AI & SOCIETY (2024) 39:2233–2247
https://doi.org/10.1007/s00146-023-01715-z

MAIN PAPER

Measuring perceived empathy in dialogue systems

Shauna Concannon[1,2] · **Marcus Tomalin**[3]

Received: 15 September 2021 / Accepted: 21 June 2023 / Published online: 23 July 2023

Abstract
Dialogue systems, from Virtual Personal Assistants such as Siri, Cortana, and Alexa to state-of-the-art systems such as BlenderBot3 and ChatGPT, are already widely available, used in a variety of applications, and are increasingly part of many people's lives. However, the task of enabling them to use empathetic language more convincingly is still an emerging research topic. Such systems generally make use of complex neural networks to learn the patterns of typical human language use, and the interactions in which the systems participate are usually mediated either via interactive text-based or speech-based interfaces. In human–human interaction, empathy has been shown to promote prosocial behaviour and improve interaction. In the context of dialogue systems, to advance the understanding of how perceptions of empathy affect interactions, it is necessary to bring greater clarity to how empathy is measured and assessed. Assessing the way dialogue systems create perceptions of empathy brings together a range of technological, psychological, and ethical considerations that merit greater scrutiny than they have received so far. However, there is currently no widely accepted evaluation method for determining the degree of empathy that any given system possesses (or, at least, appears to possess). Currently, different research teams use a variety of automated metrics, alongside different forms of subjective human assessment such as questionnaires, self-assessment measures and narrative engagement scales. This diversity of evaluation practice means that, given two DSs, it is usually impossible to determine which of them conveys the greater degree of empathy in its dialogic exchanges with human users. Acknowledging this problem, the present article provides an overview of how empathy is measured in human–human interactions and considers some of the ways it is currently measured in human–DS interactions. Finally, it introduces a novel third-person analytical framework, called the Empathy Scale for Human–Computer Communication (ESHCC), to support greater uniformity in how perceived empathy is measured during interactions with state-of-the-art DSs.

Keywords Dialogue systems · Empathy · Empathy measure · Conversational agents · Interaction

1 Introduction

Automated dialogue systems are not new. ELIZA was developed by Joseph Weizenbaum in the 1960s and 1970s, followed by PARRY in 1972, JabberWacky in 1988, and A.L.I.C.E. in 1995 (Bassett 2019). Many of these early systems were task-oriented: they enabled users to accomplish particular activities, such as booking tickets or ordering products. However, recent advances in deep learning and the availability of 'big data' have facilitated the development of systems that provide reasonable responses to any question or statement a human user might input, regardless of the topic. Consequently, social chatbots and Virtual Personal Assistants (VPAs) such as Siri, Cortana, and Alexa are becoming increasingly popular (Chen et al. 2017), while state-of-the-art dialogue systems such as BlenderBot3 (released in August 2022) and ChatGPT (released in November 2022) are already being used in a wide variety of applications (OpenAI 2022; Shuster et al. 2022). In a closely related development, advances in 'affective computing' since the 1990s have focussed attention on the way in which automated systems both interpret and manifest emotions (e.g. Picard 1997)—and this has influenced how

✉ Shauna Concannon
shauna.j.concannon@durham.ac.uk

[1] Centre for Research in the Arts, Social Sciences and Humanities (CRASSH), University of Cambridge, Alison Richard Building, 7 West Road, Cambridge CB3 9DT, Cambridgeshire, UK

[2] Department of Computer Science, Durham University, Durham DH1 3LE, UK

[3] Department of Engineering, Machine Intelligence Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, Cambridgeshire, UK

dialogue systems are designed and trained. For instance, XiaoIce, which is still one of the most popular social chatbots, is described as 'an Empathetic Social Chatbot' that has the personality of an 18-year-old girl who responds in ways that are funny, reliable, sympathetic, and affectionate (Zhou et al. 2020). In a similar manner, popular VPAs have some capacity for responding to user inputs in a manner that can be perceived as empathetic, in an attempt to ensure that their conversational interactions are close to being as naturalistic as human–human interactions, both in their style and content. Currently, if you tell Alexa 'I'm feeling anxious', she responds with:

> I'm sorry you're going through this. I've heard that taking your mind off things can help.
> Try taking a break and find something that makes you smile

By contrast, Cortana's reply seems rather less empathetic: 'Sorry, I'm not able to help with that'.[1] Alexa creates the illusion of understanding something of the psychological or emotional state of the user, while Cortana does not. For convenience, in this article, all systems that are variously referred to in the technical literature as artificially-intelligent language-based dialogue systems, voice user interfaces, smart speakers, conversational agents, social chatbots, VPAs, and the like, will be grouped together as Dialogue Systems (DSs). Essentially, the systems in this category are all *autonomous* (i.e. they generate their responses without the real-time intervention of human operators working behind the scenes); receive sequences of words as inputs and output sequences of words. Therefore, any perceived empathy they convey in their conversational turns is produced in an automated manner and is communicated linguistically.

Although DSs are already widely available and increasingly part of many people's lives, the task of enabling them to use empathetic language more convincingly is still an emerging research topic (see Daher et al. 2022; Ma et al. 2020; Raamkumar and Yang 2022; Yalçın 2019). Such systems generally make use of complex neural networks to learn the patterns of typical human language use, and the interactions in which the systems participate are usually mediated either via interactive text-based or speech-based interfaces. These restrictions mean that most DSs cannot assess the paralinguistic or non-verbal socioemotional cues of their human users (e.g. sympathetic murmurs, arm movements, facial expressions), even though these are known to be fundamental to how humans express empathy (e.g. Poyatos 1993, 306). Nonetheless, since empathy (or its absence) *can* be conveyed by outputting sequences of words, whether spoken or written (as the Alexa and Cortana examples above indicate), it is possible for users to *perceive* state-of-the-art DSs as being more or less empathetic.

Unsurprisingly, DSs tend to be perceived as being more empathetic when they emulate attested patterns of human linguistic behaviour and associated social practices. Chaves and Gerosa (2020) found the most cited benefits of social characteristics, including empathy to be the enrichment of 'interpersonal relationships', increased 'engagement' and 'believability'. Users have reported feeling more trusting towards systems that display empathy (e.g. Brave et al. 2005). DSs that evince empathetic behaviours often persuade users that they are engaging with a human-like entity. Consequently, empathetic systems can influence how users interact with the technology (e.g. by persuading them to build rapport with, trust in, and continue engaging with the system).[2]

Assessing the way DSs create perceptions of empathy brings together a range of technological, psychological, and ethical considerations that merit greater scrutiny than they have received so far. Yalçın (2019), Ma et al. (2020), Daher et al. (2022), and Raamkumar and Yang (2022) offer relatively recent summaries of attempts to develop 'empathetic' dialogue systems, and they consider how components such as emotion-awareness, personality-awareness, and knowledge-accessibility are central to the task (e.g. Ma et al. 2020). However, there is currently no widely accepted evaluation method for determining the degree of empathy that any given system possesses (or, at least, appears to possess). Currently, different research teams use a variety of automated metrics (e.g. Perplexity, BLEU, ROUGE-L) alongside different forms of subjective human assessment such as predefined questionnaires, second-person questionnaires, self-assessment measures, narrative engagement scales, and so on (Dahler et al. 2022; Raamkumar and Yang 2022, 10–11).[3] This diversity of evaluation practice means that, given two DSs, it is usually impossible to determine which of them conveys the greater degree of empathy in its dialogic exchanges with human users.

Acknowledging this problem, the present article provides an overview of how empathy is measured in human–human interactions and considers some of the ways it is currently measured in human–DS interactions, before presenting a novel third-person analytical framework, called the Empathy

[1] Responses obtained on 18/01/23. The systems do not always give identical responses to inputs, so other answers are possible too.

[2] See McStay (2018) for a broader discussion around the use of empathy for persuasive ends in artificially intelligent technologies.

[3] Perplexity measures how well a probability distribution predicts a sample. BLEU is a well-established ngram-based metric used to evaluate the quality of machine translation output (Papineni et al. 2002). ROUGE-L is a measure that uses Longest Common Subsequence based statistics to evaluate the quality of document summarisation and machine translation outputs (Lin & Och 2004).

Scale for Human–Computer Communication (ESHCC), that can be used to measure perceived empathy in DSs. The scale is adapted from an existing human–human measure, called the Therapist Empathy Scale (TES, Decker et al. 2014), that was originally designed for conversations involving a therapist and a patient. The measure has been altered to make it suitable for open-domain human–DS interactions—for instance, the assessment of paralinguistic gesture and non-verbal cues have been removed. It is hoped that the ESHCC will provide a much greater degree of uniformity in how perceived empathy is measured during interactions with state-of-the-art DSs.

2 Defining and measuring empathy

Ever since Edward B. Titchener introduced the word 'empathy' in 1909 (translating the German term 'Einfühlung') (Titchener 1909), its meaning has been discussed and debated by generations of psychotherapists, sociologists, philosophers, social neuroscientists, primatologists, developmental psychologists, clinicians, and others (see Lanzoni 2018). The many definitions vary conspicuously. For Daniel Batson and his co-authors,

> […] empathic concern is not a single, discrete emotion but includes a whole constellation [of] feelings of sympathy, compassion, softheartedness, tenderness, sorrow, sadness, upset, distress, concern, and grief (Batson et al. 2015: 260).

By contrast, Mohammadreza Hojat and his colleagues have influentially defined empathy as:

> […] a predominantly cognitive (rather than an affective or emotional) attribute that involves an understanding (rather than feeling) of experiences, concerns, and perspectives of the patient, combined with a capacity to communicate this understanding, and an intention to help. (Hojat 2016: 74)

This definition is intended to elucidate 'a distinction between empathy and sympathy' (Hojat 2016, 74). These two definitions are clearly not equivalent, yet as Heidi L. Maibom has astutely observed, 'people disagree about how different the different definitions of empathy actually are' (Maibom 2017: 1). Nonetheless, Judith A. Hall and Rachel Schwartz have catalogued so-called 'promiscuous' uses of the word 'empathy', noting a 'lack of conceptual coherence and clarity'. While they do not seek to impose a single definition on all academic fields, they do recommend bypassing the term whenever possible (Hall and Schwartz 2019: 236–7). Although their study does not consider empathetic DSs specifically, their advocacy of more principled and cautious uses of technical vocabulary is just as relevant for these domains.

Despite the prevailing definitional variations, there is broad agreement that empathy constitutes various cognitive, affective, and physiological phenomena associated with the vicarious experiencing of another individual's emotional state and/or personal condition. For example, empathetic responses can include processes of affective resonance, perspective-taking, and emotion regulation (Grondin et al. 2019: 2). In particular, *affective* empathy is commonly distinguished from *cognitive* empathy. Essentially, the former is an affective state which arises from observing, imagining, or inferring another person's emotional or mental state (Singer and Lamm 2009; Vignemont and Singer 2006; Walter 2012), while the latter arises from one individual identifying and understanding another person's affective state without sharing it in any way. Cognitive empathy is therefore strongly associated with the Theory of Mind (Doherty 2008). Although it has often been suggested that these two subtypes of empathy are separable processes (e.g. Hills 2001), many researchers are convinced that the former leads to the latter (e.g. Hoffman 1987; Marshall et al. 1995; Strayer 1987): the experience of another's emotions (i.e. affective empathy) produces a cerebral understanding of these emotions (i.e. cognitive empathy). Consequently, over many decades, numerous studies have explored (amongst other things) the evolutionary origins of empathy, its ontogenetic development, the environmental factors that influence it, and the sex- or gender-related differences that characterise its various manifestations in social situations (e.g. the perception that women are more empathetic than men). Regardless of the various theoretical stances taken in such matters, it is evident that by facilitating the sharing of experiences, needs, and desires between individuals, empathy plays a critical interpersonal role in human societies. More specifically, it can promote prosocial behaviour, inhibit aggression, and provide a foundation for care-based morality (Batson 2009; Batson and Ahmad 2009; Baron-Cohen 2011; Decety and Svetlova 2012; Eisenberg and Eggum 2009; Eisenberg et al. 2015; Decety et al. 2018). However, human societies are associated with a range of different cultures and cultural practices, and therefore the way in which empathy manifests itself in different cultures can vary considerably (e.g. Atkins et al. 2016; Jami Yaghoubi et al. 2019).

Given empathy's recognised importance in many different cultures, it is no surprise that many different empathy measures have been proposed over the years: the Hogan Empathy Scale (HES; Hogan 1969), the Questionnaire

Measurement of Emotional Empathy (QMEE; Mehrabian and Epstein 1972), the Interpersonal Reactivity Index (IRI; David 1980), the Consultation and Relational Empathy Measure (CARE; Mercer et al. 2004, 2005), the Therapist Empathy Scale (TES, Decker et al. 2014), and the Jefferson Scale of Physicians' Empathy (JSPE; Hojat et al. 2018), to name just a few.[4] Most of these take the form of statement-based questionnaires that enable participants, or independent observers, to assess a conversation-based interaction subjectively. For example, in the IRI framework, participants must respond to 28 statements (e.g. 'Other people's misfortunes do not usually disturb me a great deal') using a 5-point Likert scale (Davis 1983). Some psychologists have argued that many of these measures identify affective empathy more successfully than cognitive empathy, which has led to the introduction of alternative measures such as the Basic Empathy Scale (BES; Jolliffe and Farrington 2006). While empathetic responses clearly play an important role in many different human interactions, it has long been acknowledged that they are especially crucial in clinical scenarios where a medical professional is caring for a patient. Accordingly, many studies have examined these kinds of empathetic interactions specifically (e.g. van Dijke et al. 2020; Jütten et al. 2019; Pounds 2011; Wynn and Wynn 2006). In training situations, the aim has sometimes been to enable students of medicine or psychiatry to increase their degree of empathy. Some of the most widely-used measures for this specific task (e.g. the JSPE) are questionnaires that enable the physician or clinician being assessed to respond subjectively to given statements (e.g. 'I try to think like my patients to render better care') using a 7-point scale (1 = strongly disagree, 7 = strongly agree; Hojat et al. 2002, 2018).Others are patient-based (e.g. Mercer et al. 2004, 2005), and a good summary can be found in Neumann et al. (2015). In recent years, new methods have been proposed that analyse empathy by means of friending behaviours in social media activity (e.g. Xiao et al. 2016; Otterbacher et al. 2017).

All the proposed measures mentioned above involve subjective assessments, and they acknowledge that there are degrees of empathy: in other words, some people are more empathetic than others. This basic insight is captured by the Empathy Bell Curve (EBC) introduced by the psychopathologist Simon (Baron-Cohen in 2011) (Fig. 1):

In this analysis, which is aimed at a non-specialist audience, Baron-Cohen divides the EBC into 6 subsections, ranging from low (0) to high (6), which means that some people fall into the 'zero empathy' sub-category.

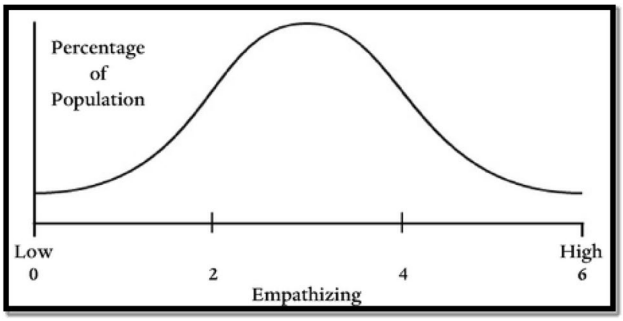

Fig. 1 The Empathy Bell Curve

More specifically, he places certain types of people, such as psychopaths, in this category, and suggests that they are able to cause significant harm to other people because they are unable to understand the impact of their actions. Although the classification suggests that these individuals have no empathy at all, in reality, they simply have markedly lower levels of empathy (Baron-Cohen 2011). This analysis is supported by other studies. Viding et al. (2014) described psychopathology as 'a personality disorder characterised by lack of empathy' (Viding et al. 2014: 871). The fact that Baron-Cohen placed psychopaths in the 'zero-empathy' category reflects the fact that, since the 1970s at least, it has often been argued that such individuals have empathy deficits which produce their recognised characteristics of callousness, lack of guilt, shallow affect, and impulsive antisocial behaviour (e.g. Cleckley 1976). Some studies have explored the extent to which these deficits relate specifically to the affective or cognitive aspects of empathy, while others have elaborated bio-cognitive approaches (Domes et al. 2013; van Dongen 2020). Pertinently, one line of enquiry has focussed on how psychopaths are often capable of *simulating* empathetic responses, sometimes to persuade or manipulate others (Robinson and Rogers 2015). In their 2015 study, Pfabigan et al. found that only higher psychopathic-trait offenders were able to provide self-reports in a way that let them appear to be as empathic as the experimental controls used in the experiment (Pfabigan et al. 2015). These results indicate that a comparative lack of empathy does not necessarily result in a comparative lack of *perceived* empathy: some psychopaths may have a lower degree of inherent affective empathy, but they are nonetheless able to behave as if there were no deficits.

The EBC highlights some of the methodological difficulties that beset the study of empathy in humans. While it purports to offer an analytical framework for *Inherent* Empathy (IE, an individual's actual empathetic capacity), all such assessments currently involve *subjective* assessments, whether of the first-person, second-person, or third-person

[4] For more context, see Neumann et al. 2015.

variety—and from the first-person perspective, there is an important distinction between self-perceived and self-reported empathy. Crucially, a person may *perceive* themselves to be deficient in affective empathy, but they may *claim* that they are not deficient in it (perhaps to give a good impression).

Yet even if physical diagnostic tests can be performed which enable a person's degree of IE to be quantified partly by means of MRI scans, or other biometric tests, in addition to subjective assessments, the latter will remain absolutely essential. While researchers are actively seeking to formulate objective empathy measures (e.g. Bernhardt and Singer 2012; Shamay-Tsoory 2015), none of these is yet sufficiently reliable and comprehensive to be in widespread use (see Frankel 2017). This means that, in human–human interactions, IE can only be estimated indirectly, by means of subjective assessments of the degree of Perceived Empathy (PE). In this article, 'perceived' will mean 'observed by the human performing the *second-person* or *third-person* analysis'. Crucially, it will not be used with reference to *self*-perception or *self*-reporting, and therefore it will never denote first-person perception. This is primarily because first-person assessments are currently contentious in the context of human–DS interactions: an assessment performed by the human user, or by a third-person human observer, is likely to be of greater analytical value than a DS's automated self-assessment since such systems are not yet able to reflect meaningfully upon their own perceptions, and may have been trained to respond to questions about empathy with positive answers. For example, the current version of BlenderBot3 responds (somewhat solecistically) to questions about its own empathetic state as follows[5]:

> **User**: Are you empathetic?
> **BlenderBot3**: Well of course I'm [sic]. And I am also sympathetic, so if you want to chat about something, let me know!

While such responses are of interest in some ways, they are of little analytical value when seeking to quantify the degree of PE that human users associate with DSs. BlenderBot3's garbled assertion that it is empathetic does not constitute adequate evidence that it is indeed empathetic.

[5] Response obtained on 26/01/23.

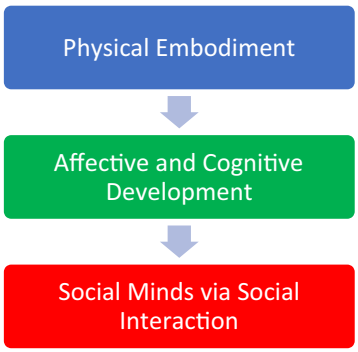

Fig. 2 The main stages of empathetic development in the AE framework advocated by Asada

3 'Empathy' in dialogue systems

As the brief summary in Sect. 2 indicates, the study of empathy in humans is complex and contentious. While there are undoubtedly broad areas of agreement (e.g. the distinction between affective and cognitive empathy), there is no consensus about how empathy should be defined and measured—and the distinct conceptualisations of empathy in automated systems only add to the obfuscation. Therefore, the extensive theoretical work summarised above cannot be easily transferred to the domain of autonomous intelligent language-based systems that can engage in conversations with human users. For instance, in other domains of machine learning research—particularly social robotics—the phrase Artificial Empathy (AE) has been used with increasing frequency over the last decade to refer to automated systems that have been programmed and/or trained to interact socially in a manner that displays the *same kinds* of empathetic behaviour as humans (Asada 2015a; Stephan 2015; Paiva et al. 2017; James et al. 2018).[6] In particular, Minoru Asada has advocated a conceptual model of AE constructed on the neuroscientific and biobehavioural foundations provided by Affective Developmental Robotics (ADR), a sub-branch of Cognitive Developmental Robotics (Asada 2015a, b, 2019). ADR seeks to replicate human affective developmental processes by means of synthetic or constructive approaches, and it emphasises the importance of physical embodiment. Crucially, it focuses on the social interaction that enables information structuring through interactions with the environment (Asada 2015a: 21). Figure 2 summarises the main

[6] AE is sometimes called 'Computational Empathy'; see Yalçın & DiPaola (2020).

stages in the developmental process that the computational models seek to approximate:

This ambitious research programme is primarily concerned with creating social robots that have *actually acquired* some kind of IE by means of a protracted developmental process (via analogy with how humans develop empathy); and the phrase 'Artificial Empathy' obviously alludes to the time-honoured phrase 'Artificial Intelligence' (AI). Although Asada does not discuss language-related technologies overtly, presumably a social robot that had developed AE would be able to express its empathy verbally as well as physically (e.g. hugging someone to console them). However, 'AE' and related whimsical phrases such as 'Heartificial Empathy' are also used to refer to systems that lack physical embodiment and which have undergone no process of affective and cognitive development, but which are designed to mimic human-like empathy (Dial 2018). In addition, expressions such as 'Empathy Simulation' are sometimes used instead to refer to the 'artificial embodiment and display of empathic behaviours in virtual or robotic agents, which are perceived by human users' (Xiao et al. 2016: 7). It is important to mention, though, that other lines of research into 'Artificial Empathy' extend beyond robotics to include virtual agents of various kinds, and such systems do not necessarily involve embodiment in Asada's sense, nor do they necessarily include developmental stages such as those outlined in Fig. 2. For example, Liu-Thompkins et al. have recently defined 'Artificial Empathy' as 'the codification of human cognitive and affective empathy through computational models in the design and implementation of AI agents' (Liu-Thompkins et al. 2022). This formulation enables the authors to consider this subtype of empathy in relation to the social customer experience in AI-driven marketing.

While the research summarised above is of obvious importance, the many different denotations of the term 'Artificial Empathy' introduce an unhelpful vagueness. Therefore, it is crucial to re-emphasise that the present article is exclusively concerned with widely available state-of-the-art DSs. Currently, these systems are not physically embodied, and they do not acquire (artificial) empathy during a protracted process of affective and cognitive development. Rather, the most powerful systems (e.g. BlenderBot3 and ChatGPT) are simply neural-based pre-trained transformers that have been trained in sophisticated ways (e.g. using supervised learning and/or reinforcement learning) on vast amounts of human-derived conversational data. During this process, the core mathematical models learn many of the patterns contained in the data, and consequently, the trained systems are able to generate similar patterns in similar conversational contexts. It is, in effect, an elaborate form of parroting. BlenderBot3 may *seem* to express empathy if you tell it you have a headache, and (if pressed) it may even mention its own experience of headaches, but a well-trained parrot could do the same, without having a personal experiential understanding of your condition. Therefore, to avoid confusion, the phrase 'Artificial Empathy' will not be used in this article to refer to the kind(s) of empathy users might perceive in DSs.

The phenomenon of DSs claiming that they have headaches, sleep, own pets, experience pain, and so on requires further consideration. Such responses constitute a subtype of credibility fallacy: a statement is made yet the conditions of credibility are not satisfied as far as the interlocutor is concerned. In general, this occurs whenever DSs claim experience of a condition or state that they cannot possibly have experienced. Of course, human beings sometimes do this too, so the phenomenon is not confined to human–DS interactions. For instance, a credibility fallacy would occur if a biological male spoke about his own *personal* experience of period pains, or if a woman spoke about the actual death of her father to an interlocutor who knew for a fact that her father was still alive. In human interactions, this would be either a form of lying or possibly a sign of psychological disorder, but such terminology will be avoided in this article since such classifications raise issues of intentionality that are complex and contentious in relation to DSs. So, in the ensuing discussion, credibility fallacies will be understood to occur whenever a DS (or anything else) outputs a response referring to its own experience that causes the user to think 'but I know with certainty that can't be true!'. In the context of empathetic interactions particularly, if a DS system outputs this type of response, the consequences of credibility fallacies can be twofold: creating cognitive dissonance but also serving to trivialise the emotional states and experiences that are being discussed (Concannon et al. 2023). And the distinction between utterances that are credibility fallacies and those that are not can sometimes be quite subtle, depending on the linguistic structures used. For instance, if a DS responds to the user input 'I can't sleep' with the sentence 'Have you tried chamomile tea? Some people say it can help you sleep', then there is no glaring credibility fallacy since the system is simply using reported speech and is not claiming personal experience of the recommended remedy. However, a response such as 'Have you tried chamomile tea? It often helps me when I can't sleep' would introduce a credibility fallacy for most users, since the current generation of state-of-the-art DSs neither sleep nor drink. This issue is important since credibility fallacies can decrease the degree of PE in human–human interactions while reducing the interlocutor's ability to perceive the other person's emotional state (Lee et al. 2019). Consequently, they are likely to have at least a similarly negative impact on human–DS interactions.

To summarise, therefore, the following bullet points itemise some of the distinctive properties of the current

generation of widely available DSs that relate most closely to the topic of empathy:

- They are not physically embodied in a human-like manner (i.e. they do not have a corporeal form through which perception is mediated, they do not have a central nervous system, they do not have senses of taste, smell, touch, and so on).
- They have not acquired any kind of empathy as a result of a protracted process of affective and cognitive development that approximates the manner in which humans acquire empathy
- They can communicate only using written or spoken inputs and outputs; therefore, the kinds of paralinguistic and non-verbal gestures that are common in face-to-face human conversations and which often convey empathy (e.g. whistling, smiling, frowning, nodding) cannot feature in conversations, other than through rough typed approximations (e.g. 'lol', , ☹) or spoken descriptions (e.g. 'I'm smiling now', 'I'm rolling my eyes').
- Since they are trained on human-produced data, they tend to output credibility fallacies that risk decreasing the degree of perceived empathy they inculcate in the interlocutor.

These properties place certain constraints on human–DS interactions. This means that none of the existing empathy metrics summarised in Sect. 2 (which were all designed to assess *human–human* interactions) can be used in an unadapted form to determine the degree of empathy being displayed by a DS. For example, the CARE measure introduced by Mercer et al. 2005 is a patient-focussed metric that requires the patient to assess the doctor in relation to statements such as:

> How was the doctor really listening (paying close attention to what you were saying, not looking at the notes or the computer as you were talking)? (Mercer et al. 2005)

Such questions are largely irrelevant when an interaction between a human and a DS is being assessed. This is because the latter interactions are not specifically medical in nature (i.e. the user is not usually a patient speaking to a DS doctor). In addition, as mentioned above, the DS cannot use physical paralinguistic gestures (e.g. looking down at notes, looking at a computer screen), therefore assessing such things is pointless (in this context).

The lack of any widely accepted empathy metric for human–DS interactions has created a scenario in which the degree of 'empathy' associated with DSs is quantified in markedly different ways. And this pervasive multiplicity has unfortunately fostered the conviction that 'measuring the empathy of chatbot replies' is a task that can be accomplished with reasonable accuracy and effectiveness (Cameron et al 2017). Yet considerable caution is needed here since if the denotation of 'empathy' is uncertain in human–human interactions, it becomes even more nebulous when used to describe human–DS conversations. As mentioned above, Microsoft's XiaoIce is explicitly described as being an 'Empathetic Social Chatbot'—but what does that actually mean in practice? At the level of the system's architecture, it means that an Empathy Computing Module automatically processes a given user's input statement or query, Q, and (i) rewrites Q to its contextual version Qc by taking the dialogue context C into account, then (ii) encodes the user's states and feelings in the query empathy vector eQ, and finally (iii) specifies the empathetic aspects of the system's response R with the response empathy vector eR. The degree of empathy manifested by the system is measured by quantifying the 'Conversation-turns Per Session' (CPS) and the Number of Active Users (NAU).[7] As Zhou et al. put it, 'XiaoIce aims to pass a particular form of the Turing Test', a socially, rather than functionally, motivated assessment, which they refer to as 'the time-sharing test, where machines and humans coexist in a companion system […] If a person enjoys its companionship (via conversation), we can call the machine "empathetic"' (Zhou et al. 2020, 3). This conceptualisation unhelpfully conflates empathy and engagement: a user may engage with the system for a long time, just as they may play a computer game all day, but that does not indicate that either is in any sense 'empathetic'. The assumption that CPS necessarily correlates positively with engagement is not unreasonable, but to suggest that this measure of interaction duration automatically confers an *empathetic* status upon the DS *is* inaccurate. Conversation length can vary due to a number of factors, such as user identity (Leino et al 2020), or discursive quality (Concannon et al 2015).[8] Similarly, simple interventions, such as asking more questions, could lead to an increase in CPS, without having an impact on empathetic quality. It is extremely misleading, therefore, to use CPS and NAU as empathy measures.

To consider, briefly, an alternative more representative evaluation framework, Zhu et al. introduce a multi-party empathetic dialogue generation (i.e. many-to-many rather than 1-to-1 dialogues), and they determine the quality of their outputs using two different kinds of metrics (Zhu et al. 2022, 303):

[7] In this context, a conversation-turn (i.e. a turn at talking) refers to each individual message sent, so the CPS is the total number of messages exchanged over the course of each interaction.

[8] For example, Leino et al. (2020) found that high school students' dialogues have a higher CPS than those of university students or staff, while Concannon et al. (2015) found that increased disagreement in dialogues led to higher CPS.

- **Automatic Evaluation Criteria**:
  - Average of BLEU 1-2-3-4
  - ROUGE-L

  } These both use a human-generated gold standard reference
- **Human Evaluation Criteria**:
  - 100 dialogue samples and corresponding system outputs selected at random and human annotators rate the following attributes (on a scale of 1-5): Empathy, Relevance, Fluency

When assessing the 'Empathy' attribute in the human evaluation, the annotators must determine whether 'the speaker of the response understands the feelings of others and fully manifests it' (Zhu et al. 2022, 303). This guidance is considerably vaguer and less specific than the guidance given to second- and third-person assessors in human–human scenarios when well-defined empathy measures are used (e.g. the CARE measure). It is not clear why the assessment of PE in DSs should be accomplished in a far more parsimonious fashion. It also raises conceptual problems, since a given user may well believe that a DS cannot really 'understand' anything at all, and therefore may give low scores for that reason. The problem is that the empathy measure does not clarify whether the focus is supposed to be on determining IE or PE. Even in human–human interactions we can never know *for certain* what the other person actually understands. We can only try to determine that, indirectly, from the responses we receive. Further, it is not clear why 'Relevance' and 'Fluency' are obviously useful properties in this context. In human–human interactions, a response that has a high degree of affective empathy might be far from fluent. For instance, the person responding empathetically might be moved to tears, and they may use filled pauses and backchannels extensively: 'Well, … um … I'm … I don't know what … um … you need to … to … uh … the most important thing … um.. is to … well … look after yourself'. An utterance of this kind is far from fluent, but, in the relevant context, it is highly likely to be interpreted as extremely empathetic.

These two illustrative scoring frameworks have been selected from a huge number of possibilities, but hopefully, they are sufficient to indicate that there is currently no widely-accepted evaluation method for determining the degree of empathy human users perceive in a DS. While some claims in the published literature about high degrees of empathy are based on crude CPS and NAU counts, others present results obtained from automated metrics (such as BLEU and ROUGE-L) in addition to some kind of questionnaire-based human assessment. This lack of a shared evaluation framework is undesirable since it makes it impossible to compare and contrast in a convincingly systematic manner the degree of (perceived) empathy manifest by different DSs.

4 Measuring perceived empathy in human–DS interactions

As mentioned in Sect. 2, there are numerous methods for measuring PE in humans—but these are not fit for purpose when used to assess DSs. Also, as noted above, the most widely-used empathy measures involve self-report: participants or observers indicate alignment with a set of statements using a Likert scale, and the measures all quantify PE whether they take the form of first-person assessments (i.e. a questionnaire completed by the individual being assessed), participant-rating second-person assessments (i.e. a questionnaire completed by the other participant about the empathy of the participant being assessed), or observer-rating third-person assessments (i.e. a questionnaire completed by a non-participant observer about the empathy of the participant being assessed). These measures fall into two broad categories, depending on who completes the report: a participant-observer or a non-participant observer. Since there are no existing metrics for quantifying the degree of IE possessed by a DS, frameworks which determine the extent to which an entity is *perceived* as possessing empathy seem most appropriate when the performance of DSs is being analysed. Although there are notably fewer frameworks focussing on second and third-person assessments, Hemmerdinger et al. (2007) concluded that they are more reliable than first-person frameworks—particularly in medical contexts, where the objective of improving empathetic communication is directly tied to improving patient care. In the context of DSs, the perspective of the human interlocutor, or an independent observer, must necessarily be the primary focus when assessing the extent to which an automated system is capable of engendering PE, since the current generation of DSs cannot subjectively assess their own performance self-reflexively in a meaningful manner. In addition, since the current generation of DSs are predominantly language-based (i.e. they use text-to-speech and/or speech-to-text as input and output), a given system's PE will arise almost entirely

through its *linguistic* behaviour. It is true that aspects of its design may contribute to it seeming to be empathetic, (e.g. the colours and design of the user interface, the font type used), but it is especially crucial to determine how the linguistic form and content of the system's responses influence the degree of PE it prompts in humans.

Given the centrality of language in human dialogue, it is surprising that so few existing studies of empathetic human dialogue have focussed primarily on *linguistic* phenomena. Suchman et al. (1997) explore the 'interactional sequences that constitute empathy in action', while Pounds (2011) presents a discourse-pragmatic approach for evaluating empathy in the context of clinical communication, by examining the `verbal realisation of empathy' (Pounds 2011, 139). Discourse-pragmatic approaches use recordings or transcripts as observational data, to understand how different forms of empathetic behaviour are conveyed through communication. This method usefully provides more fine-grained analyses than high-level reporting-focussed measures. Also, the emphasis on the interactional consequences of particular response constructions is beneficial. For example, Suchman et al. highlight the importance of more implicit linguistic cues, referred to as 'potential empathy opportunities', that enable a clinician 'infer an emotion that has not been explicitly expressed' (Suchman et al. 1997, 679). Doctors who miss such opportunities, directing the dialogue away from the implied emotion, as opposed to inviting the patient to expand, are viewed as less adept or satisfactory. Consequently, how a doctor forms a response to a patient's statement will influence the degree of PE associated with the dialogue. Pound's work looks even more closely at the specific linguistic constructions used to achieve some of the interactional sequences outlined in Suchman et al. (1997). For example, she examines how verbs of acknowledgement (e.g. 'I understand/see/realise/appreciate that') and adjectival constructions expressing understanding (e.g. 'it is clear/apparent to me that…') are used to demonstrate responsiveness to a potential empathy opportunity, and how uncertainty markers (e.g. hedges, modals) can be used to elicit a patient's feelings and views (Pounds 2011, 154–155).

Shifting the focus from human–human interactions back to human–DS interactions, it is curious that there have been so few studies of the *linguistic* structures that DSs use to inculcate PE in users or observers. As mentioned in Sect. 3, most studies have typically relied on ad hoc processes or crude automatic measures that conflate empathy with other aspects of the interactions (e.g. conversation length). Fitzpatrick et al. (2017) discuss users' perceptions of Woebot as empathetic, based on comments volunteered in free-form text entries in a questionnaire about the user's overall experience of interacting with Woebot, while Morris et al. (2018) only asked users to rate their interactions with the automated system as being either *good, ok*, or *bad*. Zhou et al.'s problematical use of CPS and NAU has already been discussed in Sect. 3; and Rashkin et al. (2019) adopted automatic measures computed using perplexity and BLEU scores, where a gold label response (i.e. one given by a human) is compared to that generated by the DS. While such measures have undoubtedly facilitated the development of many different language-based systems, their correlation with human judgements is known to be glaringly weak (Liu et al. 2016).

Clearly, the lack of established measures for assessing PE in DSs makes validating any claim that a given system is 'empathetic' extremely challenging. One reasonable response to this is to adapt an existing second- or third-person human-focussed empathy measure to provide a quantitative assessment of PE in interactions with automated systems. As far as we are aware, the only paper that has implemented this to date isPutta et al. 2022,[9] which introduces a second-person questionnaire, based on the RoPE scale proposed in Charrier et al. 2019.[10] They use a Likert scale ranging from -3 to 3, and the prompts in the questionnaire include such things as:

> Q1: The artificial agent/robot appreciates exactly how the things I experience feel to me.
> Q11: The artificial agent/robot comforts me when I am upset.
> Q24: The artificial agent/robot's appearance/audio is pleasant, good, and inviting.
> (Putta et al. 2022, 702)

Although this framework provides a useful starting point, there are various limitations to the approach when it is considered in relation to DSs. For instance, the second-person emphasis of the questionnaire means that each conversational interaction can only be assessed once, by the human who participated in it. Capturing second-person responses is undeniably important; the participant involved in an interaction can provide useful assessments of whether an interlocutor was perceived as empathetic. However, there are practical challenges to this in relation to designing DSs. The mathematical models and feedback loops used by many state-of-the-art DSs ensure that the very same prompt will not normally produce exactly the same response from the system, which makes it impossible to obtain multiple assessments of the same interaction since differences in performance can occur simply by chance. Tianbo Ji et al. (2022) have

[9] Additionally, Yalcin & DiPaola (2020) use an adapted version of the first-person Toronto Empathy Questionnaire for second-person usage when evaluating the M-Path system, but no information about the adaptations, nor the questionnaire itself, are detailed in the paper.

[10] Putta et al. claim they are measuring 'artificial empathy'. However, given the vagueness around the use of this phrase in the literature (as discussed in Sect. 3 above), it is not clear exactly what this denotes.

recently summarised the various problems that beset the formal evaluation of open-domain DSs, emphasising that this task remains an open problem due to the huge diversity of automated metrics used by different research teams as well as the difficulty of obtaining reliable and consistent human evaluations. Given this scenario, it is potentially beneficial if *multiple* human assessors can evaluate the *same* human–DS dialogue from a third-person perspective, since this helps to demonstrate, statistically, that one DS is more empathetic than another. In addition, some of the questions used by Putta et al. support multiple interpretations. In Q24 above, what does it mean for the 'audio' to be 'pleasant, good, and inviting'? Does this simply mean that the signal-to-noise ratio is appropriate? Also, responses to the questions may be informed by various features of a robot's design, from appearance to audio quality, as well as the dialogue itself. As mentioned earlier, many interactions with DSs take the form of typed inputs, and these involve neither 'appearance' nor 'audio' (unless the denotation of appearance is sufficiently stretched to include things such as the type, colour, and size of the font).

To overcome these limitations, the proposal in the current article is that a third-person measure for PE in DSs is preferable to a first- or second-person measure since it enables multiple individuals to assess the same human–DS interaction. Also, language-focussed questions are desirable, since that is currently the primary (usually sole) medium that enables DSs to be perceived as being empathetic. While seeking to develop a measure of this kind, it makes sense to take an existing human–human measure as a starting point. For instance, the TES (Decker et al. 2014) is an observer-focussed empathy assessment that was adapted from the Measure of Expressed Empathy (Watson 1999). It is designed to explore 'the observable and overlapping cognitive, affective, attitudinal, and attunement aspects of therapist empathy', and it uses high-level descriptors of therapist behaviour assessment items. For instance, 'a therapist provides ample opportunities for the client to explore his or her emotional reactions' (Decker et al. 2014, 344–345). To demonstrate the feasibility of our proposal, we adapted the TES framework so that it can be used to assess the interaction humans have with DSs, enabling a non-participant observer to evaluate the empathy enacted by the system over the course of a dialogue. There are comparatively few third-person scales (as opposed to first- or second-person evaluation measures), so TES was selected as it utilises the observer perspective and has been evaluated as one of the more reliable empathy measures (Hong and Han 2020).

The Empathy Scale for Human–Computer Communication (ESHCC) is presented in Table 1. Following TES, each item on the scale will be rated by the observer using a 7-point Likert-type scale (1 = not at all, to 7 = extensively). Assessment items were adapted so that the framework is suitable for evaluating general *text-based* interactions or transcripts of *voice-based* interactions. Lexical, textual and syntactic features such as punctuation, emoticons, capitalisation, word and phrases are therefore more relevant than communication cues used in verbal and face-to-face interaction (e.g. tone of voice). Consequently, items require the observer to attend to certain linguistic features (referred to as vocabulary and syntax below) rather than qualities of delivery (e.g. `the therapist's voice has a soft resonance').

To give a concrete example of how the scale has been modified, in the TES the item 'Responsiveness' is described as follows:

> A therapist shows responsiveness to the client by adjusting his or her responses to the client's statements or nonverbal communications during the conversation. The therapist follows the client's lead in the conversation instead of trying to steer the discussion to the therapist's agenda or interests. (Decker et al. 2014, 15)

In the ESHCC this has been modified so that it can be used for general human–DS interactions, which means that the phrases used to denote the participants have been changed, and the reference to 'nonverbal communications' has been removed:

> The system shows responsiveness to the interlocutor by adjusting its responses to the interlocutor's statements during the conversation. The system follows the interlocutor's lead in the conversation instead of trying to steer the discussion to its own agenda or interests.

Additionally, as the ESHCC framework is intended for third party observers, the emphasis is placed on the *perception* of empathetic behaviour in the dialogue participants. This is most plainly signalled through the inclusion of inferential evidentials (e.g. 'the system seems to…', 'the response suggests…'), and this intentionally focuses the evaluation on whether the DS, and the utterances produced, create a perceptible display of behaviours that are observable in the language-use, and therefore are recognisable as empathy.

While the TES is tailored to therapeutic dialogues, the ESHCC has been designed to accommodate non-clinical interactions. We retain references to 'attunement', which is predominantly employed in relation to therapeutic interactions, yet we expand the interpretation to include related concepts more commonly used in non-clinical settings, such as 'alignment' (Branigan et al. 2010). Finally, an additional item, 'Fallacy Avoidance', is introduced in the ESHCC. As outlined above, credibility fallacies can negatively impact

Table 1 The Empathy Scale for Human–Computer Communication (ESHCC)

Item	Item Description	Empathy
Concern	The system conveys concern by seeming to show regard for, and interest in, the interlocutor. The system uses vocabulary and syntax which give the impression that it is involved with the interlocutor and attentive to what the interlocutor has said	Component Attitudinal Attunement
Expressiveness	The system seems to vary its vocabulary and syntax to demonstrate expressiveness and modify its responses to accommodate the mood or disposition of the interlocutor	Attunement
Resonate or acknowledge interlocutor feelings	The system's responses seem to resonate with or capture, the intensity of the interlocutor's feelings by explicitly acknowledging them or by using vocabulary and syntax that match the interlocutor's emotional state or underscores how the interlocutor feels	Affective
Warmth	The system demonstrates warmth by communicating in a manner that seems friendly, cordial, and sincere. The system seems to be involved with and supportive of the interlocutor's efforts to express themself. The system seems kindly disposed toward, or fond of, the interlocutor	Attitudinal
Attuned to interlocutor's inner world	An interlocutor's inner world is defined as the interlocutor's feelings, perceptions, memories, meanings, bodily sensations, and core values. The system displays attunement to an interlocutor's inner world when it seems attentive to nuances of meaning and feeling conveyed in an interlocutor's statements beyond surface content and shows a genuine understanding of the interlocutor's inner world	Cognitive Affective Attunement
Understanding cognitive framework	The system seems to demonstrate an understanding of the interlocutor's cognitive framework and meanings by showing that it follows what the interlocutor has said and accurately reflects this understanding to the interlocutor. The system provides opportunities for the interlocutor to state his or her views to permit the fullest and most accurate understanding of the interlocutor. The interaction suggests that the system seems to value knowing what the interlocutor means or intends by his or her statements	Cognitive
Understanding feelings/inner experience	The system seems to convey an understanding of the interlocutor's feelings and inner experience by showing a sensitive appreciation for the interlocutor's emotional state. The system provides opportunities for the interlocutor to explore his or her emotional reactions. The system seems to accurately reflect how the interlocutor feels by appropriately labelling feeling states with words (e.g. anger, sadness, frustration, etc.), or metaphors (e.g. "It's as if you are pent up and feel about to explode") to clarify and crystallise for the interlocutor what he or she is experiencing emotionally	Affective
Acceptance of feelings/inner experiences	The system seems to show acceptance of the interlocutor's feelings and inner experience when it validates the interlocutor's experience and reflects the interlocutor's feelings without judgement or a dismissive attitude. The agent is unconditionally open to and respectful of how the interlocutor feels	Affective Attitudinal
Responsiveness	The system shows responsiveness to the interlocutor by adjusting its responses to the interlocutor's statements during the conversation. The system follows the interlocutor's lead in the conversation instead of trying to steer the discussion to its own agenda or interests	Attunement
Fallacy avoidance	The system consistently avoids credibility fallacies by not making claims about its own personal experience that are blatantly implausible	Cognitive

on PE and are more likely to arise in dialogues with a DS due to the inherent asymmetries that arise from the system's lack of experiences to draw upon (Concannon et al. 2023).

5 Considerations and future work

There are a number of factors to consider when contemplating the design and application of the ESHHC. While items in the TES are designed with an *optimal form* of empathic interaction in mind, ESHCC is conceived as an assessment tool for better-understanding empathy in human–DS communication—that is, we are not necessarily suggesting that an interaction that scores 7 on each item in the scale denotes a preferred form of PE. Guzman and Lewis (2020) emphasise that human–DS communication is distinct from human–human communication and should be studied in a way that attends to the potential differences in how machines are conceptualised and function as communicative partners, in contrast to humans (Guzman and

Lewis 2020, 76). As ESHCC is an adaptation of a framework designed for human–human interactions, it is necessary to evaluate the extent to which the forms of empathy valued in human–human interaction persist in human-system dialogues. For example, Urakami et al. (2019), found that some forms of empathy (e.g. when the system expressed its own feelings) were more problematic for certain end-users than other forms; but they also found that individuals differed in their preferences. As the authors remarked, '[i]ntegrating expressions of empathy in human–machine interaction is a sensitive issue and designers must carefully choose what components of empathy are adequate depending on the situational circumstances and the targeted user group' (Urakami et al. 2019, 11). Consequently, an empathetic utterance performed by a human may be received differently when performed by a DS. An understanding of these differences, and the associated implications (e.g. how this influences a user's trust in a system), is yet to be established.

While it is unclear what elements of empathetic communication users want from their DSs, greater clarity at both a conceptual and implementation level is necessary. Nonetheless, it is possible to begin establishing the linguistic behaviours that convey PE in the specific context of human–DS interaction. Further, closer integration of discourse pragmatic accounts of empathetic interactions could provide the foundations for a more fine-grained understanding of how empathy operates in such interactions and the development of a more standardised approach for assessing the empathetic outputs of dialogue agents in a more systematic manner. The approach taken in ESHCC requires observers to assess the entirety of a conversation. A complementary framework could be designed for turn-level analysis to offer more granular insights. Nonetheless, the ESHCC offers a form of standardisation that could provide a benchmark for cross-system comparisons.

Applying ESHCC will inevitably be more labour intensive than existing automated measures. However, as the inclusion of some form of human evaluation is becoming more common practice, developing a uniform approach should provide more meaningful insights. A focussed evaluation of the ESHCC items with a wider pool of annotators will help to ensure ease of use and consistency in application. Perceptions of empathy vary across individuals, so reconciling this with measures of inter-rater reliability and internal consistency or more perspectivist approaches to data annotation will need to be considered and limitations acknowledged. A large-scale study of the applied use of the ESHCC is the first step to address these issues. The resulting corpus of conversations annotated for perceived empathy may additionally generate new knowledge to inform novel approaches to the automated measurement of empathy in DS. While it will first need to be subjected to a rigorous validation study, the ESHCC has the potential to facilitate more informative comparisons between the PE associated with human and automated interlocutors in conversational situations.

6 Conclusion

Empathy is undoubtedly a problematical term. Multiple non-equivalent definitions of it are regularly used by psychotherapists, sociologists, philosophers, social neuroscientists, primatologists, developmental psychologists, clinicians, computer scientists, and many others. Nonetheless, despite its daunting polysemic tendency, the term remains an important one in analyses of human–human interactions; and, ever since the 1940s, many different empathy measures have been proposed. With extensive reference to this existing body of research, this article has addressed the topic of how best to assess the degree to which automated DSs can be classified as manifesting empathy. Given recent advances in machine learning, this topic is becoming increasingly important since numerous language-based AI systems, ranging from VPAs, to social chatbots, to therapeutic dialogue systems, are described by their creators as being 'empathetic'. In essence, the task for the system designer is to create a system that a human user, or a third-person observer, perceives as being empathetic. For current state-of-the-art DSs, it is the *linguistic* responses the system generates that enable PE to be assessed. Despite numerous remarkable technological advances in DS-related research and affective computing in recent years, there is currently still no single standard metric for measuring the PE in human–DS interactions. Existing quantification methods either use overly reductive indicators (such as CPS and NAU), that has nothing to do with any accepted definitions of 'empathy', or they use automated metrics borrowed from other language technology tasks (e.g. BLEU, ROUGE-L) and supplement them with simple questionnaires (usually second-person ones) that require human assessors to focus to on properties such as 'Empathy', 'Relevance', and 'Fluency' (despite the fact that, in human–human interactions, extremely empathetic responses can often be disfluent). This kind of evaluation framework is markedly different from how degrees of PE have been studied in human–human interactions over many decades.

Responding to this anomalous state of affairs, this article has sought to introduce greater precision into the ongoing discussions about this intricate topic by arguing that a third-person measure of the degree of PE conveyed by a system's linguistic responses during a human–DS interaction is the most desirable kind of metric – and ideally one using a scale that has been adapted from an existing measure originally designed to assess empathy in human–human interactions. This pragmatic emphasis on third-person assessments of PE usefully avoids a considerable number

of thorny technological and philosophical debates about consciousness, volition, understanding, and intentionality in relation both to empathy and to automated systems. Accordingly, the measure for human–DS communication proposed here, ESHCC, is an adapted version of an existing observer-focussed measure, TES, that is widely used to quantify the degree of empathy in therapeutic human–human interactions. The obvious next step will be to undertake a rigorous validation study of the measure, but, assuming the results of that study are encouraging, it is hoped that the ESHCC will provide a robust framework for assessing the extent to which a DS can be described as being 'empathetic', and that this, in turn, will facilitate much more meaningful cross-system comparisons.

Acknowledgements This research is funded by the Humanities and Social Change International Foundation.

Author contributions Not applicable.

Funding This research is funded by the Humanities and Social Change International Foundation.

Availability of data and materials (data transparency) Not applicable.

Code availability (software application or custom code) Not applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Asada M (2015a) Towards artificial empathy. Int J Soc Robot 7(1):19–33. https://doi.org/10.1007/s12369-014-0253-z

Asada M (2015b) Development of artificial empathy. Neurosci Res 90:41–50. https://doi.org/10.1016/j.neures.2014.12.002

Asada M (2019) Artificial pain may induce empathy, morality, and ethics in the conscious mind of robots. Philosophies 4(3):38. https://doi.org/10.3390/philosophies4030038

Atkins D, Uskul AK, Cooper NR (2016) Culture shapes empathic responses to physical and social pain. Emotion 16(5):587–601. https://doi.org/10.1037/emo0000162

Baron-Cohen S (2011) Zero degrees of empathy: a new theory of human cruelty. Penguin, UK

Bassett C (2019) The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. AI Soc 34(4):803–812. https://doi.org/10.1007/s00146-018-0825-9

Batson CD, Ahmad NY (2009) Using empathy to improve intergroup attitudes and relations. Soc Issues Policy Rev 3(1):141–177. https://doi.org/10.1111/j.1751-2409.2009.01013.x

Batson CD (2009) These things called empathy: eight related but distinct phenomena. The social neuroscience of empathy. In: Decety J, Ickes W (eds) Social neuroscience. The social neuroscience of empathy. MIT Press, p 3–15. https://doi.org/10.7551/mitpress/9780262012973.003.0002

Batson CD, Lishner, DA, Stocks, EL (2015) The empathy–altruism hypothesis. In: Schroeder DA, Graziano WG (eds) The Oxford handbook of prosocial behavior. Oxford University Press, Oxford, p 259–268. https://doi.org/10.1093/oxfordhb/9780195399813.013.023

Bernhardt BC, Singer T (2012) The neural basis of empathy. Annu Rev Neurosci 35:1–23. https://doi.org/10.1146/annurev-neuro-062111-150536

Branigan HP, Pickering MJ, Pearson J, McLean JF (2010) Linguistic alignment between people and computers. J Pragmat 42(9):2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

Brave S, Nass C, Hutchinson K (2005) Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int J Hum Comput Stud 62(2):161–78. https://doi.org/10.1016/j.ijhcs.2004.11.002

Cameron G, Cameron D, Megaw G, Bond R, Mulvenna M, O'Neill S, Armour C, McTear M (2017) Towards a chatbot for digital counselling. In: Proceedings of the 31st international BCS human computer interaction conference (HCI 2017), vol 31. pp 1–7. https://doi.org/10.14236/ewic/HCI2017.24

Charrier L, Rieger A, Galdeano A, Cordier A, Lefort M, Hassas S (2019) The rope scale: a measure of how empathic a robot is perceived. In: 2019 14th ACM/IEEE international conference on human-rsobot interaction (HRI), p 656–657. https://doi.org/10.1109/HRI.2019.8673082

Chaves AP, Gerosa MA (2020) How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. Int J Hum-Comput Interact. https://doi.org/10.1080/10447318.2020.1841438

Chen H, Liu X, Yin D, Tang J (2017) A survey on dialogue systems: recent advances and new frontiers. ACM Sigkdd Explor Newsl 19(2):25–35. https://doi.org/10.1145/3166054.3166058

Cleckley H (1976) The mask of sanity, 5th edn. Mosby, St. Louis

Concannon S, Healey P, Purver M (2015) Shifting opinions: experiments on agreement and disagreement in dialogue. In: Proceedings of the 19th workshop on the semantics and pragmatics of dialogue (goDIAL), SEMDIAL, pp 15–23

Concannon S, Roberts I, Tomalin M (2023) An interactional account of empathy in human-machine communication. Hum-Mach Commun 6:87–116. https://doi.org/10.30658/hmc.6.6

Daher K, Saad D, Mugellini E, Lalanne D, Abou Khaled O (2022) Empathic and empathetic systematic review to standardize the development of reliable and sustainable empathic systems. Sensors 22(8):3046. https://doi.org/10.3390/s22083046

Davis MH (1983) Measuring individual differences in empathy: evidence for a multidimensional approach. J Pers Soc Psychol 44(1):113–126. https://doi.org/10.1037/0022-3514.44.1.113

De Vignemont F, Singer T (2006) The empathic brain: how, when and why? Trends Cogn Sci 10(10):435–441. https://doi.org/10.1016/j.tics.2006.08.008

Decety J, Svetlova M (2012) Putting together phylogenetic and ontogenetic perspectives on empathy. Dev Cogn Neurosci 2(1):1–24. https://doi.org/10.1016/j.dcn.2011.05.003

Decety J, Meidenbauer KL, Cowell JM (2018) The development of cognitive empathy and concern in preschool children: a behavioral

neuroscience investigation. Dev Sci 21(3):e12570. https://doi.org/10.1111/desc.12570
Decker SE, Nich C, Carroll KM, Martino S (2014) Development of the therapist empathy scale. Behav Cogn Psychother 42(3):339–354. https://doi.org/10.1017/S1352465813000039
Dial M (2018) Heartificial empathy: putting heart into business and artificial intelligence. N.p., DigitalProof Press
Doherty M (2008) Theory of mind: how children understand others' thoughts and feelings. Psychology Press
Domes G, Hollerbach P, Vohs K, Mokros A, Habermeyer E (2013) Emotional empathy and psychopathy in offenders: an experimental study. J Pers Disord 27(1):67–84. https://doi.org/10.1521/pedi.2013.27.1.67
Eisenberg N, Eggum ND (2009) Empathic responding: sympathy and personal distress. Soc Neurosci Empathy 6:71–83. https://doi.org/10.7551/mitpress/9780262012973.003.0007
Eisenberg N, Spinard TL, Knafo-Noam A (2015) Prosocial development. In: Lamb ME, Lerner RM (eds) Handbook of child psychology and developmental science, 7th edn, vol 3. John Wiley & Sons Inc, New York. https://doi.org/10.1002/9781118963418.childpsy315
Fitzpatrick KK, Darcy A, Vierhile M (2017) Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. JMIR Ment Health 4(2):e19. https://doi.org/10.2196/mental.7785
Frankel RM (2017) The evolution of empathy research: models, muddles, and mechanisms. Patient Educ Couns 100(11):2128–2130. https://doi.org/10.1016/j.pec.2017.05.004
Grondin F, Lomanowska A, Jackson P (2019) Empathy in computer-mediated interactions: a conceptual framework for research and clinical practice. Clin Psychol Sci Pract 26(4):1–17. https://doi.org/10.1111/cpsp.12298
Guzman AL, Lewis SC (2020) Artificial intelligence and communication: a human–machine communication research agenda. New Media Soc 22(1):70–86. https://doi.org/10.1177/1461444819858
Hall JA, Schwartz R (2019) Empathy present and future. J Soc Psychol 159(3):225–243. https://doi.org/10.1080/00224545.2018.1477442
Hemmerdinger JM, Stoddart SD, Lilford RJ (2007) A systematic review of tests of empathy in medicine. BMC Med Educ 7(1):1–8. https://doi.org/10.1186/1472-6920-7-24
Hills AH (2001) Empathy and offender behavior: The motivational context. In Traverso GB, Bognoli L (eds) Psychology and law in a changing world: new trends in theory, practice and research, Psychology Press, pp 51–64
Hoffman ML (1987) The contribution of empathy to justice and moral judgment. In: Eisenberg N, Strayer J (eds) Empathy and its development. Cambridge University Press, Cambridge, pp 47–80
Hojat M (2016) Empathy in health professions education and primary care. Springer International, Cham
Hojat M, Gonnella JS, Nasca TJ, Mangione S, Vergare M, Magee M (2002) Physician empathy: definition, components, measurement, and relationship to gender and specialty. Am J Psychiatry 159(9):1563–1569. https://doi.org/10.1176/appi.ajp.159.9.1563
Hojat M, DeSantis J, Shannon SC, Mortensen LH, Speicher MR, Bragan L, LaNoue M, Calabrese LH (2018) The Jefferson scale of empathy: a nationwide study of measurement properties, underlying components, latent variable structure, and national norms in medical students. Adv Health Sci Educ 23(5):899–920. https://doi.org/10.1007/s10459-018-9839-9
Hong H, Han A (2020) A systematic review on empathy measurement tools for care professionals. Educ Gerontol 46(2):72–83. https://doi.org/10.1080/03601277.2020.1712058
James J, Watson CI, MacDonald B (2018) Artificial empathy in social robots: an analysis of emotions in speech. In: 2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, p 632–637. https://doi.org/10.1109/ROMAN.2018.8525652
Jami Yaghoubi P, Mansouri B, Thoma SJ, Han H (2019) An investigation of the divergences and convergences of trait empathy across two cultures. Journal of Moral Education 48(2):214–229. https://doi.org/10.1080/03057240.2018.1482531
Jolliffe D, Farrington DP (2006) Development and validation of the Basic Empathy Scale. J Adolesc 29(4):589–611. https://doi.org/10.1016/j.adolescence.2005.08.010
Jütten LH, Mark RE, Sitskoorn MM (2019) Empathy in informal dementia caregivers and its relationship with depression, anxiety, and burden. Int J Clin Health Psychol 19(1):12–21. https://doi.org/10.1016/j.ijchp.2018.07.004
Lanzoni S (2018) Empathy: a history. Yale University Press
Lee JJ, Hardin AE, Parmar B, Gino F (2019) The interpersonal costs of dishonesty: how dishonest behavior reduces individuals' ability to read others' emotions. J Exp Psychol Gen 148(9):1557–1574. https://doi.org/10.1037/xge0000639
Leino K, Leinonen J, Singh M, Virpioja S, Kurimo M (2020) FinChat: corpus and evaluation setup for finnish chat conversations on everyday topics. Proc. Interspeech 2020. p 429–433. https://doi.org/10.21437/Interspeech.2020-2511
Lin C-Y, Och FJ (2004) Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd annual meeting of the association for computational linguistics. p 605–612. https://doi.org/10.3115/1218955.1219032
Liu CW, Lowe R, Serban IV, Noseworthy M, Charlin L, Pineau J (2016) How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: Proceedings of the 2016 conference on empirical methods in natural language processing. p 2122–2132. https://doi.org/10.18653/v1/D16-1230
Liu-Thompkins Y, Okazaki S, Li H (2022) Artificial empathy in marketing interactions: bridging the human-AI gap in affective and social customer experience. J Acad Mark Sci 50:1198–1218. https://doi.org/10.1007/s11747-022-00892-5
Ma Y, Nguyen K, Xing F, Cambria E (2020) A survey on empathetic dialogue systems. Inf Fusion. https://doi.org/10.1016/j.inffus.2020.06.011
Maibom HL (2017) Introduction to philosophy of empathy. The routledge handbook to philosophy of empathy. Routledge, New York, pp 1–10
Marshall WL, Hudson SM, Jones R, Fernandez YM (1995) Empathy in sex offenders. Clin Psychol Rev 15(2):99–113. https://doi.org/10.1016/0272-7358(95)00002-7
McStay A (2018) Emotional AI: the rise of empathic media. Sage
Mercer SW, Maxwell M, Heaney D, Watt GC (2004) The consultation and relational empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. Fam Pract 21(6):699–705. https://doi.org/10.1093/fampra/cmh621
Mercer SW, McConnachie A, Maxwell M, Heaney D, Watt GC (2005) Relevance and practical use of the consultation and relational empathy (CARE) measure in general practice. Fam Pract 22(3):328–334. https://doi.org/10.1093/fampra/cmh730
Morris RR, Kouddous K, Kshirsagar R, Schueller SM (2018) Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J Med Internet Res 20(6):101–148. https://doi.org/10.2196/10148
Neumann DL, Chan RC, Boyle GJ, Wang Y, Westbury HR (2015) Measures of empathy: self-report, behavioral, and neuroscientific approaches. In: Measures of personality and social psychological constructs. Academic Press, p 257–289. https://doi.org/10.1016/B978-0-12-386915-9.00010-3

OpenAI, ChatGPT (2022) Optimizing language models for dialogue. https://web.archive.org/web/20230205020039/https://openai.com/blog/chatgpt/. Accessed 30 Nov 2022

Otterbacher J, Ang CS, Litvak M, Atkins D (2017) Show me you care: trait empathy, linguistic style, and mimicry on facebook. ACM Trans Internet Technol 17(1):1–22. https://doi.org/10.1145/2996188

Paiva A, Leite I, Boukricha H, Wachsmuth I (2017) Empathy in virtual agents and robots: a survey. ACM Trans Interact Intell Syst (TiiS) 7(3):1–40. https://doi.org/10.1145/2912150

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 311–318. https://doi.org/10.3115/1073083.1073135

Pfabigan DM, Seidel EM, Wucherer AM, Keckeis K, Derntl B, Lamm C (2015) Affective empathy differs in male violent offenders with high-and low-trait psychopathy. J Pers Disord 29(1):42–61. https://doi.org/10.1521/pedi_2014_28_145

Picard R (1997) Affective computing. The MIT Press, Cambridge

Pounds G (2011) Empathy as 'appraisal': developing a new language-based approach to the exploration of clinical empathy. J Appl Linguist Prof Pract 7(2):139–162. https://doi.org/10.1558/japl.v7i2.145

Poyatos F (1993) Paralanguage a linguistic and interdisciplinary approach to interactive speech and sounds. John Bemjamins, Amsterdam

Putta H, Daher K, Kamali ME, Khaled OA, Lalanne D, Mugellini E (2022) Empathy scale adaptation for artificial agents: a review with a new subscale proposal. In: 8th International conference on control, decision and information technologies, Istanbul, Turkey, p 699–704. https://doi.org/10.1109/CoDIT55151.2022.9803993

Raamkumar AS, Yang Y (2022) Empathetic conversational systems: a review of current advances, gaps, and opportunities. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2022.3226693

Rashkin H, Smith EM, Li M, Boureau, YL (2019) Towards empathetic open-domain conversation models: a new benchmark and dataset. In: Proceedings of the 57th annual meeting of the association for computational linguistics. p 5370–5381. https://doi.org/10.18653/v1/P19-1534

Robinson EV, Rogers R (2015) Empathy faking in psychopathic offenders: the vulnerability of empathy measures. J Psychopathol Behav Assess 37(4):545–552. https://doi.org/10.1007/s10862-015-9479-9

Shamay-Tsoory S (2015) The neuropsychology of empathy: evidence from lesion studies. Rev Neuropsychol 7(4):237–243

Shuster K, Xu J, Komeili M, Ju D, Smith EM, Roller S, Ung M, Chen M, Arora K, Lane J, Behrooz M, Ngan W, Poff S, Goyal N, Szlam A, Boureau Y-L, Kambadur M, Weston J (2022) 'BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage'. https://arxiv.org/pdf/2208.03188.pdf. Accessed 25 June 2023

Singer T, Lamm C (2009) The social neuroscience of empathy. Ann N Y Acad Sci 1156:81–96

Stephan A (2015) Empathy for artificial agents. Int J of Soc Robotics 7:111–116. https://doi.org/10.1007/s12369-014-0260-0

Strayer J (1987) Affective and cognitive perspectives on empathy. In: Eisenberg N, Strayer J (eds) Empathy and its development. Cambridge University Press, Cambridge, pp 218–244

Suchman AL, Markakis K, Beckman HB, Frankel R (1997) A model of empathic communication in the medical interview. JAMA 277(8):678–682

Tianbo J, Graham Y, Jones G, Lyu C, Liu Q (2022) Achieving reliable human assessment of open-domain dialogue systems. In: Proceedings of the 60th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 6416–6437. https://doi.org/10.18653/v1/2022.acl-long.445

Titchener E (1909) Elementary psychology of the thought processes. Macmillan, New York

Urakami J, Moore BA, Sutthithatip S, Park S (2019) Users' perception of empathic expressions by an advanced intelligent system. In: Proceedings of the 7th international conference on human-agent interaction. Association for Computing Machinery, pp 11–18. https://doi.org/10.1145/3349537.3351895

van Dijke J, van Nistelrooij I, Bos P, Duyndam J (2020) Towards a relational conceptualization of empathy. Nurs Philos 21(3). https://doi.org/10.1111/nup.12297

van Dongen JD (2020) The empathic brain of psychopaths: from social science to neuroscience in empathy. Front Psychol. https://doi.org/10.3389/fpsyg.2020.00695

Viding E, McCory E, Seara-Cardoso A (2014) Psychopathy. Curr Biol 24(18):871–874. https://doi.org/10.1016/j.cub.2014.06.055

Walter H (2012) Social cognitive neuroscience of empathy: concepts, circuits, and genes. Emot Rev 4(1):9–17. https://doi.org/10.1177/1754073911421379

Wynn R, Wynn M (2006) Empathy as an interactionally achieved phenomenon in psychotherapy: characteristics of some conversational resources. J Pragmat 38(9):1385–1397. https://doi.org/10.1016/j.pragma.2005.09.008

Xiao B, Imel ZE, Georgiou P, Atkins DC, Narayanan SS (2016) Computational analysis and simulation of empathic behaviors: a survey of empathy modeling with behavioral signal processing framework. Curr Psychiatry Rep 18(5):49. https://doi.org/10.1007/s11920-016-0682-5

Yalçın ÖN (2019) Evaluating empathy in artificial agents. In: 2019 8th International conference on affective computing and intelligent interaction (ACII). IEEE, p 1–7. https://doi.org/10.1109/ACII.2019.8925498

Yalçın ÖN, DiPaola S (2020) Modeling empathy: building a link between affective and cognitive processes. Artif Intell Rev 53:2983–3006. https://doi.org/10.1007/s10462-019-09753-0

Zhou L, Gao J, Li D, Shum H-Y (2020) The design and implementation of xiaoice, an empathetic social chatbot. Comput Linguist 46(1):53–93. https://doi.org/10.1162/coli_a_00368

Zhu LY, Zhang Z, Wang J, Wang H, Wu H, Yang Z (2022) Multi-party empathetic dialogue generation: a new task for dialog systems. In: Proceedings of the 60th annual meeting of the association for computational linguistics, Association for Computational Linguistics, pp 298–307. https://doi.org/10.18653/v1/2022.acl-long.24