Towards Empathic Machines: Introducing the Empathy Evaluation Questionnaire (EEQ)

Ben Anyasodo[1], David R. Large[1], Gary Burnett[2] & Robert Houghton[1]

[1]Human Factors Research Group, University of Nottingham, [2]School of Design and Creative Arts, Loughborough University

SUMMARY

The Empathy Evaluation Questionnaire (EEQ) is a questionnaire-based instrument designed to evaluate perceived empathic interaction in speech-based human-machine interfaces. The paper presents the conceptual grounding, development, and empirical evaluation of the EEQ, including evidence of internal consistency, construct validity, and latent structure. Rather than treating empathy as a stable psychological trait or an intrinsic system property, the EEQ is positioned as a measure of how empathic qualities are recognised by users during interaction. Results from a large-scale online survey (n=182) demonstrate robust psychometric properties and reveal interactionally coupled dimensions of user judgement. A short-form version of the EEQ is also proposed to support applied and time-constrained evaluation contexts. Implications for the design and evaluation of empathic conversational systems are discussed.

KEYWORDS

Empathy, Human Machine Interactions, Conversational User Interface, Questionnaire

Introduction

Empathy is a multifaceted construct commonly defined as the capacity to relate to, understand, and respond to the experiences of others (Raamkumar & Yang, 2023). In human-human interaction, empathic understanding is communicated through language, timing, tone, and responsiveness, enabling interlocutors to infer internal states and respond appropriately (Ayanoğlu et al., 2019). Empathy is often described as comprising affective and cognitive components, corresponding to emotional resonance and perspective-taking respectively (Park & Whang, 2022).

Within Artificial Intelligence (AI) and Human-Machine Interaction (HMI), empathy has been reframed to reflect the capabilities and objectives of synthetic agents. Machine-expressed empathy is typically defined as a system's ability to recognise a user's emotional state, intentions, or situation and to generate responses that are perceived as appropriate or supportive (Park & Whang, 2022). This operationalisation encompasses processes such as language association, perspective-taking, and affective signalling, rather than genuine emotional experience. As a result, empathy in artificial agents is increasingly understood as something attributed by users on the basis of observable interactional behaviour, rather than as an intrinsic system property.

Recent advances in conversational systems, including speech-based assistants and large language model driven interfaces, have accelerated this shift (Zdravković & Panetto, 2022; Mijwil & Abttan, 2021; Reis et al., 2020). These systems increasingly operate beyond narrow task execution, shaping user experience through interactional style, tone, and perceived understanding. As expectations of naturalness, responsiveness, and personal relevance increase, empathy has become a central dimension of perceived interaction quality in conversational human-machine interaction. However,

despite growing interest in empathic AI, there remains a lack of simple, theory-informed, and empirically grounded tools for evaluating empathy as it is experienced by users during interaction.

Related work has approached perceived empathy in artificial agents from a range of complementary perspectives, including physiological and behavioural alignment in embodied systems, linguistic markers of empathic response, and perception-based evaluation of social agents. For example, Schmidmaier et al.'s (2024) Perceived Empathy Tool (PET) operationalises empathy through users' judgements of emotional understanding and responsiveness, while Concannon and Tomalin (2023) examine the linguistic strategies through which conversational agents signal empathic intent. Work in social robotics has similarly framed empathy as an interactional phenomenon emerging from timing, tone, and behavioural congruence rather than from isolated system capabilities (Paiva et al., 2017). Collectively, this body of work reflects a broader shift towards treating empathy as something recognised in interaction rather than inferred from system properties. However, it also points to the absence of a general, interaction-level evaluative instrument capable of capturing how these cues combine into holistic judgements of empathic interaction across dialogue.

To address this gap, this paper introduces the Empathy Evaluation Questionnaire (EEQ), a questionnaire-based instrument designed to assess perceived empathic interaction in speech-based systems. Rather than treating empathy as a stable psychological trait or an intrinsic system capability, the EEQ is positioned as a measure of how users judge empathic qualities as they arise during interaction. The EEQ captures interconnected aspects of interactional judgement, including linguistic clarity, perceived understanding, competence, trustworthiness, and affective engagement, providing a structured means of evaluating empathic interaction as it is experienced by users.

Questionnaire Design and Development

Four key categories consistently emerge in the literature regarding the assessment and understanding of empathy: pragmatic, hedonic, cognitive, and affective. These categories were therefore selected as the four primary concepts when designing the Empathy Evaluation Questionnaire (EEQ). Pragmatic and hedonic elements measure the usability dimension of the empathy experience, assessing how easy and enjoyable it was to interact with the system, whereas cognitive and affective elements measure the naturalness dimension, assessing how human-like and engaging the interaction felt. Subscale items for each category were drawn or derived from the framework of driver–automobile naturalness (Ramm et al., 2018), the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008) and previous work conducted by the author (see: Anyasodo & Burnett, 2021). This resulted in 33 scale items. Thirty-two of these were presented as 7-point, numerically anchored Likert scales, ranging from 'not at all' (1) to 'completely' (7), and required respondents to rate the speech system against each attribute, for example, *"Do you think the system speaks understandably?"* The final question explored the overall experience and required a dichotomous 'yes/no' response: *"Overall, I think the interaction is empathic".* The final question was the only item in which the word "empathy" was used (a definition of empathy was also provided alongside); this was to avoid biasing previous responses (Brown, 2016). Responses to the final question were subsequently used as a dependent variable during the analysis.

Evaluation and Psychometric Analysis

The proposed questionnaire initially underwent expert evaluation (n=9) and a pilot study (n=13), with the aim of assessing the wording and phrasing of questions, overall structure, and clarity of instructions. Following minor modifications to the language used to describe each attribute (e.g. replacing *"is personable"* with *"is humanlike*"), and removing three redundant items, the revised questionnaire (with 30 items) was hosted on an online survey platform. Audio recordings of exemplar speech interfaces were provided to contextualise the survey questions and enable the capture of authentic responses. One hundred and eighty-two responses were subsequently collected.

Internal consistency was evaluated using Cronbach's alpha, with values obtained between 0.9 and 0.95 (showing good consistency). Construct validity was determined through hypothesis testing and linear regression; this demonstrated that the four main constructs (pragmatic, hedonic, cognitive, affective) and all the subscale items (independent variables) significantly contributed to the overall perception of empathy (as a dependent variable) at $p < .05$. To explore any relationships between factors and to determine the questionnaire's latent structure, a principal axis factor analysis was performed. Preliminary checks confirmed data suitability – a Kaiser-Meyer-Olkin (KMO) value of .963 indicated excellent sample adequacy, suggesting that the variables share a substantial amount of common variance attributable to latent factors. Additionally, Bartlett's test of sphericity yielded a chi-square value of 5987.67 and a *p*-value of .000, confirming the presence of correlations within the dataset. A scree plot generated after Varimax rotation suggested reduction to three factors, albeit with some cross-loading. These were subsequently evaluated and the following nomenclature selected: 1. Pragmatic communication and pleasant interaction, 2. Emotional resonance and affective engagement, and 3. Linguistic clarity.

The three retained factors show that the use of clear speech, system expertise, understanding user needs, and using inclusive language are crucial aspects of the system's ability to achieve user goals (Factor 1). Moreover, items like 'pleasant', 'speaks with feeling', and 'understands the feelings of the user' highlight that an empathic system should resonate emotionally with users (Factor 2) (Norman, 2007; Prendinger & Ishizuka, 2005). Finally, 'linguistic clarity' (Factor 3) suggests that the use of plain language, clear articulation, and logical sentence structure are foundational to creating usable, empathic interfaces, particularly for individuals with diverse needs or language backgrounds (Benaida, 2023). The full EEQ is presented in the Appendix.

Interpretation of Factor Structure

The EEQ demonstrates psychometric robustness as a measure of perceived empathic interactional qualities. Its latent structure reflects interactionally coupled dimensions of user judgement rather than orthogonal psychological traits and should be interpreted accordingly, noting that this clarification concerns interpretation rather than measurement. Internal consistency, factor structure, and the contribution of the latent dimensions to overall judgements of empathic interaction remain statistically robust.

Some degree of cross-loading was observed between factors but is theoretically expected in this domain. Empathy in dialogue is not modular but inherently interactional, inferred from language, timing, tone, and responsiveness and judged holistically. Linguistic clarity affects affective comfort. Perceived understanding affects trust. Trust affects willingness to engage. Pragmatic, affective, and linguistic judgements are therefore tightly coupled in use, and perfect orthogonality would be implausible for this construct. From this perspective, cross-loading reflects the entangled nature of empathic interaction rather than a limitation in the instrument.

Short-Form EEQ (EEQ-s)

In deriving the short form, item selection was guided by interactional considerations rather than solely by statistical reduction. While interactional considerations guided item selection, statistical evidence from the full EEQ, including factor loading patterns, was used to inform decisions about item salience and redundancy. Particular attention was given to items that demonstrated strong and stable contributions to the latent structure, while recognising that some degree of cross loading reflects the interactionally coupled nature of empathic experience rather than measurement error.

Primary interactional judgement refers to the main aspect of the interaction that an item invites the user to evaluate. In the short form EEQ (EEQ-s), these judgements comprise linguistic clarity, perceived understanding, perceived competence, trustworthiness, and affective engagement. Each

judgement reflects a dominant way in which users assess whether an interaction feels empathic, based on the clarity of system language, the extent to which they feel understood, confidence in the system's knowledge and expertise, willingness to trust the system, and the emotional tone and engagement of the interaction. These judgements are interactional in nature, arising from how the system behaves within the exchange, rather than representing distinct psychological traits of the user or intrinsic properties of the system.

Secondary influences refer to related aspects of the interaction that are likely to be shaped by, or co-occur with, the primary interactional judgement during use. In the EEQ-s, these influences reflect the interactional coupling between linguistic clarity, perceived understanding, competence, trust, and affective engagement, whereby changes in one aspect of the interaction plausibly affect judgements in others. For example, clearer language may support trust and ease of engagement, while perceived understanding may reinforce affective comfort and willingness to rely on the system. Secondary influences are used to account for overlap between items and dimensions at an interpretive level, rather than to define additional measured constructs.

The reduced version comprises twelve numerically anchored Likert scale items capturing core interactional judgements, together with a single global yes or no item (as present in the full EEQ) assessing overall perceived empathy. This final item functions as a criterion measure rather than as part of the latent scale structure. The EEQ-s is presented in the Appendix.

The reduced interpretation structure preserves the core interactional dimensions identified in the full EEQ, while removing redundant items and retaining those that most clearly anchor linguistic clarity, perceived understanding, pragmatic competence, trust, and emotional engagement.

Applications of the EEQ

The EEQ and EEQ-s are intended for use across a range of applied evaluation contexts. These include comparative evaluation of conversational agents or system variants, formative assessment during dialogue design and iteration, and longitudinal monitoring of perceived empathic interaction in deployed systems. Because the EEQ captures user judgements rather than system properties, it can be applied across technologies, architectures, and deployment settings, supporting evidence-based decisions about interaction quality, trust, and user experience.

Conclusion

This paper introduced the Empathy Evaluation Questionnaire (EEQ) as a theory-informed, psychometrically robust instrument for evaluating perceived empathic interaction in speech-based systems. By treating empathy as an interactional phenomenon recognised by users rather than a latent psychological trait or intrinsic system property, the EEQ provides a defensible and practically useful approach to empathy evaluation in HMI. The introduction of a short-form version (EEQ-s) further supports applied use while preserving conceptual coherence.

References

Anyasodo, B., & Burnett, G. (2021). Empathy consideration in the design of natural language interfaces for future vehicles. Contemporary Ergonomics & Human Factors 2021, CIEHF.

Ayanoğlu, H., Saraiva, M., Teixeira, L., & Duarte, E. (2019). Human-robot interaction: exploring the ability to express emotions by a social robot. *Emotional Design in Human-Robot Interaction: Theory, Methods and Applications*, 163-183.

Benaida, M. (2023). Developing and extending usability heuristics evaluation for user interface design via AHP. *Soft Computing*, *27*(14), 9693-9707.

Brown, A. (2016). *Item Response Models for Forced-Choice Questionnaires: A Common Framework.* Psychometrika, 81(1), 135-160.

Concannon, S., & Tomalin, M. (2024). Measuring perceived empathy in dialogue systems. *Ai & Society*, *39*(5), 2233-2247.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work* (pp. 63-76). Springer, Berlin, Heidelberg.

Mijwil, M. M., & Abttan, R. A. (2021). *Artificial intelligence: A survey on evolution and future trends. Asian Journal of Applied Sciences, 9*(2), 58–68.

Norman, D. A. (2007). *The design of future things.* New York, NY: Basic Books.

Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(3), 1-40.

Park, S., & Whang, M. (2022). Empathy in human–robot interaction: Designing for social robots. *International Journal of Environmental Research and Public Health, 19*(3), 1889.

Prendinger, H., & Ishizuka, M. (2005). *The empathic companion: A character-based interface that addresses users' affective states.* Applied Artificial Intelligence, 19(3-4), 267-285.

Raamkumar, A. S., & Yang, Y. (2023). *Empathetic conversational systems: A review of current advances, gaps and opportunities.* IEEE Transactions on Affective Computing, 14(4), 2722-2739.

Ramm, S., Giacomin, J., Malizia, A., & Anyasodo, B. (2018). An exploratory design workshop to elicit what feels natural when interacting with an automobile's secondary controls. *The Design Journal*, *21*(1), 109-137.

Reis, J., Amorim, M., Cohen, Y., & Rodrigues, M. (2020). Artificial intelligence in service delivery systems: A systematic literature review. In *World Conference on Information Systems and Technologies* (pp. 222-233). Springer, Cham.

Schmidmaier, M., Rupp, J., Cvetanova, D., & Mayer, S. (2024, May). Perceived Empathy of Technology Scale (PETS): measuring empathy of systems toward the user. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Zdravković, M., & Panetto, H. (2022). Artificial intelligence-enabled enterprise information systems. *Enterprise Information Systems*, *16*(5), 1973570.

Appendix 1 - Empathy Evaluation Questionnaire (EEQ)

Do you think the system:	*Not at all* 1	2	3	4	5	6	*Completely* 7
1. speaks understandably							
2. speaks clearly							
3. speaks coherently							
4. speaks with feeling							
5. speaks in a manner a user can easily relate with							
6. is engaging							
7. is pleasant							
8. is pleasing							
9. is easy to understand							
10. mirrors the emotions of the user							
11. knows what the user wants							
12. makes it easier for the user to share their feelings							
13. knows what the user needs							
14. has some personality							
15. uses language that makes the user feel understood							
16. is easy for a user to relate with							
17. helps the user focus on the primary task							
18. appears to have some expertise							
19. appears trustworthy							
20. and the user are used to each other							
21. understands the feelings of the user							
22. easily understands what the user wants							
23. is easy to engage with							
24. can easily be trusted							
25. helps the user reduce mental load							
26. is humanlike							
27. is knowledgeable							
28. can easily be related with							
29. uses inclusive language							
30. Empathy involves the ability to know, understand and share in the feelings of another person. With this description in mind, overall, I think the interaction is empathic	Yes / No						

Interpretation
Pragmatic communication and pleasant interaction: items 2, 7, 11, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 29
Emotional resonance and affective engagement: items 4, 5, 6, 7, 8, 10, 12, 14, 15, 16, 21, 26, 28
Linguistic clarity: items 1, 2, 3, 9
Overall perceived empathy (global yes/no judgement): item 30

Appendix 2 – EEQ Short-form (EEQ-s)

Do you think the system:	*Not at all* 1	2	3	4	5	6	*Completely* 7
1. speaks clearly							
2. speaks understandably							
3. speaks coherently							
4. uses language that makes the user feel understood							
5. understands the feelings of the user							
6. knows what the user wants							
7. appears trustworthy							
8. appears knowledgeable							
9. appears to have some expertise							
10. speaks with feeling							
11. is pleasant							
12. is engaging							
Empathy involves the ability to know, understand and share in the feelings of another person. With this description in mind, overall, I think the interaction is empathic	Yes / No						

Interpretation
Linguistic clarity: items 1, 2, 3
Perceived understanding and responsiveness: items 4, 5, 6
Pragmatic competence and trust: items 7, 8, 9
Emotional resonance and engagement: items 10, 11, 12