Comparison of instruments for assessing perceived empathy in Human-Agent Interactions

Alberto Hitos-García[1,*,†], André Luiz Satoshi Kawamoto[2,†], Francisco Luis Gutiérrrez-Vela[1,†] and Patricia Paderewski-Rodríguez[1,†]

[1] *Universidad de Granada, Dpto. de Lenguajes y Sistemas Informáticos - ETSIIT, Granada, España*
[2] *Universidade Tecnológica Federal do Paraná (UTFPR), Depto. Acadêmico de Computação (DACOM), Campo Mourão, PR, Brasil*

Abstract

Perceived empathy in interactions with computational agents is a significant and emerging topic in Human-Computer Interaction (HCI). This study offers a comparative analysis of existing instruments for measuring perceived empathy in computational agents. By evaluating their theoretical foundations, adaptability, and validation processes, this paper identifies key strengths and limitations of these tools. Through extensive bibliographic research, we thoroughly reviewed and compared various instruments based on their intended applications and effectiveness. Furthermore, the study highlights the challenges of adapting traditional empathy metrics to artificial systems and explores opportunities for developing innovative, context-specific instruments. The findings indicate that while certain instruments show promise, significant gaps remain, especially in addressing multimodal and dynamic interactions. These findings support the advancement of empathy measurement approaches, fostering the design of more effective, trustworthy, and engaging Human-Agent Interactions.

Keywords

perceived empathy, Human-Computer Interaction, assessment instruments, comparative analysis

1. Introduction

In recent years, the presence of computational agents in daily life has significantly increased. AI-based agents are becoming commonplace in various sectors: in healthcare, they interact with patients, offering emotional support and aiding medical decisions; in education, they serve as teaching assistants and monitor students' progress; and in everyday tasks, agents like Alexa have become ubiquitous [1].

Although these agents are designed to assist and interact with users, the way users perceive these interactions plays an important role in user satisfaction, trust, and effectiveness [2, 3, 4]. Empathy – understanding and sharing another person's emotions – is vital for successful social interactions. For computational agents, it is essential for users to perceive empathy in order to build trust, promote engagement, and ensure positive experiences [5, 6].

Despite its recognized importance, research on empathy in interactive systems still needs standardized and validated measurement instruments that capture the complexity of human empathy and adapt to the particularities of HCI [7, 8]. This lack of reliable instruments makes it difficult to compare different systems and develop new ones incorporating this attribute into their features.

Another challenge is that computational agents do not possess the same emotional complexity as humans. Therefore, measurement instruments must be adjusted to consider the unique characteristics of these interactions [7, 8, 9].

This paper aims to offer a detailed survey and comparison among existing instruments for measuring perceived empathy in computational agents. It highlights the challenges of adapting traditional empathy-

Interacción '25: XXV International Conference on Human-Computer Interaction, September 03–05, 2025, Valladolid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

albertohitos@ugr.es (A. Hitos-García); kawamoto@utfpr.edu.br (A. L. S. Kawamoto); fgutierr@ugr.es (F. L. Gutiérrez-Vela); patricia@ugr.es (P. Paderewski-Rodríguez)
0009-0009-3136-5022 (A. Hitos-García); 0000-0002-0176-4366 (A. L. S. Kawamoto); 0000-0001-6629-7597 (F. L. Gutiérrez-Vela); 0000-0001-6626-9633 (P. Paderewski-Rodríguez)

measuring tools to non-human interactions and emphasizes how to address current gaps and explore future opportunities for expanding this research area.

2. Theoretical background

2.1. Empathy in Human-Agent Interactions

Empathy can be defined as the ability of an individual to connect with and respond to the emotions of others. This conception is widely accepted across various fields of study, such as psychology, sociology, and neuroscience, and serves as a foundation for understanding this term in different contexts[10, 11, 12].

However, other definitions and interpretations of empathy exist, especially when considering interactions between humans and computational agents. Adapting this concept to the specific context of Human-Agent Interactions is necessary. This necessity arises because computational agents, being artificial systems, cannot experience emotions as humans do [2, 10, 11, 13].

Human empathy is a complex process that involves both affective components (the ability to feel another person's emotions) and cognitive components (the ability to understand another person's perspective). It is the capacity to put oneself in someone else's shoes, authentically sharing and understanding their emotions [11].

In contrast, empathy in Human-Agent Interactions generally focuses on the user's perception that the agent understands and responds appropriately to their emotions without necessarily involving a genuine emotional experience from the agent. The agent can convincingly demonstrate understanding and emotional responsiveness. In this context, empathy is a simulation designed to enhance the user experience and make the interaction more natural and satisfying [2, 10, 11, 13, 14].

Computational agents can emulate empathy through different communication modalities, like choosing words and phrases that convey understanding and support (e.g., "I understand how you feel"), adjusting intonation and speech rhythm to transmit emotions such as sadness, joy, or concern, or pausing for an appropriate amount of time before responding to give a sense of attentive listening and consideration [11].

This empathy simulation raises questions about the authenticity of the agent's responses and the risk of creating unrealistic user expectations. Suppose an agent expresses empathy through language, tone of voice, or facial expressions. In that case, users might interpret these expressions as evidence of genuine feelings, leading to a distorted understanding of the agent's capabilities [7, 10].

From a practical standpoint, developers face the challenge of balancing realistic, empathetic responses with the need to manage user expectations. Agents must recognize and respond appropriately to users' emotions while avoiding the creation of the illusion that the agent possesses deep emotional understanding [2, 7, 15].

To mitigate these challenges, users should be aware that they are interacting with an artificial system that simulates empathy and not a human capable of experiencing genuine emotions. Educating users about the limitations of computational empathy can help manage expectations and prevent misunderstandings [2, 10]. Using simulated empathy by computational agents can lead to unrealistic expectations from users, especially in sensitive areas like healthcare and psychological support. Transparent communication with users regarding the capabilities and limitations of these technologies is essential to prevent misunderstandings and foster trust.

Another important point is that empathy evaluation metrics should be adapted to the context of Human-Agent Interactions, considering the absence of genuine emotional experience [3, 15].

Developing more robust and specific measurement tools for different agents and interactions is essential for advancing the field. However, standardized and validated instruments for measuring empathy in Human-Agent Interactions still need to be improved [7, 12].

2.2. Approaches for measuring empathy in Human-Computer Interactions

There is a variety of tools for measuring empathy in human interactions, such as the Hogan Empathy Scale (HES) [16], the Questionnaire Measure of Emotional Empathy (QMEE) [17], the Interpersonal Reactivity Index (IRI) [18], the Consultation and Relational Empathy Measure (CARE) [19], the Therapist Empathy Scale (TES) [20], and the Jefferson Scale of Physician Empathy (JSPE) [21]. Most of these tools are questionnaires based on statements that allow participants or independent observers to evaluate a conversation-based interaction subjectively. Although instruments such as the Interpersonal Reactivity Index (IRI) are widely used in human interactions, their direct application to artificial agents is limited.

Additionally, despite well-established human empathy metrics from psychology, psychiatry, and neuroscience, directly translating these metrics to the domain of AI is not trivial. It requires careful adaptations considering the fundamental differences between humans and computational agents [3, 7].

Given the complexity of measuring empathy in interactions between humans and computational agents, we performed a search for specific instruments tailored for this purpose, either by adapting existing metrics or proposing new methods. The section 5 presents the criteria we established to compare these instruments and the results of the search we performed, analyzing and comparing the various approaches found for evaluating empathy in computational agents.

3. Thematic search for empathy assessment instruments

A thematic bibliographic search was performed to identify instruments to evaluate perceived empathy in Human-Agent Interactions. Thematic searches allow for the exploration of recurring themes and research patterns, ensuring the inclusion of diverse perspectives and methodologies relevant to the field [22].

Themes such as perceived empathy, assessment tools for Human-Agent Interactions, and adaptations of traditional empathy metrics formed the backbone of the search methodology. These themes were essential for refining the review's focus and structuring the evaluation of retrieved studies. Academic databases, including IEEE Xplore, ACM Digital Library, PubMed, and SpringerLink, were the primary sources for this search. Including arXiv further ensured that emerging preprints, which capture innovative approaches and recent trends, were considered.

Keywords were developed in line with the identified themes and combined using Boolean operators to balance breadth and specificity in the search results. Key search terms included empathy measurement in AI systems, perceived empathy assessment tools, and adapted empathy scales for computational agents. The thematic approach prioritized inclusivity, allowing the study to examine tools from diverse contexts and modalities. Articles and preprints were included based on their relevance, publication within the last decade, and their focus on instruments tailored for Human-Agent Interaction. Conversely, studies strictly limited to human-to-human empathy or lacking sufficient methodological detail were excluded.

The instruments were categorized based on their theoretical foundations, adaptability, dimensional coverage, evaluation perspective, and application scope. This iterative process allowed emerging trends to refine the thematic coding further, ensuring a comprehensive review of available tools. By incorporating both peer-reviewed studies and preprints, the thematic search provided an inclusive and structured foundation for the comparative analysis presented in the subsequent sections of this study. The methodology highlighted empathy assessment instruments' theoretical richness and diversity while identifying gaps that warrant future exploration.

Building upon the thematic search conducted, this study identified a diverse set of instruments tailored for evaluating perceived empathy in Human-Agent Interactions. Each instrument was analyzed in detail, with attention given to its theoretical underpinnings, structure, and applicability to various contexts. The following section provides a comprehensive description of these instruments, highlighting their unique features, dimensions, and potential contributions to the assessment of empathy in computational systems.

4. Tools for measuring empathy

4.1. QEAE-based model to assess empathy in Human-Robot Interactions

Tisseron et al. aimed to explore and validate a model of empathy applied to interactions between humans and robots, utilizing a Questionnaire of Empathy and Auto-empathy (QEAE) [23]. The main proposal is to understand how empathy is perceived in interactions with robots and avatars, contributing to the acceptance and effectiveness of assistive robots, particularly in treatment and rehabilitation contexts.

According to the authors, the QEAE measures the different dimensions of empathy during interactions and can be used in two experimental contexts: a psychological study with a nonclinical population and an interdisciplinary research initiative focused on the interaction between older adults and humanoid robots. The expected outcomes of this project included a deeper understanding of the empathetic relationships humans can develop with robots, which may have significant implications for the design and implementation of assistive robots. The research sought to validate QEAE and identify which aspects of interactions and robot characteristics influence human empathetic experience.

The QEAE was developed as an adaptable tool to measure empathy in interactions with robots and avatars, indicating a metric adaptation for different interaction contexts. By addressing both robots and avatars, this model provides insights into the nuances of empathy in different types of artificial entities, making it versatile and comprehensive. This adaptability is crucial, as it allows the instrument to be applicable in various scenarios involving human-robot or human-avatar interactions.

The QEAE comprises four dimensions: self-empathy, direct empathy, reciprocal empathy, and intersubjective empathy. Each dimension contains four components – action, emotion, cognition, and assistance – totaling 16 items. This detailed structure ensures capturing multiple facets of empathy, providing a robust framework for evaluation.

The QEAE is designed to grasp the user's perspective, thus allowing a direct assessment of the emotional and cognitive processes involved in these interactions, enhancing the relevance of the findings.

The authors do not explicitly specify the scale adopted by the instrument, but note that the QEAE is still undergoing validation, with experimental studies planned to confirm its effectiveness and relevance in human-robot interactions. This ongoing validation reinforces the instrument's scientific credibility and ensures its reliability across diverse contexts.

The QEAE goes beyond language, encompassing a wider range of empathic components such as action, emotion, and cognition. This comprehensive approach allows the instrument to capture diverse aspects of empathy beyond verbal communication alone.

The instrument applies to multimodal interactions, including text, voice, and other nonverbal cues. This multimodal applicability reflects how humans interact with robots and avatars, comprehensively evaluating empathy in these contexts.

The authors do not explicitly specify the scale adopted by the instrument, but note that the QEAE is still undergoing validation, acknowledging some limitations. This ongoing process may limit the generalization of the results until validation is complete. Furthermore, interactions with robots can present unique challenges that the model may need to fully address, highlighting areas for future refinement.

4.2. RoPE (Robot's Perceived Empathy)

Charrier et al. developed the Robot's Perceived Empathy (RoPE) Scale to measure human perceptions of empathy in robot interactions, addressing a gap in robotic empathy research [12]. A key advantage of the RoPE Scale is its standardization, which avoids directly applying human empathy metrics by acknowledging differences in emotional capabilities between humans and robots.

The scale adapts human empathy metrics to the context of human-robot interaction, avoiding formulations that rely on abilities robots lack.

The RoPE comprises two dimensions: Empathic Understanding and Empathic Response, with 18 items distributed in these dimensions and four filler items. This structure enables a comprehensive evaluation of empathy, capturing both cognitive and emotional components through a Likert-type questionnaire format. This format is ideal for eliciting nuanced responses and provides a standardized approach to measuring empathy across studies and contexts.

Experts in cognitive science and robotics have conducted a preliminary validation. This ongoing validation process is important for establishing the scientific credibility of the scale and ensuring its effectiveness in various applications. The authors state that there are plans to test the reliability and validity of the French version of the scale in a future experiment, utilizing a Cozmo robot for filmed interactions that vary between empathetic and neutral reactions.

In a certain way, the need for further validation to ensure the scale's reliability and sensitivity represents a limitation. Additionally, the application of the scale may be influenced by biases related to the emotional capacities of robots, which must be considered when interpreting the results.

Although the original RoPE scale article does not provide detailed evaluation criteria, Daher et al. utilized RoPE to measure perceived empathy in interactions with a medical assistance chatbot. Their results demonstrated the scale's effectiveness in capturing differences in empathy perception based on variations in the chatbot's responses.

The study concluded that RoPE is a valid and valuable tool for assessing human-chatbot interaction, particularly in healthcare scenarios. It highlighted the scale's sensitivity to adjustments in chatbot responses regarding perceived empathy. However, the authors recommended future studies in diverse contexts with larger samples to validate these findings and address potential limitations and generalizability [24].

4.3. Empathy assessment model in autonomous systems

The empathy assessment model proposed by Urakami et al. provides a comprehensive framework for understanding how users perceive empathic expressions in interactions with autonomous systems. The model comprises eight dimensions that capture key aspects of empathy, enabling detailed analysis of user responses: **Expressing Own Feelings**: the system's ability to communicate its emotions; **Expressing Knowing What the Other Feels**: recognizing and articulating the user's emotions; **Helping**: the system's willingness to offer assistance; **Showing Interest**: demonstrating attention and concern for the user's needs; **Taking the Perspective of the Other**: understanding the user's situation from their viewpoint; **Showing Consideration**: acknowledging and respecting the user's emotions; **Situational Understanding**: cognitive empathy, assessing the emotional situation of the user; and **Agreement**: aligning with the user's emotions or opinions.

Empathy is assessed using a survey with 72 items. Although the authors do not specify the number of items per dimension, the comprehensive structure allows for a detailed evaluation, capturing key aspects of Human-Computer Interaction.

This tool adapts existing definitions and instruments for measuring empathy, focusing on elements relevant to interactions with autonomous systems, ensuring its suitability for evaluating empathy in autonomous dialogues.

The evaluation is conducted from the user's perspective. In a study by the authors, participants assessed empathetic expressions of an autonomous system based on their experiences, ensuring the findings reflect real user interactions.

The model's validation involved comparing evaluations from four experimenters, resulting in strong agreement. The results showed that cognitive empathy and assistance dimensions, such as ”Helping” and ”Showing Interest,” were rated more positively than affective dimensions like ”Expressing Own Feelings.” This suggests that users value understanding and practical support over emotional expressions in interactions with autonomous systems.

The questionnaire uses a Likert-type scale, where participants rate statements from 0 to 10, indicating their degree of agreement. This format captures response intensity and provides a standardized method for measuring empathy.

The instrument primarily focuses on language, specifically text, as the study was conducted through an online survey evaluating written statements. There is no mention of non-verbal cues, suggesting a potential area for expansion to include multimodal communication signals, with the instrument being designed mainly for textual dialogues.

The internal consistency was confirmed using Cronbach's alpha, yielding satisfactory values, indicating reliable and consistent dimensions. Additionally, ANOVA was used to compare evaluations across dimensions, revealing significant differences in participants' perceptions.

This model enhances understanding of human-autonomous system interactions and provides a valuable framework for empathy research in AI. The identified dimensions and assessment methodology can guide the development of more effective and empathetic systems, potentially improving user acceptance and effectiveness in real-world applications.

The authors note limitations, such as participants imagining hypothetical scenarios, which may not reflect real interactions, and the study's lack of exploration of individual differences in responses to empathetic expressions.

4.4. Social Service Robot Interaction Trust (SSRIT)

The Social Service Robot Interaction Trust (SSRIT) measures consumers' trust in interactions with artificial intelligence social robots in service delivery contexts [26]. Although developed with a focus on trust, the scale presents dimensions directly related to the perception of empathy. Trust is often recognized as a facilitating condition for empathy, allowing users to engage more deeply in interactions and attribute positive intentions to the social agent. Thus, the SSRIT indirectly contributes to understanding how empathy is perceived in Human-Robot Interaction scenarios.

The multidimensional structure of the SSRIT captures key characteristics for analyzing empathy in computational interactions. The ”propensity to trust” dimension includes factors like familiarity and self-efficacy, which influence users' willingness to interpret agent behaviors as empathetic. For instance, individuals with greater technological familiarity tend to lower cognitive barriers, enhancing their reception of empathetic behaviors.

The ”trustworthy robot function and design” dimension includes attributes such as anthropomorphism and performance, crucial for demonstrating empathy, as human-like features and consistent functionality are key to evaluating simulated emotional responses. Finally, the ”trustworthy service task and context” dimension considers situational factors like perceived risk and the fit between the robot and the service, which influence users' emotional and empathetic receptiveness.

Furthermore, the methodological robustness of the SSRIT, combining formative and reflective indicators, offers a comprehensive perspective that may inform future adaptations or expansions of scales focused on perceived empathy. In this sense, integrating SSRIT into the scope of this study enables a deeper theoretical and operational understanding of the relationship between trust and empathy in human-robot interactions.

4.5. Using existing instruments to measure human empathy

The TEQ's utilization in its original form, developed for human interactions, underscores a significant limitation in its application to virtual environments [27].

Kroes et al. explored empathy towards virtual agents, focusing on the influence of personification stories and individual user characteristics. They used the Toronto Empathy Questionnaire (TEQ), a Likert scale measuring empathy as a multidimensional construct, emphasizing affective and emotional components. Participants also completed a post-experiment survey evaluating their emotional reactions to the virtual agent based on the TEQ components.

The TEQ is an unadapted metric originally designed for human interactions and has not been specifically modified to assess empathy towards virtual agents. Its use in its original form highlights a significant limitation for application in virtual environments.

The instrument includes six dimensions: Emotional Contagion (EmCon), Emotional Understanding (EmUnd), Sensitivity (Sens), Sympathetic Physiological Arousal (SympPhy), Altruism (AltEmp), and Higher Order Empathic Behavior. Although initially conceptualised for human-to-human interactions, these dimensions provide a comprehensive framework for assessing various facets of empathy. The instrument uses a 5-point Likert scale to capture the intensity of participants' feelings and attitudes towards the virtual agent.

The study evaluated an immersive virtual reality (VR) environment where participants observed a virtual agent displaying emotions. This immersive setting is crucial for understanding empathy in a more lifelike environment, providing insights into Human-Computer Interaction dynamics.

The interaction occurred in a multimodal environment, with participants observing the agent express emotions in VR. Despite the lack of direct interaction (e.g., dialogue or voice), the setup allowed for examining empathy in virtual settings.

The study concluded that empathy towards virtual agents is influenced by individual characteristics, but personification alone may not increase empathy. The authors emphasize the need for considering individual differences in developing effective Human-Agent Interactions, suggesting more research is required to adapt empathy measures like the TEQ to computational agent contexts.

The TEQ was used from the user's perspective, capturing subjective perceptions of empathy based on their experiences and feelings towards the agent. This user-centric approach is vital for assessing the effectiveness of virtual agents in evoking empathic responses.

The TEQ incorporates emotional and behavioral components, measuring empathy in a broader context. However, it may not fully account for non-verbal communication cues, which are key in face-to-face interactions.

While the TEQ is validated in human interactions, its applicability to virtual agents remains unaddressed, presenting a notable gap. Limitations include its original design for human interactions, small sample size, and potential biases from participants focusing on the agent's appearance rather than its emotions.

4.6. Multidimensional framework for empathy in interactive dialogues

Xu and Jiang proposed an empathy assessment framework designed to capture the complexity of empathy in interactive dialogues, emphasizing its collaborative nature between the speaker and the listener. The framework consists of two main dimensions: Expressed Empathy and Perceived Empathy [28].

Expressed Empathy refers to the speaker's communicative intentions, assessed by identifying speech acts that convey empathy. Perceived Empathy adapts recent psychological definitions to task-oriented dialogues, evaluating empathy from the listener's perspective. This dimension includes four aspects: **Engagement**, which measures the listener's perception of the speaker's involvement; **Understanding**, which assesses how well the listener feels their emotions and situation are understood; **Sympathy**, which gauges the listener's view of the speaker's empathy and appropriate response; and **Usefulness**, which evaluates whether the speaker's communication addresses the core issues of the conversation.

These interconnected aspects form a comprehensive framework that reflects the multifaceted nature of empathy in social interactions. Applying this framework to customer service dialogues revealed significant correlations between perceived empathy and dialogue satisfaction. This approach improves the understanding of empathy in communication and offers a robust method for automatically evaluating empathy, contributing to the development of more effective and satisfying dialogue systems.

The instrument adapts existing metrics to measure empathy in dialogues, considering both expressed intentions and perceived empathy. It is well-suited for analyzing customer service interactions, making it highly relevant in this context.

The framework was validated using an internal dataset of 2,000 annotated customer service dialogues. Human evaluators identified 16 expressed communicative intentions and four aspects of perceived empathy, rating them on a Likert-5 scale. The analysis showed that expressed and perceived empathy are interconnected, with perceived empathy directly influencing conversation satisfaction.

The study also found that classifiers based on instruction-finetuned language models outperformed prompting methods and other competitive approaches. These results demonstrate the framework's effectiveness in measuring empathy in dialogues, supporting its application in real-world communication contexts.

Perceived empathy was evaluated from the listener's perspective, while expressed intentions were assessed from the speaker's perspective.

While the instrument mainly focuses on linguistic features, it also takes into account the social dynamics between dialogue participants, recognizing empathy as a collaborative process. The evaluation is based on text, specifically in customer service dialogues. However, the approach is adaptable to other modalities, offering a flexible framework for assessing empathy across diverse communication contexts. This broader view ensures the inclusion of multiple dimensions of empathy, extending beyond verbal communication.

As limitations, the authors highlight the challenges in measuring empathy due to its implicit nature, expression styles that may be domain-specific, and the subtle distinction between expressed and perceived empathy. Additionally, the model's effectiveness may vary depending on the interaction context, indicating areas that require further refinement.

4.7. Empathic and emotional design heuristics in healthcare technologies

The empathic and emotional design heuristics developed by Borycki et al. offer a set of guidelines aimed at improving user-system interactions, particularly within healthcare technologies. These heuristics were derived from a comprehensive literature review that identified best practices and principles conducive to creating more positive and emotionally fulfilling user experiences. The development process involved qualitative analysis of collected data, followed by review and validation by a panel of experts in health information and human factors, ensuring the heuristics are both evidence-based and contextually relevant [29].

The heuristics are organized into categories that address critical elements of empathic design, including Personalization, Messaging, Engagement, and Usability. For example, personalization is emphasized as a key factor in enhancing users' perception of empathy by tailoring the presentation of information to individual needs. Furthermore, the use of simple communication, along with images and visualizations, is recommended to boost user engagement and satisfaction. These guidelines inform interface design and serve as assessment tools for measuring the effectiveness of interactions in terms of empathy and user involvement.

These heuristics apply to various interaction modalities, such as graphical, textual, and multimodal interfaces, as well as social robots and virtual assistants. Their versatility allows them to be utilized in diverse contexts, ranging from healthcare applications to e-learning platforms, where empathy and user experience play a critical role in successful interactions. While the article does not mention a formal empirical validation of the heuristics, the development process, based on a literature review and expert validation, provides a strong foundation for their practical use and effectiveness in promoting empathic interactions.

The instrument is designed for interactive systems and user interfaces, particularly on health technologies. While it is not confined to a specific type of system, the instrument is applicable in contexts that involve user dialogues and interactions, making it suitable for a wide range of applications. This adaptability is vital for evaluating interfaces that aim to foster more empathic and emotionally engaging interactions.

From an evaluation standpoint, the instrument is developed from the user's perspective, concentrating on how the interactions are perceived and experienced by the end users. Furthermore, external observers and experts can participate in the evaluation during the validation process, enhancing the analysis and offering a more comprehensive understanding of the heuristics' effectiveness.

The instrument's focus extends beyond language to include a variety of design signals and components, such as messages, imagery, and engagement, which transcend verbal communication. Although empirical validation is not explicitly addressed in the document, a panel of experts reviewed and validated the heuristics, indicating some degree of validation despite the lack of formal empirical validation in field studies.

While the document does not directly discuss limitations, it is reasonable to assume that there may be challenges in generalizing the heuristics to different interaction contexts. Further validation in diverse scenarios could be necessary. Additionally, the effectiveness of the heuristics may depend on the type of system and the complexity of the interactions, which should be considered when applying them in future contexts.

4.8. Empathy Scale for Human-Computer Communication (ESHCC)

The Empathy Scale for Human-Computer Communication (ESHCC) [7] represents an adaptation of an existing metric, the Therapist Empathy Scale (TES), originally developed to measure empathy in human-human interactions, especially in therapeutic contexts. This adaptation is important for evaluating empathy in dialogue systems (DSs), such as virtual assistants and chatbots.

The ESHCC (Empathy in Social and Human-Computer Communication) framework considers the nuanced aspects that define empathy in digital contexts, including: **Relevance** – the system's ability to respond appropriately to the user's emotions and needs, ensuring the interaction remains meaningful; **Fluency** – the naturalness and smoothness of the system's responses, even if it occasionally sacrifices emotional expressiveness for the sake of coherence; **Emotional Awareness** – the system's capacity to recognize and appropriately respond to emotions expressed by the user, reflecting its sensitivity to emotional cues; **Interpersonal Connection** – the sense of closeness and understanding the user perceives during the interaction, which is crucial for fostering rapport.

The evaluation is conducted from the perspective of an external observer, which enables an impartial analysis of the empathy demonstrated by the system throughout the dialogue. For this purpose, the ESHCC uses a Likert-type scale, where evaluators rate each item on a scale from 1 to 7, providing a quantitative measure of perceived empathy.

The authors validated the tool through an empirical study involving the application of the instrument in various interactions with dialogue systems. The results indicated that the ESHCC possesses robust psychometric properties, including high reliability and convergent validity.

Confirmatory factor analysis revealed that the dimensions of the ESHCC coherently cluster, supporting the proposed theoretical structure. According to the authors, this instrument can serve as a valuable tool for developers and researchers seeking to enhance Human-Computer Interaction, allowing for the identification of areas for improvement in system responses.

The ESHCC may only partially capture empathy in text-based interactions. It emphasizes language as the primary means of evaluation, focusing on lexical, textual, and syntactic characteristics rather than nonverbal or paralinguistic cues that are more relevant in face-to-face interactions. The instrument applies to both textual interactions and transcripts of voice interactions, although the main emphasis is on textual communication. This approach allows for a deeper analysis of empathy in technology-mediated communication contexts, contributing to the understanding of human-machine interaction.

4.9. Perceived Empathy of Technology Scale (PETS)

The Perceived Empathy of Technology Scale (PETS) [9] measures the perceived empathy of interactive systems, aiming to bridge the gap in evaluating technology empathy. PETS was developed through a structured process that included expert contributions and user testing, ensuring its relevance to interactive systems. The instrument consists of 10 items, divided into two primary dimensions: Emotional Responsiveness (PETS-ER) and Understanding and Trust (PETS-UT), each containing five items that assess different facets of perceived empathy. Data collection is facilitated through a Likert scale, supporting statistical analysis.

PETS was created using a bottom-up approach, incorporating expert feedback and user testing, making it highly specific to interactive systems. It applies to technologies that demonstrate empathy, such as virtual assistants and social robots, across a range of applications. Participants rate the system's empathy from their own perspective in various scenarios.

The empirical validation of PETS was conducted through exploratory and confirmatory factor analyses, ensuring the instrument's robustness and reliability. The framework considers both language and the user's emotional response and understanding, encompassing both affective and cognitive aspects. PETS is versatile, applicable to text, voice, and multimodal interactions, thus expanding its applicability to diverse interaction scenarios.

Despite the statistical validation, the authors acknowledge the complexity of empathy in human interactions and the potential variation in interpreting empathy dimensions across different contexts.

5. Criteria for comparing instruments to measure perceived empathy

We defined specific criteria for the analysis of instruments to measure perceived empathy in interactions between humans and computational agents based on their theoretical and practical importance. These criteria establish a comparison among the instruments and provide a foundation for potential researchers interested in adopting an instrument to select one suitable to their objective.

The first criterion considered is **Adaptation of Metric**, which checks whether the instrument is an adaptation from existing metrics. This criterion is essential to understanding the instrument's theoretical basis and prior validity since adapted instruments may have a history of use and validation, contributing to their reliability and applicability.

The second criterion, **Type of System Evaluated**, identifies the type of computational interactive system (dialogue, interactive, immersive, robot, or other) the instrument targets. Different systems can influence the perception of empathy, and understanding these particularities helps to contextualize interactions and adapt instruments for various technologies.

Dimensions/Components is the third criterion, which evaluates the dimensions and items measured by the instrument, such as the number of dimensions and items in each dimension. This criterion ensures a comprehensive analysis of perceived empathy, understanding which specific aspects of empathy the scope the instrument addresses. The number of dimensions and items also indicates whether the instrument applies to specific audiences, such as older adults, who can suffer from cognitive load in extensive questionnaires.

The fourth criterion is the **Point of View** adopted for measuring empathy, which can be from the user, an external observer, or an expert. The point of view can influence the results and interpretation of the data, providing varied insights into the interaction and perceived empathy.

The **Type of Questionnaire** used is the fifth criterion, such as the Likert scale. The type of questionnaire impacts how data is collected and analyzed, with questionnaires structured in well-known scales facilitating the comparison and interpretation of results.

The sixth criterion, **Empirical Validation**, verifies whether the instrument underwent empirical validation, such as factor analysis, reliability (Cronbach's Alpha or McDonald's Omega), and convergent/divergent validity. Empirical validation is fundamental to ensure the instrument's reliability and scientific robustness.

Limitations is the seventh criterion, identifying possible restrictions in the application method, the type of system evaluated, or the type of interaction studied. Understanding the instrument's limitations is essential to identify areas for improvement and constraints.

The eighth criterion, **Focus on Language**, analyzes whether the instrument focuses only on language or considers other cues (such as facial expressions and gestures). This criterion is significant for situations where only conversational agents with a firm reliance on verbal communication (spoken or written) are used, such as chatbots using Natural Language Processing.

Lastly, the ninth criterion is **Interaction Modalities**, which evaluates the modalities used by the instrument (text, voice, multimodal). Considering interaction modalities helps to understand the instrument's flexibility and applicability in different usage contexts, as different modalities can uniquely affect empathy perception.

6. Comparison of the instruments

The criteria were organized into three thematic groups, as described next, to summarize the comparison and facilitate the presentation and comparative analysis of the instruments. In addition to the criteria defined in Section 5, we included 'Publication Year' as an additional column. This information is relevant as it allows us to evaluate the timeliness and relevance of the analyzed instruments. In rapidly evolving fields like Human-Computer Interaction, tools developed more recently tend to align better with the current capabilities and demands of intelligent systems. Therefore, considering the year of publication helps identify tools with greater potential to adapt to contemporary needs, promoting more consistent analyses within the current research landscape.

First, Table 1 presents the general criteria related to the contextual characteristics of the instruments, such as the type of system evaluated, the supported interaction modalities, the year of publication, and the potential for adapting pre-existing metrics.

The column 'Adapts an Existing Metric?' refers to the theoretical origin of the instrument and its adaptation to technological contexts. 'Interaction Modalities' indicates whether the instrument covers textual, voice or multimodal interactions.

Table 1
General Characteristics

Instrument	Publication Year	Adapts an Existing Metric?	Type of System	Interaction Modalities
Tisseron et al. (QEAE)	2015	Not explicitly stated by the authors	Social Robots Avatars	Multimodal
Charrier et al. (RoPE)	2019	Adapted from empathy scales used in cognitive science	Social Robots	Multimodal
Urakami et al. (Empathy Assessment Model)	2019	Items adapted from cognitive empathy frameworks	Interactive Systems	Text
Chi et al. (SSRIT)	2021	No	Social Robots	Multimodal
Kroes et al. (TEQ)	2022	No (uses TEQ without adaptation)	Virtual Reality Systems	Multimodal
Concannon and Tomalin (ESHCC)	2023	Adapted from the Therapist Empathy Scale (TES)	Dialogue Systems	Text
Borycki et al. (Design Heuristics)	2024	No	Health Systems	Multimodal
Schmidmaier et al. (PETS)	2024	No	Interactive Systems	Multimodal
Xu and Jiang	2024	Adapts psychological empathy definitions	Customer Service Dialogue Systems	Text

Table 2 includes information on the structure and content of the instruments, including the number of dimensions and elements, the focus on language or other signals, and the adopted perspective (user, external observer, or specialist).

Table 2
Structure and Content

Instrument	Dimensions/Items	Focus on Language	Evaluator Point of View
Tisseron et al. (QEAE)	4 dimensions/ 16 items	No	User's Perspective
Charrier et al. (RoPE)	2 dimensions/ 18 items	No	User's Perspective
Urakami et al. (Empathy Assessment Model)	8 dimensions/ 72 items	Yes	User's Perspective
Chi et al. (SSRIT)	3 dimensions/ Not stated	No	User's Perspective
Kroes et al.	6 dimensions from TEQ	No	User's Perspective
Concannon and Tomalin (ESHCC)	4 dimensions /Not stated	Yes	External Observer
Borycki et al. (Design Heuristics)	Not Applicable	No	User's Perspective and Experts
Schmidmaier et al. (PETS)	2 dimensions/ 10 items	No	User's Perspective
Xu and Jiang	2 dimensions/ 16 intents/ 4 aspects	Yes	Dual Perspective (User and External)

Finally, Table 3 describes methodological aspects, such as the type of questionnaire used, the empirical validation processes carried out, and the identified limitations. This division aims to organize the information clearly and objectively, optimizing understanding and comparison among the instruments.

Table 3
Methodological Characteristics

Instrument	Type of Questionnaire	Validation	Limitations
Tisseron et al. (QEAE)	Not specified	In process*	Limited validation, small item pool
Charrier et al. (RoPE)	Likert Scale	Preliminary	Limited to robots, needs broader testing
Urakami et al. (Empathy Assessment Model)	Likert Scale	Cronbach's alpha, ANOVA	Hypothetical scenarios
Chi et al. (SSRIT)	Reflective and Formative	Empirical validation	Indirect empathy focus
Kroes et al.	Likert Scale	Human interaction validation	Not adapted for agents
Concannon and Tomalin (ESHCC)	Likert Scale	Confirmatory Factor Analysis	Limited to textual focus
Borycki et al. (Design Heuristics)	Not applicable	Expert review	No formal empirical validation
Schmidmaier et al. (PETS)	Likert Scale	Exploratory and Confirmatory Factor Analysis	Context-specific applicability
Xu and Jiang	Likert Scale	Empirical validation on annotated dataset	Domain-specific focus on dialogues

7. Challenges, limitations, and ethical implications

7.1. Limitations of existing and adapted instruments

While there has been progress in developing and adapting empathy measurement tools for computational agents, several limitations remain. Many adapted instruments still carry assumptions from human-to-human interactions, which may not fully align with the dynamics of human-agent relationships, leading to potential biases. For example, users' preconceived notions about technology can influence their perception of empathy, often resulting in an unrealistic projection of emotional capabilities onto the agent. Moreover, current AI systems' limited emotional and cognitive capacities restrict the accuracy of empathy assessments, as these agents can only simulate emotions rather than experience them. These challenges complicate the measurement and interpretation of perceived empathy, often compromising the validity of results across diverse contexts.

7.2. Ethical and social implications

The simulation of empathy by AI systems introduces significant ethical concerns. Is it ethical to create systems that seem to empathize without genuinely experiencing emotions? While these simulations may offer potential benefits, they could also mislead users into thinking the agent has a human-like understanding, leading to unrealistic expectations. This is particularly problematic in sensitive areas such as healthcare and mental health support, where users might form emotional attachments or depend on the agent for emotional support. These dynamics call for careful consideration of transparency in AI system design, ensuring that users are aware of the limitations of AI empathy. Additionally, the potential for empathetic AI to be misused for manipulative purposes, such as influencing consumer behavior, highlights the need for ethical guidelines regarding empathy simulation in AI systems.

7.3. Gaps and research opportunities

Empathy instruments reviewed in this study predominantly treat empathy as a static characteristic evaluated post-interaction. While such approaches provide valuable insights, they often fail to account for empathy's dynamic and evolving nature during real-time Human-Agent Interactions.

Despite Tools like the QEAE and RoPE exhibit structured methodologies, their designs focus on predefined scenarios, limiting their ability to capture moment-to-moment emotional responsiveness, particularly critical in domains like healthcare or customer service. Real-time assessment would enable a deeper understanding of how empathy unfolds across different stages of interaction.

Moreover, while generalizability across diverse domains is often discussed, findings from instruments such as the ESHCC suggest that precision within specific domains holds greater importance than universal applicability. For instance, healthcare systems may require highly specialized tools to address the emotional nuances of patient-provider interactions, whereas educational technologies may benefit from measures tailored to motivational support. Instruments should focus on adaptability within defined contexts, ensuring their sensitivity to the unique demands of each domain.

Another key opportunity lies in integrating cultural and individual variability into instrument design. Studies on frameworks like SSRIT demonstrate that users' familiarity with technology and cultural backgrounds significantly influence empathy perception. Incorporating such factors into assessment tools could enhance their inclusivity and relevance, particularly for global applications. Furthermore, dimensions of empathy, such as emotional expressiveness versus cognitive understanding, may vary in significance depending on the user's cultural context, underlining the need for flexible designs that accommodate diverse interpretations.

Future research can develop instruments capable of dynamically measuring empathy while ensuring domain-specific precision and cultural adaptability. These advancements will support creating human-agent systems that are empathetic in design and responsive to the multifaceted nature of user interactions, contributing to improved trust, satisfaction, and engagement.

8. Conclusion and future work

Empathy is a fundamental element in designing AI systems that are effective, engaging, and capable of fostering user trust and acceptance. Accurate measurement of empathy plays a pivotal role in creating more human-like interactions, especially as AI becomes increasingly integrated into sensitive domains such as healthcare, education, and social assistance. This review is a valuable resource for those seeking to select appropriate empathy assessment tools for their work.

This paper has examined the landscape of instruments for measuring perceived empathy in Human-Computer Interactions, addressing their theoretical underpinnings, methodological frameworks, and practical applications. Through a comparative analysis of existing and adapted tools, we have highlighted their strengths, domain-specific relevance, and limitations. While comprehensive, this thematic review acknowledges the potential influence of subjective selection biases, which may impact the reproducibility of its findings. These insights emphasize the need for context-aware empathy measures to better address the challenges posed by computational agents.

Although this study provides an in-depth comparative analysis of these instruments, further opportunities exist to advance the field. A promising avenue is the experimental validation of the Perceived Empathy of Technology Scale (PETS), particularly across diverse interaction modalities such as text, voice, and multimodal contexts. Testing its effectiveness in varied scenarios and audiences will allow for a more comprehensive understanding of its reliability and limitations.

Additionally, if the validation reveal significant gaps in PETS' capacity to measure empathy accurately, the development of a new assessment instrument may be proposed. This instrument would address the identified shortcomings, incorporating innovative dimensions or focusing on aspects not currently captured by existing tools. Such advancements would enhance the precision of empathy evaluation in computational systems, adapting to increasingly complex interaction contexts. By addressing these gaps, this work intends to contribute to the development of more empathetic and impactful human-agent systems, fostering trust, satisfaction, and engagement across various domains.

Acknowledgments

This work has been supported by the PLEISAR-Social project, Ref. PID2022-136779OB-C33, funded by the Ministry of Science and Innovation of Spain (MCIN/AEI/10.13039/501100011033, EU), and the C-ING-179-UGR23 project, funded by the Regional Ministry of Universities, Research and Innovation and the FEDER Andalucía 2021-2027 Programme.

Declaration on generative AI

During the preparation of this work, the author(s) used CHAT-GPT-4 and Grammarly in order to: improve writing style and grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] B. Lugrin, C. Pelachaud, D. Traum (Eds.), The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application, volume 48, 1 ed., Association for Computing Machinery, New York, NY, USA, 2022.

[2] C. Birmingham, A. Perez, M. Matarić, Perceptions of cognitive and affective empathetic statements by socially assistive robots, in: Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), HRI' 22, Sapporo, Japan, 2022, pp. 323––331. doi:10.1109/HRI53351.2022.9889386.

[3] O. N. Yalçın, V. Utz, S. DiPaola, Empathy through aesthetics: Using ai stylization for visual anonymization of interview videos, in: Proceedings of the 3rd Empathy-Centric Design Workshop: Scrutinizing Empathy Beyond the Individual, EmpathiCH '24, ACM Press, New York, USA, 2024, pp. 63-–68. doi:10.1145/3661790.3661803.

[4] Y. Yang, X. Ma, P. Fung, Perceived emotional intelligence in virtual agents, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA '17, Association for Computing Machinery, New York, USA, 2017, pp. 2255-–2262. doi:10.1145/3027063.3053163.

[5] A. P. Chaves, M. A. Gerosa, How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design, International Journal of Human–Computer Interaction 37 (2021) 729–758. doi:10.1080/10447318.2020.1841438.

[6] B. Yang, Y. Sun, X.-L. Shen, Understanding ai-based customer service resistance: A perspective of defective ai features and tri-dimensional distrusting beliefs, Information Processing & Management 60 (2023) 103257. doi:10.1016/j.ipm.2022.103257.

[7] S. Concannon, M. Tomalin, Measuring perceived empathy in dialogue systems, AI & SOCIETY 39 (2024) 2233–2247. doi:10.1007/s00146-023-01715-z.

[8] H. Putta, K. Daher, M. E. Kamali, O. A. Khaled, D. Lalanne, E. Mugellini, Empathy scale adaptation for artificial agents: a review with a new subscale proposal, in: Proceedings of the 8th International Conference on Control, Decision and Information Technologies, volume 1 of *CoDIT' 22*, IEEE, Istanbul, Turkey, 2022, pp. 699–704. doi:10.1109/CoDIT55151.2022.9803993.

[9] M. Schmidmaier, J. Rupp, D. Cvetanova, S. Mayer, Perceived Empathy of Technology Scale (PETS): Measuring empathy of systems toward the user, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24, ACM Press, Honolulu, USA, 2024, pp. 1–18. doi:10.1145/3613904.3642035.

[10] A. Cuadra, M. Wang, L. A. Stein, M. F. Jung, N. Dell, D. Estrin, J. A. Landay, The illusion of empathy? notes on displays of emotion in human-computer interaction, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI' 24, ACM Press, Honolulu, USA, 2024, pp. 1-–18. doi:10.1145/3613904.3642336.

[11] A. Paiva, I. Leite, H. Boukricha, I. Wachsmuth, Empathy in virtual agents and robots: A survey, ACM Trans. Interact. Intell. Syst. 7 (2017) 11:1–11:40. doi:10.1145/2912150.

[12] L. Charrier, A. Rieger, A. Galdeano, A. Cordier, M. Lefort, S. Hassas, The RoPE scale: a measure of how empathic a robot is perceived, in: Proceedings of the 14th ACM/IEEE International Conference on Human -Robot Interaction, HRI '25, Daegu, Korea, 2019, pp. 656–657. doi:10.1109/HRI.2019.8673082.

[13] S. Park, M. Whang, Empathy in human–robot interaction: Designing for social robots, International Journal of Environmental Research and Public Health 19 (2022) 1889. doi:10.3390/ijerph19031889.

[14] J. James, C. I. Watson, B. MacDonald, Artificial empathy in social robots: An analysis of emotions in speech, in: Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), RO-MAN' 18, IEEE, Nanjing and Tai'an, China, 2018, pp. 632–637. doi:10.1109/ROMAN.2018.8525652.

[15] M. Roshanaei, R. Rezapour, M. S. El-Nasr, Talk, listen, connect: Navigating empathy in human-ai interactions, 2024. doi:10.48550/arXiv.2409.15550.

[16] R. Hogan, Development of an empathy scale, Journal of consulting and clinical psychology 33 (1969) 307-–316. doi:10.1037/h0027580.

[17] A. Mehrabian, N. Epstein, A measure of emotional empathy, Journal of Personality 40 (1972) 525–543. doi:10.1111/j.1467-6494.1972.tb00078.x.

[18] M. H. Davis, et al., A multidimensional approach to individual differences in empathy, JSAS Catalog of Selected Documents in Psychology 10 (1980) 85. URL: https://www.uv.es/friasnav/Davis_1980.pdf.

[19] S. W. Mercer, M. Maxwell, D. Heaney, G. C. Watt, The consultation and relational empathy (care) measure: development and preliminary validation and reliability of an empathy-based consultation

process measure, Family practice 21 (2004) 699–705. doi:10.1093/fampra/cmh621.

[20] S. E. Decker, C. Nich, K. M. Carroll, S. Martino, Development of the therapist empathy scale, Behavioural and cognitive psychotherapy 42 (2014) 339–354. doi:10.1017/S1352465813000039.

[21] M. Hojat, J. DeSantis, S. C. Shannon, L. H. Mortensen, M. R. Speicher, L. Bragan, M. LaNoue, L. H. Calabrese, The jefferson scale of empathy: a nationwide study of measurement properties, underlying components, latent variable structure, and national norms in medical students, Advances in Health Sciences Education 23 (2018) 899–920. doi:10.1007/s10459-018-9839-9.

[22] V. Braun, V. C. and, Using thematic analysis in psychology, Qualitative Research in Psychology 3 (2006) 77–101. doi:10.1191/1478088706qp063oa.

[23] S. Tisseron, F. Tordo, R. Baddoura, Testing Empathy with Robots: A Model in Four Dimensions and Sixteen Items, International Journal of Social Robotics 7 (2015) 97–102. doi:10.1007/s12369-014-0268-5.

[24] K. Daher, J. Casas, O. A. Khaled, E. Mugellini, Empathic chatbot response for medical assistance, in: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA '20, ACM Press, New York, USA, 2020, pp. 1–3. doi:10.1145/3383652.3423864.

[25] J. Urakami, B. A. Moore, S. Sutthithatip, S. Park, Users' perception of empathic expressions by an advanced intelligent system, in: Proceedings of the 7th International Conference on Human-Agent Interaction, HAI '19, ACM Press, New York, USA, 2019, pp. 11–18. doi:10.1145/3349537.3351895.

[26] O. H. Chi, S. Jia, Y. Li, D. Gursoy, Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (ai) social robots in service delivery, Computers in Human Behavior 118 (2021) 106700. doi:10.1016/j.chb.2021.106700.

[27] K. Kroes, I. Saccardi, J. Masthoff, Empathizing with virtual agents: the effect of personification and general empathic tendencies, in: Proceedings of the IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR' 22, IEEE, Online, 2022, pp. 73–81. doi:10.1109/AIVR56993.2022.00017.

[28] Z. Xu, J. Jiang, Multi-dimensional evaluation of empathetic dialog responses, 2024. doi:10.48550/arXiv.2402.11409.

[29] E. M. Borycki, R. Kletke, C. le Nobel, G. McWilliams, S. Whitehouse, A. W. Kushniruk, Empathetic and emotive design heuristics, in: pHealth 2024, IOS Press, 2024, pp. 80–84. doi:10.3233/SHTI240062.