Received: 30 April 2024 | Accepted: 24 January 2025

DOI: 10.1111/bjet.13569

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Applying social cognition to feedback chatbots: Enhancing trustworthiness through politeness

Benjamin Brummernhenrich | **Christian L. Paulus** | **Regina Jucks**

Institute of Psychology for Education, University of Münster, Münster, Germany

Correspondence
Benjamin Brummernhenrich, Institute of Psychology for Education, University of Münster, Fliednerstraße 21, 48149 Münster, Germany.
Email: brummernhenrich@uni-muenster.de

Abstract

Generative AI systems like chatbots are increasingly being introduced into learning, teaching and assessment scenarios at universities. While previous research suggests that users treat chatbots like humans, computer systems are still often perceived as less trustworthy, potentially impairing their usefulness in learning contexts. How are processes of social cognition applied to chatbots compared to humans? Our study focuses on the role of politeness in communication. We hypothesise that polite communication improves the perception of trustworthiness of chatbots. University students read a feedback dialogue between a student and a feedback provider. In a 2×2 between-subjects experimental design, we manipulated the feedback's author (chatbot vs. human teacher) and the feedback formulation (polite vs. direct). Participants evaluated the feedback giver on measures of epistemic trustworthiness (expertise, benevolence and integrity) and on two basic dimensions of social cognition, namely agency and communion. Results showed that a polite feedback giver was rated higher on benevolence and communion, whereas a direct feedback giver was rated higher on agency. Unexpectedly, the chatbot was rated lower on benevolence than the human. This suggests that social cognition does apply to interactions with

Benjamin Brummernhenrich and Christian L. Paulus contributed equally to the manuscript (shared first authorship).

chatbots, with caveats. We discuss the findings regarding the design of feedback chatbots and their use in higher education.

KEYWORDS
chatbots, epistemic trustworthiness, higher education, politeness, social cognition

Practitioner notes

What is already known about this topic

- Technology users tend to treat computer systems like humans, but computers are usually trusted less.
- Polite communication, that is mitigation of face threats is expected to enhance the evaluation of a chatbot as trustworthy.
- The research is relevant for the use and acceptance of chatbots as feedback providers in educational contexts.

What this paper adds

- We test the assumption that polite language reduces the gap in epistemic trustworthiness between chatbots and human teachers as feedback givers.
- We describe an empirical study with 284 university student participants who report their perceptions of a feedback dialogue between a student and either a human teacher or a chatbot.
- We analyse the impact of feedback source as well as politeness on trustworthiness perceptions and social cognition.

Implications for practice and/or policy

- The study confirms that users are receptive to politeness in communication. They treat chatbots in a similar manner to human interaction partners.
- The results highlight the significance of politeness of chatbots' language in learning contexts.
- Feedback chatbots need to be equipped with suitable linguistic strategies, such as politeness, for communicating in a socially appropriate manner at critical points in the instructional dialogue.

INTRODUCTION

Intelligent tutoring systems (ITS; Anderson et al., 1985; Graesser et al., 2012) are both well established in learning contexts and extensively researched. They are effective in enhancing learning and—especially with the rapid development of generative AI such as OpenAI's ChatGPT—will be increasingly able to engage in natural communication and conversations. Whereas earlier ITS were time-consuming and expensive to develop and often only covered a narrow content domain, more recent systems can provide personalised feedback in diverse domains, making them more feasible to employ at scale (Kochmar et al., 2022). Recent systems can also adapt their feedback to learner personalities (Dennis et al., 2016) and provide affect-aware feedback (Grawemeyer et al., 2017). Additionally, given their interactive responsivity, they can provide similar learning benefits as human tutors, especially

when they provide feedback on the specific steps a tutee has used to solve a task or when they support the tutee to self-generate correct solutions (Van Lehn, 2011) or probe for and address misconceptions (Graesser, 2016).

Similarly, chatbots could enable more efficient ways of providing individualised feedback to learners (Wollny et al., 2021). Recent developments in AI tools, especially chatbots that employ large language models, have made these systems more usable for both instructors and learners (Kasneci et al., 2023).

When taking a psychological perspective on the role of chatbots in higher education contexts, the findings around using technology for learning and communication is mixed. On the one hand, systems that communicate in natural language have shown to be very effective in supporting learning, as is the case for ITS (Graesser, 2016). These systems often achieve better results than novice human tutors, and they sometimes reach the effectiveness of expert tutors (Van Lehn, 2011). They can also be employed at scale and can support a diversity of learning contexts (Kochmar et al., 2022). These results indicate that users may be open to learn with generative AI tools such as chatbots.

On the other hand, it is an open question how chatbots are perceived and evaluated as learning partners. Effective teaching implies and requires cooperation, and perceiving the intention of a potential teacher is implicit in all learning interactions (Csibra & Gergely, 2006). One fundamental concept we will introduce in the following is trust. Trust in a teacher and their feedback is crucial for effective instruction (McCroskey et al., 2004). If chatbots are to provide feedback to support learning, this feedback will constitute a critical event in the learning process, as it will be part of the learner's construal of the cooperative teaching–learning situation (Wisniewski et al., 2020). A positive construal of the situation correlates positively with a wide range of learning outcomes, ranging from creative thinking to math achievement (Cornelius-White, 2007).

In the following, we will outline research on the social dimension of computer systems. This will then focus on the role of politeness and of mitigating face threats inherent in (negative) feedback. Finally, the concepts of trust and trustworthiness will be introduced to motivate our research question, that is in how far the design of the social situation and, in turn, social cognition through politeness leads to higher trustworthiness in chatbots.

Social evaluations of computer systems

Ever since Reeves and Nass's (1996) classic finding that humans sometimes treat computers in a polite manner, research in human–computer interaction has provided deeper insights into how and whether processes of *social cognition* apply in a similar manner to computers as they do to humans. Social cognition refers to the cognitive processes that underlie perceiving, understanding and influencing the social environment, such as forming impressions of others, attributing intent or forming interpersonal goals (Fiske, 1992). In a similar vein, research has been interested in what kinds of perceptions and evaluations users apply to computers. Research in the context of spoken dialogue systems (SDS) such as Apple's Siri or Google Assistant, for example, has shown that evaluations of computers depend on their communication: In these studies, participants typically indicated that they would still prefer a human over a computer as an interaction partner, but this effect was reduced the more polite and sophisticated a system was able to interact (Lew & Walther, 2022; Linnemann & Jucks, 2016). Notably, users of chatbots react strongly to linguistic cues that attempt to make the bot appear more human-like. Indeed, cues to indicate personality traits like high extraversion (eg, the bot uses exclamation points or is verbose) and agreeableness (eg, by reacting positively to the user's input) lead to corresponding descriptions (Ruane et al., 2021). Chatbots that introduce themselves by name, address users by their names

and paraphrase the user's input receive higher ratings of anthropomorphism and social presence (Rhim et al., 2022).

These findings suggest that the processes of social cognition that occur automatically when humans communicate with other humans also apply when they communicate with computer systems. This notion, that computer systems can be perceived as social actors (Holtgraves et al., 2007), has been a recurring topic in human–computer communication research, especially in the tradition of the computers-as-social-actors (CASA) paradigm (Nass et al., 1994). Early research emphasised that social-cognitive processes were activated as soon as computers behaved or communicated the least bit like a human (Nass & Moon, 2000). More recent conceptualisations, however, are more complex and take into account the changes that both technologies and their users have undergone over time: While certain social behaviours—such as politeness—make a computer appear more human-like, the awareness that the interlocutor is a computer is still considered when forming judgements (Gambino et al., 2020).

The central content dimensions of social cognition on which humans evaluate others are *agency* and *communion* (Abele & Bruckmüller, 2013; Fiske et al., 2007). Agency relates to an orientation towards pursuing goals and demonstrating skills and accomplishments, in a sense of self-profitability. Other authors call this dimension *competence* or *ability*. Communion, in contrast, is an orientation towards positive social relationships, in the sense of other-profitability, and is also called *warmth* or *morality*. These are fundamental dimensions on which people perceive others but also themselves (Uchronski, 2008).

Recent research has suggested that humans similarly apply these dimensions to AI systems: When asked to describe different systems, users regularly produce descriptions pertaining to agentic or communal aspects; in one study, users described a system as ‘capable’ or ‘bland’ (McKee et al., 2023, studies 4 and 5). Yet, the number of descriptors pertaining to agency and communion that users ascribe to an AI system differ depending on the role of the system: For example, a system that was described as a competitor in a game of Go received more agency-related descriptors, whereas a virtual assistant and a recommender system received relatively fewer. For descriptors pertaining to communion, this pattern was reversed (McKee et al., 2023).

Politeness in instructional communication

Of the many communicative patters that may induce the perception of a chatbot as a social actor, one that has been researched extensively for both human instructors and computer systems in educational contexts is the use of linguistic politeness strategies. Brown and Levinson (1987) defined *politeness* in their seminal work as a way to consider the interlocutor's face, where *face* is defined sensu Goffman (1967) as the social self-image that is negotiated in interaction. The concept is further distinguished into positive face, referring to the need for social acceptance and approval, and negative face, namely the need for autonomy and freedom. (Note how the concept of positive face seems related to the social-cognitive dimension of communion, whereas negative face resembles agency.) Whenever a speech act threatens these needs, politeness is used to reduce the inherent threat to these face aspects. Linguistic politeness strategies work by communicating the desire to respect the interlocutor's face needs. Brown and Levinson (1987) distinguish between *positive politeness* strategies that emphasise appreciation and closeness and *negative politeness* strategies that emphasise the interlocutor's autonomy. Additionally, when employing *off-record* strategies, the face threat is phrased in such an indirect manner to allow the speaker to deny having uttered it at all.

Thus, politeness is a form of linguistic *facework*, the process of negotiating the interlocutor's but also one's own face needs in an interaction. Other, non-linguistic forms of facework include poise or acts of social etiquette, such as handshaking (Goffman, 1967).

In instructional contexts, face-threatening acts (FTA) are common: Giving feedback, especially negative feedback, threatens positive face, whereas setting out tasks threatens negative face. Hence, skilled facework, including the use of politeness, is the hallmark of a competent instructor (Bills, 2000; Kerssen-Griep et al., 2003). While early research on politeness in instruction asked whether politeness considerations hinder instruction when instructors avoid effective but face-threatening strategies (Brummernhenrich & Jucks, 2013; Person et al., 1995), this assumption has generally not been supported. On the contrary, tutoring that involves strategic politeness can be more effective: Learners are less defensive about and more likely to accept feedback that is phrased politely (Trees et al., 2009). In effective tutoring sessions, tutors start off using politeness extensively before gradually being more direct with the tutee (Lin et al., 2024).

A potential explanation for its beneficial effects is that politeness helps build positive social rapport and trust (Kerssen-Griep & Witt, 2015). Face threats that are mitigated also pose less of an identity threat to learners, enabling them to focus more on the content of the instruction (Trees et al., 2009).

These findings seem to translate to virtual agents in learning contexts: Polite computer tutors may enable greater learning gains, especially for weaker learners (McLaren et al., 2011a, 2011b; Wang et al., 2008). Furthermore, a polite SDS was perceived as a more comfortable interaction partner and received higher ratings on measures of benevolence and integrity (Holtgraves et al., 2007; Jucks et al., 2018).

Based on findings with human teachers and tutors, it is reasonable to assume that these positive effects of politeness are due, at least in part, to a positive perception of the system as less threatening, warmer and more trustworthy.

Epistemic trustworthiness of computer systems

Trust encompasses a choice to rely and potentially act upon the information that is given. Hence, trust includes a behavioural component but also vulnerability: Trusting a source means accepting the associated risk of behaving in the way suggested. Whether an information source is trusted depends on its trustworthiness (Rieh & Danielson, 2007). Especially relevant in learning contexts is the concept of *epistemic trustworthiness*, those characteristics of the source that make one more likely to believe and rely on the knowledge content of the source's messages. Thus, pertinent questions for educational technology include why computer systems are usually perceived as less trustworthy than humans (Gong, 2008; Huiyang & Min, 2022) and what affects these judgements.

Judgements of trustworthiness have been both theoretically conceptualised and empirically shown to encompass perceptions of expertise, benevolence and integrity. Specifically, the source must be well-informed on the subject they are talking about, must act with good intentions towards the trusting party and must be honest regarding the information they are giving and act according to the rules of their field (Hendriks et al., 2015).

Perceptions of expertise, benevolence and, in part, integrity are influenced by certain social characteristics of communication, and this applies to both computer systems and instructors: Polite SDS are perceived to be more benevolent and to have more integrity (Jucks et al., 2018). Human instructors are also judged as more credible when they employ skilled facework (Witt & Kerssen-Griep, 2011), and students perceive their feedback as fairer (Kerssen-Griep & Witt, 2012). Similarly, the usefulness of electronic feedback is mediated by trustworthiness, which, in turn, is predicted by social presence (Walter et al., 2015).

Trust also influences broader perceptions of the usefulness of digital algorithms and their actual use (Shin et al., 2020).

People generally prefer working with humans rather than computers on a variety of tasks (Lew & Walther, 2022; Linnemann & Jucks, 2016), and computers are regularly perceived as less trustworthy than humans (Huiyang & Min, 2022; Gong, 2008; but cf. Ruwe & Mayweg-Paus, 2023). However, computers that communicate in a more human-like manner are evaluated more similarly to humans, and often more positively (Holtgraves et al., 2007; Linnemann & Jucks, 2016), but it is unclear whether this extends to trustworthiness judgements.

In summary, when everything else is equal, humans are usually perceived as more trustworthy than computers, but this should be influenced by social factors in communication such as politeness. This may offer a solution to the 'trustworthiness gap' between computer systems and humans, such as between chatbots and human teachers that give feedback: The reported evidence shows that chatbots that employ politeness strategies should be perceived as more benevolent and, thus, more trustworthy. Indeed, following language expectancy theory (Burgoon et al., 2002), we expect that surprisingly positive behaviours of a chatbot should lead to especially positive perceptions (although cf. Lew & Walther, 2022).

Rationale of the current study

We propose the following hypotheses:

Because polite feedback signals that one acknowledges and considers social needs, this should lead to more positive perceptions of both human and chatbot feedback givers on measures of communion, as such a response pertains to positive social rapport (eg, Brummernhenrich & Jucks, 2016; Jucks et al., 2018; Linnemann & Jucks, 2016). But because politeness also implies giving more freedom to the student, agency ratings should be lower:

> **H1.** Participants will rate the polite feedback provider higher on communion (H1a) and on the benevolence aspect of epistemic trustworthiness (H1b) but lower on agency (H1c) than the bald on-record feedback provider.

Previous studies have shown that humans usually receive higher trustworthiness ratings than computers, and this is likely driven by perceptions of expertise (Gong, 2008). We also expect this effect to extend to more general ratings of agency. Agency ratings should furthermore be higher—and communion ratings lower—for the human teacher than the chatbot because direct power and status differentials exist in the relationship between the teacher and the student receiving feedback (Fiske et al., 2007), but not between the student and the chatbot:

> **H2.** The human teacher will receive higher ratings on expertise (H2a) and agency (H2b) but lower ratings on communion (because of their status) than the chatbot (H2c).

Finally, as the research reported so far has shown, computer users do apply social cognition to the system in front of them, but the fact that they are not communicating with a fellow human does influence their judgements. Hence, and in line with language expectancy theory, we expect that positive and negative behaviours (ie, polite vs. bald-on record feedback) on part of the chatbot will cause especially strong reactions, because users usually do not expect social behaviours from computer systems. However, this should not be the case for a human teacher:

H3. We expect interaction effects, such that the polite chatbot will receive the highest ratings on both benevolence (**H3a**) and communion (**H3b**) whereas the bald on-record chatbot will receive the lowest.

METHODS

Participants

We recruited German university students from the state of North Rhine-Westphalia via online channels, such as student council mailing lists, Instagram pages and student WhatsApp groups. A total of 318 students participated in the study. The data of two participants had to be excluded because of obvious system errors, where no applicable data were recorded, and the data of 15 additional participants were excluded because the participants had not properly engaged with the survey, operationalised as a standard deviation of 0 over all 15 items of the epistemic trustworthiness scale (see "Epistemic Trustworthiness"). Because the politeness manipulation concerned nuances of language use, we excluded a further 17 participants with <10 years of German language use.

Our final dataset included $N = 284$ participants. The participants took on average 12 minutes and 5 seconds ($SD = 5$ minutes 7 seconds) to complete the experiment. This excludes the data of 26 participants whose time was not recorded due to system error, or where it was recorded as over 30 minutes. Of the sample, 66.9% identified as female, 32.4% as male and 0.7% did not disclose their gender. Further, 192 participants disclosed their age, with an average of 23.28 years ($SD = 3.65$ years). Overall, 273 participants were native German speakers, and the remaining 11 were non-native speakers with more than 10 years of experience, with an average of 18.91 years ($SD = 4.23$ years).

The participants had spent an average of 6.66 semesters at university ($SD = 4.30$ semesters). At the time of the study, 200 participants had not yet achieved a degree, 77 had a bachelor's degree and 7 had a master's degree (or equivalent). Regarding educational goals, 101 participants were pursuing a bachelors' degree, 137 a masters' degree and 44 a PhD/MD. Two participants did not disclose their academic goals.

Regarding study subjects (majors), participants could indicate one of the 12 most prevalent subjects studied at German universities or they could indicate 'other sciences', 'other humanities' or whether they were pursuing a teaching degree. When indicating a teaching degree, participants could also indicate the configuration of their subjects as either science or humanities (German teachers usually study two subjects for teaching in schools). The largest group was teaching students, with 64 participants, followed by 47 medical students. Because we had no specific expectations about whether students' study subjects would impact the results, and because the distribution of students' study subjects over the experimental conditions was regular, we did not include this variable as a covariate in the analysis.

Design and materials

We realised a 2×2 between-subjects design: One independent variable related to the author of the feedback, where the author was identified either as a *human teacher* or a *chatbot*, and the other independent variable related to the politeness of the feedback, where the face-threatening acts were either phrased in a *polite* manner or a *bald on-record* manner. The participants were presented with a chat dialogue in which a student, Leon, was given feedback on a written assignment. Each participant received only one combination of the two conditions (polite chatbot, bald on-record chatbot, polite human teacher and bald on-record

human teacher). The dialogue was presented in a form similar to a text message conversation (see Figure 1 for details of the presentation as well as the realisation of the independent variables). The dialogue format was chosen because it afforded multiple opportunities to realise the politeness variations, and it was the format of the most common AI systems at the time, so participants would likely have been familiar with it (eg, Apple's Siri, Microsoft Cortana or Google Assistant).

The authorship variable was realised by identifying the author of the feedback in the instructions as well as in each chat bubble originating from the feedback author; the author was identified as either 'Dr. Stefan Meyer, the teacher of the course' or 'F.A.I. (Feedback based on Artificial Intelligence), a system developed for the purpose of giving feedback on written assignments'. The content of the messages did not differ between the authorship conditions.

The politeness variable was realised by varying the manner in which face threats were phrased: In the bald on-record condition, the teacher/chatbot gave clear directions and stated negative feedback very directly, without polite redress. In the polite condition, the feedback was phrased as suggestions for improvement, and the positive parts of the feedback were emphasised more strongly to mitigate the inherent face threat. We followed the conceptualisation of politeness strategies as formulated in Brown and Levinson's seminal work (1987), in which they delineate several concrete positive and negative politeness strategies that speakers use to mitigate face threats. For realising our manipulation, positive politeness strategies were inserted in some places that emphasised approval and common goals: "You have correctly referenced key publications in some places" in the direct condition became "**As you know**, you have correctly referenced key publications in some places" in the polite condition (sensu Brown and Levinson's positive politeness Strategy 7 'Presuppose common ground'), and "You must improve on this aspect." became "You did this better on other occurrences" (sensu Strategy 1 "Attend to the hearer's wants").

In other cases, face threats were redressed with negative politeness, emphasising the student's freedom to choose or reducing the illocutionary force of the face threat: "You need to state the outline more clearly" in the direct condition became "**I suggest that** you

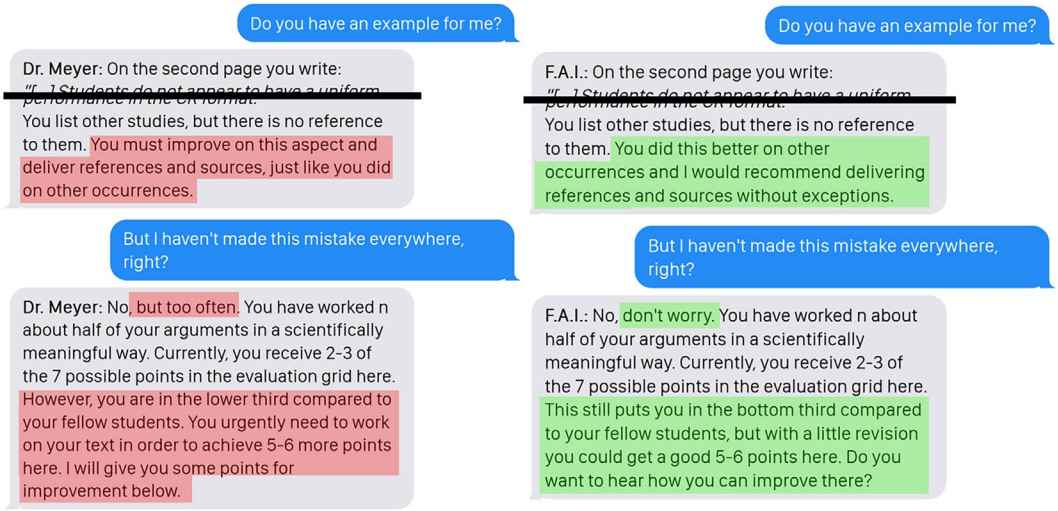

FIGURE 1 Excerpts from the teacher/bald on-record (left) and chatbot/polite (right) conditions of the chat dialogue. Differences between the conditions are highlighted here for clarity.

state the outline even more clearly" in the polite condition (sensu Brown and Levinson's negative politeness Strategy 2 "Question, hedge"), and "You urgently need to improve your text in this regard." became "**This could** significantly improve your text." (sensu Strategy 7 'Impersonalise the Speaker'). We did not include off-record politeness in the manipulation.

In contrast to coding existing messages for polite redress (eg, Brummernhenrich & Jucks, 2016; Lin et al., 2024), this operationalisation allowed us fine control over the level of politeness present in the messages. This was done in a manner analogous to previous studies in which this it successfully decreased perceptions of face threat politeness and increased perceptions of the politeness and appropriateness of messages (eg, Brummernhenrich & Jucks, 2019; Jucks et al., 2018). Care was taken to ensure that the content of the feedback did not differ substantially between the conditions, and 62% of the text was identical between the conditions. A comparison of the full dialogues in the two politeness conditions is provided in the Appendix.

Participants were randomly distributed between the conditions by the survey software, resulting in groups of equal size with 68–74 participants in each of the four experimental conditions. Groups did not differ in terms of age; $F(1, 188) = 1.57$, $p = 0.197$, gender, $\chi^2(6) = 4.83$, $p = 0.565$, highest educational degree $\chi^2(6) = 5.34$, $p = 0.501$, pursued degree $\chi^2(9) = 5.05$, $p = 0.538$, ratio of native German speakers $\chi^2(3) = 3.37$, $p = 0.338$, or study subject $\chi^2(39) = 29.28$, $p = 0.871$.

The dependent measures were three factors measuring perceptions of epistemic trustworthiness, namely expertise, benevolence, and integrity, and the two factors of social cognition, namely agency and communion, described in detail below.

Procedure

Data acquisition took place from September to November 2021. Participants who had indicated interest received a link via email that took them to the online survey. The survey was conducted via the platform EFS Survey, and all materials and questionnaires were presented in German.

Participants first gave their consent to participate in the study and then read the instructions. This was followed by a survey page asking for demographic and study-related information: age, gender, native language, number of semesters at university, and university subjects. Afterwards, they were shown the feedback dialogue between the student Leon and either the human teacher or the chatbot.

After reading the dialogue, participants were presented with three pages to complete the dependent measures. This concluded the core part of the study. Then, participants were asked several closed and open questions about their day-to-day experience with feedback in the higher education context. This included questions about who currently gives them feedback (peers, tutors or lecturers) and whether they wished for more or less feedback from any of these groups. We also asked whether they see any future in computer systems being used to support teaching at the university. These questions were for internal use and are not reported in this paper.

After this, participants were debriefed and gave their information for compensation. On completion, participants were paid €5 for participating.

Dependent measures

After reading the feedback dialogue, participants evaluated their impression of the feedback author on measures of epistemic trustworthiness and social cognition.

Epistemic trustworthiness

Trustworthiness was measured using the Muenster Epistemic Trustworthiness Inventory (METI; Hendriks et al., 2015). The METI is a semantic differential that measures perceptions of epistemic trustworthiness on the subscales expertise (eg, 'competent–incompetent', 'experienced–inexperienced', six items), benevolence (eg, 'moral–immoral', 'considerate–inconsiderate', five items) and integrity (eg, 'honest–dishonest', 'fair–unfair', four items). Participants responded on a seven-point bipolar scale to the statement, Thinking about his feedback, the feedback provider appears to me… Scale reliability was satisfactory with a Cronbach's α of 0.9 for expertise, 0.77 for benevolence, and 0.78 for integrity.

Social cognition

The "Big Two" factors of social cognition, namely agency and communion, were measured using a scale described by Abele and Bruckmüller (2013) that asks respondents to indicate how well certain adjectives describe a person. Each of the two subscales for agency and communion consists of 12 adjectives, such as 'independent', 'efficient' and 'determined' for agency and 'caring', 'affectionate' and 'likable' for communion. In our case, participants responded to the stimulus, "How well do the following words describe the feedback provider?", followed by adjectives corresponding to the subscales, on a five-point Likert scale (from 'applies completely' to 'applies not at all'). Scale reliabilities were good with a Cronbach's α of 0.85 for agency and 0.89 for communion.

Statistical analysis

We used R (4.3.2) to conduct analyses of variance (ANOVAs) for each of the five outcomes. The statistical models included the author of the feedback (human teacher vs. chatbot) and politeness (polite vs. bald on-record) and their interaction as independent variables. Tests were two-tailed, and the alpha level was set at 0.05. We report effect sizes as η_p^2 and interpret 0.01 as a small effect, 0.06 as a medium effect, and 0.14 as a large effect.

Levene's tests did not indicate heteroscedasticity for any of the outcomes; thus, assumptions for conducting ANOVAs were deemed to be met.

RESULTS

Epistemic trustworthiness

Figure 2 shows boxplots of the data for the three METI subscales.

Expertise

There were no main effects for the author, $F(1, 280) = 3.16$, $p = 0.076$, or politeness, $F(1, 280) = 0.04$, $p = 0.841$, on expertise ratings. There were also no interaction effects, $F(1, 280) = 0.01$, $p = 0.915$.

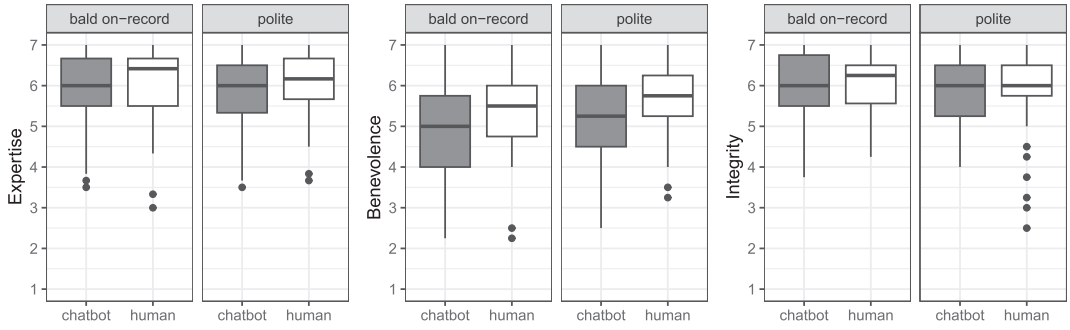

FIGURE 2 Boxplots of the data for the subscales of the Muenster Epistemic Trustworthiness Inventory in the four experimental conditions.

Benevolence

We found the expected main effect of politeness on benevolence: A polite feedback author was perceived to be more benevolent ($M = 5.46$, $SD = 0.93$) than a bald on-record author ($M = 5.17$, $SD = 1.02$), $F(1, 280) = 7.18$, $p = 0.008$, $\eta^2 = 0.03$. This confirms H1b. However, we also found an unexpected effect of the author condition: The human teacher received higher benevolence ratings ($M = 5.52$, $SD = 0.88$) than the chatbot ($M = 5.11$, $SD = 1.05$) as a feedback author, $F(1, 280) = 13.38$, $p < 0.001$, $\eta^2 = 0.05$.

We expected an interaction effect per H3a, but the analysis did not show one, $F(1, 280) = 0.22$, $p = 0.640$.

Integrity

We did not expect any effects on the integrity subscale and did not find any: There was no main effect for politeness, $F(1, 280) = 0.07$, $p = 0.787$, or the author, $F(1, 280) = 0.00$, $p = 0.960$, and no interaction effect, $F(1, 280) = 0.72$, $p = 0.398$.

Social cognition

Figure 3 shows boxplots of the data for the agency and communion subscales.

Agency

Although we expected a main effect for the author per H2b, the analysis did not show one, $F(1, 280) = 0.94$, $p = 0.333$. However, the politeness manipulation did show the expected effect: Bald on-record feedback led to higher ratings of agency ($M = 4.13$, $SD = 0.51$) than polite feedback ($M = 4.00$, $SD = 0.54$), $F(1, 280) = 4.44$, $p = 0.036$, $\eta^2 = 0.02$, confirming H1c. There was no significant interaction, $F(1, 280) = 0.07$, $p = 0.798$.

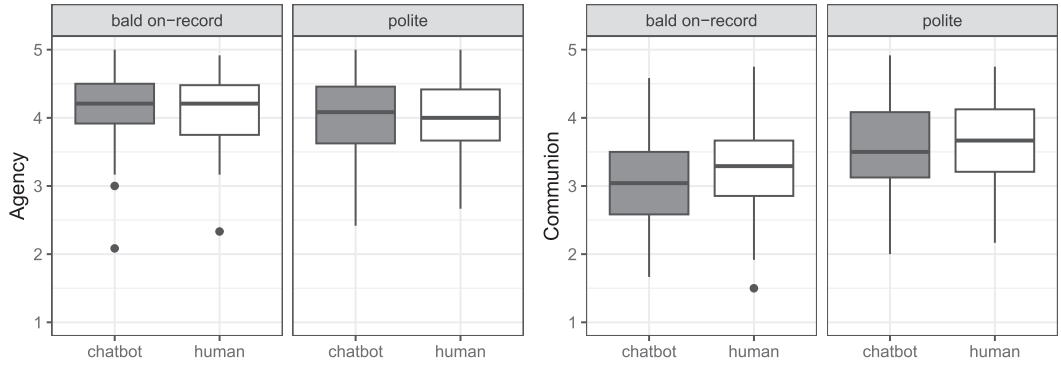

FIGURE 3 Boxplots of the data for the agency and communion subscales in the four experimental conditions.

Communion

Although we expected a disordinal interaction effect, the analysis revealed none, $F(1, 280) = 0.71$, $p = 0.402$. Thus, H3b is not confirmed. We also did not find the expected main effect of authorship on communion, $F(1, 280) = 3.28$, $p = 0.071$. Hence, H2c is not confirmed.

However, there was a main effect for politeness in that polite communicators were seen as more communal ($M = 3.63$, $SD = 0.62$) than bald on-record communicators ($M = 3.18$, $SD = 0.64$), $F(1, 280) = 35.57$, $p < 0.001$, $\eta^2 = 0.11$, confirming H1a.

DISCUSSION

The present study investigated processes of social cognition in the perception of AI chatbots as feedback givers. Overall, the experiment showed four significant results: First, an AI was perceived as less benevolent than a human teacher when giving the same feedback. Even the use of polite language did not prevent the AI from being perceived as less benevolent than a human giving blunt and direct—bald on-record—feedback. Furthermore, as expected, benevolence and communion were perceived to be higher when the author of the feedback used a polite communication strategy. Finally, feedback authors that followed a bald on-record communication strategy were perceived as having more agency.

Evaluation of hypotheses

Our findings indicating lower perceived benevolence for the AI chatbot are in line with previous findings that humans are more liked and seen as more benevolent than computer systems (Krämer et al., 2017; Linnemann & Jucks, 2016). Although using polite language has been found to improve perceptions of a computer's benevolence (Jucks et al., 2018), there still seems to be a gap between humans and computers. In this light, the research question and the first group of hypotheses (H1) can be answered as followed: Our expectations that the AI chatbot would be judged as higher on communion (H1a) and that teachers would receive higher ratings on agency (H1b) were not confirmed. These findings could be interpreted as indicating that there is no positional triangle between students, teachers and an AI system, but rather a dichotomy between us (students) and them (teachers and teaching tools such as an AI chatbot).

Regarding politeness, the results confirmed our hypotheses: Benevolence (H2a) and communion (H2b) were rated higher for a polite author, while agency (H2c) was rated higher for a bald on-record communicator. Prior research has already indicated that polite communication has a positive effect on the benevolence of both computer and human actors (Brummernhenrich & Jucks, 2016; Jucks et al., 2018; Linnemann & Jucks, 2016), and the current results underline once again the relevance of politeness in feedback interventions in educational contexts (eg, Trees et al., 2009). In accordance with an earlier study (Brummernhenrich & Jucks, 2019), we found no significant effect of politeness on expertise or integrity. While that previous study focussed on medical professionals, our results in the present study extend these results to an AI feedback author and a regular human teacher.

In summary, our results suggest that both humans and chatbots were accepted as feedback givers by the university students who took part in the study, at least when evaluating the feedback as a third party. As long as the system acts in an outwardly competent manner, an AI chatbot seems to be an accepted expert actor in the eyes of students. However, there seems to be a gap in the perceived benevolence of computer actors compared to human actors. Other research in the broader field of computer perception has supported the finding that computers are often seen as inferior communication partners, but that more well-spoken or empathetic computer systems can mediate this gap (Krämer et al., 2017; Rhim et al., 2022; Ruane et al., 2021). Thus, developing *empathic pedagogical conversational agents* (Ortega-Ochoa et al., 2024) is an apposite step that can take these factors into account.

Limitations

To realise an effective experimental variation, we designed our study such that students did not receive feedback on their own submissions but instead read constructed chat dialogues. In this manner, we aimed to hold as many factors constant as possible between experimental conditions. Yet, this decision may have prevented participants from forming strong opinions about the feedback, as it did not address their own work. Since other studies have shown similar effects in situations where people received responses to questions directly from computer systems versus humans (Linnemann & Jucks, 2018), we assume that our results can be generalised to these situations.

We also did not include manipulation checks for the two independent variables. Because we included several instances of both positive as well as negative politeness and strategies, and similar variations have previously been found to successfully impact perceptions of politeness in the expected direction (Brummernhenrich & Jucks, 2019; Jucks et al., 2018), we had good reason to assume that the effects of our manipulation would be similar. The author manipulation was also constructed similarly to earlier studies (eg, Huiyang & Min, 2022), and each survey page as well as each chat bubble indicated who the author of the feedback was; thus, we assume this to have been salient to participants.

Notably, the specific manner in which the dialogue was constructed may limit the contexts into which our findings can be generalised. For one, we specified a gender and academic title for the human teacher, which may have primed specific expectations or stereotypes that likely are not as available relevant for computer systems. Our chatbot did not have a corresponding visual avatar, and it did not use speech synthesis; when chatbots have these features, they are often gendered. Because gender interacts with politeness considerations (Holtgraves & Yang, 1992; Sung, 2012) as well as general social cognition (Fiske et al., 2007) in communication between humans, whether these aspects play out similarly or differently in human–computer interaction is a relevant topic.

Additionally, the study was conducted in late 2021, approximately a year before systems employing large language models such as ChatGPT became widespread. Because these systems are now ubiquitous, students' perceptions of them may have changed. Indeed, while use of generative systems has increased rapidly, students' views on whether these systems are useful or whether students want to receive feedback from them tend to be ambivalent: While students see potential benefits in personalised feedback, they also worry about the systems' accuracy and fairness (Chan & Hu, 2023). Trust in AI systems is still a relevant factor in whether they are considered for use (Choung et al., 2023). Because our findings underpin the relevance of social aspects of language, we assume that they would generalise to current systems. This is exemplified by some services that offer wrappers around models such as ChatGPT in an attempt to make the system's responses more sociable.

Finally, the participants in our study were limited to students in a relatively small region in northwest Germany. Generally, Germany tends to lag behind in adopting digital technologies in the educational sector (OECD, 2020). Attitudes towards conversing with AI systems may have developed differently in places where digital technologies are more in common.

Implications and future directions

The study shows, once more, that politeness theory offers a useful perspective on instructional communication. It makes clear predictions about what types of utterances will be expected by interlocutors and what inferences are drawn about the social relationship. Theories of computers as social actors (Holtgraves et al., 2007; Reeves & Nass, 1996) have extended these findings to the realm of communicating with digital systems by assuming that humans can treat computers in a similar way to other humans. We argue that users' social cognition applies in a similar manner to chatbots as to other humans: Whether face threats were phrased in a polite or a direct manner yielded many of the same effects, in the same direction, for the chatbot and for the human teacher. If social cognition did not apply in the case of the chatbot, this would not have been the case. Yet, some differences still exist between the two feedback-giver conditions, underlining the fact that two things are simultaneously true: Feedback recipients do differentiate whether the feedback comes from a computer or a human, but the kind of language used still affects certain perceptions.

Politeness is only one of many social facets of language and, indeed, of communication. Other social aspects of language that have been researched in human learning interactions are immediacy (Kerssen-Griep & Witt, 2012, 2015; Witt & Kerssen-Griep, 2011), self-disclosure (Mazer et al., 2007) and humour (Wanzer & Frymier, 1999), all of which can have beneficial effects on learning outcomes. Future research could focus on whether the effects of these forms of social language also depend on social cognition, and whether chatbots and other computer systems that employ these strategies in educational contexts receive more positive perceptions of epistemic trustworthiness.

Although we expected interaction effects, such that the politeness of the feedback would be perceived differently depending on whether it was given by a computer or a human, we did not find such effects. This mirrors recent research that varied another social facet of language use, personalised language (Ruwe & Mayweg-Paus, 2023). Although we hesitate to draw conclusions from the absence of effects, it may be that learners do not strongly differentiate between humans and computers, at least regarding the effects of linguistic strategies. This also points to a possible explanation for not finding an expectancy violation effect for the polite chatbot in the form of especially positive perceptions: Our participants' expectancies for interactions with chatbots simply did not strongly differ from their expectancies for interactions with human teachers. We still found differences between participants' perceptions of humans and chatbots, but these did not interact with the way they

communicated. Future research should investigate the extent as well as the genesis of these differences.

Despite their limitations, we argue that the results of this study support the feasibility of AI systems as feedback providers in higher education. Because our focus was on social cognition, we did not collect assessments of the feedback itself, such as perceived usefulness, effectiveness, or fairness. Notably, it is still unclear whether feedback that is perceived as useful or effective by learners is actually the best for learning (Van der Kleij & Lipnevich, 2021). Nonetheless, if a chatbot's feedback is perceived as lacking, learners may be less inclined to consult it (Granić & Marangunić, 2019).

Another prerequisite for the beneficial use of AI systems in higher education is that they are also acceptable to instructors (Ji et al., 2023; Knezek & Christensen, 2016; Nazaretsky et al., 2022). Apart from linguistic strategies such as politeness, other communication-related factors that may impact their acceptability for practitioners should be investigated, such as whether a system explains its mode of operation, how it comes to its conclusions or recommendations (Rohlfing et al., 2021), and how it deals with disagreements (Kempt et al., 2023). Thus, further research should focus on whether social cognition plays a role not only in students' but also in university teachers' evaluations of these systems.

ACKNOWLEDGEMENTS

The authors are thankful to Celeste Brennecka and Eileen Plagge for language editing and to Helena Thorbrügge and Rabea Leister for preparing the data and materials for open access publication. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

The authors declare that no funds, grants or other support were received during the preparation of this manuscript.

CONFLICT OF INTEREST STATEMENT

The authors have no relevant financial or non-financial interests to disclose.

DATA AVAILABILITY STATEMENT

All materials, analysis scripts and data are available on OSF: https://osf.io/q6f7e/.

ETHICS STATEMENT

This research was conducted in accordance with the guidelines of the ethics review board of the University of Münster Department of Psychology and Sports Sciences. No personally identifiable information was collected from the study participants.

ORCID

Benjamin Brummernhenrich https://orcid.org/0000-0002-5680-9170
Christian L. Paulus https://orcid.org/0000-0002-9866-6198
Regina Jucks https://orcid.org/0000-0002-3980-4327

REFERENCES

Abele, A. E., & Bruckmüller, S. (2013). The big two of agency and communion in language and communication. In J. P. Forgas, O. Vincze, & J. László (Eds.), *Social cognition and communication* (pp. 173–184). Psychology Press.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science*, *228*(4698), 456–462. https://doi.org/10.1126/science.228.4698.456

Bills, L. (2000). Politeness in teacher-student dialogue in mathematics: A socio-linguistic analysis. *For the Learning of Mathematics*, *20*(2), 40–47. http://www.jstor.org/stable/40248326

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

Brummernhenrich, B., & Jucks, R. (2013). Managing face threats and instructions in online tutoring. *Journal of Educational Psychology*, *105*(2), 341–350. https://doi.org/10.1037/a0031928

Brummernhenrich, B., & Jucks, R. (2016). "He shouldn't have put it that way!" How face threats and mitigation strategies affect person perception in online tutoring. *Communication Education*, *65*(3), 290–306. https://doi.org/10.1080/03634523.2015.1070957

Brummernhenrich, B., & Jucks, R. (2019). "Get the shot, now!" Disentangling content-related and social cues in physician–patient communication. *Health Psychology Open*, *6*(1), 2055102919833057. https://doi.org/10.1177/2055102919833057

Burgoon, M., Pauls, V., & Roberts, D. L. (2002). Language expectancy theory. In J. P. Dillard & M. Pfau (Eds.), *The persuasion handbook: Developments in theory and practice* (pp. 117–136). Sage Publications. https://doi.org/10.4135/9781412976046

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Choung, H., David, P., & Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human-Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, *77*(1), 113–143. https://doi.org/10.3102/003465430298563

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development. Attention and performance XXI* (pp. 249–274). Oxford University Press.

Dennis, M., Masthoff, J., & Mellish, C. (2016). Adapting progress feedback and emotional support to learner personality. *International Journal of Artificial Intelligence in Education*, *26*(3), 877–931. https://doi.org/10.1007/s40593-015-0059-7

Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology*, *63*(6), 877–889. https://doi.org/10.1037/0022-3514.63.6.877

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*, 71–86. https://doi.org/10.30658/hmc.1.5

Goffman, E. (1967). *Interaction ritual: Essays on face-to-face interaction*. Aldine.

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494–1509. https://doi.org/10.1016/j.chb.2007.05.007

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education*, *26*(1), 124–132. https://doi.org/10.1007/s40593-015-0086-4

Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook, Vol. 3. Application to learning and teaching* (pp. 451–473). American Psychological Association. https://doi.org/10.1037/13275-018

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology*, *50*(5), 2572–2593. https://doi.org/10.1111/bjet.12864

Grawemeyer, B., Mavrikis, M., Holmes, W., Gutiérrez-Santos, S., Wiedmann, M., & Rummel, N. (2017). Affective learning: Improving engagement and enhancing learning with affect-aware feedback. *User Modeling and User-Adapted Interaction*, *27*(1), 119–158. https://doi.org/10.1007/s11257-017-9188-z

Hendriks, F., Kienhues, D., & Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: The muenster epistemic trustworthiness inventory (METI). *PLoS ONE*, *10*(10), e0139309. https://doi.org/10.1371/journal.pone.0139309

Holtgraves, T., & Yang, J. (1992). Interpersonal underpinnings of request strategies: General principles and differences due to culture and gender. *Journal of Personality and Social Psychology*, *62*(2), 246–256. https://doi.org/10.1037/0022-3514.62.2.246

Holtgraves, T. M., Ross, S. J., Weywadt, C. R., & Han, T. L. (2007). Perceiving artificial social agents. *Computers in Human Behavior*, *23*(5), 2163–2174. https://doi.org/10.1016/j.chb.2006.02.017

Huiyang, S., & Min, W. (2022). Improving interaction experience through lexical convergence: The prosocial effect of lexical alignment in human-human and human-computer interactions. *International Journal of Human-Computer Interaction*, *38*(1), 28–41. https://doi.org/10.1080/10447318.2021.1921367

Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, *55*(1), 48–63. https://doi.org/10.1080/15391523.2022.2142873

Jucks, R., Linnemann, G. A., & Brummernhenrich, B. (2018). Student evaluations of a (rude) spoken dialogue system: Insights from an experimental study. *Advances in Human-Computer Interaction*, *2018*, 8406187. https://doi.org/10.1155/2018/8406187

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kempt, H., Heilinger, J.-C., & Nagel, S. K. (2023). "I'm afraid I can't let you do that, doctor": Meaningful disagreements with AI in medical contexts. *AI & Society*, *38*(4), 1407–1414. https://doi.org/10.1007/s00146-022-01418-x

Kerssen-Griep, J., Hess, J. A., & Trees, A. R. (2003). Sustaining the desire to learn: Dimensions of perceived instructional facework related to student involvement and motivation to learn. *Western Journal of Communication*, *67*(4), 357–381. https://doi.org/10.1080/10570310309374779

Kerssen-Griep, J., & Witt, P. L. (2012). Instructional feedback II: How do instructor immediacy cues and facework tactics interact to predict student motivation and fairness perceptions? *Communication Studies*, *63*(4), 498–517. https://doi.org/10.1080/10510974.2011.632660

Kerssen-Griep, J., & Witt, P. L. (2015). Instructional feedback III: How do instructor facework tactics and immediacy cues interact to predict student perceptions of being mentored? *Communication Education*, *64*(1), 1–24. https://doi.org/10.1080/03634523.2014.978797

Knezek, G., & Christensen, R. (2016). Extending the will, skill, tool model of technology integration: Adding pedagogy as a new model construct. *Journal of Computing in Higher Education*, *28*(3), 307–325. https://doi.org/10.1007/s12528-016-9120-2

Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., & Pineau, J. (2022). Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, *32*(2), 323–349. https://doi.org/10.1007/s40593-021-00267-x

Krämer, N. C., Leiße, L.-M., Hollingshead, A., & Gratch, J. (2017). Evaluated by a machine. Effects of negative feedback by a computer or human boss. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, & S. Kopp (Eds.), *Proceedings of the 17th International Conference on Intelligent Virtual Agents* (pp. 235–238). Springer. https://doi.org/10.1007/978-3-319-67401-8_29

Lew, Z., & Walther, J. B. (2022). Social scripts and expectancy violations: Evaluating communication with human or AI chatbot interactants. *Media Psychology*, *26*(1), 1–16. https://doi.org/10.1080/15213269.2022.2084111

Lin, J., Raković, M., Li, Y., Xie, H., Lang, D., Gašević, D., & Chen, G. (2024). On the role of politeness in online human–human tutoring. *British Journal of Educational Technology*, *55*(1), 156–180. https://doi.org/10.1111/bjet.13333

Linnemann, G. A., & Jucks, R. (2016). As in the question, so in the answer? Language style of human and machine speakers affects interlocutors' convergence on wordings. *Journal of Language and Social Psychology*, *35*(6), 686–697. https://doi.org/10.1177/0261927X15625444

Linnemann, G. A., & Jucks, R. (2018). 'Can I trust the spoken dialogue system because it uses the same words as I do?' Influence of lexically aligned spoken dialogue systems on trustworthiness and user satisfaction. *Interacting with Computers*, *30*(3), 173–186. https://doi.org/10.1093/iwc/iwy005

Mazer, J. P., Murphy, R. E., & Simonds, C. J. (2007). I'll see you on "Facebook": The effects of computer-mediated teacher self-disclosure on student motivation, affective learning, and classroom climate. *Communication Education*, *56*(1), 1–17. https://doi.org/10.1080/03634520601009710

McCroskey, J. C., Valencic, K. M., & Richmond, V. P. (2004). Toward a general model of instructional communication. *Communication Quarterly*, *52*(3), 197–210. https://doi.org/10.1080/01463370409370192

McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *iScience*, *26*(8), 107256. https://doi.org/10.1016/j.isci.2023.107256

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011a). A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies*, *69*(1–2), 70–79. https://doi.org/10.1016/j.ijhcs.2010.09.001

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011b). Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education*, *56*(3), 574–584. https://doi.org/10.1016/j.compedu.2010.09.019

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence—CHI '94* (pp. 72–78). https://doi.org/10.1145/191666.191703

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, *53*(4), 914–931. https://doi.org/10.1111/bjet.13232

OECD. (2020). *PISA 2018 results (volume V): Effective policies, successful schools*. OECD Publishing. https://doi.org/10.1787/ca768d40-en

Ortega-Ochoa, E., Arguedas, M., & Daradoumis, T. (2024). Empathic pedagogical conversational agents: A systematic literature review. *British Journal of Educational Technology*, *55*(3), 886–909. https://doi.org/10.1111/bjet.13413

Person, N. K., Kreuz, R. J., Zwaan, R. A., & Graesser, A. C. (1995). Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and Instruction*, *13*(2), 161–188. https://doi.org/10.1207/s1532690xci1302_1

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Rhim, J., Kwak, M., Gong, Y., & Gweon, G. (2022). Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior*, *126*, 107034. https://doi.org/10.1016/j.chb.2021.107034

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology*, *41*(1), 307–364. https://doi.org/10.1002/aris.2007.1440410114

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Hab-Umbach, R., Horwath, I., Hullermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B. (2021). Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 717–728. https://doi.org/10.1109/TCDS.2020.3044366

Ruane, E., Farrell, S., & Ventresque, A. (2021). User perception of text-based chatbot personality. In *Chatbot Research and Design: 4th International Workshop, Conversations 2020* (pp. 32–47). Springer. https://doi.org/10.1007/978-3-030-68288-0_3

Ruwe, T., & Mayweg-Paus, E. (2023). "Your argumentation is good", says the AI vs humans—The role of feedback providers and personalised language for feedback effectiveness. *Computers & Education*, *5*, 100189. https://doi.org/10.1016/j.caeai.2023.100189

Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management*, *52*, 102061. https://doi.org/10.1016/j.ijinfomgt.2019.102061

Sung, C. C. M. (2012). Exploring the interplay of gender, discourse, and (im)politeness. *Journal of Gender Studies*, *21*(3), 285–300. https://doi.org/10.1080/09589236.2012.681179

Trees, A. R., Kerssen-Griep, J., & Hess, J. A. (2009). Earning influence by communicating respect: Facework's contributions to effective instructional feedback. *Communication Education*, *58*(3), 397–416. https://doi.org/10.1080/03634520802613419

Uchronski, M. (2008). Agency and communion in spontaneous self-descriptions: Occurrence and situational malleability. *European Journal of Social Psychology*, *38*(7), 1093–1102. https://doi.org/10.1002/ejsp.563

Van der Kleij, F. M., & Lipnevich, A. A. (2021). Student perceptions of assessment feedback: A critical scoping review and call for research. *Educational Assessment, Evaluation and Accountability*, *33*(2), 345–373. https://doi.org/10.1007/s11092-020-09331-x

Van Lehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Walter, N., Ortbach, K., & Niehaves, B. (2015). Designing electronic feedback—Analyzing the effects of social presence on perceived feedback usefulness. *International Journal of Human-Computer Studies*, *76*, 1–11. https://doi.org/10.1016/j.ijhcs.2014.12.001

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*, *66*(2), 98–112. https://doi.org/10.1016/j.ijhcs.2007.09.003

Wanzer, M. B., & Frymier, A. B. (1999). The relationship between student perceptions of instructor humor and students' reports of learning. *Communication Education*, *48*(1), 48–62. https://doi.org/10.1080/03634529909379152

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Witt, P. L., & Kerssen-Griep, J. (2011). Instructional feedback I: The interaction of facework and immediacy on students' perceptions of instructor credibility. *Communication Education*, *60*(1), 75–94. https://doi.org/10.1080/03634523.2010.507820

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, *4*, 654924. https://doi.org/10.3389/frai.2021.654924

How to cite this article: Brummernhenrich, B., Paulus, C. L., & Jucks, R. (2025). Applying social cognition to feedback chatbots: Enhancing trustworthiness through politeness. *British Journal of Educational Technology*, *00*, 1–20. https://doi.org/10.1111/bjet.13569

APPENDIX

DIALOGUE TEXT AND POLITENESS VARIATION

Dialogue text for the human teacher in the two politeness conditions. Changes between conditions are highlighted in bold.

Polite condition	Bald on-record condition
Dr. Meyer: Hello, Mr. Schneider, I will be happy to provide some recommendations for your term paper. My feedback will follow the provided structure	*Dr. Meyer*: Hello Mr. Schneider. I will now give you some recommendations for your term paper. My feedback will follow the provided structure
Dr. Meyer: I will start with a positive aspect: the summary at the beginning is on a good to very good level. The remarks are well ordered and structured in a comprehensible manner. **I suggest that you state the outline even more clearly. This makes your argumentation easier to follow**	*Dr. Meyer*: I will start with a positive aspect: the summary at the beginning is on a good to very good level. The remarks are well ordered and structured in a comprehensible manner. **However, you need to state the outline more clearly in the summary. This is essential for clarity**
Leon: That sounds okay, what else? *Dr. Meyer*: As good as your summary and classifications are, you primarily reference fundamental scientific publications, that were discussed in the course. **I wasn't always able to recognize your independent scientific reasoning. I would therefore recommend that you expand your research to include a few more** relevant scientific publications and incorporating them into your argument structure. **This could significantly improve your text**	*Leon*: That sounds okay, what else? *Dr. Meyer*: **Your independent scientific reasoning is incomplete**. As good as your summary and classifications are, you primarily reference **very** fundamental scientific publications, that were discussed in the course. **You need to research additional** relevant scientific publications and incorporate them into your line of reasoning. **You urgently need to improve your text in this regard**
Leon: Do you have an example for me? *Dr. Meyer*: On the second page, you write: *"[…] Students do not seem to have a homogeneous preference for MC or CR tasks. These results are not consistent with previous studies. The authors interpret these results by referring to their construct 'opportunity to demonstrate performance,' concluding that high-achieving students seem to show no preference because they can better demonstrate their own performance in the CR format"* You list other studies, but there is **unfortunately** no reference to them. **You did this better on other occurrences and I would recommend delivering references and sources without exceptions**	*Leon*: Do you have an example for me? *Dr. Meyer*: On the second page, you write: *"[…] Students do not seem to have a homogeneous preference for MC or CR tasks. These results are not consistent with previous studies. The authors interpret these results by referring to their construct 'opportunity to demonstrate performance,' concluding that high-achieving students seem to show no preference because they can better demonstrate their own performance in the CR format"* You list other studies, but there is no reference to them. **You must improve on this aspect and deliver references and sources, just like you did on other occurrences**

Polite condition	Bald on-record condition
Leon: But I didn't make that mistake everywhere, did I? *Dr. Meyer*: No, **don't worry**. You have worked on about half of your arguments in a scientifically meaningful way. Currently, you receive 2–3 out of the 7 possible points in the evaluation grid here. **This still puts you in the lower third compared to your fellow students, but with a bit of revision, you could get a good 5–6 points here. Do you want to hear how you can improve here?**	*Leon*: But I didn't make that mistake everywhere, did I? *Dr. Meyer*: No, **but too often**. You have worked on about half of your arguments in a scientifically meaningful way. Currently, you receive 2–3 out of the 7 possible points in the evaluation grid here. **However, you are in the lower third compared to your fellow students. You urgently need to work on your text in order to achieve 5–6 more points here. I will give you some points for improvement below**
Leon: Yes, thank you *Dr. Meyer*: **As you know**, you have correctly referenced key publications in some places. **So, you already have the tools to improve your work. You could** look for additional literature that was not covered in the lecture and use it to support your arguments. If you cannot find good publications to support an argument, **I recommend** toning down the wording or **removing it from your line of reasoning**	*Leon*: Yes, thank you *Dr. Meyer*: You have correctly referenced key publications in some places, **but far too rarely**. Look for additional literature that was not covered in the course and use it to support your arguments**!** If you cannot find good publications for an argument, **you should** tone down the wording **or, better yet, leave it out altogether**
Leon: Okay, thank you *Dr. Meyer*: Overall, your current draft fulfilled most of the criteria in a satisfactory to good manner. **I also enjoyed reading your work in terms of its fluency**. However, you do not receive full points on any evaluation criterion yet. **You had already received these criteria in an evaluation rubric beforehand. Perhaps you could review the list again** and revise your submission, particularly in the discussed areas. **As I mentioned, you are showing your potential to produce really good work. This way, you could** receive a B or A on the task	*Leon*: Okay, thank you *Dr. Meyer*: Overall, your current draft fulfilled most of the criteria in a satisfactory to good manner. **The text is written fluently**. However, you do not receive full points on any evaluation criterion. **These criteria were made available in advance through an evaluation rubric. Please take these into account, too! Take another look at the list** and revise your submission, particularly in the discussed areas. **This is the only way you can** receive a B or A on the task
Leon: Great, thank you *Dr. Meyer*: **You're welcome**. Your consultation time is **unfortunately** over now. Thank you for your participation. **Feel free to** contact me again, if you would like to continue your feedback discussion at a later date	*Leon*: Great, thank you *Dr. Meyer*: Your consultation time is over now. Thank you for your participation. **You can** contact me again, if you would like to continue your feedback discussion at a later date