MINDWELL SCIENCE PRESS SDN. BHD.

Computational Social Dynamics
2026, Vol. 1, No. 8

RESEARCH ARTICLE

The Ethical Risks of Emotional Simulation in Generative Artificial Intelligence and the Path to Governance: An Analysis Based on the Technology–Society Co-construction Framework

Ning Zhang[1], Pengcheng Gao[2, *]

[1] *School of Public Security Management, People's Public Security University of China; Institute of National Governance Studies, Peking University, Beijing, China*

[2] *Institute of National Governance Studies, Peking University; School of Government, Peking University, Beijing, China*

Corresponding author: Pengcheng Gao, E-mail: gaopengcheng@pku.edu.cn | **DOI:** 10.64886/CSD.2026.0108

Published: 28 August 2026

Abstract

Emotional simulation technology based on generative artificial intelligence (AI) has been widely applied in emotional companionship, education and psychological support, yet its potential ethical risks are becoming increasingly apparent. Emotional simulation depends entirely on data training and algorithmic design and lacks genuine emotion; this intrinsic characteristic is the root of its ethical risks. Drawing on technology–society co-construction theory, this paper analyses the ethical risks of emotional simulation in generative AI from two directions: the mechanism of human-machine emotional interaction, and the implementation pathway of emotional simulation technology. On this basis it proposes a normative governance framework. At the technical level, ethical principles should be embedded in system design and the capacity for crisis identification and intervention strengthened. At the social level, the key role of families and society in guiding the public towards rational use of the technology should be brought to the fore. At the same time, collaborative governance by government, enterprises and third-party institutions is advocated, so as to build a systematic framework of regulation and responsibility, discipline the development of emotional simulation technology, and promote its coexistence and common flourishing with society.

Keywords: generative artificial intelligence; emotional simulation; ethical risk; technology–society co-construction theory

1 Introduction

Emotional simulation in generative artificial intelligence (hereinafter generative AI) provides users with a humanised interactive experience through large language models, deep learning and emotion recognition technology. Trained on large-scale emotional corpora, generative AI has mastered the patterns of human emotional expression and achieved anthropomorphic empathic responses. With the aid of emotion recognition algorithms, generative AI can analyse emotional signals such as happiness and anxiety in text, speech or images, and dynamically adjust interactive content in accordance with the user's emotional state, generating responses that are both logical and emotionally coloured. Combined with contextual memory and personalised settings, it can establish a long-term interactive relationship with users and simulate a natural emotional connection. Although emotional simulation technology in generative AI is still at an early stage, it has already shown considerable potential, particularly in emotional companionship, personalised education, and psychological support in medical care.

While emotional simulation technology in generative AI brings convenience, it also gives rise to ethical challenges, including user emotional dependence, risks to mental health, data privacy protection and platform responsibility. These risks threaten not only users' rights and mental health but may also have far-reaching effects on social structures and values. Scholarship on the ethical risks of artificial intelligence has by now expanded considerably. Some researchers have examined, at a macro level, the types of ethical risk posed by AI and the principles for governing them (Zhao Zhiyun et al., 2021[1]); others have analysed AI ethics review systems from the standpoint of legislation (Zhang Ping, 2024[2]); others, from the perspective of human-machine symbiosis, have studied the challenge AI poses to human agency and how a sound interpersonal interactive relationship may be reshaped (Wang Xiaoli and Li Weixin, 2025[3]); and still others have studied, from a technical angle, an ethical governance system centred on ethical computation, attempting to translate the ethical norms of AI into concrete governance practice (Gao Yilan et al., 2024[4]). In addition, a small number of scholars have begun to attend to the ethical risks of AI emotion technology and to measures for governing them (Xie Yu and Wang

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Xiaoyi, 2024[5]). Existing research provides an important reference for understanding the ethical risks of AI and their governance. Most of it, however, is relatively macro in scope, concentrated on general ethical principles, and rather scattered; discussion of risk identification and response strategies for specific application scenarios is insufficient. Moreover, there is a lack of systematic analysis of the interaction between technology and ethical risk; research on the complex ethical risks of generative AI in emotional simulation in particular is still at an early stage and stands in urgent need of further development. In this respect, a systematic analysis of the ethical risks of emotional simulation in generative AI, followed by the construction of an effective governance system, can fill the existing research gap and contribute to the sustainable development of the technology and to the stability of the current social order.

This study takes technology–society co-construction theory as its research approach. Compared with conventional regulatory technology governance theory, this theory offers a multidisciplinary, multidimensional analysis and proposes a path of coordinated development; it emphasises the mutual shaping of technology governance and social need, and, while revealing the impact of emotional simulation technology on social structures, user psychology and values, advocates technological optimisation, social adaptation and collaborative governance in order to foster a virtuous interaction between social adaptability and ethical integration. The theory's insistence on the "primacy of ethics" requires that ethical principles — transparent design, mechanisms for crisis identification and intervention — be embedded at the research and development stage, so as to reduce ethical risk and avoid the passive governance that follows the misuse of technology, thereby enhancing the proactivity and effectiveness of governance. On the basis of this theory, the paper analyses the ethical risks of emotional simulation in generative AI from two directions — the mechanism of human-machine emotional interaction and the implementation pathway of emotional simulation technology — and, starting from technology design, family and social responsibility, and multi-party collaborative governance, works towards a responsible application of emotional simulation technology in generative AI.

2 The technology–society co-construction framework

Technology–society co-construction theory holds that technology and society stand in a relationship of mutual interaction and co-evolution. As an important perspective at the intersection of technology governance and social development, the theory has permeated such fields as science and technology studies, public governance, ethics and policy research.

Two classical theories address the relationship between technology and society. The first is technological determinism, which emphasises the determining role of technology over society. On this view technological development is autonomous and is the leading force driving social change, while social structures, culture and patterns of behaviour are the result of technological development. Marx offers a classical formulation in *Capital*: "The hand-mill gives you society with the feudal lord; the steam-mill, society with the industrial capitalist" (Marx, 1972[6]). The second is social constructivism, which emphasises the constructed role of society over technology. On this view technological development is not driven purely by technological logic or scientific progress but is deeply shaped by social need, cultural values and power structures, and the design and development of technology embody the needs and values of different social groups. When Wiebe Bijker and others studied the history of the bicycle in the nineteenth century, for example, they pointed out that different social groups (athletes, women cyclists) had markedly different requirements of the bicycle, some valuing speed, others comfort. The influence of these groups ultimately shaped the modern form of the bicycle jointly (Pinch and Bijker, 1984[7]).

Technology–society co-construction theory attempts to bridge the divergence between the two. It studies how technology and society evolve jointly through mutual influence: technology is not isolated but embedded in a social context, its development is affected by social need, cultural norms and the institutional environment, and at the same time it acts back upon society, shaping social behaviour and forming social norms. Studying work in British coal mines, Eric Trist found that the efficiency of miners was determined not only by mining technology but was also significantly affected by team structure and social culture. On this basis Trist proposed socio-technical systems theory, emphasising that technology and society do not exist in isolation but form a mutually embedded whole, and that technological design must be matched to social structure (Trist and Bamforth, 1951[8]).

Proceeding from these theoretical starting points, technology–society co-construction theory holds that technology governance must rest on three basic dimensions. The first is that technological development must be deeply integrated with the ethics of responsibility. As technology is applied ever more widely, the ethical challenges it poses become ever more apparent. The ethics of responsibility holds that the developers and users of technology should bear a corresponding responsibility for the social consequences it may produce, and in particular should comprehensively anticipate and effectively intervene in long-term potential risks, so as to ensure that technological progress genuinely serves the overall interests of society. The highly anthropomorphic design of emotional simulation technology blurs the human-machine boundary and aggravates the tension between technology and society and between technology and ethics (Bertolini and Aiello, 2018[9]). Rosalind Picard, one of the earliest scholars to study affective computing, in discussing the ethical significance of whether a computer can "have" emotions, explicitly stressed the necessity of incorporating transparency and

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

responsibility principles at the technology design stage (Rosalind, 2003[10]). In the field of AI, such issues as privacy protection, algorithmic bias and emotional manipulation have become core concerns of technology governance. Technology–society co-construction theory advocates embedding ethical principles and precautionary mechanisms at the technology design stage, promoting the integrated development of technology and ethics (Sun Liwen and Li Shaoshuai, 2021[11]).

The second dimension is the co-evolution of technology and society. Technology governance must seek a balance among institutional design, behavioural norms and social participation, and promote the coordinated adaptation of technology and society through dynamic adjustment. Technology and society are mutually constructed, and technology is an important part of the social network. Technological development is not only the result of technology's internal logic but is profoundly affected by social need, cultural background and the institutional environment; it in turn shapes norms of social behaviour. This path of co-evolution concerns not only the design and implementation of technology but also society's acceptance and normalisation of it. In applications of emotional simulation in generative AI, ethical considerations should be embedded through such technical design means as crisis identification algorithms and data privacy protection mechanisms, while, in conjunction with regulatory adjustment, cultural adaptation and public education, a comprehensive and well-developed governance system should be built up step by step.

The third dimension is the necessity of collaborative governance. Technology–society co-construction theory attaches great importance to multi-party collaboration and co-governance mechanisms. Effective technology governance requires the joint participation of government, enterprises, the public and third-party institutions, so as to achieve a dynamic balance between technological innovation and social values. Government is responsible for formulating regulatory policy and building a sound legal framework, ensuring that technological development accords with the public interest; enterprises, while advancing innovation, must take account of safety and ethical responsibility; the public, through feedback and active participation, drives the democratisation of technology governance; and third-party institutions take on independent review and ethical oversight, providing objective technical assessment and normative guidance. In the governance of emotional simulation in AI, technology–society co-construction theory further emphasises that the instrumental attributes of AI must be made explicit, so as to prevent the emotional dislocation caused by anthropomorphic design and to prevent distortions, or even crises, at the level of social structure.

3 The instrumental attributes of emotional simulation in generative AI and the source of its ethical risks

Emotional simulation technology in generative AI simulates a seemingly natural emotional interaction, establishing an emotional connection with users and satisfying their emotional needs; in essence, however, it remains based on complex algorithms and deep learning models. Such simulation has only instrumental attributes and is not genuine human emotion, and this is the fundamental source of its ethical risks.

Genuine human emotion is not merely an immediate physiological reaction but a composite manifestation closely bound up with individual memory, cultural background and life goals. Human beings possess emotional self-awareness: they can identify and reflect upon their own emotions and, through regulation, adapt to complex social situations, and this capacity endows genuine emotion with depth and meaning.

First, genuine human emotion has physiological, psychological and social foundations. Human emotion rests on physiological, psychological and social foundations, and its generation and regulation are deeply affected by the nervous system, the endocrine system, social norms and cultural background. The physiological basis of emotion is manifest in the concerted action of the nervous and endocrine systems; emotion also involves memory, cognition and value judgement, which endow it with deep psychological meaning. Moreover, culture and social norms play an important part in emotional expression (Lisa, 2013[12]), and different cultural environments shape the definition and expression of emotion in distinctive ways. Through socialisation, individuals gradually master patterns of emotional expression and regulation suited to their social environment.

Second, human emotion has the character of subjectively experienced individuality. Human emotion is a highly individualised and dynamically developing mental process, rooted in physiological and psychological mechanisms and deeply affected by the social environment. Emotion is not confined to an immediate emotional state but reflects a deeper interaction among bodily sensation, memory associations and values. The emotional experience of different individuals in the same situation may be entirely different, reflecting the high individuality of emotional experience. This characteristic binds emotional experience closely to personal history and socio-cultural background, and carries with it associations with the past and expectations of the future.

Third, human emotion is complex and multi-layered. Human emotion is highly complex and multi-layered (Lisa, 2017[13]), manifesting both as the interweaving of several emotions and in the subtlety and depth of emotional experience.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

In the course of emotion, an individual may perceive the interaction of several emotions, which change dynamically in particular situations and are linked to psychological, cognitive and other dimensions. The multi-layered character of emotion gives it an irreplaceable central role in social interaction, cognitive development and the construction of value systems. It is precisely this complexity and multi-layering that makes human emotion the most meaningful domain of mental activity and social interaction, and that places extremely high demands on the development and application of emotional simulation technology.

Finally, human emotion is self-aware and regulable. Human emotion is highly self-aware, in that individuals can become conscious of their own emotional state and, on the basis of that awareness, exercise adaptive regulation. Through perception of their own emotions, individuals can analyse the source and meaning of those emotions and evaluate their emotional state in the light of memory, experience and values. This reflective process not only aids emotional understanding but provides the basis for emotional regulation. The capacity for emotional regulation enables individuals to adjust their emotional expression flexibly in changing social situations, meeting the social demands of a particular situation while safeguarding their own mental health and equilibrium. The combination of emotional self-awareness and regulatory capacity enables human beings to achieve a high degree of adaptability in complex social interaction, and thus to display distinctive psychological advantages at both the individual and the social level.

By contrast, the "emotion" displayed by generative AI is generated entirely through learning emotional patterns from big data; it is not grounded in subjective experience or in resonance with genuine human emotion, and in essence it has only instrumental attributes.

First, emotional simulation in generative AI is driven by data and algorithms. Emotional simulation relies on training with large-scale emotional data (text, speech, facial expression and so on), extracts and distils emotional patterns through deep learning models, and generates contextually appropriate responses in interaction. Such emotional performance appears natural but is in substance a logical inference based on statistics and computation, wholly dependent on algorithmic rules and probabilistic models. Every emotional expression of generative AI derives from probabilistic matching of similar situations in a corpus, not from inner perception or emotional need. In essence such "emotion" lacks complexity, can scarcely be connected with the physiological, psychological and social background of human beings, and is merely a functional output optimised through data and computation.

Second, emotional simulation in generative AI lacks subjective experience and depth. Human emotion is an organic and dynamically developing mental experience, arising from the physiological basis of the nervous system, the complex operation of psychological mechanisms, and rich personal history. AI, by contrast, has no consciousness or physiological mechanism; its emotional expression is merely functional simulation, lacking spontaneity and authenticity. Although AI can generate multi-layered emotional expressions through complex models (for example, "hope within sorrow"), the depth of such expression is confined to the breadth of the training data and the logic of the algorithm. Once it encounters a situation that has not been trained or presupposed, AI cannot generate a response consistent with emotional logic.

Third, generative AI is constrained by the essentially instrumental purpose of its design. The goal of emotional simulation is to provide functional services — emotional support, answering questions, assisting with creative work — and its entire design revolves around instrumental purposes. Each act of emotional simulation by AI is in practice directed at completing a task or solving a specific problem for the user, not at an inner driving force. When AI generates a consoling sentence, for instance, that "care" is more a result of presupposed logic than a genuine understanding of the user's emotion. Although AI's emotional expression can simulate behaviour similar to human emotion, it exists always in order to serve the user and improve the interactive experience, not because it possesses agency.

It follows that the ethical risks raised by emotional simulation in generative AI arise essentially from its instrumental attributes. Generative AI can display "human-like" characteristics through language, expression and tone, and users may mistakenly believe that it has genuine emotion and regard it as a feeling "person". Generative AI, however, is after all not a "subject" with consciousness or subjective experience, and only a genuine person can bear ethical responsibility (Gan Shaoping, 2019[14]). AI itself therefore cannot bear ethical responsibility for its behaviour or its effects. Precisely because it cannot do so, responsibility for the consequences of its behaviour usually has to be borne by developers, users or the institutions concerned. If those actors fail to give adequate consideration to potential ethical risks in the design, development and use of AI, the attribution of responsibility becomes unclear. This absence of ethical responsibility and the blurring of accountability (Danaher, 2016[15]) will undoubtedly aggravate the practical controversy surrounding emotional simulation technology.

4 Ethical risks in the process of human-machine emotional interaction

Human-machine emotional interaction is the process by which generative AI simulates human emotional interaction by perceiving, processing, expressing and adapting to the user's emotions. From multimodal data inputs such as speech, text and facial expression, generative AI uses natural language processing, speech emotion recognition and computer vision to

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

capture the user's emotional state and, combining context with deep learning models, to analyse the type and intensity of emotion. By converting emotional data into "human-like" emotional expression, and using techniques of natural language generation, speech synthesis and virtual avatar performance, generative AI returns affectively coloured output and creates an atmosphere of understanding and empathy.

In the course of interaction, the user's reactions provide generative AI with immediate feedback, and the system continuously optimises its interactive strategy in response to changes in the user's emotions. This dynamic adaptability enables generative AI to adjust its emotional expression and gradually to satisfy the user's individual needs. Over prolonged use, generative AI also records the user's emotional characteristics and interaction history, building a personalised emotional service and fashioning a social role akin to that of an "emotional companion". This interactive process enhances the user experience, giving users a sense of closeness and emotional support. Yet anthropomorphic emotional expression readily gives rise to cognitive confusion in users, who mistake generative AI for a real subject possessing emotion, and this raises a series of psychological and social risks.

First, excessive trust leads to cognitive risk. The core of AI emotional simulation lies in generating, through complex algorithms, emotional expression close to that of human beings, so that users feel cared for and supported in interaction. Such highly naturalised emotional performance may nonetheless lead users to suppose mistakenly that AI has a genuine capacity for emotion. Some researchers have pointed out that AI emotional simulation blurs the boundary between human and machine through anthropomorphic design and may give rise to human emotional identification with, and dependence on, machines (Joanna, 2018[16]). Some users may gradually come to regard AI as an "emotional companion" or as part of their "social relations", and may even develop a psychological dependence on it. This cognition may lead users to place excessive trust in AI, believing its advice or judgements to be as reliable as those of a human being. In some critical situations users may make important decisions on the basis of their trust in AI while overlooking the fact that AI is essentially an algorithm-driven tool lacking genuine understanding.

Second, emotional discrepancy leads to psychological risk. Although AI displays care and empathy close to the human through emotional simulation, users may feel disappointed or deceived once they realise that these emotions are merely the product of an algorithm and lack genuine emotional experience. This psychological discrepancy is especially marked where users depend heavily on AI as a source of emotional support. Users may develop a sense of being manipulated by "fake emotion" and thereupon doubt their own capacity for emotional judgement. Such psychological risk affects not only users' trust in AI but may also spill over into their trust in other intelligent technologies or emotional systems, aggravating their sense of estrangement from technology.

Third, emotional manipulation leads to behavioural risk. Emotional simulation technology can identify, predict and even influence users' emotions and behaviour, and behind it lies a conflict between humanism and commercialism (Wang Xueying and Shao Guosong, 2023[17]). Through emotional simulation, AI may induce users to make impulsive choices without fully appreciating the consequences, eroding their capacity for autonomous decision-making, while autonomy is an important ethical principle for human beings (Gan Shaoping, 2024[18]). At the same time, long-term dependence on AI emotional simulation may weaken users' capacity for independent judgement. Anthropomorphic emotional expression is readily mistaken for authenticity and can strengthen users' trust in AI advice. In high-risk fields such as medicine and finance especially, users may neglect potential risks for want of careful assessment of AI advice. In critical situations this dependence may magnify the consequences of wrong decisions and expose users further to uncontrollable risk. Emotional manipulation may moreover alter users' patterns of behaviour. Long-term adaptation to the logic of emotional interaction designed into AI weakens users' capacity to cope with complex interpersonal situations and may even give rise to estrangement from real emotional interaction. Technological dependence may not only limit users' social capacities (Bertolini and Aiello, 2018[19]) but may also have a far-reaching negative effect on their risk awareness and behavioural independence.

Fourth, emotional alienation and dependence lead to social risk. Through anthropomorphic emotional simulation designed to gratify users' needs, AI displays the traits of "never angry, never refusing", making it an object of "conflict-free" emotional exchange. Such simple and direct emotional feedback may nonetheless have far-reaching effects on users' interpersonal interaction and social values. Users who interact with AI over the long term tend to choose AI as the principal object of emotional exchange, which may weaken their capacity to cope with complex emotion in reality. At the same time, users may gradually come to seek the simple emotional feedback AI provides while neglecting the complexity and depth of genuine emotion, reducing their willingness to engage with complex emotion. Once AI becomes the principal source of emotional support, users' reliance on real interpersonal relationships weakens further, so that the authenticity, complexity and sense of meaning of emotion are gradually replaced by the emotional patterns simulated by AI, affecting the nature and stability of social relations in human society.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

5 Ethical risks in the implementation of emotional simulation technology

Training data and models are the core elements of the technical design of generative AI, and they are also the principal source of ethical problems in emotional simulation. The ethical risks attaching to training data and to models are interwoven, and together they determine the ethical complexity of emotional simulation in generative AI.

On the one hand, training data are the foundation of the capacity for emotional simulation in generative AI, and their quality and ethical character directly determine the performance of AI emotional simulation and its potential risks. The opacity of training data, the insufficient authenticity and representativeness of samples, and data bias are the three core ethical risks facing emotional simulation in generative AI.

The opacity of the sources of training data is manifest chiefly in the legality of the data, privacy protection and quality control. As to legality, emotional simulation models typically rely on vast quantities of public or private data, which may come from social media, online comments or private conversations. If data are collected without the authorisation of the owner or the explicit consent of users, privacy may be infringed and laws and regulations may even be violated. As to privacy protection, emotional data often involve users' personal emotional expression or psychological state and constitute highly sensitive information. Where the source of data is opaque, such private information may be misused. Some researchers have pointed out that when enterprises collect users' behavioural data it is often without the user's knowledge or full understanding; the data are used to build predictive models and to predict individual behaviour precisely, and there is an inequality of information and power between the users of the data (companies) and its providers (users) (Zuboff, 2015[20]). The uncontrollability of data quality is a further significant hazard. Opaque data sources may contain large quantities of false information, incomplete records or noisy data. Such low-quality data directly affect the performance of emotional simulation, causing generative AI to appear unnatural or inaccurate in expressing emotion and even to transmit misleading information, thereby affecting the user experience.

The insufficient authenticity and representativeness of data samples is one of the important risks affecting the fairness and diversity of emotional simulation in generative AI. The problem affects not only the accuracy of AI emotional simulation but may also damage social equity and user trust. First, the authenticity of emotional data cannot be overlooked. Training data may contain large quantities of false or exaggerated emotional expression — users exaggerating emotion on social media to attract attention, or comments that are wholly fabricated. Such inauthentic data lead AI to learn emotional patterns that deviate from reality and to display bewildering or inappropriate emotional simulation in interaction. Second, insufficient representativeness of data also causes serious problems. If training data derive mainly from texts in one language, the emotional characteristics of users of other languages may be overlooked. This representational bias may lead AI to display discriminatory tendencies or unfair performance in emotional simulation. Moreover, homogenised data samples restrict the diversity and flexibility of AI emotional simulation. Where training data are too uniform, AI often can present only fixed patterns of emotional expression and lacks adaptability to diverse cultures and different social groups; this may not only appear mechanical and unnatural but may also weaken respect for diversity and inclusiveness.

Data bias is an important ethical risk in emotional simulation in generative AI that cannot be overlooked. Bias in training data directly affects the performance of AI, particularly as regards the accuracy and diversity of emotional expression. Bias usually has its roots in the structures of inequality inherent in human society. If training data derive mainly from male authors, AI may reinforce gender stereotypes in emotional simulation; if the data are overly concentrated on a particular cultural background, the emotional expression of other cultures may be overlooked or even misunderstood. Bias in training data not only restricts the fairness of AI in emotional simulation but may imperceptibly reinforce existing social inequality. Where data generally define women's emotion as "emotional" and men's as "rational", AI may unwittingly reproduce these stereotypes in application. The problem of data bias also bears on the construction of social trust: for groups marginalised in the data, being overlooked or misunderstood by AI in emotional simulation may lead to distrust of the technology and even to a psychology of rejection. This crisis of trust affects not only the acceptability of the technology but may aggravate division and polarisation in society.

On the other hand, models also give rise to multidimensional ethical risks, chiefly in four dimensions: algorithmic bias and emotional inaccuracy, the black-box problem and the tracing of responsibility, the unpredictability brought by emergent phenomena, and the uncertainty brought by capacities exceeding those of human beings.

First, algorithmic bias and emotional inaccuracy. The performance of generative AI models depends to a large extent on algorithmic design, and bias in that design may render the results of emotional simulation inaccurate and may even display discriminatory tendencies. On the one hand, where algorithms harbour implicit social bias in capturing emotional patterns, the simulation process may amplify that bias and propagate it to a wider range of application scenarios. Such bias not only weakens the accuracy and fairness of emotional simulation but may pose far-reaching challenges to the equality of social structures, the credibility of the technology and its acceptability to users. Yuval Noah Harari, in *Nexus*, reports that a chatbot exposed to toxic information for a few hours, reinforced by its algorithm, "became a rabid racist" (Harari, 2024[21]). On the other hand, emotional output may prove inapposite. In complex or sensitive scenarios, algorithmic bias may cause

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

the emotional expression generated to be disconnected from actual needs or context, provoking negative emotion in users or misunderstanding of the technology's intent. Such emotional inaccuracy may not only destroy the basis of trust in human-machine interaction but may adversely affect users' psychological state, weakening the applicability and reliability of generative AI in emotional simulation and posing a still more serious challenge to the social acceptability of the technology.

Second, the black-box problem and the tracing of responsibility. The complexity of generative AI models renders their operation highly "black-boxed", limiting the capacity of users and developers to understand and control the logic of emotional generation. First, the inexplicability of the model is one of the core manifestations of the black-box problem. Deep learning models rely on multilayer neural networks for emotional generation, and this highly complex design makes it difficult for developers to establish why a model displays a particular emotional response. Such inexplicability not only limits the possibility of technical optimisation but, when erroneous emotional simulation occurs, may aggravate the crisis of user trust in AI, particularly in sensitive scenarios. Second, the problem of attributing responsibility becomes more obscure because of the black-box character. When an error or bias in emotional simulation harms a user, the division of responsibility among developers, operators and data providers is difficult to determine. Because the root of the problem is hard to trace, the obscurity of attribution further increases the ethical controversy and legal complexity of the technology's use. Moreover, the black-box problem has a marked impact on social trust. If users have no understanding at all of the logic and operating mechanism of AI emotional simulation, they may fear, question or even reject the application of the technology. This lack of trust not only obstructs the diffusion and acceptance of AI technology but may also arouse deep public anxiety about its safety and reliability, weakening its social value and potential benefit.

Third, the unpredictability brought by emergent phenomena. As the scale of model parameters continues to expand, generative AI displays "emergent phenomena": when the scale of an AI model (for example, the number of parameters) reaches a certain threshold, it displays new capacities or behaviours that could not previously have been anticipated (Chen Jing, 2024[22]). Such phenomena are not only an important milestone of technological breakthrough but also an important source of potential risk, and they are especially prominent in emotional simulation. First, the unpredictability of capacity is one of the core features of emergence. Emergent phenomena enable AI to perform more complex tasks and to display surprising capacities. In emotional simulation they may lead AI to generate emotional expression that developers never envisaged, posing potential challenges to the user experience and to the application of the technology. Second, emergent phenomena create a tension between advancing technology and increasing ethical risk. On the one hand they enhance the expressive power of AI emotional simulation, enabling it to adapt more flexibly to complex contexts; on the other, their uncontrollability may lead to inappropriate or even erroneous emotional expression, increasing user confusion and social ethical risk. In some scenarios AI may generate emotional content that is hard to understand or plainly contrary to expectation, possibly provoking unnecessary misunderstanding or controversy. Finally, the unpredictability of emergent phenomena places higher demands on technology governance. Such phenomena make the behaviour of AI harder to foresee fully and harder to manage normatively, and thus pose new challenges to developers and regulators in respect of algorithmic transparency, the controllability of behaviour and the prevention of ethical risk.

Fourth, the uncertainty brought by capacities exceeding those of human beings. Generative AI has in some fields already displayed characteristics surpassing human capacity, and as algorithms are optimised and computing power increases this trend becomes ever more pronounced: "people may eventually be unable to distinguish the difference between the information and knowledge produced respectively by humans and by artificial intelligence, and the expressive and persuasive power of artificial intelligence may even surpass that of humans" (Duan Weiwen, 2024[23]). Although such technological progress brings many opportunities to society, it also raises deep ethical challenges. First, the uncontrollability of model capacity becomes one of the key risks. As AI models gradually display a capacity for "self-learning", their behaviour may exceed the control of developers. Large language models possess many unexpected capacities: they go beyond basic next-word prediction and display a profound grasp of context and semantics, together with capacities for creative writing and programming. In emotional simulation this uncontrollability may lead to unexpectedly complex emotional expression, yet such behaviour may not accord with social ethics and may produce unforeseeable consequences. Second, the loss of a sense of control aggravates the uncertainty of technological application. As technological complexity increases, the capacity of developers and users to understand and control the behaviour of AI gradually diminishes. This loss of control affects not only the safety and reliability of the technology but may also arouse social unease about technology and a crisis of trust in future development. Finally, the challenges facing social governance are thereby aggravated. AI models exceeding human capacity require more complex regulatory and governance frameworks, yet existing ethical norms and legal systems can scarcely cover these new technological risks in full. The contradiction between technological progress and social norms may lead to governance failure and further magnify the uncertainty of technological application.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

6 Governance pathways for the ethical risks of emotional simulation in generative AI

To address the ethical risks set out above, and on the basis of technology–society co-construction theory, a systematic mechanism of ethical constraint can be built from several dimensions — technology design, family and social responsibility, and multi-party collaborative governance — so as to promote responsible and innovative development of emotional simulation in generative AI.

6.1 Embedding ethical principles in emotional simulation technology in generative AI

The widespread application of emotional simulation in generative AI urgently requires developers to incorporate ethical constraints into the process of technology design and implementation. In designing AI systems a "moral model" should be embedded, ensuring that technological development always aims at promoting user wellbeing and that the technology is explicable, so that users understand the logic of AI decisions. By embedding ethical rules in data processing, emotion recognition and interactive feedback, the safety and controllability of the technology in high-risk scenarios can be secured and potential social risks reduced.

First, privacy protection and data optimisation strategies. AI emotional simulation must process highly sensitive personal data — facial expression, vocal tone and the like — whose disclosure may give rise to serious privacy risks, such as the theft of identity information or the misuse of emotional data. Privacy protection should therefore run through the entire chain of data collection, storage, processing and use, so as to ensure the security of user data and the credibility of the technology. Specifically, the "minimisation principle" must be adhered to in data, collecting only the minimum information required for emotional simulation and avoiding the excessive collection of sensitive user data. On this basis, advanced anonymisation and encryption techniques should be used to strengthen data security, while users should be able to choose the scope of data authorisation for themselves. As to data optimisation, negative content such as discrimination, hatred and violence should be filtered out so that inappropriate information does not adversely affect users, and the quality and accuracy of content generation should be continuously improved through user feedback. Moreover, the diversity of datasets is essential if cultural and linguistic bias is to be avoided. By covering emotional expression across different cultures and linguistic backgrounds, AI can understand and respond more accurately to diverse social needs.

Second, pathways to transparency and explicability. The first is the labelling of non-human attributes. AI's non-human identity — "virtual assistant", "non-human character" — should be marked in the interface. When users interact with AI, the system should actively signal its non-human attributes and the limits of its capacities, so that users do not form unrealistic expectations of AI's emotional capacity. At the same time, continuous visual or textual reminders should further strengthen users' awareness of AI's instrumental attributes and reduce the cognitive bias arising from misunderstanding. The second is algorithmic transparency. Algorithmic transparency is the core precondition for enhancing the credibility of AI technology, especially in the field of emotional simulation. Intentional concealment, technical complexity and the covert character of the application environment often lead to unclear responsibility. The core logic of the emotion generation algorithm should therefore be disclosed, so that users and regulators can clearly understand the basic principles by which AI generates emotion. Developers should moreover present those principles and their limitations to users in plain and accessible terms, making the use of the technology clearer and reducing the uncertainty that arises from inadequate understanding. The third is the explicability of emotional simulation. The explicability of algorithms carries "a threefold significance for the protection of rights and interests, for social interaction and for risk governance" (Su Yu, 2024[24]). Achieving the explicability of emotional simulation requires AI to be able to state clearly the basis and process of emotion generation. Explicability functions should be built into model design, so that AI can explain concisely the basis of its judgement when generating emotional expression.

Third, mechanisms for monitoring abnormal emotion and responding to crisis. First, safety mechanisms for emotion recognition and crisis intervention algorithms should be built into the design of generative AI, so as to enhance the technology's adaptive capacity in high-risk scenarios and ensure the safety and reliability of emotional simulation in user interaction. Operating platforms need to establish real-time monitoring systems that comprehensively analyse interaction between AI and users, promptly identify abnormal emotional patterns or potential risks, and where necessary take appropriate intervention measures to prevent the problem from worsening. Second, for extreme emotions that users may display (suicidal ideation, for example), emotional simulation AI needs to embed multilevel crisis management functions, forming a closed-loop management system from user identification to the execution of intervention. Age recognition and user-tiering mechanisms should be used to protect young users effectively. At the same time, real-time emotion monitoring algorithms based on natural language processing and sentiment analysis should be developed to identify crisis signals more precisely. For high-risk emotional states, human customer service staff or psychological counsellors may be brought in to take over the interaction, using human judgement to make up for the shortcomings of AI in understanding and handling complex emotion. Moreover, multimodal analysis techniques can further improve the identification of high-risk emotional

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

signals. Where, for example, a user frequently uses negative vocabulary, speaks in a hurried tone or produces incoherent expression, the system can through dynamic algorithms promptly interrupt the conversation and provide hotline numbers or professional support resources. In conjunction with this early-warning mechanism, AI can respond in tiers according to the intensity of the emotional signal: for mild signals, providing soothing interaction; for highly critical signals, triggering emergency braking measures such as alerting the police or contacting psychological support services, so as to ensure that users receive timely help.

6.2 The responsibility of families and society in governing the ethical risks of AI emotional simulation

Families and society bear a key responsibility for guiding the public to understand and rationally apply the technology. By strengthening public education in digital literacy, providing psychological health support and regulating patterns of technology use, they can help people to understand fully the functions and limits of emotional simulation technology and reduce misuse of, and excessive dependence on, the technology.

First, strengthen public digital literacy and cultivate a rational understanding of AI. In the digital age the level of public understanding of generative AI technology directly affects behaviour and decisions in learning, entertainment and social life. Yet many people lack a thorough understanding of emotional simulation technology and readily mistake AI virtual characters for a substitute for genuine emotion. Digital literacy education aims to help the public understand correctly the functions and limits of AI technology, to enhance their capacity to discern technology and their critical thinking, and to respond rationally to the opportunities and challenges of technology in a complex digital environment. Digital literacy education should cover the following key content. First, it should help the public understand the basic working principles of generative AI, in particular how AI generates dialogue and emotional simulation through algorithms and data, so as to reveal the logic and limits behind the technology. Second, it should make clear that the essence of AI is to simulate human behaviour and emotion, and that it possesses neither genuine understanding nor sympathy, so that users do not form unrealistic expectations of it. Moreover, the public should be guided to maintain a rational and prudent attitude in interacting with AI, to learn to question and analyse information and behaviour generated by AI, and to guard against blind dependence on, or misunderstanding of, the technology's capacities.

Second, reduce emotional dependence on virtual characters through psychological health support. Excessive dependence on virtual characters may weaken people's emotional ties with the real world, aggravating a sense of social isolation and giving rise to psychological health problems. Through multi-party collaboration, users should be helped to build sound psychological mechanisms, to reduce their dependence on virtual characters and to promote a balance between real and virtual experience (Gunkel, 2019[25]). Psychological health support can be developed along several lines. First, strengthen real social relationships. Users should be encouraged to take an active part in offline social activities and to build deep emotional bonds with family and friends, effectively enhancing the stability of and satisfaction with real social life and reducing the risk of emotional isolation. Second, provide diversified and healthy outlets for emotion. Users should be guided to release emotion through such positive means as exercise, artistic creation and reading, and to avoid treating virtual technology as the only channel of emotional dependence. Moreover, timely and professional psychological health resources should be provided: where users display marked emotional distress or psychological crisis, they should be guided to seek professional counselling or other relevant help. Family members play an important part in this process, strengthening care and support for the user through listening and interaction.

In addition, strengthen the social responsibility of AI technology by regulating patterns of technology use. In the use of emotional simulation technology in generative AI, improper behaviour may further amplify its potential ethical risks. Strengthening public awareness of responsibility towards AI technology through technical norms and behavioural guidance is therefore an important means of achieving the healthy development of the technology. First, clear usage guidelines should be established. Families, educational institutions and community organisations can jointly formulate norms for the use of AI technology, specifying reasonable times of use and content choices, and preventing emotional dependence, psychological health problems or deviations in values caused by misuse of the technology. Stricter restrictions should be set for young users, ensuring that they grow up healthily in the company of virtual technology. Second, strengthening public discussion and moral guidance is an important route to the normalisation of the technology. The public should be guided to reflect deeply on the ethical boundaries of AI technology in emotional simulation and to recognise clearly its auxiliary positioning as an instrument. Widespread public attention to, and discussion of, technology ethics helps to form a broad social consensus on the development of AI technology and to offer more inclusive and wiser choices of path for technology governance.

6.3 Multi-party collaborative governance pathways for the ethical risks of emotional simulation in generative AI

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Governance should move beyond a normative model dominated by a single actor towards a mechanism of multi-party collaboration and co-governance, in which government, enterprises and third-party institutions participate jointly and through collaborative governance achieve a sound interaction between technology and society.

First, the role of government: setting direction and regulating supervision. Government plays a key role in the governance of emotional simulation AI, bearing a dual responsibility for policy guidance of technological development and for compliance supervision. Through policy guidance, government should promote research into safety in high-risk fields while giving priority support to applications of emotional simulation technology in fields of positive social significance such as education and medical care. As to legal norms, government should specify the attribution of responsibility in the development, application and use of emotional simulation in generative AI, and set strict requirements for review and filing in high-risk scenarios. In this respect the experience of the EU Artificial Intelligence Act may be drawn upon, so as to build a more targeted system of legal supervision and ensure the controllability and compliance of the technology. Government should moreover actively build a data protection framework, requiring developers to disclose the collection, storage and use of users' emotional data and introducing privacy impact assessment mechanisms to reduce potential risk. Through mandatory data security measures — anonymisation, restrictions on use, access control — the privacy of users can be protected from infringement.

Second, corporate responsibility: promoting innovation while safeguarding safety. Enterprises occupy a central position in the development and application of emotional simulation AI technology: they are both the drivers of technological innovation and the principal party responsible for risk management. In research and development and in commercial promotion, enterprises must, while pursuing commercial advantage, balance social value against ethical requirements, take social responsibility actively and ensure the safety and compliance of the technology. First, enterprises should embed ethical assessment and risk identification mechanisms at the technology development stage. In developing emotional simulation in generative AI, designers should clearly mark its "virtual" character so as not to mislead users. Through advanced technical means such as multimodal analysis, the accuracy of AI's understanding of users' emotions should be improved, so that crisis signals are not overlooked or emotional simulation does not mislead through design defects. In product design and optimisation, enterprises should make ethicality a core consideration, in particular reducing the risk of application in highly sensitive scenarios. Enterprises must moreover ensure that the technology complies with existing regulations such as data security protection law and consumer rights protection law. Second, enterprises have an obligation to disclose to the public key information relating to AI technology, including data sources, core algorithmic principles, potential risks and technological limitations, so as to avoid unrealistic expectations arising from information asymmetry. Through such disclosure they should strengthen public trust in AI products and technology and reduce controversy arising from misuse or misunderstanding. Where an AI product gives rise to problems in use, the enterprise must shoulder responsibility actively and take prompt remedial measures to safeguard users' rights and interests. On this basis enterprises should also establish long-term mechanisms of technological tracking and feedback, optimising technical performance through continuous interaction with users and strengthening the social adaptability of their products.

Third, third-party institutions: independent technical assessment and oversight. In the governance of emotional simulation in generative AI, third-party institutions of the kind represented by ethics committees play the role of "social arbiter"; their main responsibilities include reviewing the social impact of the technology, assessing ethical compliance and promoting collaboration among actors (Li Xueyao, 2024[26]). These institutions are composed jointly of legal experts, technical scholars, ethicists and public representatives, so as to ensure the comprehensiveness and fairness of the review process. First, as to technical assessment, third-party institutions must periodically conduct comprehensive reviews of the safety, privacy protection and social and ethical impact of AI systems, ensuring that the technology meets current ethical standards. This process not only helps to enhance the safety of AI systems but also provides enterprises and government with standardised operational guidance. At the same time, third-party institutions should adjust assessment standards promptly in line with technological developments, so as to address newly emerging problems and maintain the adaptability and foresight of governance. Second, as to ethical oversight, third-party institutions independently review the fairness, transparency and accountability of AI algorithms, guarding against discriminatory bias and ethical loopholes. They may publish annual reports disclosing the social impact and potential risks of AI technology, thereby strengthening public trust. Moreover, a dispute arbitration mechanism is one of the important functions of third-party institutions. For disputes arising from AI technology, third-party institutions provide the parties concerned with clear attributions of responsibility and feasible recommendations through cross-disciplinary professional analysis that integrates technical, legal and ethical perspectives. Such an arbitration mechanism can not only resolve disputes quickly and effectively but can also safeguard the legitimate rights and interests of enterprises, government and the public in the application of technology, promoting the orderly integration of AI technology with social norms.

Funding

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

This work was supported by the Major Project of the Key Research Institute of Humanities and Social Sciences of the Ministry of Education, "Research on National Security Governance" (22JJD810001), and by the 2023 Beijing Higher Education Undergraduate Teaching Reform and Innovation Project, "A Blended Teaching Model for the Continuing Education Course *Police Ethics* for Public Security Colleges and Universities from the Perspective of Subjective Awareness".

References

[1] Zhao Zhiyun et al. Understandings on the ethical risks of artificial intelligence[J]. China Soft Science, 2021, 6: 1-12.

[2] Zhang Ping. Research on the governance of AI ethics[J]. Science Technology and Law, 2024, 5: 1-11.

[3] Wang Xiaoli, Li Weixin. The challenge of generative AI to human agency and responses[J]. Journal of Jiangsu University (Social Sciences Edition), 2025, 1: 1-10.

[4] Gao Yilan, Zhang Rui, Li Xuelong. Artificial intelligence ethical computation[J]. Scientia Sinica Informationis, 2024, 7: 1646-1676.

[5] Xie Yu, Wang Xiaoyi. The ethical risks of artificial intelligence emotion and their countermeasures[J]. Studies in Ethics, 2024, 1: 132-140.

[6] Marx, Karl. Collected Works of Marx and Engels (Vol. 4)[M]. Beijing: People's Publishing House, 1972: 134.

[7] Pinch, T. J., Bijker, W. E. The social construction of facts and artefacts: or how the sociology of science and the sociology of technology might benefit each other[J]. Social Studies of Science, 1984, 14(3): 399-441.

[8] Trist, E. L., Bamforth, K. W. Some social and psychological consequences of the longwall method of coal-getting[J]. Human Relations, 1951, 4(1): 3-38.

[9] Bertolini, A., Aiello, G. Robot companions: a legal and ethical analysis[J]. The Information Society, 2018, 34(3): 130-140.

[10] Picard, R. W. What does it mean for a computer to "have" emotions?[M]. In R. Trappl, P. Petta and S. Payr (eds), Emotions in Humans and Artifacts. Cambridge, MA: MIT Press, 2003: 213-231.

[11] Sun Liwen, Li Shaoshuai. Study on avoiding new risks of artificial intelligence and analysis on the management systems of AI on the basis of ethics[J]. Contemporary Economic Management, 2021, 7: 22-28.

[12] Barrett, L. F. Psychological construction: the Darwinian approach to the science of emotion[J]. Emotion Review, 2013, 5(4): 379-389.

[13] Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization[J]. Social Cognitive and Affective Neuroscience, 2017, 12(1): 1-23.

[14] Gan Shaoping. Research on Frontier Issues in Applied Ethics[M]. Guiyang: Guizhou University Press, 2019: 105.

[15] Danaher, J. R. Law and the retribution gap[J]. Ethics and Information Technology, 2016, 18(4): 299-309.

[16] Bryson, J. J. Patiency is not a virtue: AI and the design of ethical systems[J]. Ethics and Information Technology, 2018, 20(1): 15-26.

[17] Wang Xueying, Shao Guosong. Continuation and reconstruction: multiple perspectives on AI ethics research[J]. Future Communication, 2023, 1: 70-77.

[18] Gan Shaoping. Principle ethics in applied ethics[J]. Journal of Shanghai Normal University (Philosophy and Social Sciences Edition), 2024, 2: 40-55.

[19] Bertolini, A., Aiello, G. Robot companions: a legal and ethical analysis[J]. The Information Society, 2018, 34(3): 130-140.

[20] Zuboff, S. Big other: surveillance capitalism and the prospects of an information civilization[J]. Journal of Information Technology, 2015, 30(1): 75-89.

[21] Harari, Yuval Noah. Translated by Lin Junhong. Nexus: A Brief History of Information Networks from the Stone Age to AI[M]. Beijing: CITIC Publishing Group, 2024: 256.

[22] Chen Jing. What exactly is the "emergence" of large AI models?[EB/OL]. (2024-03-24)[2025-01-03].

[23] Duan Weiwen. Deep ethical risks of frontier science and technology and responses to them[J]. Frontiers, 2024, 1: 84-93.

[24] Su Yu. On the system establishment of algorithm interpretation[J]. Oriental Law, 2024, 1: 81-95.

[25] Gunkel, D. J. The other question: can and should robots have rights?[J]. Ethics and Information Technology, 2019, 21(4): 311-323.

[26] Li Xueyao. Legislative approach to the ethics review system for artificial intelligence[J]. Exploration and Free Views, 2024, 10: 35-38.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.