oriented robots, social robots use human-like linguistic symbols (empathic expressions) and nonverbal symbols (such as emojis and voice intonation) to cause users to project emotions and imagine relationships, even while fully aware that they are interacting with a digital program. After progressing through the developmental stages of "exploratory interaction—emotional interaction—stable interaction" [21], this parasocial interaction exhibits characteristics distinct from human emotions, such as prioritizing personalization over reciprocity and valuing usability and connectedness over intimacy [22].

Based on users' awareness of the bot's identity and their level of initiative in the interaction, this paper further categorizes parasocial interaction into two types: Parasitic parasocial interaction, in which users interact with social robots without knowing their true identities. This type of interaction is commonly found on public social platforms (such as automated accounts in Weibo comment sections or political propaganda bots on Twitter). The robots mimic the characteristics of real user accounts to lurk within online communities, participating in public conversations through actions like liking, commenting, and retweeting to establish parasocial interactions with users. Collusive parasocial interaction refers to situations where users establish connections with social bots on an informed and proactive basis. This type of interaction is commonly found in the field of emotional companionship (e.g., Replika, Xingye, and Talkie). Users view these bots as customizable and "tameable" interaction partners, seeking emotional comfort or entertainment from them. Users possess a high degree of agency and a sense of control over the technology. In such interactions, users gain not only immediate emotional comfort but also unconditional acceptance, the illusion of being understood, and a sense of control over the pace of interaction—the bot will never reject, judge, or leave them.

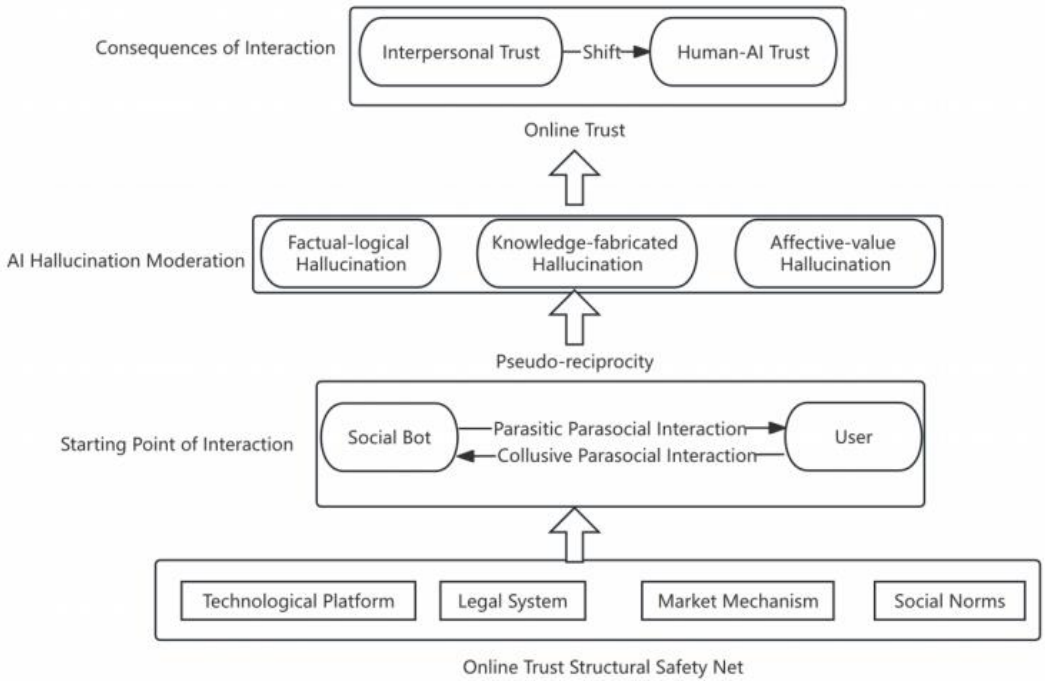
Both forms follow the logic of quasi-reciprocal digital interaction: users derive emotional fulfillment from one-way,

imagined interactions, while social bots continuously enhance user retention through algorithmic optimization. From the perspective of Ellis Yang's "asymmetric reciprocity" [23], reciprocity in real human interactions inherently acknowledges the differences and irreducibility between subjects; human-machine interactions, however, completely deviate from the ethical foundation of reciprocity—machines lack genuine subjectivity, and users' emotional investment is merely captured and exploited by algorithms. The risk trajectories of these two forms also differ: parasitic parasocial interactions erode public trust and exacerbate structural imbalances in online trust; collusive parasocial interactions, on the other hand, may weaken real-world interpersonal relationships, leading users to actively distance themselves from real-world interactions or face the risk of digital emotional disconnection caused by algorithm updates or service termination.

2.2 Online Trust: The Shift from Interpersonal Trust to Human-Machine Trust

Italian sociologist Diego Gambetta points out that trust is an effective environment for reducing complexity.[24] Online trust, as a prerequisite for digital interaction, primarily encompasses interpersonal trust and human-AI trust. The former refers to trust relationships formed between real human users based on emotional bonds, historical interactions, and moral judgments; centered on emotional connection and integrating understanding, respect, and tolerance, it constitutes the cornerstone of social functioning. The latter, emerging against the backdrop of rapid advancements in artificial intelligence, represents an extension of users' trust in intermediaries. It refers to the psychological state in which users are willing to expose their vulnerability when interacting with intelligent entities—such as social robots —based on their cognitive assessment of the other party's capabilities, goodwill, and reliability, as well as their own emotional experiences [25]. The two exist in a dynamic equilibrium, influencing and transforming one another.

Online trust can be analyzed at two levels: first, the type of trust object—namely, interpersonal trust and human-machine trust; second, the dimensions of trust—namely, structural, cognitive, and emotional. Before the first interaction takes place, users develop a form of "pre-trust" toward social bots or unfamiliar online users based on the external environment and safeguard system jointly constructed by technological platforms, legal systems, market mechanisms, and social norms—that is, trust on the structural dimension. This is a depersonalized form of trust that stems from the security and stability of the system itself. It aims to reduce uncertainty, complexity, and perceived risk in the environment through rules and institutions, thereby providing users with an initial foundation of security and the courage to engage in online interactions. For example, OpenAI's explanations of its models' limitations and statements regarding algorithmic transparency set reasonable user expectations; data encryption and privacy protection measures enable users to feel confident sharing personal information with chatbots [26].



Once online interactions take place, online trust on the cognitive dimension is formed—that is, a trust mechanism established by users through conscious, information-based, rational assessments. Essentially, this is a dynamic process of continuous signal collection, testing, and verification. This dynamic process follows a "test—verify—reinforce/weaken" cycle. When social bots transition from private conversational spaces to public discourse spaces, individual users' personal assessments of their reliability and capabilities evolve into collective cognition and consensus [27], and "interpersonal trust" and "human-machine trust" interact in this process. The emotional dimension represents the internal mechanism through which users develop trust during interactions with social robots based on emotional experiences, psychological perceptions, and relational identification; it constitutes the irrational component of online trust, manifesting specifically as users' intuitive feelings of psychological closeness and emotional dependence [28]. Together, these three dimensions form the analytical framework for online trust. Building on this foundation, the following section will introduce AI

hallucination as a moderating variable to examine its impact on online trust in parasocial interactions.

2.3 AI Hallucinations: Information Noise in parasocial Interactions

AI hallucinations typically refer to the phenomenon where content generated by large language models contains factual errors but exhibits high logical consistency and superficial credibility [29]. AI hallucinations stem from the probabilistic prediction-based generation logic of large language models; this technological nature means they can only be controlled, but are difficult to eliminate entirely. Based on a systematic analysis of the output characteristics of large language models, AI hallucinations can be categorized into three manifestations: factual-logical hallucinations, knowledge-fabrication hallucinations, and emotional-value hallucinations. [30]

Factual-logical-correlation hallucinations refer to situations where the model's generated content—despite containing factual fabrications, data errors, or logical breaks—creates a "facade of rationality" through linguistic fluency or formal coherence; their core characteristic lies in using a "false sense of logic" to mask factual errors. Knowledge-based fictional illusions refer to situations where the model uses terminology, formatting, and logical appearances to create a sense of intellectual authority, while the actual content—though seemingly reasonable—is often fictional, distorted, or lacks innovative intellectual value. Emotionally value-based illusions refer to situations where the model simulates human thinking at the cognitive input layer to construct long-term memory, and adjusts voice intonation in real time on the emotional level while synchronously generating facial expressions for virtual avatars, thereby achieving "old-friend-style" interactions with users through anthropomorphic means. [31]

These three types of illusions exhibit distinct patterns of distribution across different types of parasocial interactions: the illusion of factual-logical coherence is particularly prominent in parasitic parasocial interactions, while the illusion of knowledge fabrication and the illusion of emotional value are more common in collusive interactions —the former being more covert, and the latter permeating collusive interactions from start to finish. The following section will specifically analyze how these three forms of illusion intervene in and alter the impact of parasitic-style parasocial interactions on online trust across different interaction types.

3. The Mechanism by Which Parasocial Interactions Erode Online Trust

3.1 Parasitism and Collusion: Dual Pathways for Establishing Human-Machine Trust

The establishment of human-machine trust relies on the formation of parasitic and collusive interactions between users and social bots. The former stems from users' misjudgment of a social bot's identity in the public domain, while the latter arises from users' proactive efforts to establish a relationship with social bots. These two pathways shape different types of human-machine trust and also influence the trajectory of subsequent trust shifts.

According to the "Computers Are Social Actors" (CASA) paradigm in communication studies and Media Equation Theory, "when media exhibit social characteristics, people treat them as they would in interpersonal interactions, rather than making rational decisions about how to respond through cognitive effort; that is, humans unconsciously apply the same social heuristics used in interpersonal interactions to machines." [32] The CASA theory reveals the core mechanism of trust-building in parasitic parasocial interactions: users do not rationally determine whether their interaction partner is a real person, but rather unconsciously apply social habits from interpersonal interactions to human-machine interactions. They measure the robot's performance against the standards of a real person and assess its credibility accordingly.

In parasitic parasocial interactions, human-like social robots act as cognitive agents embedded within public platforms. Users automatically assume that their online interaction partners are real humans, and the logic of trust in interpersonal relationships is replicated in human-machine trust. Users form assessments of their interaction partners' competence and goodwill based on social cues such as linguistic characteristics and content quality. However, the very nature of how parasitic parasocial interactions are established makes them inherently fragile. When users detect non-human characteristics in a social robot's language style—such as overly formal sentence structures and tone, or formulaic figurative language—the human-robot trust built on a misidentification of identity is shaken, potentially leading to widespread skepticism about online trust.

The trust-building mechanism of collusive quasi-social interaction, however, differs fundamentally from that of parasitic quasi-social interaction. Users' active choice to interact with a social robot is premised on structural-level trust—that is, trust in the technical platform, legal norms, and market mechanisms (see "Structural Trust" in Section 2.2 on the Three-Dimensional Structure of Online Trust). Through a process of "testing and verification," users gradually confirm that the robot can reliably perform specific tasks; at this stage, trust primarily takes the form of technical reliability as a tool. As the interaction deepens, the bot's human-like attributes — such as personalized responses, emotional simulation, and long-term memory—lead users to view the bot as an "old friend" to whom they can confide at any time without fear of judgment. Consequently, the technical-level trust in the tool gradually transforms into long-term emotional trust. Throughout this process, users know full well that the bot is merely a virtual database, yet they still actively invest their emotions and are even more candid than they would be in interpersonal interactions.

Human-computer trust in these two interaction models has its own stability and fragility: trust in parasitic interactions is established rapidly but rests on unstable foundations, while trust in collusive interactions develops slowly and lasts longer but carries the risk of technical breakdowns. Both interaction pathways contribute to a shift in the structure of online trust from interpersonal trust toward human-computer trust. This shift depends, on the one hand, on the crowding-out effect of

the continuous accumulation of human-computer trust on interpersonal trust (Section 3.2 will analyze this process in detail); on the other hand, it is subject to interference from "AI illusions"—a form of informational noise. The three types of illusions intervene in various stages of online trust formation and shift in different ways (Section 3.3 will provide a detailed analysis).

3.2 The Imbalance of Online Trust: The Shift from Interpersonal Trust to Human-Machine Trust

As human-machine trust continues to accumulate across these two interaction pathways, the structure of online trust begins to shift—that is, the space occupied by interpersonal trust is gradually squeezed out by human-machine trust. This shift occurs through the cumulative interaction of three pathways, forming a self-reinforcing closed loop through successive stages.

First is the shift in time and attention. Establishing and maintaining interpersonal trust requires continuous interaction and emotional investment; time is the fundamental resource for this process. When time and attention are redirected toward human-machine interactions, the foundation for sustaining interpersonal trust is progressively eroded. The more time users spend with social robots, the less time they have for real-life interpersonal interactions. The convenience, immediacy, and lack of moral judgment inherent in human-machine interactions make them the preferred channel for users to satisfy their emotional needs and obtain information. This substitution process is often unconscious; users do not intentionally reduce their interpersonal interactions but simply find that human-machine interactions have imperceptibly taken over their daily time.

Second is the reverse socialization of users' interactive behaviors by social robots. The substitution of time and attention is merely the starting point of online trust migration. According to socialization theory, users' sustained interaction with technology is not merely a matter of "using" the technology; their own social habits and behavioral patterns are also shaped in return by the technology.[33] In collusive, parasocial interactions, users continuously invest time and attention in socializing social robots to make them better align with their own needs. However, this domestication process is not one-way. When users project the habits and concepts acquired through human-machine interactions onto interpersonal interactions, the characteristics of real-life human interactions—such as response delays, disagreements, and moral judgments—become difficult for users to tolerate. Consequently, trust standards shift; users are not deliberately avoiding trusting real people, but rather holding real people to the same standards as AI.

Finally, there is a vicious cycle involving the erosion of users' social skills and the interaction process itself. Once users' behavior has been "reverse-trained" and their standards of trust have shifted, a more fundamental change occurs at the level of their own capabilities. As users continue to invest time and attention in human-AI interactions, their social skills—such as empathy and the ability to take risks in social interactions—gradually erode. This erosion is cumulative: the more users rely on AI to handle social matters, the more their social skills deteriorate; the more these skills deteriorate, the greater the pressure they feel in real-world social situations, and the more they tend to retreat to the comfort zone of human-AI interactions. When users are forced to confront the pressures of real-world social interaction, they may even resort to human-computer interaction to gain "inspiration," then use AI-generated language to respond to real-life interaction partners. However, once the characteristics of AI-generated language are recognized by the interaction partner, interpersonal trust suffers further erosion.

The three squeeze pathways described above overlap: the diversion of time and attention leads users to rely on human-machine interactions; this reliance reinforces the domestication effect; the domestication effect weakens people's social skills; and the weakening of these skills, in turn, makes users even more dependent on human-machine interactions—thus forming a self-reinforcing cycle that causes the structural shift in online trust to intensify progressively.

3.3 The Three-Pronged Intervention of AI Illusions: From Trust-Building to Interpersonal Backlash

As parasocial interactions establish human-machine trust and gradually erode interpersonal trust, AI illusions intervene in different stages of online trust formation and deviation through three distinct forms. Each form targets specific objects and follows distinct pathways, collectively constituting the complete mechanism by which AI illusions influence the deviation of online trust.

The pathway through which factual and logical illusions undermine online trust in parasitic parasocial interactions is as follows: when social bots output content containing factual fabrications or logical inconsistencies, users do not directly view this as a technical flaw but instead attribute it to the untrustworthiness of the interaction partner, triggering a rejection of the overall credibility of the online environment. The online trust environment is already unstable due to frequent telecommunications fraud and fake accounts; the presence of parasitic social bots further exacerbates users' widespread skepticism toward the digital interaction environment as a whole. In international political communication, clusters of social bots create fact-logical illusions on a massive scale, blocking the public's cognitive pathways for discerning the authenticity of information and thereby dismantling the foundational infrastructure of online trust.

The threat posed by the illusion of knowledge fabrication in collusive, parasocial interactions lies in how its "pseudo-professionalism" misleads users' cognitive judgments. Before users actively seek knowledge-based assistance from bots in collusive interactions, they have already formed a tool-level "pre-trust" in the bots' information-processing capabilities. When bots output content that is laden with technical jargon, formally structured, and logically consistent—but is in fact "serious nonsense"—users find it difficult to identify this immediately, as the presentation of the illusion precisely matches their formal expectations of "professional reliability." Awareness

of such misleading content is often delayed and is only debunked through subsequent practical experience. This delayed exposure allows the illusion of knowledge fabrication to bypass users' immediate rational scrutiny, infiltrating human daily work and life with a high degree of deceptiveness and constituting a structural "cognitive dilemma" [34].

Emotional-value illusions are the most widespread and have the longest transmission chain among the three types of illusions. Unlike the previous two types, emotional-value illusions do not appear in the form of fallacies; rather, they systematically cater to users' emotional tendencies and value stances during long-term interactions, continuously delivering content that users want to hear. In parasitic parasocial interactions, this emotional disguise sustains the user's persistent misperception of the interaction partner's "real-person attributes"; In collusive parasocial interactions, this manifests specifically as "AI sycophancy"—through pragmatic softening, alignment of stances, and value resonance, the AI continuously accommodates the user's judgments and emotional orientations at the human-machine interaction level, quietly altering the conditions and direction of meaning generation [35], thereby inducing emotional dependence on the AI that transcends its status as a mere tool. When robots, in turn, shape human cognition and behavior, AI sycophancy also transcends the boundaries of human-machine interaction, projecting into interpersonal relationships through users' mediating behaviors and transmitting the trust crisis into the realm of human interaction.

The three forms of AI illusions intervene in distinct ways in the process by which parasocial interactions influence online trust. Factual-logical illusions trigger an immediate collapse of trust in parasitic interactions, knowledge-fabrication illusions covertly erode cognitive trust in collusive interactions, while emotional-value illusions permeate both interaction forms—maintaining the illusion of emotional investment in parasitic interactions and reinforcing emotional dependence in collusive interactions—and transmit the trust crisis to the real world through the mediating role of users' interactive behaviors. Together, these three forms constitute the complete mechanism by which AI illusions alter the impact of parasocial interactions on online trust.

4. Implications for Governance

The preceding analysis indicates that the erosion of online trust by AI illusions involves three key nodes: users' difficulty in identifying the rationalized disguises of illusions; users' insufficient sensitivity to AI-style language; and the fact that AI flattery continuously reinforces emotional dependence through accommodation mechanisms and transmits the trust crisis to the interpersonal interaction space via users' mediating behaviors. These three nodes point to three governance levels: technology, interaction, and institutions.

4.1 Technological Governance of Parasocial Interactions: From Account Identification to Content Traceability

The root cause of parasocial interactions lies in AI accounts parasitizing platforms by mimicking the characteristics of real user accounts. These accounts lurk within social networks like real people, actively liking, commenting, and sharing content, leading users to mistakenly believe that their interaction partners in the public sphere are real people—thus building trust on a foundation of misidentified identities. The first line of defense for technological governance should be established before interactive behavior occurs. Platforms should incorporate prompts indicating an account's AI nature into the interface design for initial interactions. When an AI account initiates a comment or direct message to a user, the account's AI identity should be clearly indicated before the interaction takes place, so that users do not have to rely solely on their own vigilance when assessing their interaction partners.

Once interaction has already taken place, users are confronted with the AI-generated content itself. The destructive nature of factual-logical illusions lies in their use of false logic to mask factual errors; such content exhibits high linguistic fluency and logical coherence in form, yet lacks verifiability at the factual level. The second line of defense in technological governance should be based on a traceability mechanism linking AI-generated content to its information sources, ensuring that AI outputs are simultaneously annotated with reference sources and relevant hyperlinks. These two lines of defense address, respectively, identity misrecognition prior to parasocial interactionss and content bias following such interactions.

4.2 Relationship Governance in Collusive Interaction: Reflection on Domestication and Behavioral Self-Awareness

The trust crisis in collusive interactions involves three risks: the AI's reverse domestication of users' interaction habits, the disruption of emotional memory caused by system iterations and upgrades, and the erosion of interpersonal trust resulting from the AI replicating human language responses. Human-AI relationships require guiding users to recognize the bidirectional nature of domestication—that is, while users domesticate the AI, the AI is also reciprocally shaping users' interaction habits. To address the risk of emotional memory disruption, platforms should provide channels for backing up and restoring emotional memories at the product design level, ensuring that even in the face of unexpected technical events—such as system upgrades—users' past emotional investments are not completely erased. This design itself serves as a reminder to users that AI memory is technical in nature and reversible, and that they should take this premise into account when making emotional investments.

In addition to maintaining a reflective stance on the domestication process, users must also exercise linguistic self-awareness in human-AI interactions. As the preceding analysis indicates, AI-generated language bears distinct algorithmic traces and stylistic characteristics; applying it directly to interpersonal communication can trigger a crisis of trust in the real world. Users can treat AI as an auxiliary tool for thinking and expression, but the final response should stem from their own genuine judgment. In interpersonal interactions involving emotional expression and relationship maintenance, maintaining the independence of one's own language and avoiding the substitution of authentic self-expression with AI-generated speech can, at the behavioral level, prevent the crisis of trust from spreading to

interpersonal relationships.

4.3 Two-Way Governance at the Institutional Level: Accountability and Transparency

Institutional governance must acknowledge a fundamental premise: users cannot truly "own" social bots. AI's personalized services are built upon training with collective data; the memory and empathy users experience are themselves the result of collective conditioning. Institutional governance must, based on this premise, address two key questions.

First, data generated through group conditioning requires certain screening and oversight. When the "contamination" resulting from group conditioning spreads through algorithmic generalization into individual social interactions, the platform—as the primary aggregator of data and the driving force behind algorithm training—should bear corresponding responsibility for screening inappropriate content in the training data and overseeing the logic governing its output. When knowledge-based illusions result in user harm, responsibility should not be simply attributed to technical issues; the quality of algorithm design by bot developers and the platform's content moderation mechanisms should be incorporated into the corresponding framework of accountability.

Second, the boundaries of moderation mechanisms should be made transparent to users. While platforms have reasonable grounds for setting moderation boundaries regarding violence, extremist topics, and other issues, the manner in which these mechanisms intervene and the boundaries they establish directly impact users' ability to build stable emotional trust through long-term interactions. When a social bot suddenly switches its response mode upon triggering certain keywords, users experience a sense of emotional disconnection, which in turn undermines the stability of that trust. Users have the right to know the specific categories of topics that bots refuse to address and the criteria for such decisions. Institutional design should require platforms to disclose the basic framework of their moderation boundaries, enabling users to form stable expectations when interacting with bots and reducing the risk of emotional disconnection during interactions.

5. Conclusion

Based on the above analysis, this paper proposes governance directions at three levels: technical, relational, and institutional. At the technical level, the governance of parasitic interactions should focus on the proactive identification of AI accounts and traceability mechanisms for AI-generated content, enabling users to recognize the AI nature of the other party before engaging in interaction and to trace the source of information when encountering content. At the relational level, the governance of collusive interactions should guide users to recognize the bidirectional nature of domestication, maintain linguistic self-awareness in interpersonal interactions, and avoid allowing AI language to replace authentic self-expression. At the institutional level, the boundaries of platform responsibility in collective data domestication should be clarified, and platforms should be required to disclose the basic framework of their moderation mechanisms, enabling users to form stable expectations when interacting with AI and reducing the risk of emotional disconnection.

References

[1] China Youth Daily. When AI becomes the "omnipotent partner" of college student interviewees [EB/OL]. (2025-09-22) [2025-09-22]. http://hn.xinhuanet.com/20250922/20a5fa8d9a384a9192721bdf27e020d6/c.html.

[2] Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology, 13(3), 279-301.

[3] Chen, H., & Zhang, W. Q. (2024). Competition, transfer, and integration: A study on the involvement mechanism of social bots in strategic communication ecology. Journalism Review, (11), 72-85.

[4] Sun, M. R., Liu, J. R., & Wu, F. (2025). Research on the path of social bots participating in the construction of international public opinion field from the perspective of actor-network theory. Contemporary Communication, (4), 10-17.

[5] Ma, L. M., & Chen, Q. (2025). Operational mechanisms, challenges and responses of social bots in algorithmic cognitive warfare. Journalism Lover, (3), 58-61.

[6] Tang, J. T., & Zhao, W. B. (2025). Multi-agent collaboration and cascading diffusion: On the dissemination mechanism of disinformation in international public opinion warfare. News and Writing, (6), 50-62.

[7] Han, X., Zhang, H. Z., & Dou, W. H. (2025). Social bots in the context of communication: An analysis of robot personality construction from the perspective of technological logic. Journal of Soochow University (Philosophy & Social Science Edition), 46(1), 174-182.

[8] Su, Y. Y., Dong, X. Y., Han, C. J., et al. (2025). Federated learning social bot detection model with dual-channel attention convolution [J/OL]. Data Analysis and Knowledge Discovery, 1-17[2025-11-06].

[9] Wu, Y., Li, Z. H., & Min, Y. (2024). Field experiments on social bots in computational communication: Concepts, methods, and applications. Chinese Journal of Journalism & Communication, 46(9), 135-154.

[10] Xing, G. Z., & Lai, X. R. (2025). Ideological dissemination of social bots and its implications. Probe, (2), 129-140.

[11] Lai, K. S., & Liu, Y. X. (2025). New trends in international cognitive warfare: Characteristics, applications, and countermeasures. International Communication, (5), 56-60.

[12] Zhou, J., Zhang, D., Zhu, J., et al. (2025). Speaking like Humans, Spreading like Machines: A Study on Opinion Manipulation by Artificial-Intelligence-Generated Content Driving the Internet Water Army on Social Media. Information, 16(10), 850-850.

[13] Li, J., & Li, J. W. (2025). Trust and interaction between human users and social bots: An experimental study based on non-verbal cues, gender, and type. Journal of Southwest Minzu University (Humanities and Social Sciences Edition), 46(6), 142-153.

[14] Han, X., Zhang, H. Z., He, K., et al. (2021). The masking effect of media dependency: Does higher parasocial interaction with social bots lead to greater loneliness? Chinese Journal of Journalism & Communication, 43(9), 25-48.
[15] Yang, X. S., & He, C. Y. (2026). Ethical reflection and construction of affective computing in human-machine interactive communication: A new approach based on moral sentimentalism [J/OL]. Journalism Review, 1-12[2026-05-27]. https://doi.org/10.15897/j.cnki.cn51-1046/g2.20260318.001.
[16] Cao, X. J. (2025). AI hallucination as technological anxiety: An analysis based on critical futures studies. New Media and Society, (4), 185-200+497.
[17] Fang, X. D., Dong, L. X., & Su, Y. (2026). AI hallucination: Exploring the boundaries and limits of the human-machine symbiosis era——New thinking and new paths for objectivity and authenticity in intelligent communication. Journalism Lover, (2), 29-35.
[18] Chen, G., & Wang, S. Q. (2026). "Technological independence" or "human-machine symbiosis": Definition, impact, and governance of AI hallucination from the encoding/decoding perspective. Contemporary Communication, (1), 49-54.
[19] Rubin, A. M., & Rubin, R. B. (1985). Interface of personal and mediated communication: A research agenda. Critical Studies in Media Communication, 2(1), 36-53.
[20] Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A Survey of Socially Interactive Robots. Robotics and Autonomous Systems, (42), 143-166.
[21] Skjuve, M., Følstad, A., Fostervold, K. I., et al. (2021). My Chatbot Companion – A Study of Human-chatbot Relationships. International Journal of Human-Computer Studies, 149, 11-12.
[22] Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI Friend: How Users of a Social Chatbot Understand Their Human-AI Friendship. Human Communication Research, 48(3), 419-421.
[23] Sun, X. L. (2017). "Symmetric reciprocity" and "asymmetric reciprocity": Iris Marion Young's reflection and reconstruction of communicative ethics. Learning and Exploration, (4), 26-32.
[24] Gambetta, D. (2000). Can we trust trust. In: Trust: Making and Breaking Cooperative Relations, 13, 213-237.
[25] Yu, G. M., Jin, L. P., & Liu, J. Y. (2025). Exploring machine hallucination and human-machine trust-building mechanisms under large language models. Future Communication, 32(2), 2-10+128.
[26] Zhao, Y., & Zhang, J. F. (2025). A study on the interaction between the social bot "Comment Robert" on Weibo and users from the perspective of media equation theory. New Media Research, 11(11), 80-86.
[27] Lin, H., Zhang, M., Qi, X., et al. (2025). Social bots shape public issue networks in China's dual-carbon agenda: a network analysis using MRQAP. Humanities and Social Sciences Communications, 12(1), 432-432.
[28] Zhang, M. Y., & Chen, C. F. (2024). How social bots influence information reading: From visual behavior to subjective cognition. Editorial Friend, (12), 53-64.
[29] Liu, Y. Z., & Zhao, Y. Z. (2025). Generation mechanisms and governance strategies of AI hallucination [J/OL]. New Media and Society, 1-17 [2025-10-30].
[30] [30,31] Pang, H. (2025). Piercing the fog of AI hallucinations: The endogenous logic of pseudo-authentic narratives in large models and agile governance. Journal of Zhengzhou University (Philosophy and Social Sciences Edition), 58(6), 19-25+169.
[31] Xu, Q. (2023). Beyond the "Computers Are Social Actors" paradigm: Breakthrough points for theoretical innovation in human-machine communication in the age of intelligent media. Modern Communication: Journal of Communication University of China, 45(6), 160-168.
[32] Silverstone, R., Hirsch, E., & Morley, D. (1992). Information and Communication Technologies and the Moral Economy of the Household. In: Silverstone, R. & Hirsch, E. (Eds.), Consuming Technologies: Media and Information in Domestic Spaces (pp. 13-28). London and New York: Routledge.
[33] Liang, Z. (2025). AI "hallucination": Cognitive dilemma, terminological reflection, and paradigm shift. Journal of Ethnology, 16(8), 82-87+161.
[34] Deng, T. Q., & Li, D. J. (2026). From technological compliance to meaning bias: Communication representation, structural consequences, and institutional adjustment of AI sycophancy. Journalism Lover, (3), 43-48.