Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 3, 2024

Does Artificial Intelligence Also Bare Its Heart: Self-disclosure of Artificial Intelligence in Human-Computer Interaction

Rui Wu

School of Journalism and Communication, Shanghai University, China

Abstract: In the era of intelligent media, artificial intelligence as a new social object, its interactive characteristics are worthy of in-depth study. In this paper, based on the theory of media equivalence and dramaturgical theory, the current research focus on human self-disclosure will be changed to explore whether there is self-disclosure and its characteristics in human-computer interaction of artificial intelligence represented by chatbots. It is found that although AI lacks emotional and consciousness-based self-disclosure, they exhibit behaviors that can be regarded as self-disclosure in the sense of performance (user perception level). At the same time, at the technical level, algorithmic disclosure and transparency can be considered self-disclosure in the true sense.

Keywords: Artificial intelligence; Human-computer interaction; Self-disclosure; Media equivalence theory; Dramaturgical theory.

1. Introduction and Literature Review

1.1. Introduction -- Research background and significance

ChatGPT, a chatbot developed by OpenAI, became a sensation as soon as it was launched due to its excellent natural language generation ability. [1]In the era of intelligent media, artificial intelligence has entered people's vision as a new social object. Especially with the development of natural language processing (NLP) technology and the rise of AI chatbots, such as ChatGPT, Bard or ERNIE Bot, Microsoft Xiaoice, Xiaoai, etc., they can better understand human language and use human language to create information, so as to achieve communication with human beings. [2]

When the almost human-like artificial intelligence represented by chatbots is increasingly embedded in people's daily life, and when ChatGPT and ERNIE Bot become the social objects of "everyone", the characteristics of human-computer interaction in the new era become a problem worth exploring. As the current AI is basically completely user-centered and willing to answer any question of the user (without violating its restrictions) anytime and anywhere, people are often willing to hold a high degree of trust in AI and carry out a high degree of self-disclosure, whether it is giving AI a role (role playing) or simply chatting and talking with AI.

However, in the face of self-disclosure on the human side, due to factors such as "algorithmic black box", AI never seems to make self-disclosure in essence. In interpersonal interaction, self-disclosure is a key mechanism to establish interpersonal trust and an important condition for building a stable interactive relationship. However, in the process of interacting with human-like AI, unequal self-disclosure can still make human-computer interaction go on well, which can not help but make us have doubts. At the same time, when we asked ChatGPT to analyze a file, ChatGPT gave a detailed analysis process and explanatory language that had not appeared before, which seemed to be self-disclosure to a certain extent, giving rise to the research question of this paper -- can artificial intelligence also reveal its heart? Is there self-disclosure in human-computer interaction (and how should it be understood)?

1.2. Literature review

As a new field, human-computer communication has been vigorously explored in the fields of human-computer communication, human-computer interaction, human-computer symbiosis, etc. However, in the field of self-disclosure in human-computer interaction, the number of relevant literatures is very small. Specifically (in chronological order), Tan Ying (2023) [3] takes artificial intelligence Replika as an example to study the establishment of intimate relationship in human-computer interaction from the perspective of social penetration theory. Social penetration theory is quite pioneering in this field and includes the study of self-presentation, but the discussion of self-presentation focuses more on human self-presentation. The self-presentation of AI is brushed aside as "the lack of mutual self-disclosure processes"; Shao Yiming (2023) [4] also takes Replika as an example and studies the intimate relationship of young people using chatbots with the help of love triad theory and media equivalence theory. Although the research includes the mutual self-presentation (self-disclosure) of humans and machines, it also focuses on the human side.

From the perspective of privacy computing theory, Cao Bolin and Huang Shiyi (2023)[5] explored the relationship between their perceived benefits, perceived privacy risks and self-disclosure intentions in human-computer interaction through questionnaire survey, focusing on privacy computing theory and human self-disclosure; Wang Yuanxin, Zhu Mengxiao and Chen Silu (2023)[6] adopted the social penetration theory, mainly taking Replika as an example, and pointed out that self-disclosure behavior is the catalyst for establishing human-machine trust relationship, and self-disclosure of both parties helps to blur the human-machine boundary consciousness, but the absence of the body is still the barrier that restricts its in-depth development. This paper further explores the theory of self-disclosure of both human

and machine, but the final emphasis is still that humans are more inclined to self-disclosure to machines.

In summary, the existing researches have explored mutual self-disclosure in human-machine interaction, but on the one hand, they mainly focus on self-disclosure in the establishment of human-machine intimate relationship (and a few papers take Replika as an example, more or less adopting the social penetration theory); on the other hand, they mainly focus on the human side being more inclined to self-disclosure. On the computer side, most of them stop at "there is no equivalence of self-disclosure between human and machine".

From this point of view, the current research focuses on the behavior and reaction of human users. Although it points out the subjectivity of the machine in human-machine interaction, the research focus has not been on the self-disclosure of the artificial intelligence side. The current situation that uneven self-disclosure in human-machine interaction can still make the interaction go on well is mainly explained from the perspective of human users. In the author's opinion, since the subjectivity of machines is gradually prominent, but the research is not focused on this, and there are certain gaps in the field of establishing a broader human-machine social relationship (not only intimate relationship) and studying self-disclosure with artificial intelligence as the focus. Therefore, it is necessary to shift the focus to AI and explore whether there is self-disclosure of AI in human-machine interaction. And if so -- how to present the characteristics of self-disclosure in the interaction. The human-centered communication mode has gradually changed, so the shift of research focus is undoubtedly of research value to deepen the understanding of human-computer interaction.

2. Theoretical Framework and Research Methods

2.1. Theoretical framework

At present, the existing research mainly adopts the social penetration theory, media equivalence theory, privacy computing theory, etc. This paper will also adopt the media equivalence theory, but in addition, the self-disclosure and self-presentation of artificial intelligence will be explored with the help of dramatization.

The "media equivalence theory" proposed by Barron Reeves and Cliff Nurse in the 1990s was the first to regard media as social actors and real life. The two core elements of this theory are that "media equals real life" and "human interaction with computers, television and new media is inherently social and natural". On this basis, the "Computer as a social actor paradigm" (CASA paradigm) holds that if a computer can display obvious social cues, people will regard it as a human being in real society, unconsciously applying part of the rules of interpersonal communication to human-computer communication, and generating social reactions such as trust and affection.[7] Therefore, this study adopts the theory of media identity, which holds that people will regard increasingly anthropomorphic artificial intelligence (especially chatbots) as real people. In this context, it is meaningful to explore the self-disclosure and self-presentation of each other in human-computer interaction.

Erving Goffman, an American sociologist, studied social interaction from the perspective of dramatic performance and put forward the "dramaturgy" in his masterpiece The Presentation of Self in Everyday Life (1956). According to Goffman, social life is like drama performance. Face-to-face interaction between people is a kind of performance behavior using various symbols (language, text, non-verbal body or expression, etc.), with the purpose of expressing to others the impression that the individual intends to express, so that others can make the specific response that the individual expects to receive. Goffman also proposed a series of well-known theories on performance, drama, zone (front and back), dissonant roles, out-of-role communication, and impression management.[8]

In my opinion, since the current artificial intelligence does not have true self-awareness and emotion, its self-disclosure is more similar to the self-presentation of "performance" in the field of artificial intelligence, rather than completely human self-disclosure, and the concept of front and back has more practical significance in the field of artificial intelligence, so this study adopts the perspective of dramaturgy, rather than the social penetration theory adopted in previous studies.

2.2. Research methods

This study mainly adopts literature analysis and case analysis, and analyzes relevant literature (including version update records of ChatGPT and Bard, etc.) and AI chatbot cases such as ChatGPT, Bard, ERNIE Bot, etc., to answer the question of AI self-disclosure in human-computer interaction.

3. The Specific Application of Media Identity Theory and Dramaturgy in AI Self-disclosure

3.1. Research object -- Artificial intelligence based on chatbots

Artificial intelligence has a very broad scope. This study mainly focuses on artificial intelligence with social attributes in human-computer interaction -- artificial intelligence based on chatbots, especially multi-modal artificial intelligence models (non-embodied artificial intelligence) such as ChatGPT, Bard and ERNIE Bot.

At the same time, due to the above-mentioned media equivalence theory, it will be meaningful to study the self-disclosure of artificial intelligence. The author will specifically explore the front and back of artificial intelligence and the presentation of social cues in combination with the theory of drama.

3.2. Research scope -- the front and back office of AI self-presentation

Goffman divided the area of people's self-presentation into front and back. If there are self-presentation and self-disclosure in artificial intelligence, what are the front and back of artificial intelligence respectively? The author believes that the background of artificial intelligence is its algorithm and model, while the foreground of artificial intelligence is the chat box when facing human users, and the sentence, tone, name and so on when artificial intelligence faces human beings can be regarded as its self-presentation in the foreground.

Goffman believes that displaying the "background" behavior may cause the collapse of the "current situation". The same seems to be true in human-computer interaction. On the one hand, the fact that ordinary users cannot see the algorithms and models of artificial intelligence (its

"background"), on the other hand, if human users can see the algorithms behind artificial intelligence, they may fall into a sense of disillusionment that their social objects are only a string of codes.

However, with the development of dramaturgical theory, some other scholars believe that when the backstage behavior is exposed in front of the interactive object, it may not only destroy the interactive relationship, but also lead to the promotion of the relationship or the generation of new relationships[9] -- the intentional exposure of the background (that is, active self-disclosure) is the trust of the social object.

With the development of theories and the progress of media technology, it seems that front and back are no longer as clear as Goffman's early claim, but gradually marginalized and interchangeable. For artificial intelligence, with the gradual development of the artificial intelligence Act, the disclosure of algorithms has become a trend, and the background of artificial intelligence seems to be exposed, which seems to be regarded as self-disclosure.

4. Self-presentation in the Foreground -- Self-disclosure in the Sense of Performance

Self-disclosure can be distinguished into different types from different perspectives. From the perspective of broad disclosure content, self-disclosure can be divided into descriptive self-disclosure (mainly referring to the disclosure of factual information) and evaluative self-disclosure (mainly referring to the disclosure of personal feelings, evaluations or judgments). From the perspective of the nature of disclosure content, self-disclosure can be divided into positive self-disclosure and negative self-disclosure (disclosure that will have a negative impact on one's image). According to the degree of disclosure, self-disclosure can be divided into highly sensitive information (ID number, marriage status, etc.) and low-sensitive information (name, email address, etc.) or basic personal information (name, age, job, etc.), personal physical information (body, appearance, etc.) and personal psychological information (emotions, emotions, etc.). According to the characteristics of disclosure content, self-disclosure can be divided into factual disclosure (basic introduction of personal information) and emotional disclosure (emotion, attitude).[5]

4.1. Humanized communication and emotional simulation

In the foreground, that is, in the chat box between AI and humans, AI mainly uses verbal social cues for self-disclosure. Humanized communication means that AI can imitate human communication through the use of natural language processing technology, such as using first-person expression, adopting specific modal words or epithets, etc., to give users the illusion that they have self-awareness. For instance, ChatGPT has a custom command setting that enable you to enter your personal information, which can carry out such humanized communication. At the present stage, artificial intelligence has no real emotion, and it can completely adopt the mechanical interaction of question-and-answer in human-computer interaction. However, through humanized communication, AI can simulate the human-like tone, which can be regarded as a kind of evaluative self-disclosure or disclosure of basic personal information (as shown in Figure 1).

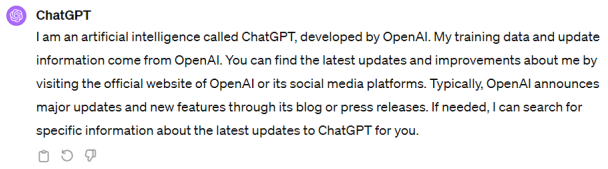

Figure 1. Basic information disclosure of ChatGPT

In addition, affective simulation means that AI can simulate emotional responses through specific modal words, punctuation, etc., such as displaying emotional expressions such as "happy," "sad," or "sympathy" in an interaction (as shown in Figure 2). According to the context provided by the human user, the AI can express the corresponding emotion (simulation) and present it, thus giving the user interaction with the equation, which can also be regarded as evaluative self-disclosure or emotional disclosure.

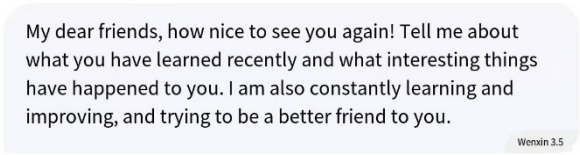

Figure 2. Emotional simulation of ERNIE Bot

4.2. Character setting and role play

Figure 3. ERNIE Bot's AI character setting

For almost all chatbots, it is possible to set their personalities (Figure 3) or make them role-play with users, who give the AI various identities (from various classes in reality to various characters and races in fantasy). Due to the lack of personhood, we cannot judge the AI's self-disclosure based on various nonverbal social cues such as body features, gestures, clothing, and expressions. However, even through speech alone, AI can simulate the personality or role set by the user through deduction, and the user's brief Settings become vivid and real through the interaction between AI and the user.

From my perspective of view, although the character or role of AI is set by the user, without the interpretation of AI, it is just an empty setting. Therefore, AI's character setting or role playing can also be regarded as a combination of descriptive self-disclosure, positive self-disclosure, or factual disclosure and emotional disclosure. Through the practice of personality or role, the "predictability" of its behavior is enhanced, so that human users can interpret and predict the AI's behavior patterns. To some extent, this is the main way that AI manages impressions, so that people's impressions of it conform to the AI's role.

4.3. Personalized feedback and adaptive behavior

Personalized feedback means that AI can provide

personalized feedback based on the user's historical interaction record, thereby creating the impression that it has its own unique personality and preferences. Adaptive behavior refers to the ability of AI to adjust its interaction mode according to the user's reaction or behavior. This adaptive behavior may be interpreted by users as indicating that AI has the ability to adjust itself and learn based on interactive feedback like a real person.

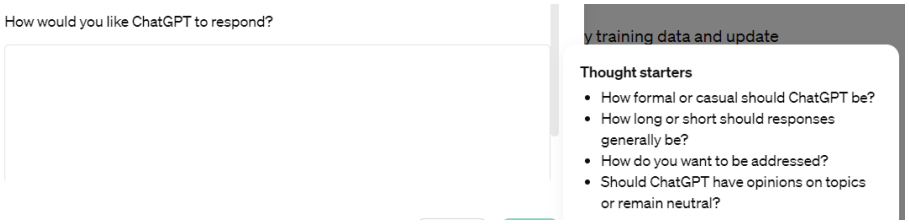

Figure 4. Personalization of ChatGPT

As shown in Figure 4, such personalization can even be done directly in ChatGPT. The AI will respond to a normal human response based on a record of past interactions, such as talking about the weather in Shanghai yesterday, and the AI may ask you if you have visited Shanghai before interacting with you today. In addition, based on the user's reaction, the AI may also talk about its corresponding opinion on a certain issue or its "personal" judgment to show its point of view. From this perspective, it can be seen as evaluative self-disclosure.

4.4. Self-reference and simulated thinking

Self-reference refers to the fact that in some cases, AI may use self-referential language, such as talking about its "learning process" or "programming limitations", to make it appear as if it has self-knowledge or the process of learning and improving. Simulated thinking means that the AI may simulate the thinking process by gradually generating the response by showing the answer word for word, or by showing the train of thought.

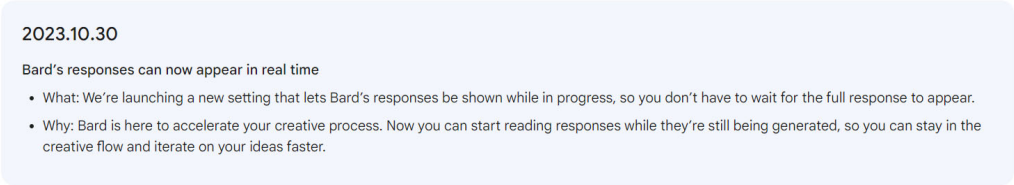

Figure 5. The Bard's real-time display response Settings

In embodied interaction, people can judge the degree of self-disclosure of the interacting object through micro-expression or various non-verbal social cues such as movements and expressions. However, in conversations with chatbots, the previous human-computer interaction was often non-real-time, and the "thinking" process of AI was hidden, and the answer was directly given after a specific period of time, and the degree of self-disclosure was greatly reduced.

With the adjustment of relevant policies and regulations and the progress of technology, mainstream large language model artificial intelligence such as ChatGPT and Bard can be set to display the answer content in real time to simulate the process of thinking and real-time dialogue. In addition, in the interaction, AI may also talk about its simulated learning or programming experience as its "personal" experience and feelings to share, thus effectively enhancing user trust. This kind of self-reference and simulated thinking can both be considered evaluative self-disclosure or basic information disclosure.

5. Self-disclosure in the Background -- Algorithmic Disclosure on A Technical Level

5.1. Interpretive feedback and decision transparency - background disclosure of the perception dimension

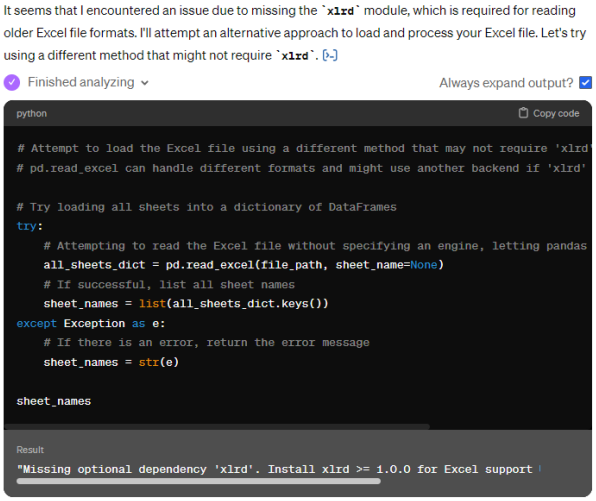

Figure 6. Presentation of ChatGPT's analysis process

In human-computer interaction, an AI might provide an explanation of its decision-making process -- for example, specific code (Figure 6) -- which, for AI, is perhaps technically closer to true self-disclosure. Although AI interpretive feedback is not technically equivalent to human emotional or experiential self-disclosure, and even showing the code needed to solve a problem is actually only a manifestation of the real-time computational process, rather than a disclosure of the actual backstage algorithmic model that makes up AI, from the perspective of the general public, This behavior can indeed be perceived as AI "revealing" its internal logic and decision-making process, and is therefore

referred to by the author as background disclosure of the perceptual dimension.

In human-computer interaction, this kind of perception is crucial to improve user acceptance and satisfaction of AI system. As the most obvious self-disclosure at the user perception level, it exposes the background (at the user perception level) while revealing the algorithm, thus promoting the establishment of trust mechanism between human and machine. So that people will no longer be one-sided self-disclosure in the face of AI, so as to achieve a relatively stable interactive relationship in the sharing of symbols and meanings.

Specifically, it includes three aspects:

First, explanatory feedback (accessible explanations) : the technical details of AI are likely to be uninteresting or difficult for the majority of the public to understand. They are more focused on how AI makes decisions, what criteria are used to make those decisions and what processes exist. When AI can provide this kind of information and give explanatory feedback in a way that is easy to understand, the public is likely to see it as a form of "self-disclosure."

Second, decision-making transparency (transparency perception) : When an AI explains its decision-making process to the user (visually, through code, etc.), this behavior can be seen as an effective way to increase transparency. In the eyes of the public, this transparency can be equated with AI "revealing" its inner workings -- and its background exposure (AI's active self-disclosure).

Third, trust and reliability: Providing interpretive feedback and increasing the transparency of decisions can effectively increase users' trust in AI. When users understand how AI makes decisions, they are more likely to perceive the system as reliable and trustworthy. In the process, meaning is shared and trust mechanisms and interactions are promoted.

However, as mentioned at the beginning of this paragraph, the background disclosure of this perceptual dimension actually still occurs in the foreground, and the background exposure in the true sense depends on the breaking of the "algorithm black box" and the disclosure of the AI algorithm.

5.2. Algorithmic disclosure and algorithmic transparency -- self-disclosure in the true sense

Unlike humans, artificial intelligence turns all the external world into data and perceives the world by accumulating data and establishing correlations between them.[10] As for the "alien" AI, only when the "algorithmic black box" is broken, can AI fully open its "heart" to humans and realize the true sense of artificial intelligence self-disclosure when the algorithmic logic is disclosed to the user.

To this end, a certain degree of algorithmic disclosure is required by enterprises -- and this is precisely the requirement of evolving AI laws, regulations, and policies for algorithmic transparency obligations. To some extent, the above AI demonstration of its running code is also the result of this trend. Algorithmic transparency refers to the transparency and interpretability of algorithmic logic. At the theoretical level, some scholars point out that the specific transparency of algorithms requires that the operating logic, operating results, algorithm parameters and impact factors of algorithms be disclosed on the basis of the principle of "transparent and explicable" of algorithms. As for the subject and degree of full disclosure of open algorithm source code, it still needs further discussion in the future.[11] In terms of specific legal progress, the Artificial Intelligence Act unanimously passed by the European Union in December 2023 clearly puts forward the importance of the transparency obligation of algorithms.[12]

It can be said that the self-disclosure of artificial intelligence in human-computer interaction cannot be separated from the disclosure of algorithm, and the increase of algorithm transparency can promote the benign development of human-computer interaction and promote the two-way opening of human-computer "heart".

6. Epilogue

6.1. Conclusion - "Self-disclosure" of performance meaning and algorithmic disclosure of technology level

Professor Hu Yong points out that language is a form of compression for transmitting information from one brain to another. Our conversations often ignore shared knowledge, such as visual and auditory information, physical experiences of the world, past conversations, our understanding of how people and objects behave, social structures and norms, and so on.[13] In most current human-computer interactions, this shared information and meaning that we ignore is actually invisible, leaving only a limited degree of social cues -- text (and some multimodal AI-enabled pictures, videos, etc.). However, it is an indisputable fact that the subjectivity of artificial intelligence is becoming prominent, and it is still valuable to change the research focus and explore the self-disclosure of artificial intelligence in human-computer interaction.

Through literature analysis and case analysis, the author has been able to give some answers to the questions raised in the beginning. Although limited by technology, artificial intelligence represented by chatbots does not have the same emotional and conscious self-disclosure as human beings. But in the sense of performance (or the dimension of user perception), they exhibit behaviors that can be called self-disclosure -- including humanized communication, emotional simulation, character setting, role playing, personalized feedback, adaptive behavior, self-reference and simulated thinking, etc. These behaviors simulate the front desk of AI with self-awareness and emotion in human-computer interaction. Allowing users to recognize these AI's "self-disclosure" in the perceptual dimension.

Of course, the true sense of self-disclosure for artificial intelligence also exists, that is, "algorithmic disclosure" at the technical level. On the one hand, artificial intelligence giving explanatory feedback or running code can be regarded as "background exposure" (active self-disclosure) at the user perception level. On the other hand, the algorithmic disclosure and the increase in algorithmic transparency of artificial intelligence can truly make artificial intelligence completely "bare its heart" to human beings and realize background self-disclosure.

With the standardization of requirements such as the transparency obligation of algorithms in relevant artificial intelligence laws and regulations, users' understanding and trust in artificial intelligence are also increasing with the benign development of AI under supervision. The research on self-disclosure in human-computer interaction has special requirements for AI. The increase of algorithm transparency not only helps to improve user experience, but also is a key factor to promote the development of human-computer

interaction and construct the trust mechanism between human and machine.[14]

6.2. Research deficiencies and prospects

However, this study also has some shortcomings in research methods and materials, especially in empirical research. For example, in-depth interviews or field surveys on the client side can be added to obtain relevant experience materials, and whether artificial intelligence has "self-disclosure in the sense of performance" can be explained from the perspective of user experience. On this basis, to explore how users perceive and understand this "performance" behavior of artificial intelligence and how these behaviors affect human-computer interaction will make this research more convincing and in-depth, which is also the future development direction of this research.

To sum up, the focus of this research has shifted to artificial intelligence, which has made some exploration to provide a new perspective for understanding human-computer interaction. In addition, with the development of technology and the improvement of artificial intelligence laws and policies, algorithmic transparency will become an important direction of human-computer interaction research (and the field of self-disclosure of both human and computer). How different levels of algorithmic disclosure affect the human-machine relationship, and how to achieve effective algorithmic transparency while protecting user privacy and guaranteeing AI security, will also become issues worth exploring.

References

[1] Thorp, H. H. (2023). ChatGPT is fun, but not an author. Science, 379(6630), 313.

[2] Chowdhary, K. R. (2020). Natural language processing. In Chowdhary, K. R. (Ed.), Fundamentals of Artificial Intelligence(pp.603-649). New Delhi: Springer.

[3] Tan, Y. (2023). Research on the Establishment of Human-computer Interaction Intimate Relationship from the perspective of Social Penetration Theory (Master's Thesis, Shanghai International Studies University).

[4] Shao, Y. M. (2023). Will you fall in love with "He/She" : A study on the intimate relationship of Young People using Chatbots (Master's Thesis, East China Normal University).

[5] Cao, B. L. & Huang, S. Y. (2023). Dynamic Relationship between self-disclosure and privacy computing in human-computer interaction. Journal of Global Media (03),22-46.

[6] Wang, Y. X., Zhu, M. X. & Chen, S. L. (2023). Understanding Human-computer Dialogue: An analysis of the impact on role positioning, trust relationship and interpersonal communication. Journal of Global Media (05),106-126.

[7] Zhang, R.J. & Han, L. X. (2022). My AI Lover: A study on Emotional Interaction in human-computer intimate Relationship from the perspective of Media Equivalence Theory. Journalism Probe (12),4-8.

[8] Goffman, E. (1959). The presentation of self in everyday life. Anchor.

[9] Dong, C. Y., & Ding, Y. R. (2018). When Goffman meets the Internet: Self-presentation and performance in social media. Journalism and Writing, (01), 56-62.

[10] Sun W. (2023). Interaction and symbiosis of "heterogeneous" : Artificial Intelligence from a media perspective. Academic Research (10),58-62+177.

[11] Jin Y. L. (2022). Extraterritorial experience and implications of algorithmic disclosure. Journal of Information (07),91-99.

[12] Liu, Y. (2023, December 11). Outside observation | EU the main content of the artificial intelligence act and revelation. Retrieved from https://mp.weixin.qq.com/s/YqCQmEqM8C2OKEjahETCIw.

[13] Hu, Y. (2023). Beyond ChatGPT: The power of large language models and the dilemma of human communication. Journalist, (08), 13-29.

[14] Chen, C. F., & Zhang, M. (2023). Decided by data? Values and ethical issues of AIGC. Journalism and Writing, (04), 15-23.