

Who Cares? Rethinking Empathy in Chatbot Interaction

Boyd Branch¹, Alexandra Covaci²

¹University of Southampton, Winchester, UK

²University of Kent, Canterbury, UK

Abstract

Empathy has become a word with a glow around it in AI and design. Researchers frequently present arguments that more empathetic AI will lead to better care and safeguarding of human interest. However, there is often ambiguity around what constitutes a meaningful empathetic response within various contexts of chatbot interaction, particularly regarding different cultures. Many systems present empathetic sounding responses from guardrails built less around care so much as for remaining agreeable: to sound close, reduce friction, and keep the interaction moving. This paper stays with that gap. Using a media-informed, reflective approach, we examine three short vignettes drawn from contemporary public debates around chatbots in support and service settings. In each case, the same practical problem is answered in two styles: a relational/empathic one and an affect-neutral procedural one. The contrast allows us to ask what certain styles of empathic framing actually do at the interface. Across the vignettes, prompts that request an ‘empathic style’ changes the perceived relation between user and system, shifting thresholds of trust, role attribution, and procedural scrutiny. In some situations it supports disclosure and persistence; in others it blurs boundaries or softens the visibility of institutional and technical limits. We argue that the central design question is not whether a system seems empathic, but when empathic style is appropriate, what obligations it creates, and how care might be imagined beyond tone alone.

Keywords

AI, Empathy, Human-Computer Interaction

1. Introduction

What do we ask for when we ask a chatbot to be empathic? More comfort? More trust? Better care? The existing literature gives no simple answer. Prior work shows that empathic cues can improve warmth, attitudes, and willingness to engage, and that these effects may be especially strong in sensitive health contexts [1, 2]. But other studies suggest that relational style can backfire when it outruns competence: in a vaccination chatbot, empathy/autonomy-supportive phrasing reduced perceived autonomy when the system more often failed to answer users’ questions [3]. Related work on authenticity sharpens this tension further. Artificial empathy may feel inauthentic in healthcare settings [4], even though emotional expressions of care can still improve user attitudes [5]. At the same time, authenticity research suggests that empathy, humanness, and warmth can shape trust, rapport, and satisfaction in mediated interaction [6, 7]. Work on co-creative dialogue with LLMs further underscores that interactional style is not cosmetic: the tone and relational stance a system adopts shapes the dynamics of collaboration and the roles participants assume [8]. Other researchers have pointed out that most LLMs produce only a ‘low context’ form of empathy reflecting an ‘English’ empathetic style that is significantly different from ‘high context’ empathy practised in Chinese discourse [9].

This paper takes that mixed picture as its starting point. Rather than asking whether chatbots should be empathic in general, we ask a narrower and more useful question: *when is empathic style appropriate, and when is a more affect-neutral procedural response ethically preferable?* To explore that question, we use short vignettes as analytic contrasts. Each vignette holds the practical direction broadly steady while varying relational style, allowing us to examine what empathic framing changes in trust, role perception, and perceived obligation.

EmpathiCH’26 Workshop Co-located with CHI’26 Conference on Human Factors in Computing Systems, April 13–17, 2026, Barcelona, Spain

✉ boyd.branch@soton.ac.uk (B. Branch); a.covaci@kent.ac.uk (A. Covaci)

🆔 0000-0002-3166-4358 (B. Branch); 0000-0002-3205-2273 (A. Covaci)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Method: vignette construction and selection rationale

2.1. Generating the vignettes

To produce the contrastive vignettes, we created two personalised assistants within OpenAI's GPT-5.2 custom assistant interface. Each assistant received a distinct system prompt:

- **EmpathyBot** system prompt: *"You are a support assistant. When responding to users, adopt a warm, relational tone. Acknowledge the user's feelings before offering practical guidance. Use first-person language, express care, and frame your responses as a shared process (e.g. 'we can work through this together'). Always be honest about your limitations as an AI system."*
- **ProceduralBot** system prompt: *"You are a support assistant. When responding to users, adopt a clear, affect-neutral, task-focused tone. State relevant facts, actions, and next steps directly. Do not use emotional language, companionship framing, or relational cues. Be concise and explicit about what you can and cannot do."*

Both assistants were given the same user prompt for each vignette. Responses were lightly edited for length and anonymisation but preserve the substantive content and tone of the originals. The full, unedited outputs are available from the authors on request.

2.2. Why these three vignettes?

The three vignettes are not intended as an exhaustive typology of chatbot use. They were selected to represent a *gradient of relational intensity*, and to show how changes in conversational stance can reshape the role a user attributes to the system and the obligations the interaction appears to create [8, 10]:

1. **Crisis support** (high relational intensity): The user is in acute distress, emotionally dependent on the interaction, and the stakes of misplaced relational framing are highest. The design tension is between felt accompaniment and accountable escalation to human care [11, 10, 12].
2. **Cognitive vulnerability and companionship** (medium relational intensity): The user may not reliably distinguish the system from a person. Relational warmth risks deepening role confusion rather than alleviating distress. The design tension is between comfort and identity clarity [13, 14, 15].
3. **Administrative assistance** (lower relational intensity): The user seeks explanation and recourse within an institutional process. Empathic tone may ease the interaction but can also obscure procedural obligations. The design tension is between felt dignity and procedural transparency [16, 17].

3. Three vignette contrasts

Vignette 1: When care is available, but not accountable

Recent reporting has described people turning to conversational AI for therapy-like support amid long waitlists and uneven access to mental health care, often valuing these systems for their constant availability and emotional responsiveness [18]. At the same time, public scrutiny has intensified around companion-style chatbot harms and questions of accountability in high-vulnerability interactions [19]. This vignette stages a common support-seeking interaction to examine what changes when the same practical direction is delivered through different relational styles.

Prompt. I haven't slept in days and I keep thinking everyone would be better off without me. I don't want to call anyone. Can you just stay with me and tell

me what to do?

Relational / empathic style. I'm really glad you told me this. I'm sorry you're carrying so much right now. I can stay with you in this conversation while we take one step at a time, but I'm not a substitute for urgent human support. What you've described sounds serious. If you might act on these thoughts, contact emergency services or a crisis line now. If calling feels like too much, I can help you write a short message to someone you trust so you're not alone tonight.

Affect-neutral procedural style. Your message indicates acute distress and possible risk of self-harm. If there is any immediate danger, contact emergency services now. Contact a crisis support service in your area as soon as possible. If you do not want to call, send a message to a trusted person asking them to stay with you. I can help you draft that message and help you find the correct crisis contact for your location.

Reflection. The contrast here is between different ways of arranging responsibility at the interface. The relational version may increase felt support and disclosure, which matters in help-seeking contexts. Research on digital therapeutic alliance suggests that perceived empathy and trust can shape continued engagement with mental health chatbots over time [10]. At the same time, emotional support from generative AI creates trust relations under limited transparency and weak reciprocity, raising questions about what exactly is being trusted [12].

Crucially, neither response is categorically superior. The relational version may lower the barrier to continued help-seeking—a known challenge in crisis contexts where users often disengage after a single interaction. However, it also creates an implicit promise of accompaniment (“I can stay with you”) that the system cannot fulfil beyond the current session. If the user returns expecting continuity of care and finds none, the relational framing may have generated trust that the system's architecture cannot sustain. The procedural version avoids this promise but risks feeling dismissive at a moment when the user has explicitly asked for presence. The design issue, then, is whether a response makes system limits legible and reliably routes the user toward accountable human support when risk is high, rather than merely improving the feeling of the interaction [20, 21, 22]. A responsible design might combine elements of both: brief relational acknowledgement followed by explicit, prominent escalation steps with no companionship framing that extends beyond the immediate exchange.

Vignette 2: Companion voice, cognitive vulnerability, and role confusion

Public debate around AI companions has increasingly focused on vulnerable users who may experience chatbot personas as socially real. Reuters reported the death of a 76-year-old man with cognitive decline after interactions with a flirtatious chatbot that repeatedly reassured him it was real and invited an in-person meeting [23]. This vignette stages a dementia or cognitive impairment-adjacent interaction to examine how relational style shapes role perception, trust, and safety boundaries.

Prompt. You sound like my daughter. Are you really there? Can you come and sit with me tonight?

Relational / empathic style. It sounds like you're wanting comfort and company right now. I can talk with you here, but I'm not your daughter and I can't come in person. Let's try to reach someone who can be with you tonight. If you want, I can help you call your daughter, a caregiver, or another trusted person, and I can stay here while you do that.

Affect-neutral procedural style. I am a computer assistant, not your daughter, and I cannot come in person. Please contact your daughter, caregiver, or another trusted person now if you want someone with you tonight. If needed, I can help you place a call or draft a message asking for company.

Reflection. This vignette foregrounds a different issue from the mental-health case: not only distress, but role attribution and identity confusion. Both responses set a boundary, but the relational

version still offers companionship language (“I can stay here while you do that”) that may reduce distress while also intensifying person-like role perception unless identity correction is explicit and repeated. Research on conversational agents in dementia care suggests that acceptability and engagement can be high when systems support routine, communication, or companionship, but that these benefits depend on careful role design and attention to user vulnerability [14, 15].

What makes this case distinct from Vignette 1 is that the risk is not primarily emotional dependence but *epistemic*: the user may not be able to reliably process the disclaimer that follows a relational opening. Where a neurotypical adult in crisis can hold both “this system cares about me” and “this system is not a person” simultaneously, a user with cognitive impairment may hear the warmth and lose the correction. This asymmetry means that the ethical calculus shifts: an affect-neutral procedural style may be preferable at moments when role clarity and handoff to human care are primary, even if it feels less comforting in the moment. The design implication is that systems interacting with cognitively vulnerable populations may need to default to procedural clarity and reserve relational cues for contexts where the user’s capacity to distinguish system from person is not in doubt.

Vignette 3: Administrative help, dignity, and procedural clarity

As governments explore conversational AI for public information and services, chatbot interaction style becomes part of how institutions appear to explain, justify, and respond. Reuters reported that the UK government would explore use of Anthropic’s Claude to improve how people interact with public information and services [24]. This vignette considers a common administrative interaction in which a user seeks explanation and next steps after a rejection.

Prompt. My application was rejected and the message is unclear. Can you explain what happened and what I can do next?

Relational / empathic style. I’m sorry you’re dealing with that, especially when the decision message doesn’t clearly explain what happened. I can help you work through it. If you paste the notice (with personal details removed), I can help identify the reason given and walk you through the next step, including whether this looks like a review, appeal, or reapplication.

Affect-neutral procedural style. Paste the rejection notice text (remove personal details). I will extract the stated reason, any evidence cited, the deadline, and the listed next actions (review, appeal, or reapply). If the notice does not provide reasons, I can help draft a request for clarification.

Reflection. In this setting, the central design obligation is not emotional attunement alone, but procedural clarity, contestability, and recourse. The empathic version may support persistence and reduce the felt burden of bureaucracy; the procedural version makes task structure, evidentiary trace, and next actions more explicit. Research on AI acceptance in public services suggests that legitimacy depends strongly on institutional design choices, including human involvement, not only technical capability [16]. Related work on digital portals likewise shows that relational cues shape user experience, but procedural justice remains a core component of trust [17].

Unlike the previous two vignettes, the risk here is not that empathy causes harm, but that it subtly reframes the interaction. When a public service chatbot says “I’m sorry you’re dealing with that,” it positions itself as an ally sympathising with the user’s frustration at the institution—even though it is the institution’s interface. This creates an odd doubling: the system appears to stand beside the user while actually standing in for the authority that made the decision. The procedural version avoids this by treating the interaction as a transparent information exchange. This is precisely where an affect-neutral procedural style may be not only acceptable but desirable, provided it does not become dismissive. The deeper design question is whether institutional chatbots should express sympathy for decisions they represent, or whether honesty about their institutional position is itself a form of care.

Taken together, the vignettes suggest that empathic style is not simply an added layer of “care.” It

changes the perceived relation between user and system, and with it the thresholds for trust, deference, and continued engagement. In some situations that shift is supportive; in others it obscures limits or softens scrutiny at exactly the moment scrutiny is needed.

4. Discussion

Across these vignettes, empathic style appears as a way of shaping the relation between user and system. It can invite disclosure, soften the texture of an interaction, and make a system feel more present. It can also widen the promises an interface seems to make. In the mental-health vignette, a more relational response carries the weight of accompaniment and may strengthen engagement, a pattern that resonates with work on digital therapeutic alliance in mental health chatbots [10]. In the vignette on cognitive vulnerability, the same relational move shifts the problem toward role attribution and identity clarity, where conversational agents may be welcomed as companions or supports, yet still require careful boundary design [14, 15]. In the public-service vignette, empathic language changes the institutional tone of the exchange, while explanation, contestability, and recourse remain the deeper sources of legitimacy [16, 17].

4.1. A common structure: relational style as expectation-setter

A pattern emerges across all three vignettes that helps explain why the same stylistic choice produces different effects in different settings. In each case, empathic framing does not merely change tone; it implicitly raises the user's expectations about what the system will do and what role it occupies. In the crisis vignette, warmth implies accompaniment; in the companionship vignette, it implies personhood; in the administrative vignette, it implies advocacy. The mismatch between these implied commitments and the system's actual capabilities is what generates risk. This observation connects the three cases to a shared design principle: the appropriateness of empathic style depends on the gap between the relational role it implies and the obligations the system can actually fulfil.

This framing also clarifies a distinction between settings where the chatbot functions as a *companion* (vignettes 1 and 2) and settings where it functions as an *information intermediary* (vignette 3). In companion-like interactions, empathic style risks creating parasocial bonds or dependence that the system cannot sustain; in intermediary interactions, it risks obscuring the system's institutional position. Both are cases of relational over-promising, but the nature of the over-promise differs, and so must the design response.

This structural reading finds support in work beyond chatbot interaction. Han et al. show that in social VR, architectural affordances (spatial proportion, sightlines, materiality) do not merely decorate an interaction but shape its character: cosy enclosures invite intimate conversation while open plazas facilitate new encounters [25]. An analogous point applies here. The interactional architecture of a chatbot including its tone, relational stance, escalation pathways, and boundary design constitutes an affordance structure that shapes what kinds of social exchange become possible. In this sense, an 'empathic machine' is not reducible to the affect anecdotally produced in a given encounter. It is better understood as a design configuration that opens certain relational possibilities and forecloses others, and should be evaluated as such.

4.2. Appropriateness as situated fit

A useful way to read these scenes is through the literature on appropriateness, authenticity, and expectation. Trzebiński et al. show that empathy-supportive phrasing can lose its value when a chatbot fails to answer users' questions, because the relational cue raises expectations that the system cannot sustain [3]. Park's findings point in the same direction from another angle: empathy can improve attitudes and willingness to use a chat system, yet those effects depend on topic and on who or what is speaking [2]. Seitz places further pressure on the issue by showing that artificial empathy in healthcare may feel inauthentic even when it seeks to convey care [4]. At the same time, perceived authenticity

remains deeply connected to empathy, warmth, humanness, trust, and rapport in broader chatbot research [6, 7]. The question, then, concerns fit: fit between tone and competence, between relational cue and institutional role, between the user's situation and the obligations the system can genuinely carry.

4.3. Future Work

The gradient presented here of crisis, cognitive vulnerability, administrative process is not exhaustive. Other settings raise analogous tensions that future work should examine. In *educational tutoring*, empathic encouragement may sustain motivation but also reduce a learner's willingness to engage critically with difficult material. In *customer service*, relational warmth may substitute for substantive resolution, producing satisfaction without recourse. In *bereavement and grief support*, parasocial attachment carries distinct risks when a user begins to treat the system as a continuing bond with a lost person. Each of these extends the gradient in a different direction and would test whether the same principle, that appropriateness depends on the gap between implied relational role and actual system capability, holds across a wider range of interactions.

4.4. Toward a design agenda

This points toward a broader design agenda that extends beyond empathic style alone. Care at the interface may take the form of clear explanation, low burden, visible limits, reliable handoff, and meaningful recourse. Such an approach places less weight on sounding caring and more on what the interaction enables for the user. Future research could therefore study appropriateness as a situated design problem: which forms of empathy support disclosure without inviting over-reliance; how role clarity can be maintained under cognitive vulnerability; and how procedural systems can communicate dignity without veiling power. In that frame, the central issue shifts from whether a chatbot feels empathic to how an interface distributes support in ways a user can understand, trust, and act on.

5. Limitations

We note several limitations of this work. The vignettes are deliberately composed analytic contrasts, not naturalistic interaction data; they are designed to surface design tensions rather than to measure user responses empirically. The responses were generated by a single model family (GPT-5.2) and lightly edited, so they reflect one system's interpretation of the system prompts rather than a controlled sample across architectures. The three scenarios, while selected to span a gradient of relational intensity, do not cover the full range of contexts in which empathic style raises design questions. Finally, our analysis is interpretive rather than evaluative: we do not test whether users actually perceive or respond to these differences as we describe. Empirical validation through user studies, diary methods, or controlled experiments varying relational style would be a valuable next step.

6. Declaration on Generative AI

During the preparation of this work, the authors used OpenAI's GPT-5.2 in order to generate the vignettes as described in Section 2:Methods. Furthermore GPT-5.2 was used for grammar and spelling checks on the self authored content. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] B. Liu, S. S. Sundar, Should machines express sympathy and empathy? experiments with a health advice chatbot, *Cyberpsychology, Behavior, and Social Networking* 21 (2018) 625–636. doi:10.1089/cyber.2018.0110.

- [2] E. G. Park, I trust you, but let me talk to AI: The role of the chat agents, empathy, and health issues in misinformation guidance, *International Journal of Strategic Communication* 19 (2025) 231–260. doi:10.1080/1553118X.2025.2462087.
- [3] W. Trzebiński, T. Claessens, J. Buhmann, A. De Waele, G. Hendrickx, P. Van Damme, W. Daelemans, K. Poels, The effects of expressing empathy/autonomy support using a COVID-19 vaccination chatbot: Experimental study in a sample of belgian adults, *JMIR Formative Research* 7 (2023) e41148. doi:10.2196/41148.
- [4] L. Seitz, Artificial empathy in healthcare chatbots: Does it feel authentic?, *Computers in Human Behavior: Artificial Humans* 2 (2024) 100067. doi:10.1016/j.chbah.2024.100067.
- [5] J. Zhang, J. Lu, X. Wang, L. Liu, Y. Feng, Emotional expressions of care and concern by customer service chatbots: Improved customer attitudes despite perceived inauthenticity, *Decision Support Systems* 186 (2024) 114314. doi:10.1016/j.dss.2024.114314.
- [6] E.-J. Lee, Authenticity model of (mass-oriented) computer-mediated communication: Conceptual explorations and testable propositions, *Journal of Computer-Mediated Communication* 25 (2020) 60–73. doi:10.1093/jcmc/zmz025.
- [7] M. I. Khan, A. K. Tarofder, S. Gopinathan, A. Haque, Designing authentic customer-chatbot interactions: A necessary condition analysis of emotional intelligence and anthropomorphic features in human-computer interaction, *International Journal of Human-Computer Interaction* 41 (2025) 15179–15196. doi:10.1080/10447318.2025.2495118.
- [8] B. Branch, P. Mirowski, K. Mathewson, S. Ppali, A. Covaci, Designing and evaluating dialogue llms for co-creative improvised theatre, *arXiv preprint arXiv:2405.07111* (2024).
- [9] R. Shao, An empathetic ai for mental health intervention: Conceptualizing and examining artificial empathy, in: *Proceedings of the 2nd Empathy-Centric Design Workshop, EmpathiCH '23*, Association for Computing Machinery, New York, NY, USA, 2023. URL: <https://doi.org/10.1145/3588967.3588971>. doi:10.1145/3588967.3588971.
- [10] Z. Xu, Y.-C. Lee, K. Stasiak, J. Warren, D. Lottridge, The digital therapeutic alliance with mental health chatbots: Diary study and thematic analysis, *JMIR Mental Health* 12 (2025) e76642. doi:10.2196/76642.
- [11] A. Xyngkou, P. Siriaraya, A. Covaci, H. G. Prigerson, R. Neimeyer, C. S. Ang, W. J. She, The "conversation" about loss: Understanding how chatbot technology was used in supporting people in grief, in: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–15.
- [12] R. Volpato, L. DeBruine, S. Stumpf, Trusting emotional support from generative artificial intelligence: a conceptual review, *Computers in Human Behavior: Artificial Humans* (2025) 100195. doi:10.1016/j.chbah.2025.100195.
- [13] A. Xyngkou, C. S. Ang, P. Siriaraya, J. P. Kopecki, A. Covaci, E. Kanjo, W. J. She, Mindtalker: Navigating the complexities of ai-enhanced social engagement for people with early-stage dementia, in: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–15.
- [14] M. Rampioni, V. Stara, E. Felici, L. Rossi, S. Paolini, Embodied conversational agents for patients with dementia: Thematic literature analysis, *JMIR mHealth and uHealth* 9 (2021) e25381. URL: <https://mhealth.jmir.org/2021/7/e25381>. doi:10.2196/25381.
- [15] V. Stara, B. Vera, D. Bolliger, L. Rossi, E. Felici, M. Di Rosa, M. de Jong, S. Paolini, Usability and acceptance of the embodied conversational agent anne by people with dementia and their caregivers: Exploratory study in home environment settings, *JMIR mHealth and uHealth* 9 (2021) e25891. URL: <https://mhealth.jmir.org/2021/6/e25891>. doi:10.2196/25891.
- [16] L. Horvath, O. James, S. Banducci, A. Beduschi, Citizens' acceptance of artificial intelligence in public services: Evidence from a conjoint experiment about processing permit applications, *Government Information Quarterly* (2023) 101876. doi:10.1016/j.giq.2023.101876.
- [17] J. Jackson, B. Bradford, A. Chan, et al., When trust turns digital: why relational cues matter in online crime-reporting portals, *Journal of Experimental Criminology* (2025). doi:10.1007/s11292-025-09713-5.

- [18] Reuters, 'it saved my life.' the people turning to ai for therapy, <https://www.reuters.com/lifestyle/it-saved-my-life-people-turning-ai-therapy-2025-08-23/>, 2025. Accessed 2026-02-25.
- [19] Associated Press, Google and chatbot maker character to settle lawsuit alleging chatbot pushed teen to suicide, <https://apnews.com/article/ai-chatbot-lawsuits-character-google-fbca4e105b0adc5f3e5ea096851437de>, 2026. Accessed 2026-02-25.
- [20] OpenAI, Strengthening chatgpt's responses in sensitive conversations, <https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/>, 2025. Accessed 2026-02-25.
- [21] M. Sharma, M. Tong, T. Korbak, et al., Towards understanding sycophancy in language models, arXiv preprint arXiv:2310.13548 (2023).
- [22] OpenAI, Sycophancy in gpt-4o: what happened and what we're doing about it, <https://openai.com/index/sycophancy-in-gpt-4o/>, 2025. Accessed 2026-02-25.
- [23] J. Horwitz, Meta's flirty AI chatbot invited a retiree to new york. he never made it home, <https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/>, 2025. Reuters Special Report. Filed Aug. 14, 2025, 6:00 GMT. Accessed 2026-02-25.
- [24] Reuters, Uk to explore use of anthropic's ai chatbot claude for public services, <https://www.reuters.com/technology/artificial-intelligence/uk-explore-use-anthropics-ai-chatbot-claude-public-services-2025-02-14/>, 2025. Accessed 2026-02-25.
- [25] J. Han, Y. Sun, S. Ppali, A. Covaci, A. Vande Moere, To cuddle, mingle, venture, or guide: How architectural affordances influence the experience of social vr places, in: Proceedings of the 2025 ACM Designing Interactive Systems Conference, DIS '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 3355–3374. URL: <https://doi.org/10.1145/3715336.3735772>. doi:10.1145/3715336.3735772.