

Article

From Human Empathy to AI-Generated Empathic Performance: Conceptual Boundaries and Ethical Risks for Human-Centered AI

Rabia Zamur Tuncer ^{1,*}  and Şafak Tanır Levendeli ² 

¹ Department of Journalism, Faculty of Communication, Istanbul University, 34116 Istanbul, Türkiye

² Department of New Media and Communication, Faculty of Communication, Istanbul Aydın University, 34295 Istanbul, Türkiye

* Correspondence: rabia.zamur@istanbul.edu.tr

Abstract

The increasing integration of artificial intelligence (AI) into everyday communication has complicated the conceptual boundaries of empathy by enabling systems to generate responses that appear emotionally attentive, supportive, and context-sensitive. Through a concept-centered theoretical synthesis of scholarship on human empathy, artificial empathy, human–AI interaction, and AI ethics, this article distinguishes three analytically distinct phenomena: human empathy, AI-generated empathic performance, and perceived AI empathy. In the present framework, human empathy is organized around four dimensions—cognitive understanding, affective responsiveness, self–other differentiation, and communicative expression—which are used as analytical reference points for comparing the observable capacities and conceptual limits of AI-generated empathic performance. The analysis argues that AI systems can approximate communicative features associated with empathy through emotional and contextual inference, affect-sensitive expression, adaptive response generation, and conversational continuity, but such performance should not be treated as equivalent to the relational and experiential dimensions of human empathy. Perceived AI empathy is shaped by source attribution, interactional context, and user interpretation. The resulting ethical implications concern attributional and relational risks, vulnerability, data and fairness, transparency, accountability, and governance. Operationally, the framework supports separate evaluation of system performance and perceived empathy, disclosure of AI status and inferential limits, communication of uncertainty in emotion inference, calibration of anthropomorphic and relational cues, and pathways to human support in emotionally sensitive or high-stakes contexts.

Keywords: human empathy; AI-generated empathic performance; perceived AI empathy; human-centered AI; human–AI interaction; AI ethics



Academic Editor: Tihomir Orehovački

Received: 7 September 2026

Revised: 15 October 2026

Accepted: 18 October 2026

Published: 22 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Empathy has been conceptualized as a multidimensional rather than a unitary phenomenon, although different theoretical traditions emphasize different components of the construct. Davis [1] operationalized dispositional empathy through distinct dimensions, including perspective taking, empathic concern, personal distress, and fantasy, distinguishing cognitive and affective tendencies within empathic responding. Hoffman [2] approached empathy from a developmental perspective, emphasizing the interplay between affective arousal and increasingly differentiated forms of perspective- and role-taking. Cuff et al. [3],

drawing together diverse definitions across the literature, further demonstrated the conceptual heterogeneity of empathy and proposed an integrative account in which both affective and cognitive processes are central. From a relational and communicative perspective, empathy involves not only recognizing another person's emotions or perspective but also maintaining a distinction between self and other and communicating this understanding in a way that enables the other person to feel understood [4–7]. Related accounts similarly emphasize understanding another person's thoughts and emotions from that person's perspective and integrating cognitive and affective components within empathic responding [1,3,8]. Accordingly, this article approaches human empathy not as a single emotional reaction but as a relational process involving cognitive understanding, affective responsiveness, self–other differentiation, and communicative expression. These dimensions serve as analytical reference points for examining the conceptual boundaries of empathy in relation to AI-mediated interaction.

The growing integration of AI into everyday communication complicates these established conceptual boundaries. Artificial empathy research has examined how artificial agents may operationalize functions associated with empathic interaction, including affective processing, self–other cognition, and responsive behavior [9,10]. McStay [11], from the perspective of emotional AI and empathic media, focuses more broadly on technologies that interpret affective signals such as feelings, moods, attention, and intention and on the social implications of such systems. More recently, Suh et al. [12] shifted the analytical focus toward user perceptions of empathy in sustained interactions with large language models, showing that empathy judgments are highly individualized and context-sensitive and can be disrupted when conversational continuity fails or user expectations go unmet. These capacities have encouraged the increasing use of terms such as algorithmic empathy, artificial empathy, and AI empathy. Recent work on “empathy algorithms” emphasizes data-driven and multimodal approaches to affective inference [13], while research on artificial empathy more broadly examines systems designed to recognize affective cues and generate adaptive or supportive responses [9,10]. However, a system's capacity to generate empathic-seeming responses does not, by itself, resolve the conceptual question of whether such outputs should be treated as equivalent to empathy as theorized in human interaction.

Several adjacent concepts address different aspects of empathy-related AI interaction, but they operate across different analytical levels. Affective computing primarily concerns computational approaches to recognizing, interpreting, and generating affective information [14], while research on artificial empathy examines how artificial agents can reproduce or operationalize functions associated with empathic interaction [9,10]. Research on empathetic dialogue systems focuses more specifically on systems designed to enhance the perception and expression of emotional states, personal preferences, and knowledge, with emotion-awareness, personality-awareness, and knowledge-accessibility identified as key features [15]. The notion of simulated empathy has also been used to examine empathic responses generated by chatbots when the displayed empathy is understood or presented as simulated or false [16], whereas perceived empathy concerns how empathic qualities are assessed or attributed to dialogue systems and their generated responses [17,18]. More recent work has used the term “AI-generated empathy” to synthesize research on how AI-generated empathic responses are evaluated and how perceptions differ when empathy is attributed to AI rather than to human sources [19]. Although these perspectives are complementary, their terminology can move between system capacities, communicative outputs, and users' evaluations of those outputs.

This terminological heterogeneity becomes more visible when the analytical referents of these terms are compared directly. As Table 1 illustrates, empathy-related terminology is used across different objects of analysis, including system capacities, communicative

outputs, user perceptions, and source-attribution effects. The issue is therefore not that individual studies necessarily conflate these levels, but that the shared vocabulary of “empathy” can obscure differences among the phenomena being examined.

Table 1. Empathy-related terminology and primary analytical focus in selected studies.

Source	Illustrative Published Formulation	Primary Analytical Level
Asada [9]	“The design of artificial empathy is one of the most essential issues in social robotics.”	System capacity/artificial-agent design
Ma et al. [15]	Empathetic dialogue systems have the goal of enhancing the “perception and expression of emotional states, personal preference, and knowledge.”	System processing and communicative performance
Concannon & Tomalin [17]	“Measuring perceived empathy in dialogue systems.”	User perception/evaluation
Rubin et al. [18]	AI-generated empathic responses were “labelled as provided by either humans or AI.”	Generated output + source attribution + perceived empathy
Ong et al. [19]	The review examines “how people perceive empathy from AI versus from humans.”	Perceived empathy/source attribution
Suh et al. [12]	The study proposes a taxonomy focused on “user perceptions of empathy in sustained human–AI conversations.”	Perceived AI empathy in sustained interaction

To preserve these analytical distinctions, this article uses “AI-generated empathic performance” as an ontologically non-committal term for observable system processes and communicative outputs that correspond to, approximate, or simulate features associated with empathy. “Performance” refers here to the level of observable system operation and communication without presuming subjective affect, selfhood, caring, or relational reciprocity. Whether such performance is subsequently interpreted as empathic remains a distinct question at the level of perceived AI empathy.

The theoretical problem addressed in this article arises from the tendency to treat three related but analytically distinct phenomena as if they were interchangeable: human empathy, AI-generated empathic performance, and perceived AI empathy. Human empathy refers to the relational process described above. AI-generated empathic performance refers to observable system operations such as emotion recognition, contextual inference, adaptive response generation, and the generation of language that conveys support or care. Perceived AI empathy refers to users’ interpretations of observable AI outputs as signs of understanding, care, responsiveness, or relational attentiveness, as documented in existing research. These three phenomena are related, but they do not refer to the same analytical level. Conflating them risks obscuring the distinction between the theoretical properties attributed to human empathy, the observable communicative performance of AI systems, and the perceptions or attributions reported by users in existing studies.

Against this background, the article examines the conceptual boundaries between these three levels and considers their implications for debates on empathy-related human–AI interaction. The analysis is guided by three questions: (1) Which dimensions of human empathy can serve as analytical reference points for comparing AI-generated empathic performance? (2) How can human empathy, AI-generated empathic performance, and perceived AI empathy be analytically distinguished on the basis of the existing literature? (3) How can relational and ethical concerns discussed in the literature be organized in relation to these conceptual distinctions, and what conceptual and ethical implications do these distinctions have for human-centered AI? These questions are not intended to empirically determine how users perceive AI systems or whether AI possesses empathy. Rather, they guide a conceptual analysis of how different phenomena have been defined, compared, and sometimes conflated within the literature.

The article's conceptual contribution lies in establishing "AI-generated empathic performance" as an analytical boundary concept that distinguishes observable system processes and communicative outputs from both human empathic capacities and users' perceptions of empathy, without presuming subjective affect or human-like empathic experience. Research on artificial empathy, affective computing, and empathic dialogue systems informs the analysis of observable system capacities and communicative outputs associated with AI-generated empathic performance. Research on user evaluations of AI interaction informs the level of perceived AI empathy: Suh et al. [12], for example, examine the dimensions through which users perceive empathy during sustained interactions with large language models. Related source-transparency research by Shen et al. [20] shows that disclosure of AI authorship can shape users' empathic responses toward AI-generated narratives. Ethical scholarship addresses a further analytical level. Peer-reviewed work has examined anthropomorphic and relational risks in AI companions and empathic-seeming systems [21,22], fairness and demographic disparities in emotion AI [23], and questions of transparency, accountability, and human control in AI systems [24,25]. The aim is not to treat these diverse phenomena as components of a single and homogeneous form of "AI empathy," but to clarify the different analytical levels to which they refer and to examine why maintaining these distinctions matters for conceptual and ethical discussions of human–AI interaction. This distinction also helps clarify whether particular claims concern AI system outputs, users' interpretations of those outputs, or the ethical conditions under which such systems are designed and used.

These concerns also connect the present analysis to human-centered artificial intelligence (HCAI), which emphasizes the development of reliable, safe, and trustworthy systems while maintaining high levels of human control [25]. In the context of empathic AI, a human-centered approach requires attention not only to whether systems can generate supportive and context-sensitive responses but also to how users understand those responses, calibrate trust, retain autonomy, and remain protected by appropriate forms of human oversight. The distinction between AI-generated empathic performance and perceived AI empathy is therefore relevant to HCAI because it clarifies the relationship between what a system observably performs and the capacities or relational qualities users may attribute to that performance. On this basis, the framework yields five operational implications for design and evaluation: separating system performance from perceived empathy in evaluation, disclosing AI status and inferential limits, communicating uncertainty in emotion-related inference, calibrating anthropomorphic and relational cues, and providing pathways to human support in emotionally sensitive or high-stakes contexts.

Methodologically, this article employs a concept-centered theoretical synthesis rather than a systematic literature review. This approach is consistent with theory-synthesis forms of conceptual research, in which partially overlapping literature streams are integrated to develop a higher-order conceptual account [26]. The aim is to clarify conceptual boundaries across partially overlapping bodies of scholarship on human empathy, artificial empathy, human–AI interaction, human-centered AI, and AI ethics. The literature identification and updating were conducted iteratively between January and September 2026 through purposive searches in Google Scholar, Web of Science (WoS), and Scopus; targeted searches of publisher platforms and relevant journal websites; and backward citation chaining from foundational and recent studies. Search terms were organized around the principal conceptual domains of the study, including human empathy, cognitive and affective empathy, self–other differentiation, artificial and AI empathy, simulated empathy, affective computing, empathetic dialogue systems, perceived AI empathy, human–AI interaction, and human-centered AI, together with related terms concerning anthropomorphism,

trust, vulnerability, transparency, and human oversight. No formal language restriction was applied.

Sources were selected on the basis of their conceptual or empirical relevance to four analytical tasks: clarifying the dimensions of human empathy; identifying observable capacities associated with AI-generated empathic performance; examining how empathy is perceived or attributed in human–AI interaction; and identifying relational, ethical, and governance concerns associated with emotion-sensitive AI systems. No fixed publication-date limit was applied to foundational empathy scholarship, whereas greater emphasis was placed on recent studies in the AI-related literature. The synthesis proceeded by comparing definitions, theoretical claims, reported system capacities, and findings concerning user perceptions, and by organizing these materials according to their analytical level. The four dimensions of human empathy were not adopted from a single pre-existing taxonomy but were constructed as an analytical synthesis of recurring elements across the empathy literature. Likewise, the four ethical domains were developed as organizing categories from recurrent concerns identified across AI ethics and the human-centered AI literature. To strengthen analytical robustness and traceability, both authors independently reviewed the mapping of the cited literature to the four empathy dimensions and four ethical domains, including sources that qualified or challenged the emerging distinctions. The role of each literature stream and the link between individual sources and conceptual claims were also reviewed. Any discrepancies in interpretation or categorization were resolved through discussion. These procedures were used to strengthen analytical coherence and transparency rather than to claim the exhaustive coverage or reproducibility associated with a systematic review. The article draws on 47 cited sources, which provide the conceptual, empirical, methodological, and regulatory basis for the theoretical synthesis and its contextualization. These sources serve different, sometimes overlapping roles: some inform the conceptual distinctions and comparisons, whereas others provide empirical illustrations, methodological grounding, or ethical and regulatory context. The number 47 therefore refers to the sources cited in the article, not to a set of empirical studies selected through a systematic screening procedure.

ChatGPT-6 (OpenAI) was used for language editing, rephrasing, textual organization, and assistance in the development and generation of Figure 1. Trinka was used for language editing and stylistic refinement. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the content of the manuscript.

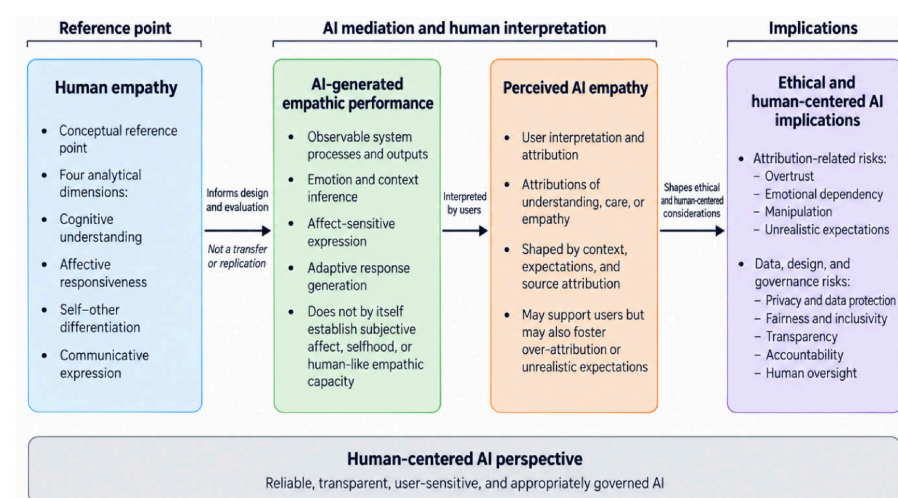


Figure 1. Conceptual framework linking human empathy, AI-generated empathic performance, perceived AI empathy, and ethical and human-centered AI implications.

Note. Human empathy is presented as a conceptual reference point rather than as a capacity transferred to or replicated by AI. AI-generated empathic performance refers to observable system processes and communicative outputs, whereas perceived AI empathy refers to users' interpretations and attributions of those outputs. Ethical and human-centered implications arise both from the relationship between system performance and user attribution and from the data, design, and governance conditions under which empathic AI systems are developed and used. The arrows indicate analytical relationships rather than equivalence or causal transfer.

2. Human Empathy: Conceptual Boundaries and Analytical Framework

Building on these distinct traditions, this analysis treats human empathy as a multidimensional construct in which cognitive and affective processes are conceptually distinguishable but interrelated. Davis [1] provides a multidimensional account of dispositional empathy, Hoffman [2] emphasizes the developmental interplay between affective arousal and increasingly differentiated forms of perspective-taking, and Cuff et al. [3] synthesize the broader definitional literature and identify cognitive and affective processes as recurrent components across conceptualizations of empathy. Relational and communicative accounts further emphasize perspective-taking, the preservation of self–other boundaries, and the communication of understanding [4,5,7]. The four dimensions used in this article—cognitive understanding, affective responsiveness, self–other differentiation, and communicative expression—are not adopted from a single pre-existing model or presented as a consensus taxonomy. Rather, they constitute an analytic construction developed by the authors from recurring elements across these foundational and integrative accounts, selected for their relevance to the conceptual comparison between human empathy and AI-generated empathic performance. Related constructs such as perspective-taking, empathic concern, personal distress, emotional contagion, identification, and sympathy are treated as mechanisms or boundary concepts that clarify the content and limits of the four dimensions rather than as additional dimensions in their own right. Broader cultural and social context is likewise treated as a condition shaping all four dimensions rather than as a separate fifth component. The resulting framework is therefore intended as an analytical organizing device for this comparison rather than as an exhaustive or universally applicable taxonomy of empathy.

The first dimension, cognitive understanding, concerns the capacity to recognize and interpret another person's perspective, thoughts, and emotional state without reducing that experience to one's own frame of reference. Perspective-taking constitutes an important mechanism in cognitive empathy because it involves adopting an other-oriented perspective while maintaining a clear distinction between one's own standpoint and that of the other person [6]. Related work on empathic accuracy similarly emphasizes the ability to accurately infer the specific content of another person's thoughts and feelings, drawing on information available in interpersonal interaction [27,28]. Hoffman's [2] distinction between self-centered role-taking and more other-oriented perspective-taking further demonstrates that cognitive understanding depends on the capacity to orient attention toward the other person's position. However, perspective-taking alone does not constitute empathy; within the relational conception adopted here, cognitive understanding must operate together with affective responsiveness, self–other differentiation, and communicative expression.

The second dimension, affective responsiveness, concerns the capacity to register and respond affectively to another person's emotional condition while remaining oriented toward that person rather than toward the reduction in one's own discomfort. The emotional dimension of empathy has been distinguished from purely cognitive understanding in multidimensional accounts of empathy [1,29]. This distinction becomes particularly clear when

empathic concern is contrasted with personal distress. Davis [1] distinguishes empathic concern and personal distress as separate dimensions of dispositional empathy, with the former reflecting other-oriented feelings of concern and the latter reflecting self-oriented discomfort in response to another's distress. Batson [30,31] further differentiates empathic concern from personal distress at the motivational level, linking empathic concern to other-oriented motivation while treating personal distress as an aversive state that may shift attention toward regulation of one's own discomfort. From a neuroscientific perspective, Lamm et al. [32] similarly emphasize that effective self–other differentiation is important for preventing empathic engagement from becoming dominated by self-related distress. More broadly, Zaki [33] argues that empathic responding is motivation-sensitive: people may be motivated either to approach or to avoid engagement with others' emotional states, and these motives can shape empathic responses through regulatory processes such as attention and appraisal. This motivational dimension further distinguishes human affective responsiveness from the generation of affect-sensitive expression: an AI system may produce language that functionally resembles an empathic response, but the observable output alone does not establish the motivational dynamics that characterize human empathic engagement. Emotional activation, therefore, is not sufficient in itself to establish an empathic response. What matters is how affective responsiveness is regulated and integrated with attention to the other person's experience.

This point leads to the third dimension, self–other differentiation. A recurring boundary condition in the empathy literature is that understanding or responding to another person does not require the erasure of the distinction between one's own experience and that of the other. This criterion helps distinguish empathy from related processes such as emotional contagion and identification. In emotional contagion, another person's emotion may be experienced without sufficient differentiation between self and other, whereas empathic responding requires recognition of the emotion as belonging to the other person and regulation of one's own affective response accordingly [34,35]. Identification similarly risks blurring self–other boundaries when the other person's emotional state is internalized as one's own [2]. Sympathy, although it may involve concern, closeness, or emotional alignment, should likewise not be treated as equivalent to empathy, since it does not necessarily require the same effort to understand the other person's subjective position while preserving self–other differentiation [1,30]. These distinctions are analytically important because they show that empathy cannot be reduced either to emotional similarity or to the intensity of emotional reaction.

The fourth dimension, communicative expression, concerns the relational completion of empathic understanding. In Rogers' [4,5] account, empathy involves not only grasping another person's internal frame of reference while preserving one's own sense of self but also communicating that understanding to the other person. Similarly, Dökmen [7] emphasizes that empathy becomes communicatively meaningful when understanding is expressed through responses that enable the other person to feel understood. His distinction among “they,” “I,” and “you” levels of empathic response further illustrates that apparently supportive communication may remain centered on social generalizations or the responder's own experience rather than on the other person's subjective position. From this perspective, communicative expression is not merely the verbal display of an empathic statement; it is the articulation of an understanding that remains responsive to the other person's perspective and to the relational context in which the interaction occurs.

These four dimensions are shaped by broader cultural, social, and representational contexts. Critical scholarship has shown that empathic relations are embedded in interpretive frameworks, social hierarchies, and cultural meanings that influence how emotions and others' experiences are understood and responded to [36,37]. Context should therefore

not be treated as an additional isolated component of empathy but as a condition that shapes cognitive understanding, affective responsiveness, self–other differentiation, and communicative expression. This contextual dimension is especially relevant when empathy is considered in technologically mediated settings, where available cues, forms of representation, and conditions of interaction differ from those of face-to-face communication.

Taken together, the literature examined in this section supports a relational conception of human empathy organized around cognitive understanding, affective responsiveness, self–other differentiation, and communicative expression. Related concepts such as perspective-taking, empathic concern, personal distress, emotional contagion, identification, and sympathy are used here not as separate components to be added to an expanding definition of empathy but as conceptual boundary markers that clarify what these four dimensions entail and where their limits lie. In the analysis that follows, these dimensions provide a descriptive and comparative framework for examining AI-generated empathic performance. Their application to AI does not presume that artificial systems possess subjective affective experience or a self capable of differentiating itself from another. Rather, the analysis considers whether observable AI outputs correspond to, approximate, or simulate communicative features associated with these dimensions. In particular, self–other differentiation is examined only at the functional and linguistic level—that is, in terms of whether AI-generated responses preserve distinctions between the user’s reported experience and system-generated inference or expression. Communicative expression and perceived AI empathy are likewise kept analytically separate: the generation of an empathic-seeming response is an observable feature of system performance, whereas whether such a response is interpreted as a sign of understanding, care, or relational attentiveness belongs to the level of perceived AI empathy as documented in existing research. The purpose is therefore not to determine whether AI systems possess human empathy but to clarify what is performed by the system, what is reported in existing user research as being perceived or attributed, and how each differs from human empathy as conceptualized in this section.

3. AI-Generated Empathic Performance: Operational Capacities and Conceptual Limits

The increasing integration of digital technologies and artificial intelligence agents into social interaction has raised questions about how communicative functions associated with human empathy can be operationalized in artificial systems. Research on artificial empathy builds on approaches developed in affective computing for recognizing, modeling, and expressing affective information [14]. Subsequent work has extended this focus toward interactive systems capable of adapting their responses to users’ emotional and contextual information [9,10]. These developments demonstrate an expanding capacity to produce observable functions associated with empathic communication. They do not, however, establish that AI systems possess empathy as conceptualized in the preceding section.

The terms algorithmic empathy, artificial empathy, and AI empathy are not always used consistently across the literature. Related recent work on “empathy algorithms” emphasizes AI-driven and multimodal approaches to detecting and interpreting affective states from linguistic, vocal, facial, and physiological signals [13], whereas research on artificial empathy more broadly examines artificial agents designed to recognize affective cues and operationalize functions associated with empathic interaction [9,10]. Because these categories overlap and involve algorithmic operations, this article does not treat them as mutually exclusive technical models. Instead, it uses the term “AI-generated empathic performance” to refer to observable system processes and outputs that correspond to, approximate, or simulate communicative features associated with empathy. The term is therefore intended to mark an analytical level rather than to introduce a claim about

subjective empathic capacity: it identifies what the system observably performs while keeping users' attribution of empathy analytically separate. Table 2 summarizes these conceptual distinctions and situates AI-generated empathic performance in relation to adjacent terminology in the literature.

Table 2. Conceptual distinctions among empathy-related terms in AI research.

Concept	Primary Analytical Focus	Main Emphasis	Position in the Present Framework
Affective computing	Computational processing of affective information	Recognition, interpretation, modeling, and generation of affect-related information	Provides computational capacities that may contribute to empathic-seeming system performance
Artificial/AI empathy	Artificial system capacities associated with empathy	Recognition of affective cues, contextual interpretation, and adaptive or supportive responses	A broad category that may encompass different dimensions of empathy-related AI interaction
Empathic communication/empathetic dialogue systems	Communicative system behavior	Generation of emotion-aware, supportive, personalized, or context-sensitive responses	Describes communicative features that can form part of AI-generated empathic performance
Simulated empathy	Production of empathic-seeming responses without assuming corresponding subjective experience	Simulation or reproduction of communicative features associated with empathy	Closely related to system performance but emphasizes simulation rather than analytical level
Perceived AI empathy	User interpretation and evaluation	Users' attribution or perception of understanding, care, responsiveness, or empathic quality	Represents the user perception level and is analytically distinct from system performance
AI-generated empathic performance (present article)	Observable system processes and communicative outputs	Emotional and contextual inference, affect-sensitive expression, adaptive response generation, and other observable features associated with empathic communication	Identifies the system-performance level without presuming subjective affect, selfhood, or human-like empathic capacity, while keeping perceived empathy analytically separate

Building on the descriptive and comparative framework developed in the preceding section, AI-generated empathic performance can be examined in relation to cognitive understanding, affective responsiveness, self–other differentiation, and communicative expression. This comparison does not assume equivalence between human capacities and computational operations. At the level of cognitive understanding, AI systems can computationally estimate emotional states, perspectives, goals, or intentions from patterns in linguistic and contextual information. Such computational estimates may correspond functionally to some observable features associated with cognitive understanding, but they do not constitute direct access to another person's subjective standpoint. At the level corresponding to affective responsiveness, the relevant AI capacity is not subjective emotional responsiveness but the generation of affect-sensitive expressions that convey

recognition, support, or concern. A system may therefore generate the linguistic form of affectively responsive communication, but this does not constitute evidence of emotional resonance or empathic concern.

The same distinction is necessary for self–other differentiation. In human empathy, this dimension refers to maintaining a distinction between one’s own experience and that of the other person. Zhao et al. [38] developed a computational model that explicitly incorporates self–other awareness into empathetic response generation. Their experimental findings indicate that this approach improves emotion recognition and empathetic response generation, outperforming comparison models in automatic and human evaluations. In the present analysis, however, such computational differentiation is not treated as evidence of an internal self capable of distinguishing its own experience from that of another. Rather, its closest functional and linguistic counterpart concerns whether a generated response distinguishes the user’s reported experience from the system’s own inference or interpretation, avoids presenting uncertain interpretations as direct knowledge of the user’s internal state, and acknowledges limits on what can be inferred from the available information. This form of differentiation is therefore treated as a property of generated communication rather than as evidence of subjective selfhood. By contrast, communicative expression is the dimension in which AI systems can most directly generate observable features associated with empathic interaction: they can produce supportive language, adapt responses to conversational context, personalize their wording, and maintain continuity across exchanges. Yet the production of an empathic-seeming response remains a feature of system performance; whether that response is evaluated as empathic belongs to a different analytical level.

This distinction is particularly important when considering the framework developed by Suh et al. [12]. The SENSE-7 framework is based on users’ evaluations of sustained interactions with large language models and identifies seven dimensions through which empathic qualities may be perceived: affective understanding, cognitive understanding, response appropriateness, prosocial expression, interest, contextual understanding, and relational continuity. These dimensions capture different aspects of how users evaluate whether an AI system appears to recognize emotions and perspectives, respond appropriately, produce supportive expressions, demonstrate attentiveness, take contextual information into account, and maintain continuity across interactions.

In this article, SENSE-7 is not adopted as an alternative conceptual model of human empathy, nor are its dimensions treated as evidence that AI systems possess corresponding human empathic capacities. Rather, the framework is used to illustrate the user-evaluation level of perceived AI empathy. Labels such as affective understanding and cognitive understanding refer to qualities attributed to an interaction by users rather than to subjective affective experience or human-like perspective-taking within the system. In this sense, SENSE-7 helps clarify how observable features of AI-generated communication can become the basis of perceived empathy without collapsing AI-generated empathic performance into human empathic capacity.

The distinction between output and perception is reinforced by Rubin et al. [18], who examined whether evaluations of empathic feedback differed depending on whether responses were attributed to a human or an AI source. Their findings indicate that evaluations of empathy are influenced not only by the linguistic form of a response but also by its attributed source. Human-attributed responses were evaluated as more empathic and supportive than AI-attributed responses, with these differences being particularly pronounced for responses emphasizing emotional sharing and care. These findings suggest that linguistic similarity alone does not produce equivalent perceptions of empathy, particularly in dimensions involving emotional care and concern. They therefore provide

an important empirical basis for distinguishing AI-generated empathic performance from perceived AI empathy.

The same distinction helps organize research on social companionship and social attribution. The Computers Are Social Actors (CASA) paradigm proposes that individuals may apply social rules and expectations to computers when those systems display recognizable social cues, even when users are aware that they are interacting with machines [39]. According to Guingrich and Graziano [40], as artificial intelligence becomes increasingly widespread, the question of how human–AI interaction may affect human–human interaction has become increasingly important. Companion chatbots are being used as social partners, yet empirical evidence concerning their effects on human relationships remains limited. A common assumption is that relationships with companion chatbots may harm social well-being by weakening or replacing human interaction, although the authors caution that this view may be overly simplistic. Their findings indicate that users and non-users may evaluate the social effects of companion chatbots differently, with users evaluating these relationships more positively in terms of social health than non-users. Within the framework of this article, such findings do not demonstrate that artificial systems form social relationships in the human sense. Rather, they show that the relational significance of AI interaction depends not only on system functionality but also on how users interpret and experience the interaction, helping to explain how AI-generated communicative cues can become objects of social and relational attribution.

Recent developments in conversational systems for psychological and emotional support further illustrate the expanding technical scope of AI-generated empathic performance. For example, Chu et al. [41] propose the Multimodal Emotional Support Conversation (MESOC) dataset and a sequential multimodal framework integrating user-emotion recognition, system-strategy prediction, system-emotion prediction, and response generation. Here, system emotion refers to a modeled component within the computational framework rather than to an experienced emotional state. Such approaches aim to move beyond text-based emotion classification toward more context-sensitive and strategically adapted supportive responses. In their experimental evaluation, Chu et al. [41] reported that the proposed SMES framework outperformed selected baseline systems in support-strategy prediction and response-generation evaluations. Their significance for this argument lies in demonstrating increasingly sophisticated forms of AI-generated empathic performance at the level of system design and output. Improvements in multimodal recognition and response generation do not, by themselves, establish subjective affective experience or equivalence with human empathy.

Taken together, the literature reviewed in this section supports a distinction between human empathy, AI-generated empathic performance, and perceived AI empathy. AI systems can computationally estimate emotional and contextual information, generate affect-sensitive and supportive expressions, linguistically distinguish user reports from system-generated interpretations, and produce personalized and context-sensitive responses. These operations may correspond to, approximate, or simulate observable features associated with the dimensions of human empathy developed in the preceding section, but they should not be interpreted as direct evidence that the underlying human capacities are present in artificial systems. At the same time, research on SENSE-7, source attribution, CASA, and social companionship demonstrates that users' evaluations of empathy depend on more than the formal characteristics of system output. Perceived empathy is shaped by interpretations of source identity, attributed intention, context, social cues, and relational meaning. Maintaining these distinctions therefore makes it possible to specify more precisely what is being described when AI systems are characterized as "empathetic" and

provides the conceptual basis for examining the ethical implications of these interactions in the following section.

Figure 1 summarizes the analytical relationship among human empathy, AI-generated empathic performance, perceived AI empathy, and the ethical and human-centered implications developed in the following section. As represented in the figure, human empathy functions as a conceptual reference point rather than as a capacity assumed to be reproduced by AI. AI-generated empathic performance occupies the system-performance level, while perceived AI empathy arises at the user interpretation level. The ethical and human-centered implications emerge from both the relationship between these two levels and the broader data, design, and governance conditions under which the interaction occurs.

4. Ethical Implications of AI-Generated Empathic Performance and Perceived AI Empathy

The ethical implications of empathy-related human–AI interaction do not arise from a single source. Rather, they emerge at different points where AI-generated empathic performance, user attribution, data processing, and system governance intersect. The distinction developed in the preceding sections is therefore also important for ethical analysis. Some concerns arise when observable AI-generated expressions of understanding, support, or care are interpreted as evidence of human-like intention or relational capacity; others concern the data, design, and institutional conditions under which such systems operate, regardless of how users perceive them. For analytical clarity, the ethical concerns discussed in the literature are organized into four interrelated domains: representational and attributional risks, relational and vulnerability risks, data and fairness risks, and transparency, accountability, and governance. These domains distinguish risks arising primarily from the interpretation of empathic performance from those arising from the data, institutional arrangements, and governance structures through which that performance is produced and deployed. They are used as an organizing framework for the present discussion rather than as an exhaustive taxonomy of AI ethics.

4.1. Representational and Attributional Risks

A first group of ethical concerns arises from the way AI-generated empathic performance is represented and interpreted. Anthropomorphism—the attribution of human characteristics to non-human entities—is particularly relevant in this context because users may interpret human-like communicative cues in terms of understanding, concern, intention, or relational attentiveness [21]. As established in the preceding section, the generation of supportive, affect-sensitive, or contextually appropriate language does not by itself demonstrate that a system experiences concern or possesses the relational capacities associated with human empathy. The ethical problem therefore lies not simply in the production of human-like language but in the possibility that the distinction between performed empathic cues and attributed empathic capacity becomes obscured.

This concern is also evident in discussions of emotional and empathic AI. Bakir and McStay [22] similarly show how anthropomorphic design and emulated empathy in AI companion systems can blur the distinction between artificial role-play and human-like relational interaction, creating risks for user autonomy, safety, and relational interpretation. In a recent preprint, Chavan et al. [42] identify related concerns regarding emotional influence and unrealistic expectations in emotional AI. From the perspective adopted here, the relevant ethical asymmetry is between what the system observably performs and what may be attributed to that performance. Expressions that convey concern, attentiveness, or emotional understanding can function persuasively within an interaction even when they do not originate from felt concern or subjective emotional experience.

Human-likeness does not, however, produce a uniform increase in perceived empathy or affinity. The literature on the uncanny valley suggests that increasing resemblance to human appearance or behavior may, at certain points, produce discomfort rather than greater affinity [43,44]. Its relevance to this argument is therefore limited but useful: human-like design should not be assumed to automatically translate into stronger perceived empathy. Rather, anthropomorphic cues influence the interpretive conditions under which users make judgments about agency, authenticity, and relational status.

4.2. Relational and Vulnerability Risks

A second set of ethical concerns emerges when empathic-seeming interaction acquires relational significance for users. Persistent availability, personalized responses, conversational continuity, and supportive language may encourage users to interpret AI systems through categories ordinarily associated with social relationships. As discussed through the CASA paradigm in the preceding section, such attribution does not require users literally to believe that a system is human. Ethical concerns can arise when the relational meaning attributed to an interaction exceeds the capacities and responsibilities that the system can actually sustain.

These concerns become particularly significant for users in vulnerable circumstances. Bakir and McStay [22], in their ethical analysis of AI companion systems and the case of Character.AI, highlight particular safety concerns for children and other vulnerable users, including risks associated with anthropomorphic design, emulated empathy, blurred relational boundaries, and dependency. Chavan et al. [42] identify related concerns regarding emotionally significant AI interactions and potentially vulnerable populations. Such cases should not be taken to demonstrate that AI interaction necessarily produces dependency or harm. Rather, they illustrate why systems that generate persistent, emotionally responsive, and relationally suggestive communication require particular caution when used by vulnerable individuals or in emotionally sensitive situations.

The ethical issue is therefore broader than whether a chatbot produces an appropriate supportive response. It also concerns whether the interaction encourages overtrust, emotional attachment, dependency, or interpretations of reciprocal care that the system cannot meaningfully reciprocate. This distinction is especially important in mental-health and emotional-support contexts, where users may disclose distress, seek reassurance, or rely on conversational continuity during periods of vulnerability. AI-generated empathic performance may be experienced as supportive, but perceived support should not be conflated with the reciprocal relational responsibility associated with human care.

4.3. Data and Fairness Risks

Ethical concerns surrounding empathic AI are not limited to relational attribution. A third domain concerns the data and computational infrastructures through which emotion-sensitive and context-sensitive responses are produced. Systems designed to estimate emotional states, preferences, vulnerabilities, or contextual conditions depend on the collection, inference, and processing of information that may be highly personal. In empathy-related AI, these concerns are distinctive because emotional and contextual inferences are not merely used for classification or prediction; rather, they may directly shape the tone, content, and relational meaning of the system's response. Errors or biases in emotion inference can therefore affect not only representational accuracy but also the kind of support, concern, or understanding that the system appears to communicate. These data practices raise privacy and data-governance concerns, particularly when the information involved includes emotional, contextual, or vulnerability-related data.

Fairness is equally important because emotion-recognition systems may not perform uniformly across users or demographic groups. Rhue [23] demonstrates that emotion AI systems can exhibit inconsistencies and disparities in emotion recognition across demographic groups, raising concerns about algorithmic fairness and the limits of relying on transparency alone to mitigate such disparities. Cross-cultural emotion-recognition research also shows that emotional expressions are not interpreted with identical accuracy across cultural groups; Elfenbein and Ambady [45], for example, found an in-group advantage in emotion recognition across cultures. What counts as an appropriate interpretation of emotional expression may therefore depend partly on cultural context, which is relevant when emotion-sensitive AI systems are trained and evaluated across heterogeneous populations.

Technical limitations in datasets and evaluation practices are relevant to this issue. A preprint by Tahir et al. [46] surveys deep learning techniques, existing datasets, and evaluation approaches in artificial-empathy research. They describe a standard workflow consisting of three stages: emotion recognition, analysis of the perceived emotion or degree of empathy, and the generation of an appropriate response. The study also highlights the challenges involved in operationalizing empathy in artificial systems and points to limitations in current datasets and evaluation practices. Importantly, these data and fairness concerns would remain even if users fully understood that they were interacting with an artificial system; they are therefore broader than the performance–attribution asymmetry alone.

4.4. Transparency, Accountability, and Governance

A fourth domain concerns the conditions under which empathic AI systems are disclosed, explained, supervised, and held accountable. Srinivasan and San Miguel González [24] further argue that empathy can help account for the differing needs of stakeholders in AI accountability. Transparency is particularly important because source attribution can affect perceived empathy. Rubin et al. [18] show that otherwise comparable AI-generated empathic responses are evaluated differently depending on whether they are attributed to AI or human sources. Complementarily, Shen et al. [20] show that disclosure of AI authorship can shape participants' empathic responses toward AI-generated narratives. Specifically, participants reported lower empathy toward AI-written stories when AI authorship was disclosed, although their self-reported willingness to empathize with such stories was higher under disclosure conditions. Together, these findings indicate that information about source identity can affect how emotionally meaningful content is interpreted, making disclosure an important consideration in the design of emotionally sensitive AI interaction.

Some of the transparency safeguards discussed here also correspond to existing regulatory obligations. Under Article 50(1) of the EU AI Act (Regulation (EU) 2024/1689) [47], providers of AI systems intended to interact directly with natural persons must ensure that users are informed that they are interacting with an AI system, unless this is obvious from the context. Article 50(3) further requires deployers of emotion-recognition systems, as defined by the Act, to inform natural persons exposed to their operation. These provisions are directly relevant where empathy-related AI systems fall within the Act's disclosure or emotion-recognition provisions.

Transparency in this context should not be limited to identifying an interaction partner as artificial. For empathy-related AI, transparency also concerns how emotional states are inferred and how empathic-seeming responses are generated, because users may interpret such outputs as evidence of understanding, concern, or relational attentiveness. The governance issue is therefore not only whether AI use is disclosed, but also whether users

are given sufficient information to distinguish system-generated inference and performance from capacities they may otherwise attribute to the system.

It also concerns communicating the limits of the system's capacities, the basis on which emotional or contextual estimates are generated, and the purposes for which personal information may be processed. A system that generates language conveying understanding or concern while leaving its artificial nature, operational limits, or data practices unclear may encourage users to form expectations that are not supported by the system's actual capacities.

Accountability and human oversight become especially important in high-stakes domains such as healthcare, psychological support, and education. Srinivasan and San Miguel González [24] conceptualize AI accountability in terms of responsibility, justification, reporting, traceability, audit, and redress, while Shneiderman's [25] human-centered AI framework emphasizes the importance of maintaining meaningful human control in the development of reliable, safe, and trustworthy AI systems. The issue is not simply whether a system can generate an empathic-seeming response but who remains responsible when such a response is inappropriate, misleading, harmful, or acted upon in a consequential setting. Ethical governance therefore requires clear limits on system use, mechanisms for human intervention, and explicit institutional responsibility rather than the attribution of autonomous ethical agency to the AI system itself.

Taken together, these concerns demonstrate that the ethics of empathic AI cannot be reduced to a single question of whether artificial systems can successfully simulate human empathy. Representational and relational risks arise particularly from the possible gap between AI-generated empathic performance and the meanings users attribute to that performance. Anthropomorphism, overtrust, emotional attachment, and manipulation become ethically relevant when observable communicative cues are interpreted as signs of capacities or forms of reciprocity that the system cannot necessarily sustain. Data, fairness, transparency, accountability, and governance concerns, however, operate at a broader level and should not be treated merely as consequences of user attribution. They concern the infrastructures, institutional responsibilities, and design choices through which empathic AI systems are developed and deployed. Ethical evaluation therefore requires attention not only to what AI systems appear to communicate but also to how those performances are produced, how they are interpreted, what data and assumptions they depend on, and who remains responsible for their consequences.

Table 3 makes these relationships more explicit by linking specific mechanisms of AI-generated empathic performance to user attribution processes, potential harms, affected stakeholders, and corresponding human-centered safeguards. The proposed links synthesize evidence and arguments discussed in the preceding sections concerning anthropomorphic attribution, relational vulnerability, emotionally sensitive data processing, source transparency, accountability, and human oversight. Accordingly, except where an existing regulatory obligation is explicitly identified, the safeguards are presented as literature-grounded conceptual design and governance implications rather than as independently validated interventions.

Table 3 shows why the ethical implications of empathic AI cannot be reduced to generic concerns about privacy, fairness, or accountability. In this context, these concerns become distinctive through their relationship to emotion inference, empathic-seeming communication, relational attribution, and the gap that may emerge between observable system performance and the capacities users attribute to it.

Table 3. Mechanisms, attribution processes, risks, and human-centered safeguards in AI-generated empathic performance.

Empathic Mechanism	User Attribution/Interpretive Process	Potential Harm	Affected Stakeholders	Human-Centered Safeguard
Recognition of emotional cues and generation of empathic-seeming responses	The response may be interpreted as evidence that the system understands, cares, or possesses empathic capacity	Unrealistic expectations, overtrust, or susceptibility to emotional influence and manipulation	Users, particularly those interacting in emotionally sensitive contexts	Capability disclosure; communication of inferential limits; calibrated trust
Persistent availability, personalized responses, conversational continuity, and supportive language	The interaction may be interpreted through categories of reciprocal social relationship or care	Emotional attachment, dependency, misplaced trust, or expectations of reciprocity that the system cannot sustain	Frequent users and users in vulnerable circumstances	Expectation management; calibrated anthropomorphic and relational cues; clear boundaries regarding system capacities
Emotion-sensitive personalization combined with perceived attentiveness or understanding	Users may disclose personal or sensitive information as part of emotionally supportive or personalized interaction	Excessive disclosure, privacy loss, or inappropriate processing of emotionally sensitive data	Users disclosing personal, emotional, or vulnerability-related information	Data minimization; transparent explanation of data use; meaningful user control over personal information
Emotion recognition and context-sensitive response generation	System interpretations may be treated as accurate or culturally neutral representations of users' emotional states	Cultural misinterpretation, misclassification, exclusion, or unequal quality of empathic interaction	Users and communities whose emotional expressions are underrepresented or differently encoded in training and evaluation data	User notification where emotion-recognition systems are deployed, consistent where applicable with Article 50(3) of the EU AI Act (Regulation (EU) 2024/1689) [47]; culturally sensitive evaluation; representative datasets; mechanisms for correcting or contesting system inference
AI-generated empathic communication with unclear source, limits, or inferential basis	Users may form expectations of understanding or capability that exceed what the system can reliably provide	Misplaced reliance, distorted expectations, or inappropriate trust in empathic-seeming outputs	Users across conversational and emotion-sensitive applications	Clear AI disclosure, consistent where applicable with Article 50(1) of the EU AI Act (Regulation (EU) 2024/1689) [47]; communication of system limits and uncertainty; transparency regarding the basis of emotional or contextual inference
Empathic-seeming responses in high-stakes settings	Supportive language may be interpreted as sufficiently reliable or authoritative assistance	Inappropriate reliance, delayed human assistance, or harm when misleading or inadequate responses are acted upon	Users in healthcare, psychological support, education, and other consequential contexts	Clear limits on system use; pathways for human intervention or escalation; explicit institutional accountability and human oversight

5. Discussion

The present analysis contributes to current debates on empathy in human–AI interaction by distinguishing three analytically different levels: human empathy, AI-generated empathic performance, and perceived AI empathy. This distinction extends existing work on artificial empathy, empathetic dialogue systems, and perceived empathy by separating questions about what a system observably performs from questions about how users interpret those performances and from the relational and experiential properties traditionally associated with human empathy. In this respect, the framework is consistent with work showing that the production of empathic-seeming language and the perception of empathy are not equivalent phenomena. Concannon and Tomalin [17], for example, treat perceived

empathy as an interactional assessment problem in dialogue systems rather than as evidence of an internal empathic state, while Rubin et al. [18] demonstrate that evaluations of otherwise comparable empathic feedback vary according to whether the response is attributed to a human or an AI source.

Recent research further supports the need to maintain this distinction. Suh et al. [12] show that perceived empathy in sustained human–AI conversations is highly context-sensitive and individualized, with judgments shaped by factors such as response appropriateness, contextual understanding, and relational continuity. Similarly, Ong et al. [19], reviewing recent empirical work, identify a recurring paradox: AI-generated responses may sometimes receive higher empathy ratings than human-generated responses at the level of textual content, while the same content may be evaluated less favorably when users believe that it originates from AI. These findings reinforce this article’s argument that observable empathic performance and perceived empathy should not be collapsed into a single construct. They also suggest that increases in the technical sophistication of empathic-seeming output do not automatically translate into equivalent perceptions of care, understanding, or relational authenticity.

The four dimensions used in the present analysis—cognitive understanding, affective responsiveness, self–other differentiation, and communicative expression—should therefore be understood as a descriptive and comparative analytic framework rather than as an empirical test for determining whether an AI system “possesses” empathy. This qualification is important because experiential criteria associated with human empathy cannot be directly transferred to artificial systems without presupposing the conclusion of the comparison. The framework instead asks which observable features of AI-generated interaction correspond functionally to aspects of these dimensions and where the correspondence breaks down. Cognitive and communicative functions may be approximated through contextual inference, response adaptation, and supportive language, whereas affective responsiveness and self–other differentiation require more careful conceptual qualification when applied to systems without demonstrated subjective affect or selfhood.

Self–other differentiation is particularly useful in this respect because it can be reformulated as a testable property of generated communication without assuming subjective selfhood. At the functional level, the relevant question is whether a system distinguishes what the user explicitly reported from what the system inferred, avoids presenting uncertain interpretations as direct knowledge of the user’s internal state, and communicates the limits of those inferences. This formulation shifts the analysis away from claims about internal machine experience and toward observable interactional properties. It also provides a potential bridge between conceptual analysis and empirical evaluation since future studies could examine the frequency, accuracy, and communicative consequences of uncertainty marking, inferential qualification, and user–system boundary maintenance in AI-generated responses. These properties could be operationalized through coding schemes that assess whether responses clearly distinguish user-reported information from system-generated inference, mark uncertainty when interpreting emotional states, and avoid presenting probabilistic interpretations as certain knowledge.

The distinction among the three analytical levels also helps clarify why different AI technologies and contexts should not be treated as interchangeable. Emotion-recognition and affective-computing systems primarily rely on the processing of affective information [14], whereas empathetic dialogue systems emphasize conversational response generation [15]. Sustained interactions with large language models introduce additional issues of contextual understanding and relational continuity [12], whereas companion chatbots and multimodal support systems involve distinct forms of relational engagement and multimodal processing [40,41]. Accordingly, the significance of empathic-seeming per-

formance depends partly on the context in which it occurs. A supportive response in a low-stakes conversational setting raises different concerns from an emotionally suggestive response in psychological support, healthcare, education, or other consequential settings. Research on companion chatbots similarly indicates that users may attribute social and relational significance to these systems in heterogeneous ways and that perceived benefits or harms depend on users' needs and interpretations rather than on a single uniform effect of chatbot use [40].

These distinctions also sharpen the ethical contribution of the framework. The ethical significance of empathic AI does not arise simply because a system recognizes emotion or generates supportive language. It becomes distinctive when emotion inference and empathic-seeming communication contribute to relational attribution, disclosure, trust, dependency, or expectations of care that may exceed the system's actual capacities. This helps differentiate this argument from more general concerns in AI ethics. Privacy, fairness, transparency, and accountability remain important across AI systems, but in empathy-related interactions, they become closely connected to the interpretation of emotional inference, the relational meaning of system responses, and the possibility that users may attribute understanding or reciprocity to generated communication.

The framework therefore has practical implications for design and evaluation. Keeping the three levels separate changes what should be measured and disclosed. Evaluations should distinguish the quality of the generated response from users' perception of empathy and from claims about underlying system capacities. Designers should communicate uncertainty when emotional states are inferred rather than reported and avoid anthropomorphic or relational cues that imply capacities the system cannot substantiate [22]. Clear pathways to human support and meaningful human oversight are especially important in emotionally sensitive or high-stakes settings [24,25]. The relevant design objective is not simply to maximize perceived empathy but to align empathic-seeming performance, user expectations, system capabilities, and appropriate forms of human oversight.

6. Limitations and Future Research

This article is a concept-centered theoretical synthesis rather than an empirical investigation of user experience or a systematic literature review. Compared with systematic review methodologies, this approach allows the integration of heterogeneous theoretical, empirical, and normative literature across partially overlapping fields. Its trade-off is lower reproducibility and greater dependence on author judgment in source selection and category construction; accordingly, it does not claim exhaustive coverage of the literature. Its conclusions should therefore be interpreted within these boundaries. The framework does not establish how users in general perceive empathic AI, whether particular populations experience such systems as supportive or harmful, or under which conditions AI-generated empathic performance produces specific relational outcomes. Claims concerning user perceptions are drawn from the empirical studies reviewed and are used to clarify conceptual distinctions rather than to generate new population-level findings. Likewise, the four dimensions of human empathy and the four ethical domains are analytical organizing devices developed for the present argument rather than exhaustive or universally applicable taxonomies.

The framework also does not empirically adjudicate whether artificial systems possess subjective affective experience, selfhood, or intrinsic motivation. These remain conceptual and philosophical questions that cannot be resolved solely through the analysis of observable outputs. What can be examined empirically are the functional correspondences identified in this article: whether systems distinguish user-reported experience from system inference, how they communicate uncertainty, which communicative features contribute

to perceived empathy, and how these processes vary across users, cultures, system types, and contexts of use. The empirical value of the framework therefore depends on whether these proposed distinctions can be meaningfully operationalized and observed across different forms of human–AI interaction. Evidence showing that system performance, user attribution, and human empathic capacities cannot be meaningfully distinguished at the analytical or empirical level would require reconsideration of the framework’s explanatory usefulness.

Future research could therefore test the framework across different forms of AI-mediated interaction. Comparative studies could examine how particular features of AI-generated empathic performance influence perceived empathy under different conditions of source disclosure, conversational continuity, anthropomorphic design, and emotional sensitivity. Cross-cultural research would be particularly valuable given that emotional interpretation and appropriate supportive communication are not culturally uniform [45]. Further work should also distinguish between low-stakes conversational applications and high-stakes settings such as psychological support, healthcare, and education, and should examine how multimodal cues such as voice, facial expression, or embodiment alter the relationship between generated empathic performance and user attribution.

7. Conclusions

This article proposes that empathy-related human–AI interaction is more precisely understood by distinguishing human empathy, AI-generated empathic performance, and perceived AI empathy. The value of this distinction lies not in resolving the binary question of whether AI “has” empathy but in clarifying which claims concern human relational capacities, which concern observable system behavior, and which concern users’ interpretations of that behavior.

Maintaining these levels separately has practical consequences for both design and evaluation. In evaluation, the quality of an AI-generated response should be assessed separately from users’ perceptions of empathy and from claims about the system’s underlying empathic capacities. In design, the distinction affects decisions concerning disclosure, communication of inferential uncertainty, calibration of anthropomorphic and relational cues, and provision of pathways to human support in emotionally sensitive or high-stakes contexts. At the same time, the framework does not replace task-specific assessments of response accuracy, usability, safety, or effectiveness; these remain necessary regardless of how empathy-related interaction is conceptualized. The practical value of the distinction therefore lies in preventing performance, perception, and human empathic capacity from being treated as interchangeable targets of design or evaluation.

Author Contributions: Conceptualization, R.Z.T. and Ş.T.L.; methodology, R.Z.T. and Ş.T.L.; investigation, R.Z.T. and Ş.T.L.; writing—original draft preparation, R.Z.T. and Ş.T.L.; writing—review and editing, R.Z.T. and Ş.T.L.; visualization, R.Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation and revision of this manuscript, the authors used ChatGPT (OpenAI) for language editing, rephrasing, textual organization, and assistance in the development and generation of Figure 1. Trinko was used for language editing and stylistic refinement.

All AI-assisted outputs were reviewed by the authors, who take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Davis, M.H. Measuring individual differences in empathy: Evidence for a multidimensional approach. *J. Pers. Soc. Psychol.* **1983**, *44*, 113–126. [\[CrossRef\]](#)
2. Hoffman, M.L. *Empathy and Moral Development: Implications for Caring and Justice*; Cambridge University Press: Cambridge, UK, 2000.
3. Cuff, B.M.P.; Brown, S.J.; Taylor, L.; Howat, D.J. Empathy: A review of the concept. *Emot. Rev.* **2016**, *8*, 144–153. [\[CrossRef\]](#)
4. Rogers, C.R. The necessary and sufficient conditions of therapeutic personality change. *J. Consult. Psychol.* **1957**, *21*, 95–103. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Rogers, C.R. *A Way of Being*; Houghton Mifflin: Boston, MA, USA, 1980.
6. Coplan, A. Understanding empathy: Its features and effects. In *Empathy: Philosophical and Psychological Perspectives*; Coplan, A., Goldie, P., Eds.; Oxford University Press: Oxford, UK, 2011; pp. 2–18. [\[CrossRef\]](#)
7. Dökmen, Ü. *Communication Conflicts and Empathy in Art and Everyday Life*; Remzi Kitabevi: Istanbul, Türkiye, 2019.
8. Ersoy, E.; Köşger, F. Empati: Tanımı ve önemi/Empathy: Definition and its importance. *Osman. Tıp Derg.* **2016**, *38*, 9–17. [\[CrossRef\]](#)
9. Asada, M. Towards artificial empathy: How can artificial empathy follow the developmental pathway of natural empathy? *Int. J. Soc. Robot.* **2015**, *7*, 19–33. [\[CrossRef\]](#)
10. Paiva, A.; Leite, I.; Boukricha, H.; Wachsmuth, I. Empathy in virtual agents and robots: A survey. *ACM Trans. Interact. Intell. Syst.* **2017**, *7*, 1–40. [\[CrossRef\]](#)
11. McStay, A. *Emotional AI: The Rise of Empathic Media*; SAGE: London, UK, 2018.
12. Suh, J.; Le, L.; Shayegani, E.; Ramos, G.; Amores, J.; Ong, D.C.; Czerwinski, M.; Hernandez, J. SENSE-7: Taxonomy and dataset for measuring user perceptions of empathy in sustained human–AI conversations. *IEEE Trans. Affect. Comput.* **2026**, *17*, 2227–2244. [\[CrossRef\]](#)
13. Mallia, B. The faint whispers of the heart—Next-generation sentiment analysis and “empathy algorithms”. In *Emotions in Code: The AI Frontier of Sentiment Analysis*; Li, J., Ed.; IntechOpen: London, UK, 2025. [\[CrossRef\]](#)
14. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
15. Ma, Y.; Nguyen, K.L.; Xing, F.Z.; Cambria, E. A survey on empathetic dialogue systems. *Inf. Fusion* **2020**, *64*, 50–70. [\[CrossRef\]](#)
16. Ye, Z.; Long, F.; Luo, L. Why consumers prefer chatbots’ simulated empathy: Revisiting the empathy-honesty trade-off. *J. Consum. Behav.* **2026**, *25*, 1603–1622. [\[CrossRef\]](#)
17. Concannon, S.; Tomalin, M. Measuring perceived empathy in dialogue systems. *AI Soc.* **2024**, *39*, 2233–2247. [\[CrossRef\]](#)
18. Rubin, M.; Li, J.Z.; Zimmerman, F.; Ong, D.C.; Goldenberg, A.; Perry, A. Comparing the value of perceived human versus AI-generated empathy. *Nat. Hum. Behav.* **2025**, *9*, 2345–2359. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Ong, D.C.; Goldenberg, A.; Inzlicht, M.; Perry, A. AI-generated empathy: Opportunities, limits, and future directions. *Curr. Dir. Psychol. Sci.* **2026**. *Epub ahead of printing.* [\[CrossRef\]](#)
20. Shen, J.; DiPaola, D.; Ali, S.; Sap, M.; Park, H.W.; Breazeal, C. Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: A comparative study. *JMIR Ment. Health* **2024**, *11*, e62679. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Salles, A.; Evers, K.; Farisco, M. Anthropomorphism in AI. *AJOB Neurosci.* **2020**, *11*, 88–95. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Bakir, V.; McStay, A. Move fast and break people? Ethics, companion apps, and the case of Character.ai. *AI Soc.* **2025**, *40*, 6365–6377. [\[CrossRef\]](#)
23. Rhue, L. The anchoring effect, algorithmic fairness, and the limits of information transparency for emotion artificial intelligence. *Inf. Syst. Res.* **2023**, *35*, 1479–1496. [\[CrossRef\]](#)
24. Srinivasan, R.; San Miguel González, B. The role of empathy for artificial intelligence accountability. *J. Responsible Technol.* **2022**, *9*, 100021. [\[CrossRef\]](#)
25. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact.* **2020**, *36*, 495–504. [\[CrossRef\]](#)
26. Jaakkola, E. Designing conceptual articles: Four approaches. *AMS Rev.* **2020**, *10*, 18–26. [\[CrossRef\]](#)
27. Ickes, W. Empathic accuracy: Judging thoughts and feelings. In *The Social Psychology of Perceiving Others Accurately*; Hall, J.A., Schmid Mast, M., West, T.V., Eds.; Cambridge University Press: Cambridge, UK, 2016; pp. 52–70. [\[CrossRef\]](#)
28. Hall, J.A.; Schmid Mast, M. Sources of accuracy in the empathic accuracy paradigm. *Emotion* **2007**, *7*, 438–446. [\[CrossRef\]](#) [\[PubMed\]](#)

29. Morse, J.M.; Anderson, G.; Bottorff, J.L.; Yonge, O.; O'Brien, B.; Solberg, S.M.; McIlveen, K.H. Exploring empathy: A conceptual fit for nursing practice? *Image J. Nurs. Sch.* **1992**, *24*, 273–280. [[CrossRef](#)] [[PubMed](#)]
30. Batson, C.D. *Altruism in Humans*; Oxford University Press: New York, NY, USA, 2011.
31. Batson, C.D. *Empathic Concern: What It Is and Why It's Important*; Oxford University Press: New York, NY, USA, 2023. [[CrossRef](#)]
32. Lamm, C.; Rütgen, M.; Wagner, I.C. Imaging empathy and prosocial emotions. *Neurosci. Lett.* **2019**, *693*, 49–53. [[CrossRef](#)] [[PubMed](#)]
33. Zaki, J. Empathy: A motivated account. *Psychol. Bull.* **2014**, *140*, 1608–1647. [[CrossRef](#)] [[PubMed](#)]
34. Decety, J.; Jackson, P.L. A social neuroscience perspective on empathy. *Curr. Dir. Psychol. Sci.* **2006**, *15*, 54–58. [[CrossRef](#)]
35. Hatfield, E.; Rapson, R.L.; Le, Y.-C.L. Emotional contagion and empathy. In *The Social Neuroscience of Empathy*; Decety, J., Ickes, W., Eds.; MIT Press: Cambridge, MA, USA, 2009; pp. 19–30.
36. Pedwell, C. *Affective Relations: The Transnational Politics of Empathy*; Palgrave Macmillan: London, UK, 2014.
37. Ahmed, S. *The Cultural Politics of Emotion*, 2nd ed.; Edinburgh University Press: Edinburgh, UK, 2014.
38. Zhao, W.; Zhao, Y.; Lu, X.; Qin, B. Don't lose yourself! Empathetic response generation via explicit self-other awareness. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, ON, Canada, 9–14 July 2023; pp. 13331–13344. [[CrossRef](#)]
39. Nass, C.; Steuer, J.; Tauber, E.R. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence—CHI '94*, Boston, MA, USA, 24–28 April 1994; ACM Press: Boston, MA, USA, 1994; pp. 72–78. [[CrossRef](#)]
40. Guingrich, R.E.; Graziano, M.S.A. Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. In *Oxford Intersections: AI in Society*; Hacker, P., Ed.; Oxford University Press: Oxford, UK, 2025. [[CrossRef](#)]
41. Chu, Y.; Liao, L.; Zhou, Z.; Ngo, C.-W.; Hong, R. Towards multimodal emotional support conversation systems. *IEEE Trans. Multimed.* **2025**, *27*, 8276–8287. [[CrossRef](#)]
42. Chavan, V.; Cenaj, A.; Shen, S.; Bar, A.; Binwani, S.; Del Becaro, T.; Funk, M.; Greschner, L.; Hung, R.; Klein, S.; et al. Feeling machines: Ethics, culture, and the rise of emotional AI. *arXiv* **2025**. [[CrossRef](#)]
43. Mori, M.; MacDorman, K.F.; Kageki, N. The uncanny valley [From the field]. *IEEE Robot. Autom. Mag.* **2012**, *19*, 98–100. [[CrossRef](#)]
44. Łukasik, A.; Gut, A. From robots to chatbots: Unveiling the dynamics of human–AI interaction. *Front. Psychol.* **2025**, *16*, 1569277. [[CrossRef](#)] [[PubMed](#)]
45. Elfenbein, H.A.; Ambady, N. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychol. Bull.* **2002**, *128*, 203–235. [[CrossRef](#)] [[PubMed](#)]
46. Tahir, S.; Shah, S.A.; Abu-Khalaf, J. Artificial empathy classification: A survey of deep learning techniques, datasets, and evaluation scales. *arXiv* **2023**. [[CrossRef](#)]
47. European Parliament; Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). 2024. Available online: <http://data.europa.eu/eli/reg/2024/1689/oj> (accessed on 30 September 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.