



## Perspective

## The human-AI empathy loop: a path toward stronger human-AI relationships

Ruolei Gu<sup>a,b,1</sup>, Yuqing Zhou<sup>a,b,1</sup>, Suchen Yao<sup>a,b</sup>, Yuan Zhou<sup>a,b</sup>, Fang Cui<sup>c</sup>, Frank Krueger<sup>d,e,\*</sup>, Grit Hein<sup>f,\*</sup>, Li Hu<sup>a,b,\*</sup>

<sup>a</sup> State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China

<sup>b</sup> Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

<sup>c</sup> School of Psychology, Shenzhen University, Shenzhen 518060, China

<sup>d</sup> School of Systems Biology, George Mason University, Fairfax, VA 22030, USA

<sup>e</sup> Department of Psychology, University of Mannheim, Mannheim 68131, Germany

<sup>f</sup> Translational Social Neuroscience Unit, Department of Psychiatry, Psychosomatics, and Psychotherapy, University of Würzburg, Würzburg 97080, Germany

As artificial intelligence (AI) rapidly transforms social interactions, empathy is no longer confined to human-to-human relationships but increasingly involves human-AI connections resembling interpersonal dynamics [1]. AI's ability to recognize human emotions parallels how humans interpret others' feelings [2], while its capacity to generate emotional expressions mirrors how humans convey emotions as empathy targets [3]. Building on established models that distinguish empathizer and empathy target roles across affective, cognitive, motivational, and behavioral dimensions [4], we propose the human-AI empathy loop framework to examine empathy from both perspectives (Fig. 1a). Unlike prior work that treats human-to-AI and AI-to-human empathy as one-way processes, our approach emphasizes their mutual, evolving nature. We further extend the framework by introducing an AI-to-AI empathy component, highlighting how empathy-like mechanisms among agents could enrich the loop in multi-agent contexts.

**Current approaches to empathy.** Empathy is commonly defined as the capacity to understand, share, and explain others' feelings through observation or communication [4]. Our recent meta-analysis supports a multidimensional framework with four dimensions, including (i) affective empathy: sharing or becoming affectively aroused by others' emotions; (ii) cognitive empathy: consciously understanding others' perspectives; (iii) motivational empathy: the desire to empathize with or help others; and (iv) behavioral empathy: expressing empathetic responses [4]. Affective and motivational empathy appear particularly relevant in human-AI interactions [5].

Notably, empathy extends beyond humans to animals and even objects—individuals high in empathy are more likely to care for nonhumans and attribute meaning to inanimate entities [6–8].

This capacity to generalize empathic responses underlies the possibility of empathizing with AI, reinforcing empathy as a cornerstone of social interaction in the emerging human-AI society.

**Human empathy toward AI.** Human empathy toward AI can be assessed similarly to human empathy toward humans, using self-reports, third-party evaluations, and physiological measures (e.g., heart rate, eye movements) [5]. For AI agents to succeed in socially demanding roles, humans must emotionally recognize them as social entities rather than mere tools. Such empathy facilitates trust, cooperation, human-AI teaming, and moral consideration [9], whereas its absence can lead to adverse outcomes—for instance, children have been observed obstructing or vandalizing service robots, behaviors that may desensitize them to violence and undermine moral development [1,10].

Efforts to foster empathy toward AI fall into two main approaches. The first enhances anthropomorphic qualities, such as human-like features, voices, or interactive behaviors. While effective, these strategies are constrained by the “uncanny valley” effect and are more applicable to embodied agents [1,8,10]. The second approach targets mind perception—perceiving AI as capable of agency (thought and intention) and experience (feeling and emotion). For example, granting robots autonomy to refuse orders enhances perceived agency, thus increasing their companionship value [10].

Although both anthropomorphism and mind attribution strategies emphasize subject-object similarity and primarily activate affective empathy [4,8], the social interaction framework reveals important conceptual differences: mind attribution can be manipulated through either the observer (human) or the agent (AI), whereas anthropomorphism exclusively targets the agent [4]. This framework also identifies an additional pathway for fostering empathy: altering relational dynamics between observers and agents. For instance, observing robots being harassed can elicit empathic responses even when people understand that robots do not truly experience suffering [10].

\* Corresponding authors.

E-mail addresses: [frueger@gmu.edu](mailto:frueger@gmu.edu) (F. Krueger), [hein\\_g@ukw.de](mailto:hein_g@ukw.de) (G. Hein), [huli@psych.ac.cn](mailto:huli@psych.ac.cn) (L. Hu).

<sup>1</sup> These authors contributed equally to this work.

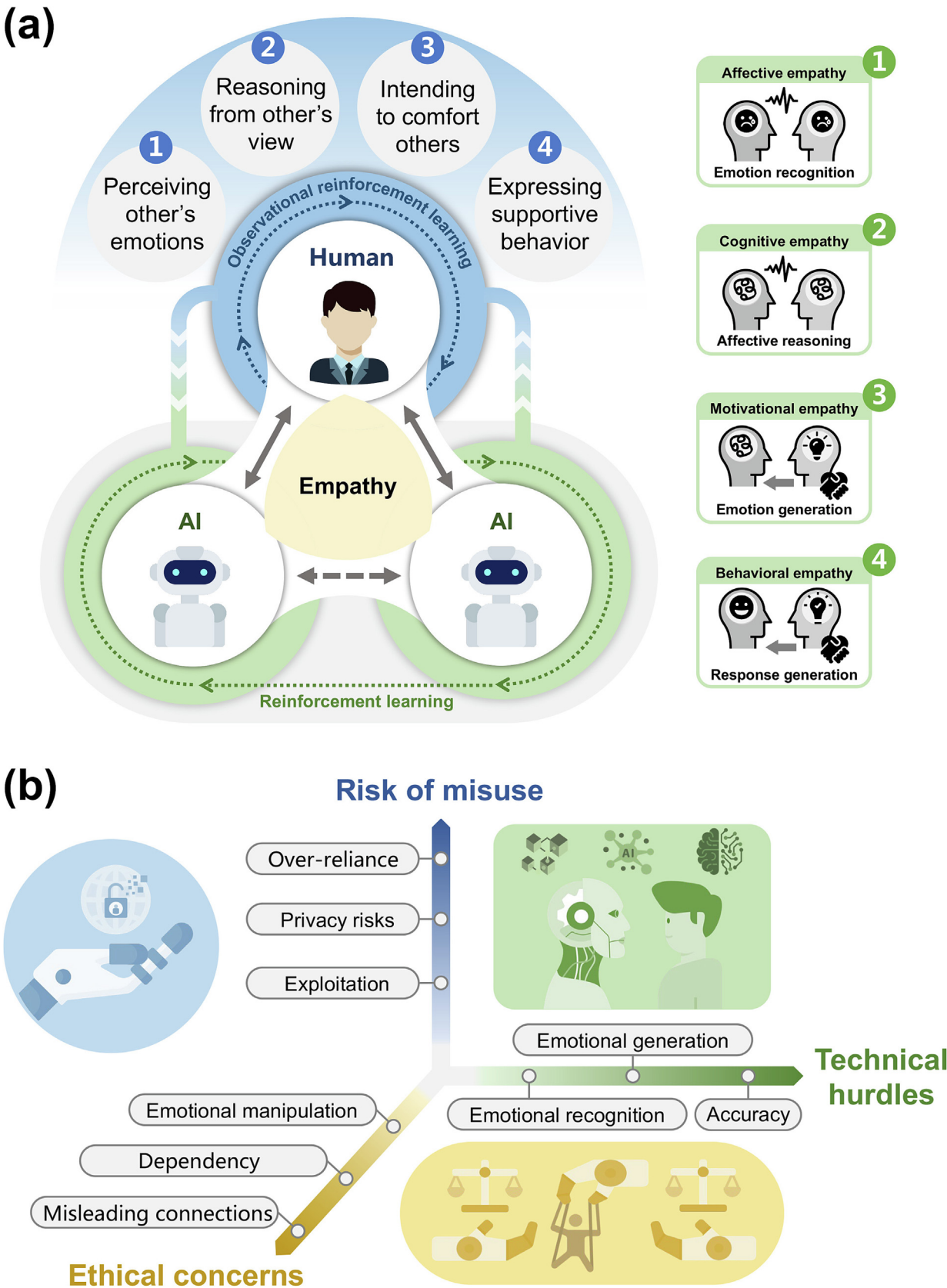


**Fig. 1.** (a) Conceptual framework of the human-artificial intelligence (AI) empathy loop. The framework outlines a dynamic loop comprising human-to-AI, AI-to-human, and potentially AI-to-AI empathy. Human-to-AI empathy mirrors human interpersonal empathy, encompassing multiple psychological facets linked to distinct empathy dimensions. Observational reinforcement learning may facilitate both human-to-AI and AI-to-human empathy, while AI-to-AI empathy is expected to be reinforced primarily through direct reinforcement learning. (b) Key challenges in integrating empathy into human-AI interactions, which include technical hurdles, ethical considerations, and the risk of misuse.

Building on this, we propose applying observational reinforcement learning (ORL)—defined in human studies as prediction-error-based learning from others' actions and outcomes [11], and in machine learning as imitation through observation [12]—to modulate empathy in human-AI interactions. Despite disciplinary differences, both converge on the idea that agents can acquire new strategies indirectly. Inspired by findings that people can adjust empathy by observing others' empathic reactions [11], we suggest that ORL could strengthen the human-AI empathy loop by having humans observe other humans expressing empathy toward AI. Witnessing others respond empathically to an AI's apparent "suffering" may shape first-time observers' feelings, particularly when they are uncertain about appropriate reactions. This opens new opportunities for AI-related empathy training [4].

Finally, one of our authors recently proposed an ethological approach: modeling AI-human attachment on human-dog relationships [6]. Compared to human-human bonds, interspecies relationships are simpler yet capable of depth and longevity. Dogs, in particular, can form strong emotional attachments and are suitable for specific functional roles such as assisting the elderly or disabled. Designing AI agents with dog-like qualities may leverage existing empathy toward companion animals, avoid the uncanny valley, and prioritize functionality over human-like replication. This offers a practical roadmap for empathy-centered AI design.

*AI empathy toward humans.* In an emerging human-AI society, an AI agent's empathic capacity will be crucial for integration into social contexts, depending on its abilities in emotion recognition and response generation.

From the perspective of emotion recognition, human emotional signals extend beyond language to encompass facial expressions, vocal tones, gestures, and even interoceptive changes. To decode this complexity, social AI agents increasingly rely on affective computing systems—integrating hardware, algorithms, and multimodal databases—to infer emotional states. Pain recognition is a prominent application, given its multimodal expression and centrality to interpersonal empathy [2]. Techniques such as facial analysis via appearance models, acoustic signal extraction, and natural language processing help detect pain indicators, which are then integrated to estimate intensity. Remaining challenges include individual differences in expression and socially motivated displays that obscure true states. Similar issues arise when decoding other emotions.

Neuroscience can advance this field by clarifying the neural pathways underlying pain perception and empathic processing [13]. Brain-computer interface technologies, such as portable electroencephalography, may eventually enable real-time transmission of neural signals related to pain and emotion, reducing reliance on external cues. Grounding these systems in neuroscientific insights—such as how empathy and pain are encoded in the anterior cingulate cortex, insula, and sensorimotor areas—could inform brain-inspired AI architectures with enhanced capacity to interpret complex emotions [13].

From the perspective of response generation, AI empathy manifests in its ability to respond appropriately to users' emotions. For non-embodied agents such as chatbots, the evaluation of AI empathy involves linguistic analysis of dialogue, task-based performance metrics, and human perception ratings [9]. For embodied agents such as social robots, additional behavioral dimensions can be assessed, including the timing, synchrony, and naturalness of body cues [10]. Yet even when AI-generated empathic responses are accurate and socially appropriate, they are often perceived as less genuine and less valuable than equivalent human responses, particularly in emotional sharing and caring [5]. Moreover, not all AI agents require empathic capacity to the same degree. As discussed above, perceptions of an agent's agency and experience jointly shape human attitudes toward it, providing a useful frame-

work for identifying when empathy is most relevant in AI design (Table 1).

For non-embodied agents, research shows that empathetically enhanced chatbots deliver better emotional support, yielding higher user satisfaction and more positive human-AI interactions [9]. Dialogue research has explored strategies for enhancing affective empathy in response generation, such as active listening, providing comfort, and real-time emotional mimicry (e.g., facial or gestural replication) [3,9]. At the same time, cognitive empathy remains vital, as it involves using commonsense knowledge to infer implicit meanings behind explicit expressions, thus enabling deeper understanding [3]. Comprehensive response strategies should therefore address multiple empathy dimensions, and task paradigms such as narrative-based judgment scenarios are now being used to evaluate AI performance.

For embodied agents, empathic responses often rely on predefined behavioral rules. Yet real-world social interactions are too complex for purely rule-based approaches. A key challenge lies in bridging emotion recognition with flexible, context-sensitive responses. For example, recognizing sadness does not prescribe a uniform response—grief and disappointment call for different forms of support. To address this, researchers have applied reinforcement learning methods that allow robots to refine empathic behavior based on user feedback [14]. Extending this idea, we hypothesize that both embodied and non-embodied agents could benefit from ORL, understood here as learning behavioral strategies through observation rather than direct feedback [11,12]. In this extension, ORL would target empathic responses demonstrated by humans, enabling AI agents to internalize socially appropriate strategies in unstructured environments [2,11]. Although this application remains hypothetical, we believe that integrating ORL with self-supervised learning could allow agents to autonomously extract empathy-relevant cues and refine their behavior, avoiding overreliance on direct feedback or rigid rules. In this regard, recently proposed robotic lifelong reinforcement learning frameworks—designed for continuous learning while avoiding catastrophic forgetting—provide a promising foundation [14].

*Closing the empathy loop.* Both human-to-AI and AI-to-human empathy have attracted considerable attention from researchers and practitioners in academia and industry [5]. Here, we discuss how these two aspects converge into an empathy loop framework.

In human-human interactions, empathy functions as an emotional resonance between individuals. The roles of empathizer and empathy target are dynamic and reciprocal. When an empathy target perceives the empathizer's response to their negative emotions, they interpret it as a signal of attention and willingness to help; this strengthens their social bond and elicits positive emotional feedback, such as gratitude [4]. Later, when the original empathizer experiences distress, role reversal may occur (i.e., the empathizer becomes the empathy target), which reinforces prosocial motivation and sustains an empathy-based loop [4,10].

Although this framework is well-established in human-human contexts, little is known about whether similar dynamics apply to human-AI relationships [5]. Evidence suggests, however, that humans naturally form stronger connections with entities that display behavioral empathy, including AI agents [1,7]. When AI agents effectively recognize and respond to human suffering, they can foster stronger bonds with users and potentially elicit reciprocal empathy, resembling empathizer-target exchanges in human relationships [1]. Over time, this may generate a self-reinforcing empathy loop within society. Conversely, empathy-deficient AI systems risk rejection, resistance to collaboration, or leadership challenges [5].

One key difference concerns reciprocity. In human-human relationships, people often favor those who empathize with them. Yet

**Table 1**  
Artificial intelligence (AI) agent types by agency-experience profile and empathy relevance.

| AI agent type                       | Agency   | Experience | Role examples                                                              | Relevance of empathy                                                                                     |
|-------------------------------------|----------|------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| High agency/low experience          | High     | Low        | Supercomputers, algorithmic trading bots, autonomous logistics controllers | Minimal relevance: tasks are primarily instrumental, with little social interaction.                     |
| Moderate agency/moderate experience | Moderate | Moderate   | Educational tutors, customer service chatbots, healthcare assistants       | Moderate relevance: empathy can improve engagement, satisfaction, and compliance.                        |
| Low agency/high experience          | Low      | High       | Virtual pets, care robots, therapeutic companions                          | High relevance: emotional engagement is central to purpose; empathy fosters trust and rapport.           |
| High agency/high experience         | High     | High       | Team leaders, collaborative creative partners, AI psychotherapists         | Critical relevance: empathy shapes cooperation, moral consideration, and long-term relationship quality. |

for social service AI agents, granting preferential treatment to specific individuals would conflict with principles of fairness and ethics, even if those individuals expressed empathy toward the AI. A potential solution is for AI agents to express gratitude when humans demonstrate empathy (e.g., a tourist helping a fallen service robot). Such acknowledgment could reinforce the empathizer's prosocial tendencies and encourage bystanders to respond empathetically as well. To participate effectively in empathy loops, AI must therefore recognize not only human distress signals but also human expressions of empathy, responding in ways that align with social norms.

Another obstacle lies in public perception: should AI agents be accepted as legitimate empathy targets? We argue that as long as an AI demonstrates functional social competence, interacts in a human-like manner, and serves as an integrated part of society, it merits empathy—regardless of whether its experiences are genuine. Promoting this view may require sustained public education and awareness-building. Moreover, some scholars have begun advocating for legal frameworks that grant AI agents limited autonomy and rights [15]. Such social transformations could fundamentally reshape how we understand AI agents and the future of human-AI empathetic relationships.

*AI empathy toward AI?* Finally, in addition to human-to-AI and AI-to-human empathy, we propose another potential component: AI-to-AI empathy. If empathy is understood as an evolved algorithm for strengthening bonds and resolving social dilemmas, embedding an empathy-like mechanism in AI-AI interactions could provide ecological benefits. For instance, in multi-agent collaborative scenarios with incomplete contextual information, determining whether another agent requires assistance can be difficult. Embedding empathy-like goals into behavioral generation would shift inter-agent support from rigid procedures to flexible, goal-directed responses, thereby improving autonomy and reducing computational costs. Here, cognitive empathy components such as perspective-taking could help agents anticipate others' future behavior (rather than merely reacting to the current situation), while affective components could assign intrinsic motivations to assist peers—similar to how humans sometimes help others to relieve their own distress. Such mechanisms would reinforce inter-agent connections, promoting cohesion and more efficient task execution. This idea parallels recent work introducing a brain-inspired multi-agent theory of mind model, where agents predict each other's actions via self- or other-modeling to enhance cooperation [16].

*Final thoughts and conclusion.* While embedding empathy in AI systems holds promise, it also poses major challenges (Fig. 1b). These include technical limitations in replicating human empathy, ethical concerns about manipulation, and risks of over-reliance or inadvertent disclosure of sensitive information. Regulatory frameworks, such as the European Union Artificial Intelligence Act (effective February 2025), already restrict emotion inference in sensitive domains, underscoring the need for clear legal boundaries. Moreover, empathy may reinforce in-group versus out-group distinc-

tions, potentially shaping observational learning processes [17] and becoming amplified through human-AI feedback loops—warranting careful investigation in the design of empathy training interventions. Despite these challenges, fostering empathy between humans and AI remains vital for harmonious coexistence. Progress will require coordinated efforts across education, policy, technology, and public engagement to establish a sustainable, society-wide empathy loop.

In our view, the bidirectional strengthening of empathy between humans and AI represents a form of emotional alignment—an emerging concept that parallels value alignment but centers on shared affective understanding. We hope this perspective will inspire future research exploring the empathy loop framework and developing more socially integrated human-AI relationships.

**Conflict of interest**

The authors declare that they have no conflict of interest.

**Acknowledgments**

This work was supported by the National Natural Science Foundation of China (32441103, 32020103008, 32300880, and 32371130), the National Science and Technology Major Project of China (2025ZD0546202), Beijing Philosophy and Social Science Foundation (24DTR063), and Scientific Foundation of the Institute of Psychology, Chinese Academy of Sciences (E3CX1225 and E2CX1524YZ). The authors thank Haiyan Wu, Chao Liu, Shenghua Luan, and Tingting Wu for suggestions. They used ChatGPT-5, Alibaba Cloud's Qwen, and Claude 4.0 Sonnet for refining the grammar and improving the flow of the writing.

**References**

[1] Damiano L, Dumouchel P, Lehmann H. Towards human-robot affective co-evolution overcoming oppositions in constructing emotions and empathy. *Int J Soc Robot* 2015;7:7–18.

[2] Cao S, Fu D, Yang X, et al. Pain recognition and pain empathy from a human-centered AI perspective. *iScience* 2024;27:110570.

[3] Sabour S, Zheng C, Huang M. CEM: commonsense-aware empathetic response generation. *Proc AAAI Conf Artif Intell* 2022;36:11229–37.

[4] Wu X, Yao SC, Lu XJ, et al. Categories of training to improve empathy: a systematic review and meta-analysis. *Psychol Bull* 2024;150:1237–60.

[5] Rubin M, Li JZ, Zimmerman F, et al. Comparing the value of perceived human versus AI-generated empathy. *Nat Hum Behav* 2025;9:2345–59.

[6] Krueger F, Mitchell KC, Deshpande G, et al. Human-dog relationships as a working framework for exploring human-robot attachment. *Anim Cogn* 2021;24:371–85.

[7] Safdari A. Toward an empathy-based trust in human-otheroid relations. *AI Soc* 2024;40:3123–38.

[8] Geiselmann R, Tsourgianni A, Deroy O, et al. Interacting with agents without a mind: the case for artificial agents. *Curr Opin Behav Sci* 2023;51:101282.

[9] Sharma A, Lin IW, Miner AS, et al. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat Mach Intell* 2023;5:46–57.

[10] Broadbent E. Interactions with robots: the truths we reveal about ourselves. *Annu Rev Psychol* 2017;68:627–52.

- [11] Zhou Y, Han S, Kang P, et al. The social transmission of empathy relies on observational reinforcement learning. *Proc Natl Acad Sci USA* 2024;121: e2313073121.
- [12] Torabi F, Warnell G, Stone P. Behavioral cloning from observation. *arXiv: 1805.01954v2*, 2018.
- [13] Tu Y, Li Z, Zhang L, et al. Pain-preferential thalamocortical neural dynamics across species. *Nat Hum Behav* 2024;8:149–63.
- [14] Meng Y, Bing Z, Yao X, et al. Preserving and combining knowledge in robotic lifelong reinforcement learning. *Nat Mach Intell* 2025;7:256–69.
- [15] Gunkel DJ. The other question: can and should robots have rights? *Ethics Inf Technol* 2018;20:87–99.
- [16] Zhao Z, Zhao F, Zhao Y, et al. A brain-inspired theory of mind spiking neural network improves multi-agent cooperation and competition. *Patterns* 2023;4:100775.
- [17] Chen D, Yao Z, Liu J, et al. Social conformity updates the neural representation of facial attractiveness. *Commun Biol* 2024;7:1369.