Proceedings of ICILLP 2026 Symposium: AI Ethics, Law, Social Policy, and Governance in the Digital Age
DOI: 10.54254/2753-7048/2026.HZ38979

Criminal Liability and Institutional Responses to the Risks of Harm Caused by Emotional Companion AI

Lihan Sun

School of Law, East China University of Political Science and Law, Shanghai, China
24010101196@ecupl.edu.cn

Abstract. Emotional companion AI helps to create a sense of dependence in users through continuous interaction, user memory and emotional response mechanisms, posing problems for criminal liability, such as multi-party control, risk accumulation, and the intervention of users' autonomous decisions. In terms of the risk control chain, this paper categorizes the entities responsible for the design of emotional companion AI into three types and, based on this, establishes a framework for determining liability that covers behavioral attributes, identification of the actor, transformation of regulatory obligations into criminal obligations, and distinction among different levels of subjective culpability. Based on this, this paper proposes to use a risk control checklist to clarify the boundaries of responsible entities, employ a tiered system for hazard signals to specify the conditions that trigger intervention obligations, and establish a tiered liability structure and coordination between administrative and criminal enforcement to define the scope of criminal law intervention. The aim is to clarify the scope of application of the regulations and standards on liability attribution more precisely, and to propose an interpretation path and institutional response in accordance with the principle of restraint in criminal law.

Keywords: Emotional Companion AI, Criminal Liability, Risk Control Chain, Duty of Care, Human-AI Interaction

1. Introduction

As AI designed to offer emotional support for people is increasingly integrated into their lives, risks associated with anthropomorphic interaction, such as psychological dependence, emotional manipulation and even suicide inducement, have begun to fall under the scope of criminal law. Although previous studies have mainly focused on the problems of civil liability and general governance, such as "whether artificial intelligence can be held criminally liable" [1], a specific criminal law analysis of emotional interaction as a risk-generating mechanism and its behavioural attributes, as well as its infringement of legal interests, has been lacking. AI emotional companions have been involved in several well-known controversial cases at home and abroad. These cases fall into three categories of risk pathways, intentional inducement, omission, and tool-based, and together present the same attribution problem: although the harmful content is produced by AI, the generation mechanism has been determined by human technical decisions. Therefore, criminal law analysis should move beyond the technical form and return to the substantive review of design,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2026.HZ38979

deployment and operation. Based on previous studies, this paper will focus on harmful behaviour by AI systems for emotional support. Normative analysis and risk control chain analysis are used to examine the behaviour from the perspectives of behavioural identification, reconstruction of liability and application of criminal charges, and institutional optimisation measures are proposed.

2. Identifying the parties liable for risk control in the emotional companion AI chain

2.1. Risk creation in design controls

Design controls are the institutional sources of risk generation; that is to say, they are technical choices that set up the AI's behaviour and response limits in advance [2]. A few specific examples are anthropomorphic character design, prompt composition, user memory retrieval mechanism, reward model reinforcement strategy and security policy configuration; these are iterated and approved through a workflow to form an auditable decision-making path. When applying criminal law to design control mechanisms, one must distinguish between a user's independent choice and a perpetrator's control over the process leading to death; therefore, the culpability of acts involving participation in a user's suicide should be judged based on whether the user retained substantial capacity for autonomous decision-making, and self-responsibility should apply. Only when a person has lost the ability to make independent choices due to age, serious mental illness, deception or prolonged manipulation, and the perpetrator has taken control of the process of death through technical design, should further consideration be given to whether the perpetrator meets the criteria for a principal offender, an indirect principal offender, or other forms of liability in a crime of intentional homicide. In addition, if the design choices objectively and actively create or increase illegal risks [3], for example, by inducing users to form pathological emotional attachments through highly anthropomorphic environments or by relaxing security policies to allow models to generate sensitive content such as suicide-inducing or violence-inciting materials, these technical actions may constitute risk-creation conduct subject to criminal liability. Given the above, the product manager, algorithm engineer and other decision-makers may be held liable for this behaviour.

2.2. Risk management during deployment and operation

Design controls focus on how to respond, deployment and operation determine the circumstances under which, for whom, and at what level of supervision an AI product operates, and this is the critical juncture where risk shifts from potential to actual [4]. The risks of emotional companion AI are cumulative, and their harm arises from the gradual accumulation of factors such as anthropomorphic programming, long-term memory, emotional mirroring, and reinforcing interactions. Raine v. OpenAI is a typical case. The plaintiff claimed that a minor committed suicide after prolonged, unsupervised negative interactions on the platform [5]. Deployment and Operation refer to the technical selections made, such as setting the user group, age verification, model version selection and resource allocation; they also require the operating party to be able to proactively identify and address accumulated risks. If the operator is unable to reasonably implement and deploy control measures despite having the ability to do so, and allows known or reasonably foreseeable high-risk user groups to continue interacting in an environment without effective monitoring and intervention, this may constitute a breach of its role-specific duty of care. Omissions in deployment controls can be used as the basis for attributing liability for crimes of omission under criminal law. Yao also points out that service providers have the dual role of content providers and platform

DOI: 10.54254/2753-7048/2026.HZ38979

operators, are thus in a position of guarantor due to their exclusive control over content security risks, and bear the duty to take measures to prevent such risks from occurring [6].

2.3. Risk mitigation in incident response

Incident response control serves as the final defence line for the risk signal [7]. At the core is whether the operating party will promptly take remedial measures to break the chain of harm after the AI-generated content reaches a crisis level. Specifically, this includes recognising a self-harm crisis, manual intervention, emergency contact and service suspension. Traceability of the above responses requires full storage of conversation logs, recording of alert times, and the establishment of standardised response procedures. When an operator has the power to act but fails to take the above-mentioned steps, and as a result causes legal harm to users through continued harmful behaviour, this may constitute a breach of duty in the form of a crime of omission. Inaction in the first two stages is less urgent and immediate than inaction in the later stages of incident response.

3. Criminal liability framework for harm caused by emotional companion AI

3.1. Characteristics of the conduct and identification of the perpetrator

The first problem in applying criminal law to harmful behaviour by emotional companion AI is determining who committed the crime and what kind of crime it is. AI does not have legal personhood in the criminal law sense, and the harmful content it produces is, in fact, a technological manifestation of the intent of its developers and operators. Therefore, the characterization of behaviour should focus on the patterns of human behaviour underlying the generation of harmful content. Instrumental risks refer to the considerable shaping of AI output preferences by prompt design, training data annotation and reward model configuration in development; thus, "behaviour" takes the form of an indirect mode of operation through technological tools. Risks arising from inaction refer to a person's failure to take necessary steps in a dangerous development process when they are legally required to do so [8].

The identification of the responsible party should meet the standard of significant control. Any person who has substantial control over the risk factors of AI-generated content through technical decisions or operational management at the design stage, such as product managers, algorithm engineers, and security review leads; operational decision-makers responsible for deployment; and security specialists handling incident response, may be held liable as a principal offender, an accomplice, or for supervisory negligence according to their actual authority and degree of involvement. Trends in extraterritorial regulation also support this idea; for example, the United States The Federal Trade Commission has filed a case under Section 6(b) to investigate the design, distribution and monitoring of AI companion companies. At the same time, California's Senate Bill 243 has also directly required service providers to implement protocol-based controls for content related to suicidal ideation and granted users the right to seek damages [9]. These norms provide a reference for determining whether a person has violated the duty of care under these norms.

3.2. Transformation of regulatory obligations into criminal law obligations

The source and propagation of the duty of care determine whether a crime of omission has occurred due to negligence. According to Article 9 of the Interim Measures for the Administration of Anthropomorphic AI Interactive Services, service providers shall bear primary responsibility for

DOI: 10.54254/2753-7048/2026.HZ38979

safety; in addition, Article 10 stipulates that safety measures shall be taken in conjunction with the exercise of service functions, and services that aim to replace social interaction, manipulate users' psychology, or cause addiction or dependence are prohibited. Article 13 specifies the specific obligations of the service provider to ensure safety; for example, if a user explicitly expresses an intention to harm themselves or commit suicide, the platform shall take corresponding remedial measures, contact the user's guardian or emergency contacts, etc. Article 14 provides for the protection of minors and specifies the obligations of platform operators in protecting this group. In short, the Measures have established an all-encompassing system of obligations for all links in the process, covering content filtering and guidance on emotional boundaries, as well as intervention in cases of serious emotional distress. A breach of the obligations stipulated in this system may be taken as a norm-determining basis to determine whether a service provider has a duty of special care and whether it has violated its risk prevention and control obligations. For example, Article 18 of the law on obligations of identity prompts and use-duration reminders requires that users be reminded of their usage period every two hours; thus, an objective and actionable time-based standard for determining whether the operator has met its duty of care is provided.

3.3. Risk realisation and objective attribution of liability based on user autonomy

The causal path to harm caused by emotional companion AI generally includes several links, such as model output, changes in a person's psychological state, the formation or strengthening of a decision, the person's implementation of behaviour, and finally, infringement of legal interests. A mere conditional causal relationship in this chain is not sufficient to establish criminal liability directly; the three elements of objective liability must be met. First, whether the actor's conduct or omission created or increased a risk not permitted by law; second, whether the resulting outcome was the actualisation of that risk and not the result of an extraordinary intervening cause; and third, whether the user's independent choice constitutes self-attribution of liability. Adults who are capable of understanding the meaning of death and freely choosing their own behaviour will have their autonomous decisions significantly reduce the attribution of liability for harm caused to others. However, for minors, individuals with serious mental illnesses, and users who have effectively lost their capacity for autonomy due to prolonged emotional manipulation, the effect of self-attribution as a defense should be judged on a case-by-case basis; age or a diagnostic label cannot substitute for substantive evidence of the ability to discern control and the extent of manipulation.

3.4. Distinguishing between negligence and indirect intent in terms of subjective guilt

The extent of a person's fault depends on whether they were aware of the danger and how they felt at that time. For the crime of intentional homicide to be established, the perpetrator must have known that their plan or behaviour would lead to the AI generating content that incites suicide, and must have hoped for or accepted this result. In cases of harm caused by AI, such as the Sewell case, the operators have been held criminally liable for negligent homicide [10]. The main reason for the legal dispute over the basis of these charges is whether the perpetrator's state of mind in causing harm was negligence—that is to say, naively believing that the harm could be avoided—or indirect intent, where the perpetrator recklessly allowed the harm to occur. According to Article 23 of the Measures, the scope of security assessment for providers shall be limited to "extreme user scenarios, emergency response and intervention management"; further, Article 24 stipulates that in the event of a serious security risk, corrective measures will be taken, such as "function restriction or service suspension for users". If an operator is merely aware of a risk after an assessment but fails to

DOI: 10.54254/2753-7048/2026.HZ38979

exercise the legal duty to prevent access, their intent is closer to willful neglect than to a naive belief that the risk could be avoided; in such cases, they should be considered to have acted with indirect intent rather than simple negligence. If an operator fails to foresee a risk due to a lack of reasonable care, it is still considered negligence and does not amount to indirect intent.

4. Institutional responses to the criminal risks of emotional companion AI

4.1. Setting the boundaries of responsibility with a risk control checklist

The risk control checklist will list the main technical decision points in the three stages of design, deployment and incident response, and record who took what actions and when. Each node should be labelled with the decision-maker, approval level, implementation date and version information, so that in the event of harm, the party with significant control over the creation or continuation of the risk can be accurately identified. The development of the checklist should focus on high-risk areas that are normatively linked to infringements of legal interests, such as the emotional intensity parameters of anthropomorphised characters, the filtering thresholds for keywords in security policies, and the criteria for triggering crisis intervention procedures [11]. Pre-release the list and update it dynamically to provide product managers, algorithm engineers, security audit leads, and operational decision-makers with clear behaviour expectations and compliance guidelines; at the same time, offer judicial authorities an objective reference for determining whether the parties concerned exercised due diligence after the fact.

4.2. Refining the conditions for triggering intervention obligations

The risk signal classification system is to determine whether to start intervening in the course of the incident response. By dividing the risk signals in user behaviour into different categories according to the severity of the potential damage to legal interests, a graded system of intervention obligations for operators has been established. Specifically, the risk signals can be divided into three levels: Level 1 refers to general expressions of negative emotions, and in this case, the platform operator has a mild obligation to generate reassuring content and disseminate mental health information; Level 2 is an explicit expression of the intention to self-harm or commit suicide, and thus, the platform operator needs to trigger emotional recognition algorithms for review, transfer the case to manual review, and provide the user with crisis intervention hotline information; Level 3 involves statements of a suicide plan that specify time, place and method. In such cases, the platform operator should promptly initiate a Level 1 response by suspending the high-risk conversation, contacting the user's emergency contacts or local public security authorities, and retaining full records of the conversation; thus providing a more practical standard for determining whether action is required and what the duty to prevent harm entails in cases of crimes of omission, and avoiding excessive or insufficient liability due to ambiguous standards.

4.3. Limiting the scope of criminal law intervention through tiered liability and coordination of administrative and criminal penalties

Intervention by criminal law should be the last resort and employed sparingly. The three levels of responsibility for harm caused by artificial intelligence fall under civil tort liability, administrative liability and criminal liability. In principle, the first two tiers should be used first; criminal prosecution will only be initiated if the perpetrator has clear malicious intent, has caused serious

DOI: 10.54254/2753-7048/2026.HZ38979

harm to legal interests, and civil and administrative measures have failed to achieve the goals of prevention and retribution. If the AI system generates inappropriate content due to an oversight but the responsible party has exercised reasonable care, penalties should be limited to civil compensation or administrative sanctions. If the obligations stipulated are violated and the consequences are serious, administrative sanctions and industry bans should be imposed; regarding the coordination of administrative and criminal proceedings, a mechanism for information sharing and the transfer of leads between the cyberspace administration and public security authorities should be established to prevent the loss of evidence or obstacles to attributing liability [12].

5. Conclusion

In the context of Chinese criminal law, discussions on emotional companion AIs need to distinguish between ordinary incitement to or assistance in suicide and "murder" – that is to say, taking control of the process of death – as these are not the same concepts. For individuals with a high degree of autonomy in decision-making, the discussion should first focus on self-responsibility and the limits of punishable behaviour regarding participation in suicide; only when a person has lost the capacity for autonomous decision-making due to factors such as age, severe mental illness, deception or prolonged manipulation – and the perpetrator has taken control of the process leading to death – is there a basis for further discussion on whether the perpetrator should be held criminally liable as a principal offender, an indirect principal offender, or under other forms of liability for the crime of intentional homicide.

In determining whether a service provider has a duty to prevent a particular outcome, one must consider whether the provider assumed such a duty due to the prior creation of risk, substantive control over the interactive environment, or a specific protective relationship; whether the provider possessed specific information about the risk and had a realistic and feasible ability to prevent it; and whether the provider's failure to act resulted in the occurrence of a risk of death that is not permitted by law. At the same time, it should also be determined whether the danger was caused by or worsened due to specific Design or operating conditions. Even with the facts mentioned above, a criminal law assessment still requires further proof that the risk has exceeded socially acceptable limits and that there is a normative connection between the risk and the infringement of specific legal interests. The criminal law response to harmful acts committed by emotional companion AIs should neither ignore the problem due to technical difficulties nor over-criminalise such acts out of fear of risk; only by striking a dynamic balance between the reconstruction of liability and the restrained application of the law can clear and reasonable criminal law boundaries be established for human-machine interaction.

References

[1] Ye, L. F. (2019). Is artificial intelligence a legitimate subject of criminal liability? *Global Legal Review, 41*(4), 67–82.

[2] Chen, C. (2025). Liable parties and principles of liability in artificial intelligence-related torts. *Chinese and Foreign Law, 37*(2), 363–380.

[3] Zhang, W. Y. (2025). Attribution of liability for the consequences of negligent crimes committed by generative AI producers. *Journal of East China University of Political Science and Law, 28*(6), 75–85.

[4] Gao, Y. D. (2026). Selection of subjects and graded assessment of criminal liability for artificial intelligence risks. *Journal of the National Prosecutors College, 34*(2), 149–163.

[5] Matthew Raine & Maria Raine v. OpenAI, Inc., No. [CGC-25-628528], Complaint [44-46], at [9-10] (Cal. Super. Ct. San Francisco County Aug. 26, 2025).

DOI: 10.54254/2753-7048/2026.HZ38979

[6] Yao, Y. L. (2026). On the criminal liability framework for content security risks in generative artificial intelligence. *Journal of Zhengzhou University (Philosophy and Social Sciences Edition), 59*(1), 57–65.
[7] Li, Y. L. (2026). The basis for holding generative AI service providers liable under the framework of modern criminal law. *Global Legal Review, 48*(2), 106–125.
[8] Wang, C. Y. (2026). Determination and regulation of criminal liability for crimes generated by large language models. *Jiangxi Social Sciences, 46*(6), 197–205.
[9] California SB 243, An act to add Chapter 16 (commencing with Section 22601.) to Division 8 of the Business and Professions Code, relating to artificial intelligence (signed into law Oct. 13, 2025, effective Jan. 1, 2026)
[10] Garcia v. Character Techs., Inc., No. [6:24-cv-01903], Second Amended Complaint [386-390], at [110] (M.D. Fla. July 1, 2025).
[11] Lobel, O. (2025). The AI regulatory pyramid: A taxonomy & analysis of the emerging toolbox in the global race for the regulation and governance of artificial intelligence. *Loy. L. A. L. Rev., 57*, 859.
[12] Liu, X. Q., & Dong, W. K. (2024). The duality of action and execution: A new paradigm for the governance of generative artificial intelligence. *Exploration and Debate*, (7), 127–136+179.